BULGARIAN WORD ORDER AND THE ROLE OF THE DIRECT OBJECT CLITIC IN LFG

T. Florian Jaeger
Veronica A. Gerassimova

Linguistics Department, Stanford University

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

http://csli-publications.stanford.edu/

ABSTRACT – This paper provides an LFG account of the Bulgarian direct object clitic's interaction with information structure (i.e. topic-focus structure) and word order. We show that the direct object clitic has at least two functions (it is both a topical object agreement marker and default pronoun) and then demonstrate how our account correctly predicts in which syntactic environment which of the two functions can be chosen. In order to achieve this we allow for two different ways to identify a 'topic' in LFG – a move, which reduces the necessary claims about the direct object clitic's behaviour to the most general principles of LFG (i.e. Uniqueness, Completeness, Extended Coherence). The proposed analysis is based on extensive evidence (our own online experiment, Leafgren 1997a,b, 1998, and Avgustinova 1997), and incorporates recent findings on the discourse-configurationality of the left periphery in Bulgarian clauses (cf. Rudin 1997, Arnaudova 2001, Dimitrova-Vulchanova & Hellan 1998). Although covering a much broader range of data from spoken Bulgarian than other formal accounts, our account makes the right predictions about possible word orders and the optional, or obligatory presence/absence of the direct object clitic. Unlike almost all other recent accounts, our analysis does not rely on the assumption of configurationality, which has been shown to be problematic for Bulgarian (cf. Gerassimova & Jaeger 2002).

I Introduction*

Contemporary, colloquial Bulgarian allows for clitic doubling of objects in certain contexts. The object clitics can occur as the only realization of the object, as in (1), double an NP or double a long form pronoun, as in (2). Although there is also an indirect object clitic whose distribution is for the most part parallel with the direct object clitic's, we restrict ourselves to the investigation of the direct object clitic (henceforth DOC).[1] All examples given in this paper only contain the direct object clitic. The DOC can also occur in an embedded sentence from which the direct object has been extracted. An example for extraction out of an adjunct clause, is given in (3). (4) is an example of object extraction out of a sentential subject. For ease of understanding, the DOC and the coreferential object (if present) are underlined.

(1) Decata ja običat.[2]
children$_{DEF.PL}$ DOC$_{3.SG.FEM}$ love$_3$
The children love her.

(2) Decata ja običat Marija/neja.
children$_{DEF.PL}$ DOC$_{3.SG.FEM}$ love$_3$ Maria/her$_{3.SG.FEM.ACC}$
The children love Maria/her.

(3) Radioto, koeto Todor otide na plaž [bez da (go)
radio$_{DEF}$ which Todor went$_3$ on beach without SBJ DOC$_{3.SG.MASC}$
izključi], e na Elena.
switch-off is of Elena
The radio which Todor went to the beach without switching off is Elena's.

* Our special thanks go out to Peter Sells, Joan Bresnan, Elizabeth Traugott, Chris Manning, Arnold Zwicky, and Tracy H. King for their advice and support all throughout the progress of our research. We want to thank Mary Dalrymple, Tracy H. King, and Jonas Kuhn for their help with some formal aspects of LFG. We also want to very much thank Ruth Kempson for making us aware of several interesting questions and providing good ideas how to approach them, as well as Iskra Iskrova for discussing the relevant data with us. Last but not least, we benefited from the questions and suggestions from the anonymous reviewers of the LFG02 abstracts, Shiao-Wei Tham, Judith Tonhauser, Andrew Koontz-Garboden, and especially Lev Blumenfeld. We are also grateful for the great feedback we got at the LFG02 and at an earlier presentation of our work for the Linguistics Department, Stanford. Thanks to Tracy H. King (again) and Lev Blumenfeld for feedback on the final draft of this paper. All remaining mistakes remain ours and must not be reproduced without our permission, ;-).

[1] We use the term direct object clitic (DOC) to refer to the set of linguistic *forms* of the direct object clitic, not their meaning. These are the following forms: SG – 1st *me*, 2nd *te*, 3rd masc./neut. *go*, fem. *ja*; PL – 1st *ni*, 2nd *vi*, 3rd *gi*.

[2] We use the following glosses: 1, 2, 3 – first, second, and third person; DEF – definite suffix, INDEF – indefinite specific article; FEM – feminine, MASC – masculine, NEUT – neuter; PL – plural, SG – singular; REFL – reflexive pronoun, SBJ – subjunctive marker. SMALL CAPS indicate emphatic accent.

(4) Todor e jasno, če Ivan *(go) e vidjal.
Todor is clear that Ivan $DOC_{3.SG.MASC}$ is seen
Todor, it is clear that Ivan has seen him.

In this paper, we discuss different functions of the Bulgarian direct object clitic and its interaction with syntax (especially word order) and information structure. Our research relates to research on D[iscourse] F[unction] and G[rammatical] F[unction]-configurationality, an issue which has been identified as the primary, so far unresolved issue in the literature on (South) Slavic Syntax (cf. Siewierska & Uhlířová 1998:143 in their review of the recent literature on the word order of Slavic languages).

We propose an analysis of the DOC, which - we argue - accounts for a whole range of data that so far have not been explained, including examples of free word order in a non-dependent marking language. We argue that the DOC has not one but several functions, one of which has not been recognized at all in the literature. This is at least partially the reason why the issue of the DOC's functions is still unresolved in the literature. First, in the clitic doubling construction (we explain what we mean by this in the next paragraph), the DOC is a non-anaphoric direct object TOPIC-agreement-marker. Second, the DOC is the default direct object pronoun. Third, the DOC is an intrusive direct object pronoun in extractions (cf. Sells 1984). Due to lack of space, we only discuss the first two functions here. We argue that some of the confusion about these functions in the literature is due to different notions of topic and suggest a way to resolve this issue within LFG. Furthermore, we account for the range of possible word orders given the presence or absence of the DOC.

The last point is especially important since – to the best of our knowledge – all existing accounts either hardly, if at all, capture the generalizations relating to possible word orders, or only account for a relatively small subset of them. To guarantee a broad coverage of data, we test and compare the predictions of our analysis with the data provided in Avgustinova (1997; elicited question-answer pairs) and Leafgren (1997a,b, 1998, 2001; corpus studies of written/spoken, informal/formal Bulgarian). Moreover, we use the case of island violations to show how the distribution of the DOC as default pronoun or topic marker is correctly predicted.

Before we provide an outline of the structure of this paper, we will briefly clarify our use of the term 'clitic doubling'. With clitic doubling (henceforth CD) we refer to the overt doubling of a constituent, usually an argument (here the direct object), by a phonologically weak, syntactically non-projecting[3] lexical element, i.e. a clitic (here the DOC). CD is a prominent topic in the literature on Slavic and Balkan linguistics (e.g. Franks & King 2000, Rudin 1990/1991, 1996, 1997, Dyer 1992, Guentcheva 1994, a.o.), the typology of pronouns, agreement (e.g. Bresnan & Mchombo 1987), configurationality (e.g. Baker's (1991) *pronominal object hypothesis*), and case assignment (e.g. Rudin 1997). The aspects of CD that are addressed here include the following. First, how is coreference between the clitic and the doubled NP established? In MP/GB this comes down to the question whether, for example, fronted objects are moved or anaphorically bound by the DOC. In LFG terms, this corresponds to the issue of functional control vs. anaphoric binding. Second, does the clitic mark a grammatical function (GF) or a discourse function (DF)[4] or both? Third, is the clitic and/or the lexical object NP the object argument? This is interesting since according to some theories (e.g. GB, MP) only one constituent can be assigned CASE. The LFG framework is less restrictive in this respect. As long as UNIQUENESS (cf. Bresnan 2001:47) is fulfilled, information belonging to the same GF can be distributed among several syntactic constituents. Nevertheless, translated into LFG, the above-mentioned question remains, namely whether the clitic provides information on OBJ PRED (i.e. the PRED value of the object).

In the remaining sections, we proceed as follows. In section II, we introduce some basic facts about Bulgarian, including some phrase structure rules describing the internal order of the predicate clitic cluster and capturing the fact that Bulgarian is not configurational. In section III, we briefly

[3] See Toivonen (2001:chapter 3) for a typology of non-projecting words.
[4] Note that, with 'discourse function', we do not refer to discourse function as defined in Schiffrin (1988) or Fraser (1988). We comply to the naming convention of LFG and use the term discourse function (DF) to refer to what more precisely could be called f-structure correlate of an information structural role.

describe some earlier analyses of the DOC's function. In section IV, we introduce recent findings on the discourse-configurationality in Bulgarian, incorporate them into our analysis, and formalize the direct object clitic's (DOC) properties in CD. In this context, we discuss our proposal in the light of the known data and show how the interaction of the proposed lexical entry of the DOC and the proposed phrase structure rules make the right predictions about grammaticality of certain word orders and their information structural correlate (we will elaborate on this below). We also use section IV to introduce our model of the information structure (henceforth IS) component and its interface with other components (e.g. f-structure). In section V, we discuss a second function of the DOC, which has so far been ignored in the literature, namely its use as the default pronoun of Bulgarian. In section VI, we briefly survey islands in Bulgarian to show how our account makes the right predictions about the distribution of the different types of DOCs. Last, we will summarize the conclusions and mention some open issues in section VII.

II An introduction to some aspects of Bulgarian

Bulgarian is a South-Slavic language spoken by approximately 9 million speakers[5] world wide. If not mentioned otherwise, we will use the term Bulgarian to refer to contemporary, colloquial, spoken Bulgarian. Bulgaria has a strong prescriptive tradition and the differences between written vs. spoken and formal vs. informal Bulgarian seem to be immense.[6] Clitic doubling (henceforth CD) is very rare in formal and written Bulgarian. Leafgren (2001:4) shows that the frequency of CD in formal written texts (0.5% of all object occurrences) contrasts sharply with the 10% frequency of CD in informal oral texts. Furthermore, we restrict ourselves to those dialects of Bulgarian which make productive use of the object clitics, i.e. mostly the Western dialects (cf. Leafgren 1997a:119).

Since Bulgarian is in many respects the most atypical Slavic language and has some typologically uncommon properties, we sketch those characteristics of Bulgarian that will turn out to be relevant for understanding the analysis presented in section IV.

Unlike all other Slavic languages (except for Macedonian) Bulgarian has lost its case marking system. Some scholars have argued that the definiteness suffix (singular: masc. *-a*, fem.: *-ta*, neut.: *-to*; plural: *-te*/*-ta*) identifies the subject. This is wrong since the definiteness suffix can also be attached to an object. The only dependent-marking device in Bulgarian is the preposition *na* which among other things identifies the indirect object. In certain environments even this last bit of dependent-marking can be dropped (cf. Vakareliyska 1994).

Despite the almost complete lack of dependent-marking, Bulgarian allows very free word order. With different requirements on the context, the intonation and morpho-syntactic marking, all theoretically possible word orders can actually be observed (cf. Siewierska & Uhlířová 1998:107-10 for ditransitives and implicitly Avgustinova 1997:112). While we provide more details on the effect of the DOC on word order in section IV, it is generally true that some word orders are not possible without the DOC. In other words, the DOC seems to 'license' certain word orders. Two examples for alternative word orders with the DOC are given below (based on Avgustinova 1997:112).

(5) Parite *(gi) vze Olga.
money$_{DEF}$ DOC$_{3.PL}$ took$_{3.SG}$ Olga

(6) vze *(gi) Olga parite.
took$_{3.SG}$ DOC$_{3.PL}$ Olga money$_{DEF}$
Olga took the money.

Note, however, that Bulgarian shows a clear preference for a SUBJ-V-DO-IO surface order, a tendency noted by several scholars (cf. Leafgren 2002:1, Dyer 1992:63, Avgustinova 1997:114, a.o.). Leafgren (2002:1) argues that averaged over all registers and genres about 80.5% of all

[5] Data gathered in 1995. For more information, refer to Ethnologue, Barbara F. Grimes, eds. 13th Edition.
[6] During an online experiment that we designed to get native speaker judgments on contemporary, colloquial, spoken Bulgarian we first ran into problems since our informants were so strongly influenced by the idea that they had to judge the prescriptive correctness instead of 'what they actually say'.

sentences are SVO. Dyer (1992) shows that SVO is not only statistically the most common constituent order but also *stylistically neutral*.

The lack of stringent word order and case marking is – at first – surprising. However, Bulgarian has other means to identify grammatical functions, namely intonation and head-marking. Here, we focus on head-marking, more precisely one kind of head-marking in Bulgarian, clitic doubling by the direct object clitics. Before we turn to the interaction of the direct object clitic and word order, we want to briefly mention other morphosyntactic means of Bulgarian. First, the sentence predicate agrees with the subject in person and number, and participles (i.e. subjunctives) agree also in gender. The predicate combines with the clausal clitics into the predicate clitic cluster. Because there is an extensive literature on the internal order of the predicate clitics cluster (e.g. Avgustinova 1997, Siewierska & Uhlířová 1998; see Franks & King 2000:234ff. for a summary), we do not discuss this issue here. To understand the examples given later in this paper it is sufficient to bear in mind the following, simplified schema for the internal order of the predicate clitic cluster, where IOC stands for 'indirect object clitic' and DOC for 'direct object clitic' (cf. Englund 1977:109-19). For our purpose, the annotated phrase structure rule in (8) captures the generalization in (7).

(7) aux (not 3.SG) > IOC > DOC > aux (3.SG)

(8) $V \rightarrow$ (V_{CL}) (N_{CL}) (N_{CL}) (V_{CL}) V'
(↑SUBJ PERS)≠3 (↑OBJ2)=↓ (↑OBJ)=↓ (↑SUBJ PERS)=3 ↑=↓
(↑SUBJ NUM)≠SG (↑SUBJ NUM)=SG

The clitic cluster as a whole is preverbal except for the cases where this would cause the clitics to be clause-initial. In those cases, the verb is preposed to the clitic cluster. In other words, the positioning of the Bulgarian clitic cluster is subject to the Tobler-Mussafia effect (cf. Tomić 1997, 1996, Rudin et al. 1998:566; for an OT account to typology of clitic positioning, see Billings 2000) and not to Wackernagel's Law (unlike the clausal clitics in almost all other Slavic languages). The object clitics belong to the clausal clitics. In the case of clitic doubling, the object clitic(s) agree in person, number and gender (only for 3.SG) with the reduplicated object. Unlike the object clitics, which can only occur in the clitic cluster, the second kind of pronouns in Bulgarian, namely the long form pronouns, have the same syntactic distribution as full lexical NPs. The long form pronouns, when occurring alone, mark contrastive or emphatic focus (cf. Avgustinova 1997:116 Vakareliyska 1994:125; see Leafgren 1997a:118 for a table of all clitic pronouns and long form pronouns), in which case they always receive stress (compare (9) and (10) below).

(9) Decata обичат НЕЯ.[7]
children$_{DEF.PL}$ love$_3$ her$_{3.SG.FEM.ACC}$
The children love HER.

(10) Decata ja обичат neja.
children$_{DEF.PL}$ DOC$_{3.SG.FEM}$ love$_3$ her$_{3.SG.FEM.ACC}$
The children love her.

To sum up what has been said so far, Bulgarian is a non-case marking, partially head-marking, free word order language with optional clitic doubling of objects. Another important aspect of Bulgarian that has been ignored in the literature so far is the lack of evidence for G[rammatical] F[unction]-configurationality. Although already Rudin (1985) mentions that there seems to be no such evidence, GF-configurationality plays a crucial role in most recent analyses of Bulgarian syntax (including those on CD). We have shown elsewhere (cf. Gerassimova & Jaeger 2002) that it is difficult if not impossible to find evidence *for* GF-configurationality. More precisely, some tests, such as weak crossover tests, variable binding tests, extraction tests, etc., clearly argue for non-configurationality of Bulgarian. Therefore we do not assume GF-configurationality here.

[7] In our examples throughout the paper, we mark emphatic accent/stress with SMALL CAPS. Although only a part of the word receives emphatic accent we will just mark the whole word as prosodically emphasized.

The annotated phrase structure rule in (11) captures this and describes a flat VP with unordered constituents (c.f. Kiss 1995:11 for Hungarian).[8]

(11) VP → (XP) , (PP), V'
(↑GF)=↓ (↑OBJ2)=↓ ↑=↓

In section IV, we show that the flat VP hypothesis is necessary for or at least highly compatible with the formal account of CD and its interaction with possible word orders presented here. Before we turn to our own analysis of the DOC in CD and of its use as default pronoun of Bulgarian, we briefly summarize previous analyses of the DOC.

III Previous analyses of the DOC

All of the accounts discussed here have exclusively dealt with C[litic] D[oubling] (sometimes also referred to as clitic replication in the literature) and ignored other uses/functions of the DOC. To the best of our knowledge, the function of the DOC as default pronoun (cf. section V) and its interaction with the use of the DOC in the CD construction have not been described by anyone yet. The existing accounts of the DOC can be distinguished according to their basic hypothesis. We will discuss each of them in the order they are listed below.

(H1) The object clitics mark non-canonical word orders.
(H2) The object clitics mark the case (of the doubled constituent).
(H3) The object clitics mark definite objects.
(H4) The object clitics mark specific objects.
(H5) The object clitics mark topical objects.

Both (H1) and (H2), i.e. the word order marker and the case marker hypotheses, suggested in AG (1983,3:187-188, 282-283), Popov (1963:166, 229-230), Cyxun (1968:110) and Georgieva (1974:75), have in common the claim that CD together with word order serves to disambiguate case roles. Leafgren (1997a:124) concludes that under this view sentences with CD should be unambiguous even if both subject and object have the same gender, number, etc. However, this is not the case. Sentences with CD can be ambiguous. For example, as shown below both VOS and VSO word orders are possible with the same stress assignment as long as the clitic is present.

(12) Parite gi VZE Olga.
money$_{DEF}$ DOC$_{3.PL}$ took$_{3.SG}$ Olga
(13) VZE gi parite Olga.
took$_{3.SG}$ DOC$_{3.PL}$ money$_{DEF}$ Olga
Olga took the money.

Furthermore, the word order marker hypothesis cannot explain why the DOC is optional and why it can occur in both the unmarked and the marked word order, and the case marker hypothesis fails to account for the optionality of the object clitics. The definiteness-marker hypothesis, (H3), as proposed in Cyxun (1962:289-290), Minčeva (1969:3), Ivančev (1957:139), Georgieva (1974:75),[9] has been shown to be wrong by Ivančev (1968:164) and Kazazis & Pentheradoukis (1976:399-400), since indefinite *specific* NPs can be doubled (cf. Leafgren 1997a:122), as shown in (14). *Edno* is an instance of the Bulgarian indefinite, specific article.[10]

[8] We use the XP annotated with (↑GF)=↓ to express that all kinds of core arguments can occur in this position (including e.g. COMPs).
[9] Also see Popov & Popova (1975:48) and Popov (1973:173), who, probably aiming at specificity, require the doubled NP to be 'articulated' (cf. Leafgren 1997a:121).
[10] The specific, indefinite article has the following paradigm: Singular: masc. *edin*, fem. *edna*, neut. *edno* 'a certain, a particular'; Plural: *edni* 'certain' (cf. Vakareliyska 1994:122). More precisely, this article requires an

(14) Edno dete go vidjax da pluva.
a-cerain child DOC$_{3.SG.NEUT}$ saw$_{1.SG}$ SBJ swim$_{3.SG}$
I saw a (certain) child swimming.

Avgustinova (1997:92-95) is a recent proponent of the specificity-marker hypothesis, (H4).[11] She distinguishes between [+/-limited] nominal material and further divides [+limited] nominal material into [+/-specific] and [-limited] nominal material into [+/ -generic]. In her terminology only [+limited, +specific] objects can be doubled. The specificity-marker hypothesis is motivated by the contrast between (14) and (15). In (15) the fronted, [-specific] object cannot be doubled although the corresponding sentence (16) with neutral word order and without CD is grammatical.

(15) *Njakoja po-nova kola iskam da si ja kupja.
some$_{-SPEC}$ newer car want$_{1.SG}$ SBJ REFL$_{1.SG}$ DOC$_{3.SG.FEM}$ buy
Intended: *I want to buy (for myself) some newer car.*

(16) Iskam da si kupja njakoja po-nova kola.
want$_{1.SG}$ SBJ REFL$_{1.SG}$ buy some$_{-SPEC}$ newer car
I want to buy (for myself) some newer car.

This point is further supported by the exceptions to the generalization that *edni* is [+specific] (cf. footnote 10 above). In (17) *edni po-iziskani drexi* is [-specific] (cf. Avgustinova 1997:95) and the fronted object cannot be doubled.

(17) *Edni po-iziskani drexi gi dadoxa na Ivan.
some$_{-spec}$ stylish clothes DOC$_{3.PL}$ gave$_{3.PL}$ to Ivan
Intended: *Some stylish clothes, they gave (them) to Ivan.*

However, (H4) has also proven to be insufficient since generics *can* and in some cases even *must* be doubled, as illustrated in (18). Independently of our observations, Alexandrova (1997) and Guentchéva (1994), too, point out that generics and interrogatives can be doubled (for the doubling of interrogatives, cf. also Jaeger 2002).

(18) Slonovete *(gi) obučavat xorata.[12]
elephants$_{DEF}$ DOC$_{3.PL}$ train$_{3.PL}$ people$_{DEF}$
The elephants, (the) people train.

So far we have shown that [-limited, +generic], e.g. (18), and [+limited, +specific] object NPs, e.g. (14), can be doubled while [+limited, -specific] object NPs cannot be doubled, as shown in (15) and (17). This raises the question if [-limited, -generic] object NPs can also be doubled. As for the examples above, we use the object fronting construction to test this.[13] The examples (19) and (20) are taken from Avgustinova (1997:92). The corresponding CD examples, (21) and (22), are ungrammatical.

(19) Tuk kupuvam knigi.
here buy$_{1.SG}$ books$_{-DEF}$
I buy books here.

NP not marked by the definiteness suffix. For a formal description of the semantics of *edin*, see Izvorski (1994) who, among other things, shows that, in her terminology, *edin* is not always [+specific].

[11] See also Kazazis & Pentheradoukis (1976) and Vakareliyska (1994:122).

[12] Actually, (18) is grammatical without the DOC if *slonovete* is realized with emphatic stress and thus receives the exclusive focus. This is what we would expect since this is a case of FOCUS-fronting (see section IV). In this paper, we are only interested in non-focus object fronting, i.e. object fronting without emphatic stress on the object. Therefore, whenever we star an example with a fronted object that is not given in small caps, we always mean that this example is ungrammatical for fronted non-focused objects.

[13] This will become clearer in section IV. In short, a fronted object without focus intonation must be doubled by the corresponding object clitic if this is possible at all. If doubling is not possible (like for e.g. [+limited, -specific] object NPs) the resulting clause is ungrammatical.

(20) Târsja prijateli.
look-for$_{1.SG}$ friends$_{-DEF}$
I am looking for friends.

(21) *Knigi tuk gi kupuvam.
books$_{-DEF}$ here DOC$_{3.PL}$ buy$_{1.SG}$
Intended: *Books, I buy here.*

(22) *Prijateli gi târsja.
friends-$_{DEF}$ DOC$_{3.PL}$ look-for$_{1.SG}$
Intended: *Friends, I am looking for.*

To sum up, we have shown that [-limited, +generic] and [+limited, +specific] objects can be doubled, whereas [-limited, -generic] and [+limited, -specific] objects cannot be doubled. In the following, we adopt a slightly different but equally common classificatory system where nominal material is [+/-generic], and the [-generic] NPs are further divided into [+/-specific]. Then, the generalization is captured as follows: [-generic, -specific] NPs can *not* be doubled.[14] Note that is is typologically common that [+specific] and [+generic] NPs pattern together (Shiao-Wei Tham, p.c.). Our observations, like those of Alexandrova (1997) and Guentchéva (1994), contrast with Avgustinova's (1997) claim that only [+limited, +specific] NPs can be doubled. Our data also rejects Rudin's (1997) analysis that the DOC only doubles (topical) [+specific] NPs.

Now consider the topic-marker hypothesis, (H5), as formulated in Leafgren (1997a,b, 1998), Avgustinova & Andreeva (1999), and to some extent Ivančev (1974), Georgieva (1974), Minčeva (1969), Popov (1963:167) and the AG (1983,3:188). According to this hypothesis, the above-mentioned restriction on the doubled object is an indirect effect of the requirement that the doubled object has to be topical. Leafgren (1997a:136ff.) further shows that topicality marking in Bulgarian cannot be reduced to agentivity or subjecthood, two scales that correlate with the scale of topicality in many languages (for a discussion of those hierarchies, cf. Givon 1976). However, Leafgren (1997a,b, 1998) does not show how his proposal (i.e. (H5) as stated above) accounts for the contrast between (14) and (15) or the ungrammaticality of (17), (21), and (22). In fact, (H5) turns out to be to drastic in its formulation. Consider examples (23) and (24). In our classificatory system, *njakolko* is a [-definite; -generic, +specific] quantifier, *malko* a [-definite; -generic, -specific] quantifier. *Njakolko*, unlike *malko*, is compatible with and sometimes even requires CD.[15] However, there is no apparent reason why *njakolko spisanija* in (23) should be a topic and *malko spisanija* in (24) not. Thus it seems hard to explain the difference between (23) and (24) by (H5).[16]

(23) Ima njakolko spisanija koito mnogo xora (gi) xaresvat.
have a-few$_{+SPEC}$journals$_{-DEF}$ which$_{3.PL}$ lots people DOC$_{3.PL}$ like$_{PL}$
There are a few (certain) journals that a lot of people like (them).

(24) Ima malko spisanija, koito mnogo xora (*gi) xaresvat.
have a-few$_{-SPEC}$ journals$_{-DEF}$ which$_{3.PL}$ that people DOC$_{3.PL}$ like$_{PL}$
There is a small number of journals that a lot of people like.

There are two ways out of this problem. One is to take typological evidence as, for example, sketched in Lambrecht (1994:155-56) who claims that topics have to be "referring expressions" to

[14] Thanks to Shiao-Wei Tham for discussing different classificatory systems for the semantics of nominal material with one of the authors (F.J.). Remaining mistakes are, of course, due to the authors.

[15] We are thankful to Ruth Kempson for pointing us to this data and helping us to gather it. We also are very grateful for the patience of Iskra Iskrova who explained and discussed (23) - (25) (and other material) with one of us (F.J.) in detail.

[16] Interestingly, one of our informants pointed out that for her (24) is only grammatical if either only *koito* 'which' or only the DOC *gi* 'them' is realized. This relates to the third use of the DOC as an intrusive pronoun in extractions (cf. Sells 1984), which we cannot discuss here due to lack of space. For V.G. and another informant, (24), as given above, is grammatical.

show that there are universal restrictions on the semantics of topics.[17] This approach will result in a notion of topic that will be qualitatively quite different from that of Leafgren (1997a:127) who defines the topic to be 'what the clause is about'. Second, one could claim that CD in Bulgarian has more than one constraint on the semantics and information structural role of the doubled object, namely a) doubled objects have to be topical, and b) doubled objects cannot be [-generic, -specific]. As the comparison between (23) and (25) shows, this is necessary anyway to explain why CD is *obligatory* in some cases and *optional* in others.

(25) Njakolko spisanija mnogo xora *(gi) xaresvat.
a-few$_{+SPEC}$ journals$_{-DEF}$ lots people DOC$_{3.PL}$ like$_{PL}$
A few (certain) journals, a lot of people like (them)

Here we are mainly interested in the differences between cases of obligatory and optional CD and therefore do not care to commit ourselves to either of the two ways. The account presented here (cf. section IV) is compatible with additional constraints on the semantics (e.g. specificity). Although we are aware that the inherently vague and widely varying definition of topic is problematic for (H5), we take this hypothesis as the starting point for a formalization of the properties of the DOC in the CD construction, which we introduce in the next section. In other words, we adopt an approach similar to that in Lambrecht (1994): topics cannot be [-generic, -specific]. We leave the details open to future research. Finally, note that none of the above-mentioned approaches captures the fact that the DOC can also be the default pronoun. It is exactly the interaction between this use and its use as a topical object agreement marker that provides interesting evidence for our analysis. We come back to this issue in section VI. Next, we present our analysis of the DOC in the CD construction and in its use as the default pronoun.

IV DF-configurationality and the DOC in clitic doubling

There is good evidence from the extensive literature on the left periphery of the Bulgarian clause that Bulgarian is DF-configurational (see Dimitrova-Vulchanova & Hellan 1998, Rudin 1994, 1990/1991, 1985, Arnaudova 2001, Lambova 2002, Dyer 1992, 1993, Leafgren 1997c, a.o. on Bulgarian; Kiss 1995, 2001 for DF-configurationality). Bulgarian allows hanging topics (cf. Cinque 1977), or EXTERNAL-TOPIC (cf. Aissen 1992, Kiss 1994:80; also King 1995 for Russian), for which we account by the following annotated phrase structure rule:[18]

(26) EP → ({NP, PP, AP, SubjP}) CP
(↑E-TOPIC)=↓ ↑=↓

Also, there is extensive evidence for fronted TOPICs[19] in a position preceding the complementizer (in principal an arbitrary number of TOPICs can be fronted; cf. Rudin 1994, 1990/1991, 1985:24-25). Consider example (27), which is accounted for by the proposed phrase structure rule (28).

(27) Toj kaza Marija če šte ja vidi.
He said Marija that will her see
He said that he will meet Maria.

(28) CP → {NP, PP, AP, SubjP}* C'
↓∈(↑TOPIC) ↑=↓

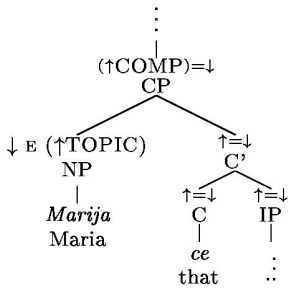
[17] See also Givon (1992:308-309) who claims that contrastive topics can be [+referring, +definite], or [–referring, –definite] but not [+referring, –definite].
[18] We use the abbreviation SubjP to refer to a subjunctive phrase.
[19] Throughout the paper, we use capital letters for DFs, which are part of the f-structure, and non-caps for IS-roles.

Finally, Bulgarian has a FOCUS-position following the TOPIC-position. In subordinate clauses the FOCUS – unlike the fronted topic(s) – follows the complementizer, as in example (29). We thus propose the two phrase structure rules presented in (30) and (31). We apply the annotated phrase structure rules (28), (30), and (31) to the examples in (27) and (29), and present the resulting partial c-structures under (31).

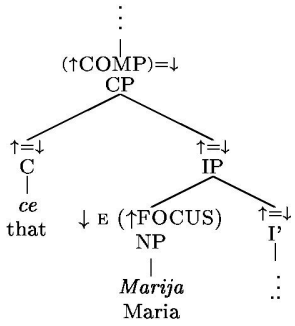
(29) Toj kaza če MARIJA šte vidi.
He said that Marija will see
He said that he will meet MARIJA.

(30) C' → C IP
↑=↓ ↑=↓

(31) IP → {NP, PP, AP, SubjP}* I'
↓∈(↑FOCUS) ↑=↓



c-structure 1



c-structure 2

The preliminary results of an online experiment[20] designed by us suggest that fronted, topical objects (i.e. not the hanging EXTERNAL TOPICs) are *always* doubled. Note that this still allows for *non-topical* fronted objects (i.e. FOCUS objects). Without going into further detail here, we assume that focused fronted objects can be distinguished from topical fronted objects by the different stress assigned to them. Our results are supported by the observations in Dimitrova-Vulchanova & Hellan (1998:xviii), and implicitly Avgustinova (1997:112). In order to capture this fact and Leafgren's (1997a,b, 1998) claim that CD always marks topicality of the doubled object, we propose that the syntactic topic position is assigned the following outside-in functional uncertainty equation. The rule in (32) is the updated rule from (28).[21]

(32) CP → {NP, PP, AP, SubjP}* C'
↓∈(↑TOPIC) ↑=↓
(↑XP* [GF])=↓

The DOC is identified as the direct object by its lexical semantics and the phrase structure rule for the predicate clitic cluster (see (8) above on p. 5). The agreement between the DOC and the doubled object guarantees that no spurious ambiguities are predicted, even in the case of multiple object fronting. Below we give a representative lexical entry for *ja*, the 3.SG.FEM form of the DOC.

[20] Human Subjects Application #0102-655, approved by the Human Subjects Panel, Stanford. The experiment can be found at http://symsys.stanford.edu/experiment/. In this experiment subject where asked to judge Bulgarian sentences after being primed for colloquial spoken language. All judgments were elicited using magnitude estimation, i.e. subjects were asked to assign a gradual value for the "goodness" of each sentence in respect to an always present reference sentence.

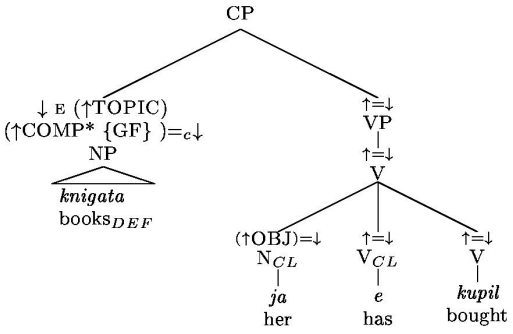
[21] The squared parentheses are a convention used to express that the bracketed part of the equation is not defining (Dalrymple, p.c.). Note that (↑XP* [GF])=↓ ≡ (↑[XP* GF])=↓. A similar rule seems to be necessary for FOCUS-fronting but in that case the whole equation is defining.

```
ja: N_CL - DOC
    (OBJ ↑)
    (↑PERS) = 3
    (↑NUM) = SG
    (↑GEN) = FEM
```

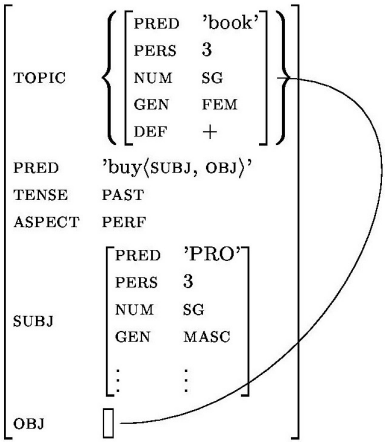
Figure 1 Simplified lexical entry for the DOC *ja* (preliminary version)

The proposal predicts that fronted objects must be doubled since the DOC is the only way to *define* the object function without violating UNIQUENESS (ignoring UNIQUENESS, one could wrongly generate a second object in the VP to satisfy COHERENCE and COMPLETENESS).[22] As an example, consider the sentence in (33) with subject pro-drop and a fronted topical object. The corresponding c- and f-structure are given below.[23] We leave it to the reader to convince herself that the f-structure is the only predicted one in our account.[24]

(33) Knigata ja e kupil.
book$_{PL.DEF}$ DOC$_{3.SG.FEM}$ AUX$_{3.SG}$ bought
The books, he has bought.



c-structure 3



f-structure 1

Crucially, our proposal captures the intuition that it is the absence or presence of a fronted topical object that causes obligatory CD. However, Leafgren argues that the following two generalization hold (the second point is also supported by Vakareliyska 1994:125):

(34) *All* doubled objects are topics.
(35) Object doubling is *always* optional.

In other words, CD is just one option of identifying an object as topical.[25] Unfortunately, Leafgren (1997a,b,c, 1998, 2001, 2002) does not formalize his working definition of 'topic' any

[22] If required we can rule out generation of the DOC in the normal object location by phonological rules like the Tobler-Mussafia effect (see section II).

[23] Due to formatting reasons, we use curly brackets in the tree where we use the standard notation, i.e. square parentheses, in the phrase structure (32) rule above.

[24] Note that the subject function is defined through the verbal subject agreement morphology (including an optional PRED PRO since subject drop is common in Bulgarian), so that subjects, too, can be in the fronted position.

[25] Leafgren (2001:4) shows that in 1200 object occurrences, 0% of the non-topical objects are doubled. This contrasts with 10.8% doubled topical objects in spoken Bulgarian. Alternative means of topical object marking depend on the register. In informal, spoken Bulgarian, speakers may also use marked word order (i.e. object-topic fronting) or intonation or just not mark the topicality of the object when the context unambiguously identifies the object to be the topic (cf. Leafgren 1997b:128). In more formal registers, passivization or impersonal reflexive constructions can be used to mark that the semantic object is topical (cf. Leafgren 2001).

more precisely than 'What a clause provides or requests information about' (cf. Leafgren 1997a:127, referring to Sgall 1975:303; see also Sgall 1993). Leafgren gives the following example to illustrate his topic definition:[26]

(36) Vanja$_i$ ne ja$_i$ vâlnuvat tezi nešta ...
Vanja NEG DOC$_{3.SG.FEM}$ worry$_{3.PL}$ these things
These things don't worry Vanja ...

The generalization in (35) conflicts with Dimitrova-Vulchanova & Hellan's (1998) and our own observations. Leafgren's work is based on a corpus study of more then 7,000 object occurrences in written texts (1997a,b; including ~200 cases of CD), more than 3,000 object occurrences in spoken texts (1998: including ~200 cases of CD), and a comparative study of 1,200 object occurrences each in informal oral, formal oral, and formal written texts (Leafgren 2001). In light of such extensive evidence, we should try to resolve the mismatch between Leafgren's and our observations. There are two main sources for this mismatch aside from the apparent problem with informal topic definitions. First, although Leafgren (2001) considers informal oral texts, (35) is based on Leafgren's (1997a,b) work on written corpus (consisting of 2 novels and 2 short stories). The online experiment done by us (cf. above) aims at judgments about informal contemporary spoken Bulgarian. Secondly, and more importantly, Leafgren does not control for fronted *focused* phrases. Actually, Leafgren (1997a:132) explicitly allows for topics to be "focused" (in his terminology). Although this admittedly has to be done at some point, it is not the purpose of this paper to determine the exact semantic and/or pragmatic function of what we have called 'topic' so far (for Bulgarian). Here the crucial point is that Bulgarian seems to have two sentence initial positions, here labeled TOPIC and FOCUS (see above) that can be distinguished in terms of the stress contours that go along with them. Thus we have an independent motivation for those two positions[27], which we label TOPIC and FOCUS. One of those two positions, namely TOPIC, requires CD if it is filled by an object. Thus the distinction of TOPIC and FOCUS allows us to capture a generalization, which Leafgren misses, without additional stipulation. For simplicity's sake, we will assume that the TOPIC and FOCUS position each encode at least their corresponding I[nformation] S[tructural] roles, namely topic and focus (again, here we are not concerned with the meaning of the two IS-roles). Somewhat more formally, this constraint can be stated as in (37), where DF is the set of f-structure features that encode discourse functions, and for a given input DF the function IS-role(DF) yields the corresponding IS-role (e.g. topic for TOPIC).

(37) $X \in \text{DF} \Rightarrow x \in$ IS-role(DF), where X is the f-structure correspondence of a linguistic form w, and x is the denotation of w.

Similarly to EXTENDED COHERENCE (cf. Bresnan 2000), we can formulate a constraint INFORMATION PACKAGING COHERENCE that guarantees that the generalization in (37) holds for all DFs of an f-structure.

INFORMATION PACKAGING COHERENCE (IPC) - preliminary version
(38) An F-structure FS fulfills IPC iff every discourse function DF in FS fulfills (37).

Similarly to other authors (e.g. Choi 1999), we assume an IS-component which has interfaces not only to f-structure but also to the Prosodic Structure (PS) and the Lexical Structure (LS), see Figure 2. Here we are not interested in the interface between PS and IS but in the interface between

[26] Note one important detail in Leafgren's definition. The topic is defined on the level of a clause, not a sentence. This allows for topics in, for example, subordinate clauses. Examples like (27) above clearly show that this is necessary.
[27] More precisely, we have a motivation for a formal distinction, which we choose to capture in terms of c-structure position.

LS and IS.[28] Note that the one-way implication of (37) works to our advantage. While we want every phrase that is fronted to TOPIC to be part of the IS-topic, we want to allow for non fronted constituents to bear the role of the IS-topic, too.

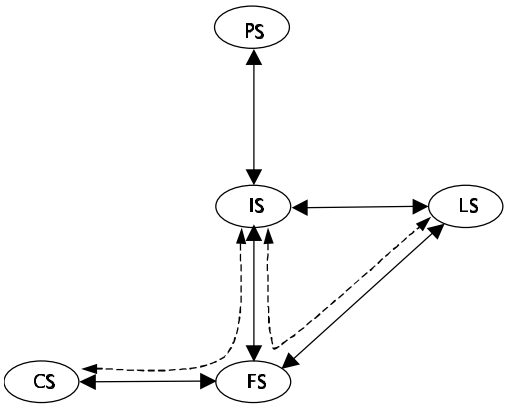


Figure 2 Relations between parts of the grammar that are relevant for IS.

Now with the proposed model of IS, LS, and FS interaction in mind, we can restate Leafgren's generalizations in (34) and (35) more precisely as (39) - (41).

(39) *All* doubled objects are *IS*-topics.
(40) TOPIC objects *must* be doubled.
(41) IS-topic objects *can* be doubled.[29]

In order to predict that CD implies that the doubled object is part of the IS-topic, i.e. to guarantee (39), we have to slightly modify the lexical entry of the DOC(s). Again, the 3.SG.FEM form *ja* is given as a representative example in Figure 3. The upwards-pointing arrow with the subscript 'IS' indicates that the referent identified by the DOC is mapped onto information structure (where it is identified as a part of the IS-topic).

ja: N_{CL} - DOC
($\uparrow_{IS} \in topic_{IS}$)
(OBJ↑)
(↑PERS) = 3
(↑NUM) = SG
(↑GEN) = FEM

Figure 3 Revised lexical entry for the topic-marking DOC *ja*.

The lexical entry in Figure 3 together with the revised annotated phrase structure rule for the TOPIC position in (32) captures all of the above-mentioned generalizations, (39) - (41), and therefore resolves the apparent conflict between Leafgren's work and e.g. Dimitrova-Vulchanova & Hellan's claims. Moreover, our account predicts optional CD for fronted FOCUS-objects as long as they are part of the IS-topic. If we adopt a two-dimensional IS-component[30], following Choi (1999), implicitly Leafgren (1997a,b), a.o., this is not surprising at all. Indeed, reduplication of fronted FOCUS objects can be observed in Bulgarian. First, CD of fronted object wh-phrases (cf. Jaeger

[28] In a model like the one presented here, encoding of IS through CS (and therefore within LFG through F-structure, FS) corresponds to what is commonly called discourse configurationality (henceforth, DF-configurationality).
[29] Here we do not address the pragmatic factors which determine in which contexts speakers tend to make use of this mechanism (CD to mark IS-topicality of the object). See Givon (1987) for a general discussion of this.
[30] By two dimensional, we mean that there is not only one dimension along which information structural roles differ e.g. topic-comment or link-tail-focus (cf. Vallduví 1993, 1992). Instead informational structural roles differ along two dimensions, e.g. they can be [+/- prominent] and [+/- given] (cf. Choi 1999).

2002; see also Dimitrova-Vulchanova & Hellan 1998:xxi-xxii), which are usually considered to be in FOCUS (see, for example, Rudin et al. 1998), is possible.

(42) Kogo kakvo go iznenada?
whom what DOC$_{3.SG.MASC}$ suprised$_{3.SG}$
Whom did what surprise?

Second, of all non-wh, focused constituents, only contrastive topics can be doubled. The proposal presented here therefore accounts, among other things, for the fact that contrastive topics can be doubled, a fact that Avgustinova's (1997) analysis of CD cannot straightforwardly account for since she employs the one dimensional IS-component proposed in Vallduví (1992, 1993).

As mentioned in the introduction, one aim of this paper is to provide a formal account for CD and its interaction with word order and IS. The current section has done exactly this. Second, we wanted to resolve the discrepancy between the different empirical approaches to Bulgarian CD and the theoretical literature. For one part, we have already done this by resolving the mismatch between our own empirical studies, Leafgren's work and the theoretical literature on CD and DF-configurationality in Bulgarian. We did this by distinguishing between two independently motivated phrase structural positions and their correspondences in the IS-component. The analysis resulting from this is able to capture both the generalization from the extensive empirical work and predicts the right restrictions resulting from certain word orders (i.e. obligatory CD of TOPIC objects). Next, we use the second source of data for spoken Bulgarian mentioned above, Avgustinova's (1997) elicited question-answer pairs, to briefly test if the presented proposal makes correct predictions about possible word orders beyond the fronted TOPIC construction.

It is beyond the scope of this paper to provide a detailed analysis for all of the patterns (i.e. word order-intonation-information structure mappings) described by Avgustinova (1997:112). Although this issue is open for further research, we suggest that Bulgarian has some kind of 'default ordering' within the flat VP (see above, phrase structure rule (11) in section II). Among other features, such as definiteness, person, referentiality, etc., topicality of a phrase seems to be one – maybe the major – determining factor for the constituent order with the VP.[31] Leafgren (1997c:5ff.) shows that topic-before-comment seems to be the more important ordering mechanism in Bulgarian than subject-before-object or agent-before-patient, both in terms of frequency[32] and in that all violations of the two other conditions serve to satisfy the topic-before-comment condition or another discourse or information structure constraint (e.g. CD and object fronting). The assumption of a default order similar to the one suggested by the Prague school (cf. Functional Sentence Perspective, henceforth FSP; Sgall 1993) but only applied to the flat VP instead of the whole clause explains why a certain default constituent order can be observed in Bulgarian while, at the same time, only a few strict rules (like the above-mentioned TOPIC object fronting) seem to hold. We ask the reader to keep in mind the notion of default ordering as just described during our discussion of Avgustinova's (1997) data.

Apart from direct object fronting, which results in OSV and OVS orders (for the sake of simplicity, we only consider transitive verbs here), there is one other word order that usually requires the DOC, namely VOS. According to Avgustinova (1997), VOS is possible with either $V_{FOCUS}O_{topic}S_{topic}$ or $VO_{topic}S_{focus}$.[33] Here, $VO_{topic}S_{focus}$ is predicted by FSP default word ordering working on a non-configurational VP. The same reasoning applies to $V_{FOCUS}O_{topic}S_{topic}$. Given this, we should expect $V_{FOCUS}S_{topic}O_{topic}$ to be equally acceptable, if the object and the subject are equally

[31] Note that this is not uncommon at all. It has long been known that scrambling in languages like e.g. German or Japanese is sensitive to the above-mentioned categories. Furthermore, especially topicality of phrases has been shown to play a role in determining the word order in several languages (cf. Choi 1999 for German and Korean; Ishihara 2000 for Japanese).

[32] Topic-before-comment ordering holds in 91.0%, subject-before-object in 89.5%, and agent-before-patient in 88.3% of the cases. The correlation between the three scales explains why the numbers are so close.

[33] Recall our convention to use capital letters for DFs (as part of the f-structure) and lowercase letters for IS-roles. Since Avgustinova (1997) does not make a comparable distinction, the annotation is our translation of her classification.

topical. Indeed, Avgustinova's data set contains examples for this word order. For both cases of a FOCUSed verb, topical subject and topical object, the clitic marks which of the NPs is the object. Furthermore, for both SOV and VSO the clitic is at least possible (if not preferred) *if and only if* the object is part of the topic.

Thus, in addition to what we said above, the proposal presented here accounts for the experimentally elicited word orders listed in Avgustinova (1997). Although a detailed syntactic analysis of all possible word orders has to be left to further research, we have sketched an analysis of the DOC in CD and its interaction with word order and information structure. We will refer to this use as '(direct object) topic agreement marker' usage. This label makes reference to Bresnan & Mchombo (1987) who distinguish grammatical and anaphoric agreement markers. We now turn to a second function of the DOC, its use as 'default' pronoun, and then show that our proposal makes the right predictions about the occurrences of those two different functions of the DOC.

V The DOC as default pronoun

Although this is not a salient topic in the literature on the Bulgarian object clitics (for an exception see Vakareliyska 1994:125), there is no doubt that the DOC has another use as the default pronoun. To further clarify what we mean by *default* and to illustrate the relation between the two types of pronouns, consider the following dialogue, where (44) but not (45) is a possible continuation of (43) if no contrast is intended:

(43) "Karl sreštna onazi tancjorka včera"
Karl met$_{3.SG}$ that dancer yesterday
Karl met that dancer yesterday.

(44) "Ivan sâšto ja (*neja/*NEJA) poznava."
Ivan too DOC$_{3.SG.FEM}$ her/HER knows$_{3.SG}$
Ivan knows her, too.

(45) #"Ivan sâšto poznava neja/NEJA."
Ivan too knows$_{3.SG}$ her/ HER
Intended: Ivan knows her too.

Since this has not been done by others, we tested for the possibility that all cases of the DOC as alleged default pronoun might be due to (topic) object drop. For some more details on the test, we refer the reader to our handout (Jaeger & Gerassimova 2002:10). Here we will just mention that, like English, Bulgarian allows specific and unspecific object drop (depending on the verb, cf. Fillmore 1986). We found that there are still cases left where object pro-drop is not possible and the DOC is the only realization of the object in the sentence. Thus we are forced to assume that there is one variant of the DOC with a PRED PRO. For a formal LFG analysis, this raises the question whether there are two entries for each DOC or one with an optional PRED PRO. Consider the hypothesis that there is one DOC with an optional PRED PRO. In that case, the default pronoun use of the DOC would always result in the object (i.e. the clitic itself) being marked as topic. It is not clear whether this is desirable, although one could argue that all pronouns have to be topical in some sense anyway, since their referent is 'salient' (cf. Chafe 1976) most of the time (in order to be identifiable). For now, it may be better to think of two separate lexical entries for the DOC, one with an optional PRED PRO (the default pronoun) and one with the topic equation. Again, this is illustrated for *ja*.

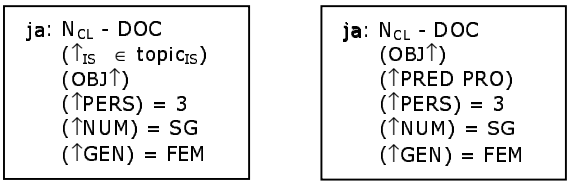


Figure 4 Revised lexical entries for the DOC *ja*.

The existence of two lexical entries poses the question of how our proposal can guarantee the right use of DOC for a given sentence. So far, because of the functional control established by the TOPIC position, the optional PRED PRO use is ruled out by UNIQUENESS whenever a fronted (object) constituent sits in a TOPIC position. Whenever the object is realized within the VP, UNIQUENESS again rules out two PRED values for the object, since the DOC *defines* the object (instead of just constraining it). With no other object constituent being realized, the DOC is interpreted as object (pronoun). In our account, this is guaranteed by (EXTENDED) COHERENCE. Next, we show that data from topicalization out of islands further support this analysis.

VI Island data: When can which type of DOC occur?

In this section, we show how our proposal makes the right predictions about the distribution of the two uses of the DOC (i.e. as default pronoun and as topic agreement marker). Rudin (1985) shows that NPs and PPs (whether complex or not) are islands to any kind of extraction in Bulgarian. Most of the other classical islands, however, do not seem to be islands in Bulgarian. This is supported by the preliminary results of our still ongoing online experiment (see above). Consider, for example, the following cases of topicalization:

(46) Todor e jasno, [$_{CP}$ če Ivan *(go) e vidjal].
Todor is clear that Ivan DOC$_{3.SG.MASC}$ is seen
Todor it is clear that Ivan has seen him.

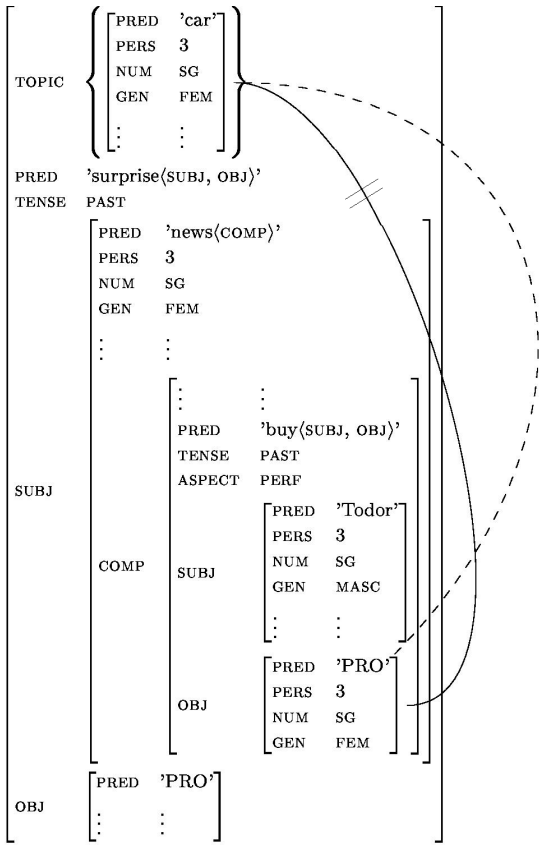
(47) Jabălkite [$_{NP}$ čovekyt, [$_{CP}$ kojto *(gi) donese]], e pilot.
apples$_{DEF}$ man$_{DEF}$ who DOC$_{3.SG.NEUT}$ brought is pilot
The apples the man who brought (them) is a pilot.

Sentences (46) and (47) show that topicalization out of a sentential subject, in (46), and a relative clause, in (47), is possible. Just as in the case of simple fronting, the DOC is obligatory. Now consider topicalization out of an island (here, an NP):

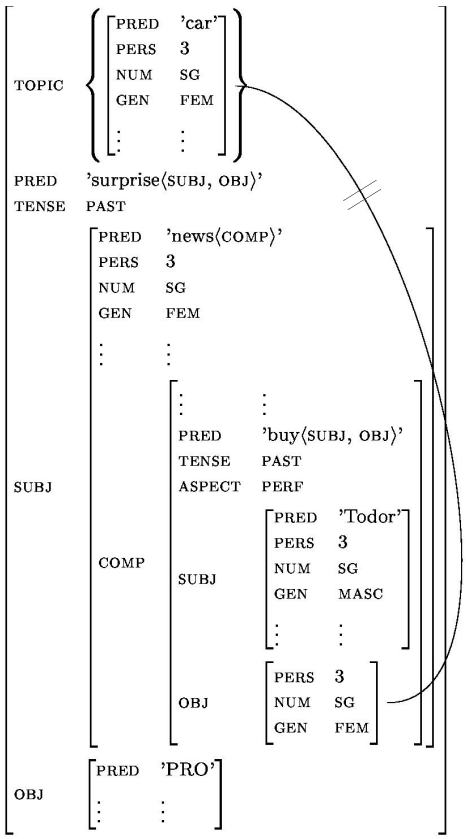
(48) *Kolata [$_{NP}$ novinata, [$_{CP}$ če Todor (ja) e kupil]], ni učudi.
car$_{DEF}$ news$_{DEF}$ that T. DOC$_{3.SG.NEUT}$ is bought us surprised
Intended: The car the news that Todor has bought (it) surprised us.

Regardless of whether the DOC is realized, sentence (48) is ungrammatical. According to Bresnan & Grimshaw (1978), filler-gap dependencies (i.e. functional control within LFG), but not anaphoric binding, obey island constraints. The DOC does not repair island-violations. In our account, the ungrammaticality of (48) is explained as follows. The fronted constituent can only satisfy the outside-in functional uncertainty equation (and thereby EXTENDED COHERENCE) if it is functionally controlled by a GF-bearing constituent further down in the f-structure. The fronted object cannot be functionally controlled by a constituent with a PRED value because this would violate UNIQUENESS. The DOC cannot be realized in the embedding sentence to bind the fronted object since the object function of the embedding sentence already has an object (with a PRED value). Finally, the DOC with a PRED PRO (i.e. the default pronoun) could be realized in the embedded clause in order to satisfy COMPLETENESS and COHERENCE. However, the outside-in functional uncertainty equation of the fronted object would still have to be resolved. This is not possible since the embedded GF (i.e. the direct object) is not accessible – it is in an island (cf. f-structure 2)[34]. F-structure 3 is out for the same reason – because the functional control violates the island condition (cf. above, Bresnan & Grimshaw 1978).

[34] Anaphoric binding is indicated by dotted lines, functional control by solid lines. The doubled crossed line stands for an island violation, which results in an invalid f-structure.



f-structure 2



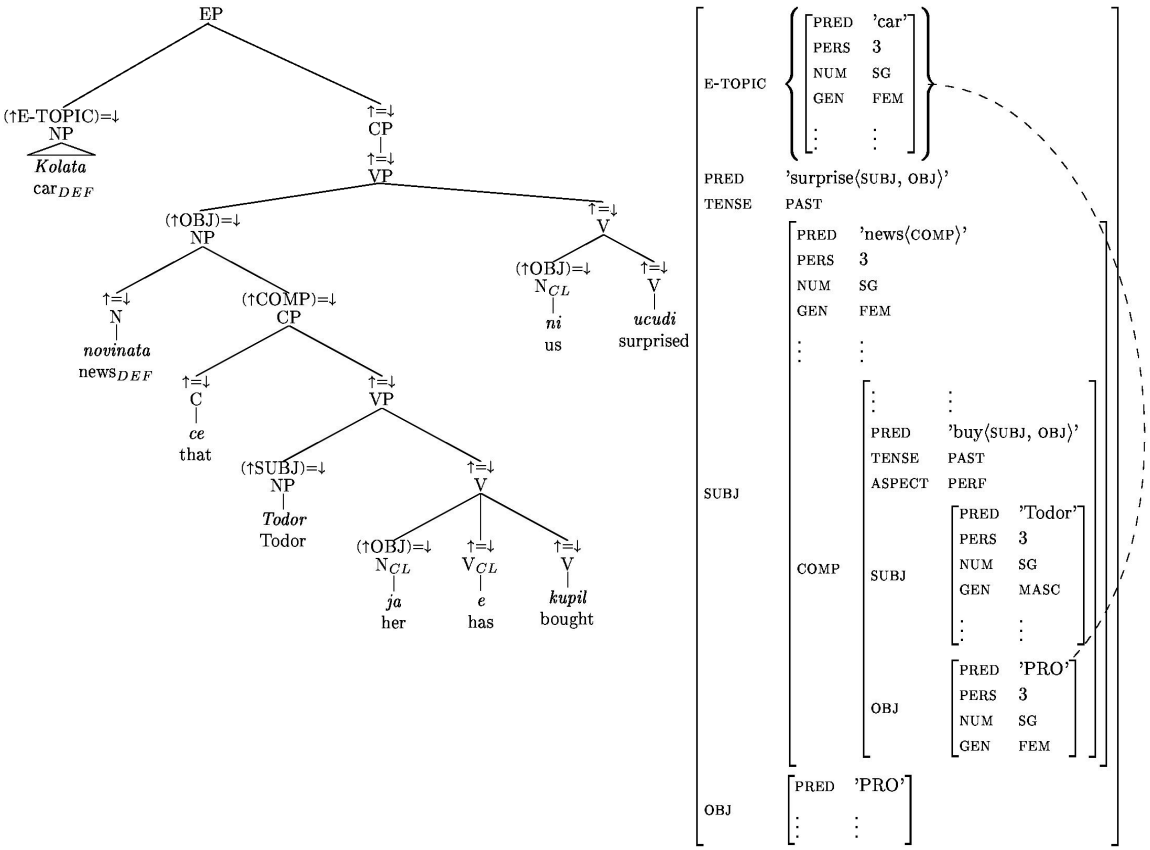
f-structure 3

The grammaticality of (46) and (47) is predicted, too. The fronted object is functionally controlled by the DOC, which has to be realized because it is the only constituent that agrees in person, number, and gender with the fronted object. The DOC with a PRED PRO cannot be chosen because this would violate UNIQUENESS.

There is one more phenomenon that supports our analysis: EXTERNAL TOPICs. Example (49) – if uttered with a clear pause between the fronted object and the following sentence – is grammatical. This kind of a detached constituent fulfills the criterion of a hanging topic (cf. Cinque 1977) or EXTERNAL TOPIC.

(49) Kolata PAUSE novinata, če Todor ja e kupil, ni učudi.
car$_{DEF}$ PAUSE news$_{DEF}$ that Todor DOC$_{3.SG.NEUT}$ is bought us surprised
The car, .. the news that Todor has bought (it) surprised us.

In the account presented here, the grammaticality of (49) follows from the fact that the phrase structure rule for EXTERNAL TOPICs, see (26) above, is not annotated with an outside-in functional uncertainty equation. Therefore, no functional control violates the island constraint and the DOC with a PRED PRO is realized in the embedded clause. To satisfy EXTENDED COHERENCE, the PRED PRO anaphorically binds the EXTERNAL TOPIC. This is illustrated by the c- and f-structure given below.



c-structure 4 **f-structure 4**

To sum up, the island data presented above is not only compatible with our theory but also predicted by it. In the next and final section, we summarize our analysis and list some open questions.

VII Conclusions and Outlook

We have shown how two functions/uses of the DOC interact. The DOC is a grammatical (direct object) agreement marker and the default pronoun of Bulgarian. In contrast to the object marker in Chicheŵa (cf. Bresnan & Mchombo 1987: 745), the Bulgarian DOC does not mark an f-structure TOPIC but an IS-topic. This insight helps to position the DOC within a typology of (object) markers. The DOC's object topic-marking function, in interaction with the proposed annotated phrase structure rules (i.e. especially the functional control of fronted topic), accounts for both obligatory TOPICalized object doubling and optional doubling of topical objects in general.

Our account stresses that linguistic forms can have several (independent) functions. This is even more evident when we consider that the DOC has a third function as intrusive pronoun, as mentioned in the introduction. First results of an ongoing online experiment on the intrusive pronoun DOC in extractions support our analysis. Those results will have to be fully incorporated into a complete account of the DOC.

The optionality of CD in many contexts shows that speakers have different options of coding e.g. a topical object depending on the register, genre and maybe other factors (see Leafgren 1997a, 2001, 2002 for a similar thought). Possible generalizations relating the choice of forms to their functions and other factors, such as register, merit further investigation. For example, does the absence of intonation in written language enforce the use of alternative linguistic means (such as more strict case marking in the case of otherwise optional case marking, or more strict word order) to identify GFs and DFs.

Also, we have provided one further example of a (non-dependent-marking) language which seems to compensate lack of GF-configurationality by morpho-syntactic means (head-marking). Although subject to further testing, the presented analysis is supported by a broad empirical basis. In addition to the native speaker intuitions of one of the authors (V.G.), the analysis accounts for data from Leafgren's (1997a, 1997b) corpus-based studies, Avgustinova's (1997) elicited question-answer data (more than 20 word order-prosody mappings for a transitive verb), and the data collected in our online experiment. To the best of our knowledge, unlike all other formal accounts so far (e.g. Rudin 1997, 1996, 1990/1991, 1985, Dyer 1992, Avgustinova 1997, Dimitrova-Vulchanova & Hellan 1998, Franks & King 2000), the account presented above predicts the obligatory CD in the case of fronted objects and provides a possible explanation for the optionality of CD in other cases. For example, Rudin's (1997) MP analysis of the DOC as a AgrO-head cannot predict why the DOC is obligatory in certain cases yet optional in others. Furthermore, we explicitly addressed the relatively free word order of Bulgarian and predicted the resulting word order depending on the IS-roles assigned to the different phrases. Although empirically attested, many of the word orders discussed at the end of section IV are ignored in most of the theoretical literature on Bulgarian.

While our analysis accounts for all observed word orders (including predictions about prosody via proposed constraints on the IS, e.g. via IPC, cf. (38) in section IV), it does not predict spurious parses or ambiguities arising from the lexical ambiguity of those two uses. The account presented here could therefore close the gap between the work on DF-configurationality and free word order in Bulgarian. It also is a first step to resolve the mismatch between the broad-coverage empirical work on Bulgarian and the literature on formal aspects.

Further research is necessary in order to see how the different functions of the DOC relate to each other. We also think that it is worth to investigate if there are further restrictions on the optional or obligatory presence of the DOC in certain contexts. For example, there are still possible mismatches in the observations made by Avgustinova (1997) and Leafgren (1997a,b) regarding the question in exactly which contexts the DOC is obligatory. Once we have a better picture of all the factors that determine the possible word orders for a given context, a (stochastic) OT account may be able to combine those factors into a formal description of the data. Related to this, it is very interesting that those dimensions which are strict factors in Bulgarian CD (i.e. specificity and topicality), seem to have occurred subsequently in the diachronic development of the much more general Macedonian CD and show up as statistic preferences in contemporary Macedonian CD (as a careful reading of Čašule 1997 suggests). Finally, another phenomenon that needs further research is the CD of quantified NPs. While we have shown how quantified NPs confirm that [-specifics] cannot be doubled (cf. section III), the details of CD of quantified NPs are yet to be worked out.

VIII References

AG, ACADEMY GRAMMAR (edited by Popov, Konstantin)

- (1983): *Grammatika na sâvremennija bâlgarski knižoven ezik, t. 3. sintaksis*; Sofia, Bâlgarska Akademija na Naukite.

AISSEN, JUDITH

- (1992): *Topic and Focus in Mayan*; in: *Language 68,* p. 43-80.

ALEXANDROVA, G.

- (1997): *Pronominal Clitics as g(eneralized) f(amiliarity)-licensing AGR^0*; in Browne, W. / Dornisch, E. / Kondrashova, N. / Zec, D. (eds.): *Formal Approaches to Slavic Linguistics: The Cornell Meeting 1995*, pp. 1-31; Ann Arbor, Michigan Slavic Publications.

ARNAUDOVA, OLGA

- (2001): *Prosodic movement and information focus in Bulgarian*; in: *Proceedings of Formal Approaches to Slavic Linguistics 9*; Michigan Slavic Publications.

AVGUSTINOVA, TANIA

- (1997): *Word Order and Clitics in Bulgarian*. in: *Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 5*. DFKI, Saarbrücken.

AVGUSTINOVA, TANIA & BISTRA ANDREEVA
- (1999): *Thematic intonation patterns in Bulgarian clitic replication*; in: *ICPhS99, San Francisco*, pp. 1501-4.

BAKER, MARK C.
- (1991): *On some subject/object asymmetries in Mohawk*; in: *Natural Language and Linguistic Theory 9*, pp. 537-576.

BILLINGS, LOREN A.
- (2000): *Phrasal Clitics*; in: *Slavic Linguistics 9(2), special issue: Festschrift for Leonard Babby*.

BRESNAN, JOAN
- (2001): *Lexical Functional Syntax*; Blackwell Publishing, Cambridge.

BRESNAN, JOAN / GRIMSHAW, JANE
- (1978): *The Syntax of Free Relatives in English*; in: *Linguistic-Inquiry* 9(3), pp. 331-391.

BRESNAN, JOAN / MCHOMBO, SAM A.
- (1987): *Topic, Pronoun, and Agreement in Chichewa*; in: *Language, Volume 63, Number 4*, pp. 741-782.

ČAŠULE, ILIJA
- (1997): *Functional Load of Short Pronominal Forms*, in: *Journal of Slavic Linguistics 5(1)*. 3-19.

CHAFE, WALLACE, L.
- (1976): *Givenness, Contrastiveness, Definiteness, subjects, Topics, and Point of View*; in Li, Ch. N. (ed.): *Subject and Topic*, pp. 27-55; New York, Academic Press.

CHOI, HYE-WON
- (1999): *Optimizing Structure in Context. Scrambling and Information Structure*; in: *Dissertations in Linguistics*; Stanford, CSLI Publications.

CINQUE, GUGLIELMO
- (1977): *The Movement Nature of Left Dislocation*; Linguistic Inquiry, 8, 2, pp. 397-412.

CYXUN, GENADIJ A.
- (1962): *Mestoimennata enklitika i slovoredât v bâlgarskoto izrečenie*; in: *Bâlgarski ezik 12(4)*, pp. 283-91.

DIMITROVA-VULCHANOVA, MILA / HELLAN, LARS
- (1998): *Introduction*; in Dimitrova-Vulchanova & Hellan (eds.): *Topics in South Slavic Syntax and Semantics*, pp. ix-xxvii; Amsterdam & Philadelphia, John Benjamins Publishing.

DYER, DONALD L.
- (1993): *Determinedness and the Pragmatics of Bulgarian Sentence Structure*; in: *Slavic and East European Journal, Volume 37, Number 3*, pp. 273-292.
- (1992): *Word Order in the simple Bulgarian Sentence: A Study in Grammar, Semantics and Pragmatics*; Amsterdam – Atlanta, Rodovi B. V.

EMBICK, DAVID / IZVORSKI, ROUMYANA
- (1994): *On long head movement in Bulgarian*; in: *Proceedings of the Eleventh Eastern States Conference on Linguistics 1994*, pp. 104-115; Ithaca , Cornell University.

FILLMORE, CHARLES J.
- (1986): *Pragmatically Controlled Zero Anaphora*; in: *BLS 12*, pp. 95-107.

FRANKS, STEVEN / KING, TRACY HOLLOWAY
- (2000): *A Handbook of Slavic Clitics*; Oxford, Oxford University Press.

FRASER, BRUCE
- (1988): *Types of English discourse markers*; in: *Acta Linguistica Hungarica 38*, pp. 19-33.

GEORGIEVA, ELENA
- (1974): *Slovored na prostoto izrečenie v bălgarskija knižoven ezik*; Sofia, Bălgarska Akademija na Naukite.

GERASSIMOVA, VERONICA A. / JAEGER, T. FLORIAN
- (2002): *Configurationality and the Direct Object Clitic in Bulgarian*; in Nissim, M. (ed.): *Proceedings of the Seventh Student Session of the ESSLLI 2002 in Trento, Italy, August 5th – 16th, 2002*, Chapter 6.

GIVÓN, TALMY
- (1992): *On Interpreting Text-Distributional Correlations. Some Methodological Issues*; in: Payne, D. L. (ed.): *Pragmatics of Word Order Flexibility*, pp. 305-320; Amsterdam & Philadelphia: John Benjamins Publishing.
- (1987): *The Pragmatics of Word-Order: Predictability, Importance and Attention*; in Hammod, M. et al (eds.): *Studies in Syntactic Typology* (=*Typological Studies in Language 17*), pp. 243-284; Amsterdam: J. Benjamins Publishers.
- (1976): *Topic, Pronoun, and Grammatical Agreement*; in Li, Ch. N. (ed.): *Subject and Topic*, pp. 149-188; New York & London, Academic Press Inc.

GUENTCHEVA, ZLATKA
- (1994): *Thématisation de l'objet en bulgare*; Bern: Peter Lang.

ISHIHARA, SHINICHIRO
- (2000): *Stress, Focus, and Scrambling in Japanese*; in: *MIT Working Papers In Linguistics 39*, pp. 142-175.

IVANČEV, SVETOMIR
- (1978): *Prinosi v bălgarskoto i slavjanskoto ezikoznanie*; Sofia, Nauka i izkustvo.
- (1968): *Problemi na aktualnoto členenie na izrečenieto*; in Ivančev (1978:173-84).
- (1957): *Nabljudenija vârxu upotrebata na člena v bâlgarski ezik*; in Ivančev (1978:128-52).

IZVORSKI, ROUMYANA
- (1994): *On the Semantics of the Bulgarian "Indefinite Article"*; in: *Formal Approaches to Slavic Linguistics: The MIT Meeting 1993*, pp. 235-254; Ann Arbor: Michigan Slavic Publications.

JAEGER, T. FLORIAN
- (2002): *Multiple Wh-questions in Bulgarian*; draft, available at http://www.stanford.edu/~tiflo/ as by 10/2002.

JAEGER, T. FLORIAN & VERONICA A. GERASSIMOVA
- (2002): *Bulgarian word order and the role of the direct object clitic in LFG*; handout for the *LFG02, Athens, July 3rd-5th, 2002*, available at http://www.stanford.edu/~tiflo/ as by 10/2002.

KAZAZIS, KOSTAS & JOSEPH PENTHERADOUKIS
- (1976): *Reduplication of indefinite direct objects in Albanian and Modern Greek*; in: *Language 52(2)*, pp. 398-403.

KING, TRACEY H.
- (1995): *Configuring Topic and Focus in Russian*. in: *Dissertations in Linguistics*; Stanford: CSLI Publications.

KISS, KATALIN E.
- (2001): *Discourse configurationality*; in Haspelmath, M. / Koenig, E. / Oesterreicher, W. / Raibler, W. (eds.): *Language Typology and Language Universals*, pp. 1442-1455; Berlin & New York, de Gruyter.
- (1995): *Introduction* in Kiss, K. E. (ed.): *Discourse Configurational Languages*, pp. 3-27; Oxford, Oxford University Press.
- (1994): *Sentence Structure and Word Order in Kiefer*; in F. / Kiss, K. E. (eds.): *Syntax and Semantics 27: The Syntactic Structure of Hungarian*, pp. 1-84; London, Academic Press.

LAMBOVA, MARIANA
- (2002): *Multiple topicalization wh-fronting in Bulgarian and the fine structure of the left-periphery*; in: *The 26th Penn Linguistics Colloquium.*

LAMBRECHT, KNUD
- (1994): *Information structure and sentence form. Topic, focus, and the mental representations of discourse referents*; Cambridge: Cambridge University Press.

LEAFGREN, JOHN R.
- (2002): *Register, Mode, and Bulgarian Object Placement*; presented at the Balkan Conference 2002, source: leafgren@email.arizona.edu.
- (2001): *Patient Packaging in Informal and Formal, Oral and Written Bulgarian*; presented at the AATSEEL 2001, New Orleans, source: leafgren@email.arizona.edu.
- (1998): *Object Reduplication in Spoken Bulgarian*; presented at the AATSEEL 1998, San Francisco, source: leafgren@email.arizona.edu.
- (1997c): *Topical Objects, Word Order, and Discourse Structure in Bulgarian*; presented at the AAASS 1997, source: leafgren@email.arizona.edu.
- (1997b): *Definiteness, Givenness, Topicality and Bulgarian Object Reduplication*; in: *Balkanistica 10*, pp. 296-311.
- (1997a): *Bulgarian Clitic Doubling: Overt Topicality*; in: *Journal of Slavic Linguistics, Volume 5, Number 1*, pp. 117-143.

MINČEVA, ANGELINA
- (1969): *Opit za interpretacija na modela na udvoenite dopâlnenija v bâlgarskija ezik*; in: Andrejčin, L. et at (eds.): *Izvestija na instituta za bâlgarski ezik 17*, pp. 3-50; Sofia: Bâlgarskata Akademija na naukite.

NORDLINGER, RACHEL
- (1998): *Constructive Case: Evidence from Australian languages*; in: *Dissertations in Linguistics*; Stanford, CSLI Publications.

PENČEV, JORDAN
- (1993): *Bŭlgarski Sintaksis: Upravlenie I Svŭrzvane*; Plodiv, Plovdivsko Universitetsko Izdatelstvo.
- (1984): *Stroež na bâlgarskoto izrečenie*; Sofia, Nauka i izkustvo.

POPOV, KONSTANTIN P.
- (1973): *Po njakoi osnovni vâprosi na bâlgarskija knižoven ezik*; Sofia: Narodna prosveta.
- (1963): *Sâvremenen bâlgarski ezik: Sintaksis*: Sofia: Nauka i izkustvo, 2nd edition.

POPOV, KONSTANTIN P. & VENČE S. POPOVA
- (1975): *Vâprosi na azikovata stilistika*; Sofia: Narodna prosveta.

RUDIN, CATHERINE
- (1997b): *Kakvo li e* li: *Interrogation and Focusing in Bulgarian*; in: *Balcanistica 10*, pp. 335-346.
- (1997a): *AgrO and Bulgarian Pronominal Clitics*; in: *Annual workshop on formal approaches to Slavic linguistics. The Indiana meeting 1996*, pp. 224-252; Michigan, Michigan Slavic Publications.
- (1996): *On Pronominal Clitics*; in Dimitrova-Vulchanova, M. / Hellan, L. (eds.): *Papers from the First Conference on Formal Approaches to South Slavic Languages, Volume 28*, pp. 229-246; University of Trondheim Working Papers in Linguistics.
- (1994): *On focus position and focus marking in Bulgarian questions*; in: *Papers from the Fourth annual meeting of the Formal Linguistic Society of Midamerica, April 15-18, 1993*, pp. 252-266, Iowa City.
- (1990/1991): *Topic and Focus in Bulgarian*; in: *Acta Linguistica Hungarica, Vol. 40 (3-4)*, pp. 429-449.
- (1989): *Multiple questions south, west and east: A Government-Binding approach to the typology of wh-movement in Slavic languages*; in: *International Journal of Slavic Linguistics and Poetics 39-40*.
- (1988b): *On multiple questions and multiple WH fronting*; in: *Natural Language and Linguistic Theory 6*, pp. 445-502.
- (1988a): *Multiple Question in South Slavic, West Slavic and Romanian*; in: *Slavic and East European Journal, Volume 32, Number 1*, pp. 1-24.
- (1985): *Aspects of Bulgarian Syntax: Complementizers and WH Constructions*; Columbus, Ohio: Slavica Publishers.

SCHIFFRIN, DEBORAH
- (1988): *Discourse markers*; Cambridge & New York: Cambridge University Press.

SELLS, PETER
- (1984): *Syntax and semantics of resumptive pronouns*; Ph.D. Thesis, University of Massachusetts at Amherst.

SIEWIERSKA, ANNA / UHLIŘOVÁ, LUDMILA
- (1998): *Word order in Slavic languages*; in Siewierska, A. (ed.): *Constituent Order in the Languages of Europe*, pp. 105-150; Berlin/New York, Mouton de Gruyter.

SGALL, PETR
- (1993): *The Czech Tradition*; in Jacobs, J. / von Stechow, A. / Sternefled, W. / Vennemann, T. (eds.): *Sytnax. Ein internationales Handbuch zeitgenoessisches Forschung – An international Handbook of Contemporary Research*, pp. 349-368; Berlin, New York, de Gruyter.
- (1975): *Conditions of the use of sentence and a semantic representation of topic and focus*; in Keenan, E.L. (ed.): *Formal semantics of natural language*, pp. 297-312; Cambridge: Cambridge University Press.

TOMIĆ, OLGA MIŠESKA
- (1997): *Non-initial as a default clitic position*; in: *Journal of Slavic Linguistics, Volume 5*, pp. 301-323.
- (1996): *The Balkan clausal clitics*; in: *Natural Language and Linguistic Theory, Volume 14*, pp. 811-872.

VAKARELIYSKA, CYNTHIA
- (1994): *"Na"-drop in Bulgarian*; in: *Journal of Slavic Linguistics, Volume 2, Number 1*, pp. 121-150.

VALLDUVÍ, ENRIC
- (1993): *Information Packaging: A Survey*; MS. Centre for Cognitive Science and Human Communication Research Centre , University of Edinborough.
- (1992): *The Informational Component*; New York, Garland.