A LOCAL GRAMMAR DISAMBIGUATOR OF COMPOUND CONJUNCTIONS AS A PRE-PROCESSOR FOR DEEP ANALYSERS

Tzvetomira Venkova
St. Kliment Ohridski University, Sofia, Bulgaria

1. INTRODUCTION[1]

The paper discusses a model for processing Bulgarian compound *da*-conjunctions, built on the principles of finite-state techniques, namely local grammars. This model is intended for implementation as a pre-processor for a deep analyser because of its potentials to solve some basic problems of NLP thus providing the input of the deep analyser with linguistic information which can be crucial for the efficiency of the analysing procedure.

The outcome of the pre-processor can concern the following aspects of the work of the deep analyser: unambiguous treatment of the lexical semantics of a number of compound lexemes; correct marking of clause boundaries of some adjunct clauses; determining the semantic type of adjunct clauses according to the conjunctions, disambiguated by the pre-processor; greater precision in the identification of prepositional phrases due to the preceding disambiguation of some very frequent prepositions.

The choice of the compound *da*-conjunctions as an object of pre-processing is motivated by their specific position in Bulgarian. On the one hand, their analysis is a lexical-semantic task because of their compound structure and polysemy (polyfunctionality) of each of their elements. On the other hand, their function as subordinators in complex sentences poses a number of issues concerning syntax, such as clause boundaries, semantic type of adjunct clauses and the differentiation between prepositional phrases and compound conjunctions with non-adjacent positions of elements.

Local grammars provide a mechanism suitable for the integrated morpho-syntactic treatment of these conjunctions and their analyses can be formalised and implemented as a separate parsing module with two sub-modules - one for the elliptic forms, and the other for the discontinuous forms of these conjunctions. After pre-processing, the linguistic information can be encoded for the particular tasks of a deep analyser.

The model is based on some principles of local grammars as discussed in Brent 1993, Karlsson 1990, Koskenniemi 1990, Silberztein 1997.

The paper is organised as follows. Section 2 discusses the main problems for the computational analysis of the *da*-conjunctions and the principles of local grammars which can be applied for the solution of these problems. Section 3 presents the overall structure and the underlying principles and

[1]As the greater part of the research reported here was conducted in close co-operation with the University of Tuebingen, I am grateful to Prof. Dr. Erhard Hinrichs and the researchers at the Seminar fuer Sprachwissenshaft for the helpful seminars and discussions. I am also grateful to Prof. Eric Laporte, Frank Richter and Detmar Meurers for their comments on an earlier version of this paper. I have also benefited from the comments of the three anonymous reviewers.

parameters of the model of the compound *da*-conjunctions, based on local markers. The linguistic motivation for the application of a particular set of local rules to two specific realisations of the compound *da*-conjunctions: variable slots and ellipsis, is presented in Sections 4 and 5 respectively. Some advantages of the implementation of the suggested local grammar to NLP systems are presented in Section 6. Section 7 contains some conclusions.

2. BASIC PROBLEMS FOR THE AUTOMATIC ANALYSIS OF THE COMPOUND *DA*-CONJUNCTIONS AND THEIR POSSIBLE SOLUTION BY LOCAL GRAMMARS

The automatic analysis of *da*-conjunctions comes upon the specific difficulties of compound words: systematic ambiguity, compounds with variable slots, ellipsis of elements.

2.1. Systematic ambiguity

The main problem is the systematic ambiguity between the compound word taken as a whole and the interpretation of each of its components as a separate word. The processor has to distinguish between compounds and simple words. It is a challenging task because there is no formal marking between the elements of the compound words in the text.

When the elements of the compound word can not occur as simple words but only in combination with each other, then it is not difficult to identify them no matter if they are adjacent or not. But the task becomes really complicated when the elements are polysemantic (polyfunctional). In such cases some of the elements have a number of interpretations. If the reading of the word as whole is not explicitly indicated, then the number of alternative incorrect readings is very large, which overburdens the system.

Compound *da*-conjunctions in Bulgarian have as a first element one of the following words: *bez, za, vmesto, osven, predi* and as a second element the very frequent and highly polyfunctional word *da*, e.g.:

(1) *Predi da otkrie izlozhbata, toj pokani samo profesora.*
before to open-3sg exhibition-the
Before opening the exhibition, he invited only the professor

(2) *Stanah, bez da zapalja osvetlenieto.*
without to switch on-1sg lights-the
I got up without switching on the lights.

The first element (denoted by *x* below in 3.2) represents a class of words which can have two alternative functions: F1 - a preposition, e.g. *predi koncerta 'before the concert'* and F2 - an element of a compound *da*-conjunction, e.g. *predi da dojde* (lit. *before to come*) *'before coming'*.

The second element *da* has six alternative functions: F1 - a simple subordinating conjunction, joining the main and the subordinate clause, e.g. *iskam da otida 'I want to go'*; F2 - an element of a compound *da*-conjunction, *e.g. predi da dojde* (lit. *before to come-3p,sg, Pres.*) *'before coming'*; F3- a linking element in complex verb forms, e.g. *shtjah da otida (*lit. *I would to go-1p,sg, Pres*) *'I would go'*; F4 - *an* element of an indefinite pronoun, *e.g. kojto i da dojde* (lit. *whoever to come-3p,sg, Pres*) *'whoever comes'*; F5 - a modal particle in optative sentences and questions, e.g. *Da zhivee kralja (*lit. *To live the king*) *'Long live the king'*, *Da sedna li? 'Shall I sit down?'*; F6 - an affirmative particle, e.g. *Da, taka e 'Yes, it is"*.

Actually many compound words are among the most frequent words in the language. Their misinterpretation at the stage of tokenisation and morphological annotation can lead to completely wrong parsing results at the next stages. This concern especially the functional words such as the conjunctions discussed in this paper. If in the deep analyser semantic or syntactic information is attached to such words and they have not been correctly identified, the results be totally misleading.

2.2. Compounds with variable slots

Compound da-conjunctions are difficult to parse because they do not belong to the simplest case of compound words - the so called *frozen compounds* (Gross 1986:2). Frozen compounds are easy to process because no variation of their constituents is possible, i.e. they don't change their grammatical form and can not be separated, substituted or omitted, for example Engl. *in front of*. Among Bulgarian compound conjunctions as frozen compounds can be treated those with *che* and *kato* as a first element, for example *vypreki che, makar che, cled kato.* They can be put in the dictionary of compound words as one lexical unit *_vypreki_che_*, *_makar_che_* and can be tagged unequivocally. They can be identified by a simple recognition procedure for they do not require any lemmatisation or syntactic analysis in order to be reduced to a dictionary form.

However the status of the conjunctions formed with *da* is different. Although they do not change their grammatical form, the adjasent position of their elements can be violated because it is possible for other words to occuir between them, cf. *bez da razberat* (lit. *without to realise*) *'without realising'* and *bez izobshto da razberat* (lit. *without at all to realise*) *'without realising at all'*. This opportunity for realisation with a 'slot' between their elements is one of the reasons for their special treatment Example (3) demonstrates a sentence where the linear adjacency is violated and other words occur between the conjunction elements:

(3) *Shte to se obadja, predi tatko da se e vyrnal vkyshti.*
before father to refl have returned-3sg home
I shall call her, before father has returned home.

Their recognition is no longer a simple matching operation. And what makes the case more complicated is the fact that the slot can be filled with lexical units whose number and parts of speech can vary. The possible variations of the words in the slot are discussed below in Section 4.

2.3. Ellipsis

Another problem of the automatic analysis of the compound da-conjunctions is the possible ellipsis of their first element in compound sentences with more than one subordinate clauses, for example:

(4) *Tja reshi da pospi,*
She decided to take some sleep

predi da dojdat decata i $\#_{predi}$ *da zapochnat da tichat.*
before to come -3pl children-the and $\#_{before}$ *to start -3pl running around*
before the children come andstart running around.

In (4) the second occurrence of the conjunction *predi da* (lit. *before to*) *'before'* is realised in elliptic form, i.e. without the first element *predi.* Such cases should necessarily be registered in the formalism by rules which make the ellipsis explicit. Otherwise *da* in the elliptic form of the compound conjunction will incorrectly be tagged as a simple conjunction which will lead to incorrect syntactic interpretation of the whole sentence.

2.4. Local grammars as an approach for the solution of the parsing problems concerning the compound *da*-conjunctions

A common practice in those parsing systems of Bulgarian which do not apply local-grammars or stochastic methods of disambiguation, is to ignore the problems concerned with the compound words and to parse each sequence of characters between two separators as a word. However, statistic experiments show that ignoring compound words leads to high increase of error rate. A pre-processing module in the NLP system ought to identify them in order to ascribe the right morphological information and to prevent further syntactic analyses from going into the wrong direction. Such a module can be based on a *local grammar*.

Local grammars describe the words in concordance with their local context. They consist of *local rules* which are based on some properties of the words, belonging to the context of a word-form, which are defined as markers, relevant for its disambiguation. The context markers can be written down in simple symbolic notation and further converted into finite-state machines.

An important fact for the understanding of the text-processing procedure, based on local grammars, is that the computer does not have any previous semantic knowledge but only the text and a dictionary where all possible morphological tags of *da* are presented. If the word-form *da* is taken as an example, it has six different functions and therefore it will have six different tags in the dictionary. A decision between the six alternatives should be taken any time the processor comes upon the occurrence of this word. The choice would be motivated by characteristics of the words from the context (see Section 3 below).

The local grammar rules do not add more tags to the text. They are applied to texts which have already been grammatically tagged and in which the ambiguous words have been ascribed all possible tags from the dictionary. The aim of the local grammar is to discard the incorrect tags and thus to leave only the correct readings.

Most authors consider the sentence as the *sufficient context* for local rules and the rules normally operate within the sentence boundaries. If the clause boundaries are established by first disambiguating the conjunctions, then the clause can be used as sufficient context. In this respect, the disambiguation of conjunctions discussed in the paper can be considered as an important step in the parsing of the text as a whole because limiting the sufficient context boundaries improves the efficiency of the rules.

Local grammars can easily be described by means of *finite-state transducers (FST)* as is the case with the *INTEX* system, discussed in Silberztein 1994. Often these grammars are termed as finite-state formalisms. They can be regarded as absolute indicators of a structure (see Koskenniemi 1990, Karlsson 1990) or be treated as probabilistic (Brent 1993). Promising results, concerning compound words, are reported by Laporte 1988, Silberztein 1994, 1997.

As the systematic ambiguity is the essence of the difficulties concerned with the variable slots and ellipsis, the local grammars, which are known as especially suitable for disambiguation, can be applied to solve the NLP problems concerning compound *da*-conjunctions.

3. A MODEL OF THE COMPOUND *DA*-CONJUNCTIONS, BASED ON LOCAL GRAMMAR PRINCIPLES

3.1. Context fields

As to the disambiguation of the compound *da*-conjunctions, the local context can be limited within the sentence boundaries. Taking into account this limitation and the realisation of the conjunctions in the text, the local context can be divided into three fields: C1, C2 and C3 (see Figure 1). C1 is the context field between the beginning of the sentence and the first conjunction element, i.e., in linear terms C1 is the left-hand context of the compound *da*-conjunctions. C3 is the context field between the second element of the compound *da*-conjunctions and the end of the sentence, i.e. C3 is the right-hand context of the compound *da*-conjunctions. C2 is the segment between the two elements of the compound *da*-conjunctions. We have chosen the working term *inner context* for C2. C1, C2 and C3 can be viewed as ordered n-tuples. In a general form the compound *da*-conjunctions can be represented as consisting of two elements: *x* and *da*. *X* is a variable which belongs to list L1, consisting of the following elements: {*bez, za, vmesto, osven, predi*} or to list L2 consisting of one element - the empty word, represented by the symbol '#'. The empty word is a word without letters. It is introduced to represent the first conjunction element when it is elided.

The element *da* is constant. Sentence boundaries in Figure 1 are indicated by the symbols '〈〉'.

Figure 1

〈	***C1***	***x***	***C2***	***da***	***C3***	〉

3.2. Realisations of the model

The basic difficulties of the automatic analysis - the variable slot and ellipsis of conjunction elements - outlined above in Section 2 can be captured by the suggested context-sensitive scheme.

The variable slot concerns the representation of C2. When the two elements are adjacent, then C2 is empty (C2=∅).

The ellipsis concerns the representation of x. In cases of ellipsis *x* is an empty word (x=#). C1 is never empty in cases of ellipsis because an elliptic form of the conjunction can not occur in the first subordinate clause.

C1 can be empty when the subordinate clause is preceding the main clause and the conjunction takes an initial position. C3 can never be empty because the conjunction always introduces a subordinate clause.

These statements can be expressed by the model as it is shown in the figures below. Each figure represents a different realisation of the conjunction in the text.

Figure 2 represents Realisation 1 (R1 - Basic realisation), see (5) and Figure 2 below. The elements of the conjunction are adjacent (*C2*=∅). There is no ellipsis ($x \neq \#$).

(5) *Tja si trygva v deset, <u>vmesto da</u> raboti do dva.*
<u>in stead of to</u> work-3sg until two
She leaves at ten, in stead of working until two.

Figure 2

R1 - Basic realisation

〈	***C1*** $= \emptyset \vee$ ***C1*** $\neq \emptyset$	$x \neq \#$	***C2***$=\emptyset$	***da***	***C3***$\neq\emptyset$	〉

Realisation 2 (R2 - Realisation with a variable slot) represents the case in which the two elements of the compound conjunction are realised in the text ($x \neq \#$) but they are separated by other sentence elements ($C2 \neq \varnothing$), cf. (6) and Figure 3 below.

(6) *Po-dobre da otida az, vmesto majka ti da idva.*
in stead of mother your to come-3p,sg
I'd better go myself, instead of your mother to come.

Figure 3

R2 - Realisation with a variable slot

⟨	**$C1 = \varnothing \vee C1 \neq \varnothing$**	**$x \neq \#$**	**$C2 \neq \varnothing$**	***da***	**$C3 \neq \varnothing$**	⟩

Realisation 3 (R3 - Elliptic realisation) is illustrated in Figure 4. In R3 only the second conjunction element occurs in the text ($x=\#$) and it is adjacent to the empty position of the first element ($C2=\varnothing$), cf. (7) and Figure 4.

(7) *Po-dobre da otida az vmesto majka ti da idva i* $\#_{vmesto}$ *da se izmorjava.*
in stead of mother your to come-3sg and $\#_{in\ stead\ of}$ *to get-3sg tired*
I'd better go there myself, insted of your mother to come and get tired.

Figure 4

R3 - Elliptic realisation.

⟨	**$C1 \neq \varnothing$**	**$x = \#$**	**$C2 = \varnothing$**	***da***	**$C3 \neq \varnothing$**	⟩

Realisation 4 (R4 - Elliptic realisation with a variable slot) is illustrated in Figure 5 below. The slot is between *da* and the empty position of the first element. However, such cases are not registered in the corpus and therefore it is more or less a theoretical possibility. Because of its very low frequency rate, its automatic recognition has not been considered as an aim of the intended NLP system. It is illustrated below in (8) and in Figure 5.

(8) *Predi da otida tam i* $\#_{predi}$ *pone malko da ja opoznaja,*
before to go-1sg there and $\#_{before}$ *at least slightly to her get to know-1sg*
Before I go there and get to know her at least slightly,

ne moga da kazha nishto za neja.
I can't say anything about her.

Figure 5

R4 - Elliptic realisation with a variable slot

⟨	**$C1 \neq \varnothing$**	**$x = \#$**	**$C2 \neq \varnothing$**	***da***	**$C3 \neq \varnothing$**	⟩

Realisation 1 is the simplest case for automatic recognition. When *x* and *da* are adjacent they can function only as a composite conjunction. It can be included in the dictionary as a 'frozen' compound word '*_x_da_'* and it can be recognised by a simple string matching procedure. This is the most frequent realisation of the compound *da*-conjunctions.

Realisations 2 and 3 also represent a considerable number of occurences of the compound *da*-conjunctions in the corpus. According to the statistical analysis of the pilot corpora, for example 21 % of the occurences of the conjunction <u>*bez da*</u> (lit. <u>*without to*</u>) *'without'* and 40% of <u>*vmesto da*</u> (lit. <u>*in stead of to*</u>) '*in stead of*' are realised with other words in the variable slot. Therefore their correct identification plays an important role in the correct recognition of all compound *da*-conjunctions and, in general, of the conjunctions as a class in the corpus. The basic principles of their modelling is discussed in Sections 4 and 5 below.

3.3. Context markers

The context-dependent analysis of the compound *da*-conjunctions is based on distinctive markers from one or more context fields. It is important to include in the local grammar only those elements of the context which can serve as markers for a particular reading and thus resolve ambiguity.

Such markers are supposed to ascribe only F1 to *x* and only F2 to *da* and to ignore all other possible functions.

It is important for the understanding of the local markers definition procedure to keep in mind that it is applied only at a particular stage of corpus analysis. As the input text is already morphologically analysed by a non-deterministic parser, each word in the text has already been given all possible meanings (or functions as functional words are concerned) existing in the electronic dictionary. For example, an input sentence would look like (9)*:*

(9) *Toj*(PPron, 3p sg) *vyrveshe*(V3p sg Imp) *bez*(Prep):(E-CConjDA) *izobshto*(Adv)
da(Conj):(E-CConjDA):(E-V):(E-IPron):(MPart):(APart) *byrza*(V 3p sg Pres):(Adj f sg).
'He walked without being in a hurry at all'

The tags in (9) have the following interpretation: *PPron* 'personal pronoun', *V* 'verb form', *Prep* 'preposition', *Conj* 'simple conjunction', *CConjDA* 'compound *da*-conjunction', *Adv* 'adverb', *E-** 'element of *', *IPron* 'indefinite pronoun', *MPart* 'modal particle', *APart* 'affirmative particle', *p* 'person' *sg* 'singular', *pl* 'plural', *Imp* 'Imperfect', *Pres* 'Present'. The symbol ':' separates the alternative tags of the polysemantic (polyfunctional) words.

The word-forms *bez*, *da* and *byrza* in (9) are ambiguous. To disambiguate them means to discard the incorrect meanings and to leave only one tag.

The local markers for disambiguation can be of different types. Particular words can be used as *lexical markers*. For example the presence of *kojto i* in context field C2 indicates the realisation of F4 of *da* (see 2.1. above). Lexical markers are defined as lists of words. *Grammatical markers* indicate the presence of words belonging to a particular class. For example the presence of a word with the tag 'V' in context field C2 can exclude F2 of *da* from the possible tags. *Morphological markers* indicate particular morphological values of the grammatical categories, referring to a particular word. As such, for example, can be used the identical person and number values of the verb-forms in the main and the subordinate clauses. Some punctuation marks can function as *punctuation markers*. For example the presence of a comma immediately before *da* can be an additional marker for ellipsis.

4. BASIC LOCAL MARKERS FOR R2 - REALISATION WITH A VARIABLE SLOT

4.1. Principles of markers identification

The standard procedure in identifying markers for R2 would be to analyse the words which can be realised in the slot and to single out some specific formal features on the basis of the syntactic or morphological properties of these words. In the literature this slot has not been discussed in detail so far. The fronting of elements from the subordinate clause to the position between the conjunction parts is mentioned without further comments in Mutafchieva 1979:308 and GSBKE Vol. II:454. The normal occurrence of the subject of the subordinate clause in this position is registered in Penchev 1993:101. The task of markers definition requires extensive analysis of this slot, based in its greater part on the corpus data. The results of the analysis showed that practically any element of the subordinate clause, apart from the verb, can occur in this position (see for example (10) and (11) below.

(10) *Toj beshe gotov da deistva,*
He was ready to act,

predi chovejyt da se beshe obail.
before man-the to refl. have called-3sg.
before the man has called.

(11) *Predi na Ivan da mu kazhat, vsichko beshe svyrshilo.*
before to Ivan to him tell-3pl everything has ended-3sg
Before they tell Ivan, everything was over.

It is also possible that combinations of sentence elements such as subject plus object or subject plus adjunct or object plus adjunct occur in the slot. Fifteen combinations of types of sentence elements were counted as potentially realised there. The sentence elements can be realised in the slot as phrases whose length is not limited. A combination of a direct object and an adjunct is presented in (12) below as an example:

(12) *Toj poznavashe vsichki, bez nego izobshto da go poznavat.*
He knew-3sg everyone without him at all to him know-3pl
He knew everyone, without them knowing him at all.

Obviously the picture of this slot is so variegated that it can not be strictly defined by a set of the elements occuring in it. However, a different approach can be more efficient in this case - the so called 'forbidden factor' approach can be applied. In stead of formalising what can occur in the slot, what can NOT occur there is formalised. It is assumed that any occurrence of the word-form *da* preceded by any of the word-forms which are elements of L1, i.e. {*bez, za, predi, vmesto, osven*} is an instance of a compound *da*-conjunction, no matter how far they are from each other, UNLESS a number of forbidden factors are realised. A set of markers can be formally defined for these factors (see 4.2. - 4.6. below).

4.2. The verb as a marker

The verb is the only element of the subordinate clause that can not be fronted. It is not possible for the verb of the subordinate clause to appear in the position between the two elements of the *da*-conjunctions on any context conditions. Therefore the verb can be defined as a very strong exclusive marker. For these reasons the following rule can be defined: ***Rule 1: An occurrence of a verb between the string* x *belonging to L1 and the string* da *excludes the interpretation of* x *and* da *as parts of a composite da-conjunction.***

Actually in such cases the strings, which belong to L1, realise their alternative function as prepositions and *da* realises one of its alternative functions (cf. the list of functions of *da* above in 2.1.). The cases like this are numerous in the corpus. They are illustrated by (13) and (14) below:

(13) *Bez men njama da svyrshi svetyt.*
without me will not to end-3sg world-the
Without me the world would not end.

(14) *Zhena mu go posreshtashe bez mnogo prikazki i mu davashe da jade.*
wife his him met-3sg without much talking and him gave-3sg to eat
His wife met him without talking and gave him to eat.

However, it is necessary that Rule 1 is restricted because there are counter-examples in which a verb can occur between *x* and *da*, i.e. in C2. A verb in context field C2 does not belong to the main clause or to the subordinate clause, introduced by the compound *da*-conjunction, but to a relative clause or a verbal expression.

In the most typical case, relative clauses are introduced by relative pronouns, see (15):

(15) *Toj beshe gotov da deistva,*
He was ready to act,

predi chovekyt, ot kogoto chakashe signal, da se beshe obadil.
before man-the from whom expected-3sg. a signal to refl. has called-3sg.
before the man, from whom he expected a signal, has called.

As the relative pronouns are a limited class of words and their position is always to the left of the verb in C2, they can be included in a list and considered as restrictive markers.

Although relative clauses, linked by the conjunctions *che* and *da*, are not as common as those linked by relative pronouns, they should also be taken into account (for more details about Bulgarian relative clauses see Penchev 1993:90). According to GSBKE, vol. III:313, about 200 Bulgarian lexemes can be relativised by *da*- and *che*-clauses. J. Penchev claims that such lexemes should have a mark in their dictionary entries indicating the fact that they allow *da*- and *che*-clauses (Penchev 1993:91). Therefore, it might be considered possible that these lexemes can be included in a list and defined as markers.

Che-clauses, which correspond to some of the English *that*-clauses, are illustrated below in (16), in which a relative *che*-clause appears in C1:

(16) *Te zhiveeha spokojno,*
They lived peacefully,

predi sluhyt, che bandata e blizo, da se raznese v dolinata.
before rumour-the that gang-the is close, to refl spread-3sg in valley-the
before he rumour, that the gang is close, spread in the valley.

The nouns with a capacity to link with *che*-clauses can be included in a list and the elements of the list in combination with *che* can be introduced as markers. Both markers appear to the left of the verb and to the right of *x* and *che* precedes the noun marker.

An example of the relative clauses, introduced by *da*, roughly corresponding to some English *to*-clauses, are shown in (17) below:

(17) *Lena zaspa,*
Lena fell asleep,

bez misylta da$_1$ ispusne vlaka da$_2$ ja smushtava.
without the thought to$_1$ miss the train to$_2$ bother her.

The possibility *bez* and *da$_1$* to be identified as elements of the same compound *da*-conjunction is excluded by Rule 2 below. Unfortunately, the connection between *bez* and *da$_2$* is impossible to be identified according to purely formal criteria and therefore cases like (17) will be incorrectly discarded by Rule 1.

A different type of counter-examples is presented by the cases when the verb in C2 is an element of a predicative expression such as: *da rechem 'so to say', predstavi si 'just imagine', vidish li 'you see'* etc., see (18):

(18) *Toj trygna vchera, bez, predstavi si, da kazhe pone 'dovizhdane'.*
He left-3sg yesterday, without, imagine-Imp refl, to say-3sg at least goodbye.
He left yesterday, without, just imagine, saying at least 'goodbye'.

However, it is possible for such expressions to take such a position only in colloquial speech. The definition of a special restrictive marker for them is necessary only if a corpus of colloquial speech is to be parsed. As such expressions are always preceded and followed by commas, the presence of two commas to the left and to the right of the verb respectively, can be introduced as punctuation markers.

Verb forms in field C2 can also occur as formants of a particular type of indefinite pronouns such as: *kojto i da e (*lit. *whoever to be)* '*any*', see (19):

(19) *Onazi vecher Ani zaspa,*
That night Ani fell asleep,

bez kakvito i da bilo pritesnenija da smutjat synja i.
without any troubles to bother sleep her
without any troubles bothering her sleep.

The indefinite pronouns of this type are a very limited number and they can be included a list and defined as markers too.

The additional markers, discussed so far can be added to Rule 1 as restrictions. When any of these markers is present in the analysed text in the defined position and form, then the interpretation of *x* and *da* as parts of a composite *da*-conjunction is not to be discarded.

4.3. Some nouns as markers

The word *x* functions as a preposition and not as an element of a compound da-conjunction in sentences like (20) and (21):

(20) *Zhiveeh bez nadezhda da polucha pomosht.*
lived-1sg without hope to get-1sg help
I lived without any hope to get help.

(21) *Trygnaha predi izgrev slynce s namerenie da go hvanat.*
left-3pl before sunrise with intention to him catch-3pl
They left before sunrise with an intention to catch him.

To exclude the interpretation of *bez/predi* and *da* as parts of a compound *da*-conjunction in sentences like (20) and (21), the set of abstract nouns such as *zhelanie 'wish', nadezhda 'hope', namerenie 'intention'*, are introduced as lexical markers. These nouns can govern relative clauses, linked to the main clause by simple *da*-conjunctions. As it was discussed in 4.2. above, such nouns constitute a finite set and can be defined as a list L3. Another rule can be based on these markers: ***Rule 2: An occurrence of a noun from list L3 between the string* x *belonging to L1 and the string* da *excludes the interpretation of* x *and* da *as parts of a compound* da*-conjunction.***

It is not always the case that the noun from L3 and *x* belong to the same prepositional group. It is possible that *x* belongs to a prepositional group and the noun from L3 is linked to a different subordinate clause, e.g. (22):

(22) *Tja se sybudi predi drugite,*
She refl woke-3sg before the others
She woke up before the others

no njamashe zhelanie da stava tolkova rano.
but had not-3psg wish to get up so early
but had no wish to get up so early.

But such examples do not pose a problem to the analyser because the interpretation of *x* and *da* as parts of a compound *da*-conjunction is discarded by Rule 1 as there is a verb in C2.

However there is a different structure with a noun belonging to L3, see (23) below, which should be distinguished from (20) above. :

(23) *Zhiveeh, bez nadezhdata da me napusne nito za mig.*
lived-1sg without hope-the to me leave-3sg even for a second
I lived without the hope leaving me even for a second.

While in (20) *nadezhda* 'hope' is an element of a prepositional group in the main clause, extended by a relative da-clause, in (23) the same word is the subject of the subordinate adverbial clause of manner and *bez da* is a compound *da*-conjunction (23) is equivalent to (24):

(24) *Zhiveeh, bez da me napusne nadezhdata nito za mig.*
lived-1sg without to me leave-3sg hope-the even for a second
I lived without the hope leaving me even for a second.

Obviously some additional criteria should be considered. For instance, they can be of a morphological type. As such the person and number values of the verb in the subordinate clause, i.e. the verb in C3 can be considered. This verb is always third person singular in sentences like (23) because its subject is an abstract noun. In sentences like (20) the person and number of the verb can take all possible values in concord with the subject which can be any kind of a noun phrase. Therefore, when the verb in C3 is not in third person singular, then it is sure that it is not a structure of the type (23) and therefore *x* and *da* are not parts of a compound conjunction. However, when the verb in C3 is in third person singular, then the morphological criterion is not neutralised and therefore can not be used as a marker.

The presence of an article after the noun from L3 can be possibly considered as a different type of morphological marker. Actually, in many sentences this noun can not function as a subject only if it is not followed by an article, cf. (23) and (25):

(25) **Zhiveeh, bez nadezhda da me napusne nito za mig.*
lived-1sg without hope to me leave-3sg even for a second.

But exceptions to this are also possible, if a different modifier is used, cf (26):

(26) *Zhiveeh, bez hikakva nadezhda da me krepi.*
lived-1sg without any hope to me help.
I lived without any hope helping me.

And one more problem also arises to the possible acceptance of the article as a marker. When the noun is followed by an article, then it can function both in sentences like (23) and (20), and therefore the article is neutralised as a marker.

A different approach is to introduce a punctuation marker. When *x* and *da* are elements of a compound *da*-conjunction, they are normally preceded by an intonational pause, which is indicated by a comma. Therefore, the presence of a comma in C1, immediately to the left of *x*, can be considered as a restrictive marker to R2. However, when punctuation markers are introduced, some problems when colloquial speech is parsed might occur because of the transcription.

It can be seen from the discussion of counter-examples such as (23) that the process of decision taking is based on the analysis of different structural levels and it presupposes the weighing of a number of factors.

4.4. Some adjectives and participles as markers

Some adjectives and participles, which govern complement clauses, can be regarded as exclusive markers for the compound *da*-conjunction, e.g.

(27) *Taralezhyt nastryhvashe predi ogradata, gotov da se svie na kylbo.*
hedgehog-the bristled-3sg before fence-the ready to refl roll up-3sg into a ball.
The hedgehog bristled before the fence, ready to roll up into a ball.

List L4 contains adjectives and participles such as *gotov 'ready', sposoben 'able', reshen 'determined'*. It is used as a basis for ***Rule 3. An occurrence of an adjective or a participle from L4 between the string* x *belonging to L1 and the string* da *excludes the interpretation of* x *and* da *as parts of a compound* da*-conjunction.***

4.5. Indefinite pronouns as markers

The string *da* can also function alternatively as a part of a relative pronoun or an adverb such as *kojto i da 'whoever', kydeto i da 'wherever', kogato i da 'whenever'.*

(28) *Vecher pie po edna rakija bez meze, kydeto i da se namira.*
evening drink-3sg one rakija without an appetiser wherever to be
In the evening he drinks one rakija without an appetiser wherever he is.

Such pronouns should be explicitly marked when there is a prepositional group preceding the indefinite pronoun or adverb, otherwise they will be incorrectly tagged. The *k*-words which constitute the indefinite pronouns together with the second element '*i*' are defined as lexical markers and included in the list L5. ***Rule 4. An occurrence of a pronoun or adverb from L5 between the string* x *belonging to L1 and the string* da *excludes the interpretation of* x *and* da *as parts of a compound* da*-conjunction.***

4.6. Integrating the rules into a pre-processing module

The rules outlined so far can be further formally represented as a system of constraint rules which can be implemented in a finite-state transducer. Such a system is logically rather complex as it is built on the principles of non-contradiction and agreement: no rule should contradict another and if there is a marker shared by a number of rules it should be explicitly indicated at each rule. After the corpus tests, some of the rules were further restricted in regard to some specific cases such as, for example, parenthetical expressions. Some additional rules concerning word-order and punctuation, which seem conspicuous to human mind, had to be explicitly defined for the sake of computer processing. The enhancement of the rules increased the accuracy of the automatic tagging.

5. BASIC LOCAL MARKERS FOR R3 - ELLIPTIC REALISATION

5.1. Principles of markers identification

The basic feature of R3 is the absence of surface realisation of the first element of the compound *da*-conjunction, cf. (29)

(29) *Te ostanaha na masata, bez da jadat i da pijat.*
they remained-3pl at table-the without to eat-3pl. and to drink-3pl.
They remained at the table without eating and drinking.

The ellipsis of *bez* before *da* can be explicitly represented by the symbol for the blank word -# . The elided element is attached to it in subscript, cf. (30):

(30) $\#_{bez}$ *da pijat*
$\#_{without}$ *to drink*
'without drinking'

However, as the computer has no semantic information, it should rely on markers explicitly present in the text to tag the conjunctions correctly.

The first thing the procedure identifies is a presence of a non-tagged *da* in the sentence Schematically it is represented as the first 'brick' in the model:

Figure 6

〈 | ***da*** | 〉

A second feature which is taken into consideration when distinctive markers are identified is the obligatory presence of a compound *da*-conjunction with a realised *x* at the surface level in the left-hand context of the elliptic conjunction, cf. 31:

(31) *bez da jadat i* $\#_{bez}$ *da pijat*
without to eat-3pl and $\#_{without}$ *to drink-3pl*
without eating and drinking'

The fact that the ellipsis of the first compound conjunction element is possible only in complex sentences with more than one subordinate clause linked by the same compound conjunction to the main one should be represented in the model. An important aspect of the interpretation of the model is the fact that it is applied as a text processing procedure only after the all compound *d*a-conjunctions in realisations R1 and R2 have been identified and tagged (see Section 4). Therefore the conjunctions in R1 and R2 can be used as formal markers. Figure 7 below represents schematically what is explicitly indicated as phrase 31 is concerned ('#' - sentence boundary, '...' - non-restricted number of positions).

Figure 7

C1

〈 | ***... {x da}(CConj DA) ...*** | ***da*** | 〉

Although the model in Figure *7* represents the cases of elliptic compound *da*-conjunctions it also represents some other functions of *da* and therefore should necessarily be restricted. For example, *da*$_3$ in (32) is a simple *da*-conjunction:

(32) *bez da*$_1$ $_1$*jadat i da*$_2$ *se opitat da*$_3$ *pijat.*
without to eat-3pl and to refl. try-3pl to drink-3pl
without eating and trying to drink

As a marker distinguishing *da*$_2$ from *da*$_3$ are considered the coordinators *i 'and', ili 'or'* as well as the comma ','. They are represented by the variable y and are included in the list L6. They are regarded as markers only if they occur immediately before *da*:

Figure 8

	C1			
⟨	***... {x da}(CConj DA) ...***	***... y∈L6***	***da***	⟩

The scheme in Figure 8 can be expressed verbally as the following rule:

If within the sentence boundaries there is
a) a non-disambiguated string* da *and
b) to the left of* da *there is a string tagged as CConjDA and
c) immediately to the left of* da *there is a string which is an element of L6
then* da *is an element of an elliptic compound* da*-conjunction.

5.2. Integrating the rules into a pre-processing module

The basic markers determining the identification of the elliptic forms of the compound *da*-conjunctions have been outlined in *5.1.* However, a number of additional rules have to be added so as to attain a system of rules with high accuracy.

The ambiguity of some markers presents one of the basic problems. Such markers have to be disambiguated by additional rules. A group of rules concern the ambiguity of the elements belonging to L6 (see Figure 8 above) which are defined as a marker immediately before *da*. The string *i* which is an element of L6, apart from being a coordinator, can also function as a part of a concessive compound *da*-conjunctions such as ***'dori i da'*** *'even if'* ***'makar i da'*** *'although'* ***'i da'*** *'although'* and as a part of indefinite pronouns and adverbs, such as ***'kojto i da'*** *'whoever'*, ***'kydeto i da'*** *'wherever'*. As the alternative functions of *i* concern a small number of words, these words are defined as lists - L7 and L8 for each function respectively. The occurrence of any element belonging to these two lists to the left of the element *y*, which belongs to L6, excludes the interpretation of *da* as a part of a compound *da*-conjunction. The concessive conjunction *i da* presents more problems to the analysis because the words belonging to L7 and L8 are not relevant for its disambiguation. The punctuation marker 'comma' immediately before *i* is introduced to mark the occurences of a concessive conjunction.

Other rules concern a different case of elision - that of the word governing subordinate *da*-clauses. In (33) the verb ***moga*** ***'****can'* is elided before *da*$_3$:

(33) *Predi $\underline{da_1}$ dojdesh, mozhesh $\underline{da_2}$ kupish hljab i $\underline{da_3}$ se obadish na Sergei.*
before to come-2sg can-2sg to buy-2sg bread and to refl call-2sg Sergei
Before you come, you can buy bread and call Sergei.

The strings da_2 and da_3 are simple *da*-conjunctions joining subordinate clauses, governed by the verb *moga 'can'*. Da_2 can not be incorrectly identified as an element of an elliptic compound *da*-conjunction because the necessary markers are not present (cf. Figure 8 above). However, da_3 is problematic because it meets all conditions. A special rule is added to exclude cases like da_3. It is based on the occurrence of da_2 between them.

If the local grammar is to be implemented in NLP systems processing colloquial speech, another case of functional ambiguity should be considered. It concerns the use of modal expressions such as wishes or swear-words which also start with *da*. In them *da* realises its function as a modal particle. They are separated be comma too (the comma is an element of L5, see Figure 8 above) which might mislead the analyser, cf da_2 below:

(34) *Toja ne dava nishto, predi da_1 vidi parite, $\underline{da_2}$ go vzemat mytnite.*
He not give-3sg anything before to see-3sg money-the, to him take-3pl the devils.
He doesn't give anything before he sees the money, damn it.

A rule, based on group of markers from context field C3, can be introduced to discard such cases.

The rules outlined in this section have been incorporated into module which is suitable for implementation in a text-processing system. This module should be applied together with the module based on the local grammar section described in Section 5.

6. SOME ADVANTAGES OF THE IMPLEMENTATION OF THE SUGGESTED LOCAL GRAMMAR IN A FINITE-STATE PRE-PROCESSOR

The development of a local grammar for the recognition of *da*-conjunctions can have considerable effect on the accuracy of automatic tagging procedures in several aspects.

In the first place, the occurences of the compound *da*-conjunctions in the text will be recognised and correctly tagged. Context-sensitive rules are the only way to recognise correctly and tag automatically as parts of one discontinuous word those conjunction elements which occur in non-adjacent positions in the text (see Section 4 above) or when some elements are elided (see Section 5 above).

In the second place, the correct tagging of *da*-conjunctions is an important step towards the disambiguation of the occurences of one of the most frequent words in the corpus (see Nikolova 1987) - the word *da*. The effect of its correct tagging on the overall accuracy rate of the tagging procedure of the whole text will be tremendous. Statistical research on two pilot corpora of 50 000 tokens each, shows that 2.5% of the tokens in fiction monologue and 4.1% in colloquial speech are occurences of *da*. (Having in mind that most words appear only once, that is 0, 002% in each corpus, these percentages are very high.) As it was discussed above, the word *da* is functionally ambiguous. Defining unambiguously one of its functions will reduce the ambiguity to a considerable extent. Moreover, the model as a whole is oriented towards facilitating further exploration of *da*-functions.

In the third place, the accuracy of the whole tagging procedure will be further increased by disambiguating the first element of the *da*-conjunctions which alternatively function as a preposition. Prepositions are also among the most frequent words in the corpus. After the *da*-conjunctions are disambiguated, all the occurences of these prepositions can receive only one tag.

Based on the positive effects of the finite-state module on morphological tagging, the implementation of the suggested model can considerably enhance syntactic analysis. It can be integrated in a syntactic parser as a sentence boundaries indicator and as a detector of subordinate clauses in compound sentences, in defining the semantic types of adjunct clauses and in phrase structure analysis.

7. CONCLUSIONS

The application of a local grammar pre-processor to compound *da*-conjunctions shows promising results at managing some basic problems of automatic parsing: variable slots and ellipsis of conjunction elements. The local markers are not chosen at random but constitute a logically structured system, based on extensive corpus experiments. The 'shallowness' of the local rules is ostensible. They are defined according to the complex interaction of linguistic factors from different structural levels. The implementation of the model into NLP systems increases the accuracy of the analysis and the reliability of the result. On the other hand, the local grammar processing has its limitations determined by the total exclusion of semantic information concerning the individual items and he sentence structure. The shortcomings due to those limitations can be overcomed by the implementation of a theory-based approach in the next-step processor.

Bibliography

Brent, M. 1993. From Grammar to Lexicon. Unsupervised Learning of Lexical Syntax. In Computational Linguistics, Vol. 19, No.2, June 1993, pp.243-262.

Rechnik na bulgarskija ezik 1981., t. III, Izdatelstvo na BAN, Sofia, 1981. [Dictionary of Bulgarian Language 1981, vol. III].

GSBKE 1993. Gramatika na syvremennnija bylgarski knizhoven ezik, t. III, Sintaksis. Izdatelstvo na BAN, Sofia, 1993. [GSBKE 1993 Grammar of Modern Bulgarian Language, vol. III. Syntax]

Gross, M. 1986. The Representation of Compound Words. In COLING 86, p.1-6.

Koskenniemi, K. 1990. Finite-state Parsing and Disambiguation. In COLING 90, pp.229-232.

Karlsson, F. Constraint Grammar as a Framework for Running Text. In COLING 90, pp.168-173

Mutafchieva, Z. 1979. Otstypitelnite syjuzi v syvremennija bugarsli ezik. In Pomagalo po bylgarski sintaksis, Nauka i izkustvo, Sofia, 1979, pp. 299 - 311. [Concessive conjunctions in Modern Bulgarian Language]

Laporte, E. 1988. La reconnaissance des expressions figees lors de l'analyse automatique. In Langages 90 Danlo, L. editor, pp.117-126.

Nikolova, Tzv. 1987. Chestoten rechnik na bylgarskata razgovorna rech, Nauka I izkustvo, Sofia, 1987. [Frequency dictionary of Bulgarian colloquial speech]

Penchev, J. 1993. Bylgarski sintaksis. Upravlenie i svyrzvane. Plovdivsko universitetsko izdatelstvo, Plovdiv 1993. [Bulgarian Syntax. Government and Binding]

Silberztein, M. 1994. INTEX: a corpus processing system. In Proceedings of COLING-94, Kyoto.

Silberztein, M. 1997. The Lexical Analysis of Natural Languages. In Finite-State Language Processing, E. Roche and Y. Schabes (eds). MIT Press, 1997, pp. 175-203.