Raising and Control Constructions in a Bulgarian UD Parsebank of Parliament Sessions

Petya Osenova
Division of Bulgarian Language
Sofia University "St. Kl. Ohridski"
osenova@uni-sofia.bg

Abstract

The paper discusses the raising and control syntactic structures (marked as 'xcomp') in a UD parsed corpus of Bulgarian Parliamentary Sessions. The idea is: to investigate the linguistic status of this phenomenon in an automatically parsed corpus, with a focus on verbal constructions of a head and its dependant together with the shared subject; to detect the errors and get insights on how to improve the annotation scheme and the automatic detection of this phenomenon realizations in Bulgarian.

Keywords: control and raising verbs, Bulgarian Parliamentary Corpus, Universal Dependencies.

1 Introduction

In the Universal Dependencies (UD) syntactic guidelines the dependancy relation *xcomp* is viewed as a clause that belongs to the group of core arguments together with *csubj* and *ccomp*. It is used in two cases: a) in constructions with obligatory control (object-to-subject and subject-to-subject) and usually non-finite (for example, in the sentence 'I want to sleep', the non-overt subject of 'sleep' is determined by the overt subject 'I' of the higher predicate 'want'), and b) for the respective types of secondary predication (for example, in the sentence 'She declared the cake beautiful' the predicates 'declared' and 'beautiful' are connected through *xcomp*). In this survey I am interested in the open clausal complements only, i.e. 'a predicative or clausal complement without its own subject'. As the guidelines further say: 'That is, there should be no available interpretation where the subject of the lower clause may be distinct from the specified role of the upper clause. In cases where the missing subject may or must be distinct from a fixed role in the higher clause, *ccomp* should be used instead [...]. This includes cases of arbitrary subjects and anaphoric control.'[1]

The aim in this paper is to observe the *xcomp* types of subject-to-subject control structures in an automatically parsed parliamentary corpus for Bulgarian. I am interested in the following questions: a) what kind of control syntactic structures were realized with respect to a main and a controlled predicate; b) what kind of subjects were realized in the control structures – both formally and semantically; c) were any error types detected within the observed structures; d) how do these observations contribute to the linguistic typology of Bulgarian control structures and to their better modeling and detection. I consider the linguistic investigations over parsebanks as a way to identifying real language problematic phenomena for parsing beyond the already modeled constructions in grammars, annotation schemes and manually annotated treebanks. I also believe that they give us hints on how to improve the coverage of a treebank (for example, through the means of active learning) for better linguistic research.

The paper is structured as follows: in the next section the details on the parsed corpus as well as on the used model are given. Section 3 focuses on the relation *xcomp* with respect to the above mentioned research questions. Section 4 concludes the paper.

2 The UD parsebank of Bulgarian Parliamentary sessions

This study was performed over the Bulgarian ParlaMint corpus[2] because it has been annotated with respect to the UD schema and is freely available for research. In future, the plan is to extend the texts in the parsebank with newsmedia and social

[1] https://universaldependencies.org/u/dep/xcomp.html

[2] https://www.clarin.si/repository/xmlui/handle/11356/1431

media corpora, among others.

ParlaMint[3] is a project supported by CLARIN-ERIC[4]. Its first phase - ParlaMint I - was completed in the period of years 2020 - 2021. Parliamentary data directly correspond to the most recent events with global impact on human health, social life and economics such as the current COVID-19 pandemic. The Bulgarian ParlaMint corpus contains plenary meetings from 2014-10-27 to 2020-07-31 and includes 717 documents, or 19,096,761 words. The data is publicly available from the project website. Now in the subsequent project phase - ParlaMint II (2022 - 2023) - more data have been compiled to the current corpora, and parliamentary corpora for new countries have been added.

The Bulgarian Parliamentary data was downloaded from the official website of the Bulgarian National Assembly [5]. The sessions for each day were represented in a single html file which was relatively easy to convert to XML. The conversion was performed in an incremental way. Initially, the data was converted into a basic TEI XML format and then uploaded into the CLaRK system — (Simov et al., 2004). Afterwards, the Parla-CLARIN format [6] was used for validation. However, this turned out to be too permissive, so an additional constraint schemata were applied. Within CLaRK system the conversion was done with the help of constraints (as implemented rules) and regular grammars for inserting some elements. The speaker information (such as date and year of birth, occupation, party memberships, personal web page, etc.) and incident data (such as applause, laughing, entering or leaving the plenary room, noise, etc.) were extracted, classified and returned back into the texts with the appropriate features added. Thus the present linguistic research can be extended in future with adding more society-oriented features from the available metadata – like which member of Parliament uses what control constructions and with what a reference, etc.

The created corpora were processed with the `classla-stanfordnlp` pipeline, which annotates text on the levels of morphosyntax, lemmas, dependency syntax and named entities for Bulgarian, Croatian, Serbian, and Slovene.[7] This model is a CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages. The Bulgarian part was trained with the UD Bultreebank model and on the provided big corpus of Bulgarian data. The resulting analyzed corpus of parliamentary sessions was uploaded into the CLaRK System where it was possible to search for respective subtrees related via *xcomp* within the UD syntactic structures. The extracted patterns include the control verb, the dependant verb and the subjects when they are explicit at the higher or lower verb level (although in *xcomp* constructions an explicit subject at the lower clause is not expected). In Figure 1 an example in XML of an extracted pattern is given from the CLaRK system. The sentence is as follows: But not can-1.PL to give-1.PL more money, 'However, we cannot give more money either'. The *xcomp* relation connects the verb in the higher clause - 'can' - with the one in the lower clause - 'give'. Both subjects are not overt.

In Figure 2 three examples are graphically visualized where the head and dependant verbs are related through *xcomp*.

In the tree on the top-left the following sentence is given (here glossed, and all that follow are also glossed): *Can-2.PL to check-2.PL (You can check)*. In this subject-to-subject control both subjects are null since Bulgarian is a pro-drop language. We consider this structure as a true control one because the subject of the verb in the lower clause - 'check' - is the same as the one of the verb in the higher clause - 'can'.

In the tree on the top-right the following sentence is given: *Raynov will come to them.CLITIC take (Raynov will come to take them)*. Here the main verb 'come' has an explicit subject – the surname Raynov – in contrast to its dependant verb 'take'. I do not consider such a structure a true control one, since the verb 'come' can take dependant verbs with a different subject. One test that can be used here is the possible substitution of the marker да (to) with the subordinator за да *('for to', in order to)*. In the example the subjects of the two verbs are the same. We would like to have a way to distinguish such cases in parsebanks.

In the tree in the bottom-middle, the following sentence is given: *How would could to happen this? (How could this happen?)*. Here the explicit subject is realized to the dependant verb 'happen'

[3] https://www.clarin.eu/parlamint

[4] https://www.clarin.eu/

[5] https://www.parliament.bg/bg/plenaryst

[6] https://github.com/clarin-eric/parla-clarin

[7] https://pypi.org/project/classla/

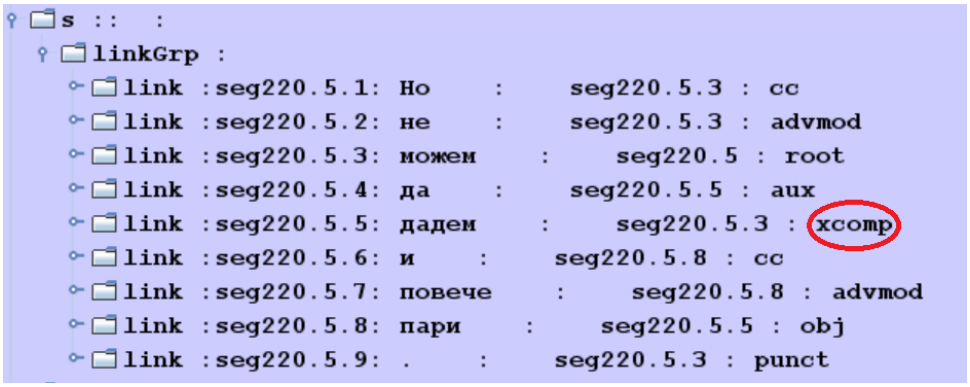

Figure 1: An extracted pattern from the CLaRK system.

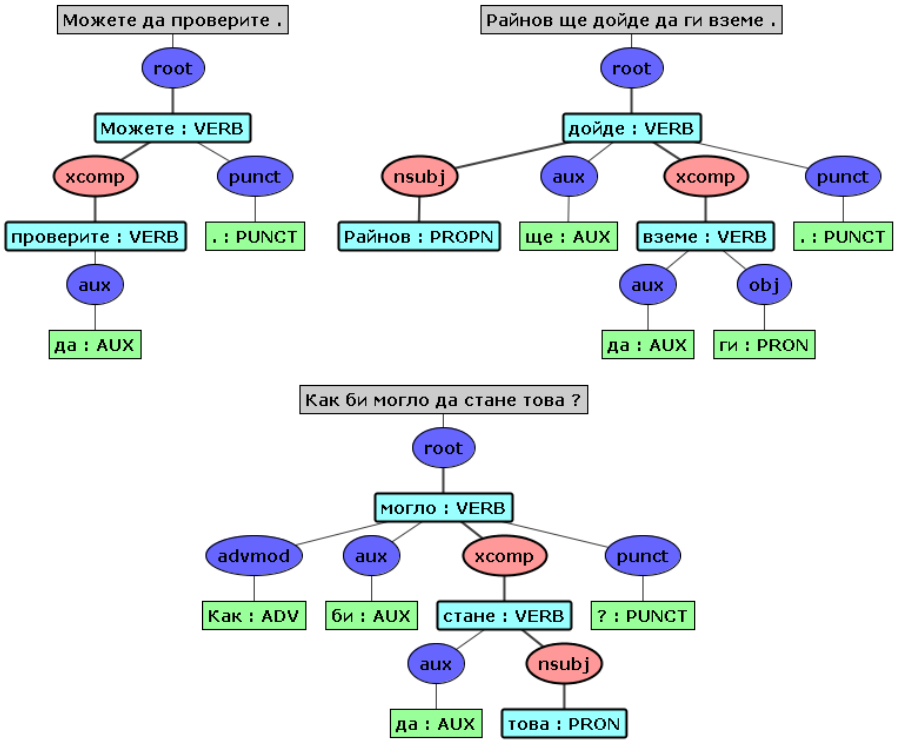

Figure 2: Visualized patterns with the *xcomp* relation.

in contrast to the main modal verb. However, here many other factors play a role. For example, the adjacency of the pronominal subject either to the main or to the dependant verb with respect to the illocutionary force - interrogative in this case. I view such patterns as formally controlling.

All the patterns presented here were used as templates in the process of extracting a subcorpus for the current study.

3 The *xcomp* realisations and their analysis

The control verbs are usually discussed on a par with the raising ones. The literature on control and raising verbs from a theoretical or a specific language point of view is very rich and sometimes controversial. For that reason I will mention now only the work on control and semantic resource sensitivity by (Asudeh, 2005). The author gives an overview of the main approaches and proposes a structure sharing alternative for both – non-finite and finite control. The analysis is based on Glue Semantics and is performed within the framework of LFG.

In the original constituency Bultreebank (later converted into the UD style), the control structures were not specially marked as such. There was a mechanism to indicate the same subject in the syn-

tactic structures via co-reference links. However, these links reflected the contextual usages of same-subject-hood, not the real control. Thus, they can be viewed as overgenerating. This means that no real distinction was made between structures of control where the predicate imposes on its dependent the same subject in all contexts, and structures where the same subject is not obligatory and thus would allow the appearance of different subjects. Making such a differentiation is not a trivial task per se. At the same time, the fact that raising verbs do not impose any restrictions to their subjects (expletive as a rule) has been reflected by assigning the referential subject to the lower clause verb.

3.1 Structures of control in Bulgarian: a brief overview

In the traditional Bulgarian grammar literature the control verbs are viewed as imposing argument sharing. These verbs are modal (with some exceptions) or phasal. They are considered auxiliaries and thus constitute the so-called 'complex verbal predicate' forming a simple sentence where both verbal subjects are co-indexed. See an overview of the various points of view in (Viktorova, 2005). These verbs are: *(*мога *(can),* трябва *(have to),* започвам *(start),* продължавам *(continue),* спирам *(stop))* with their synonyms. The exceptions include the verb искам *(want)* because it can take various subjects.

Among the modals there exist also raising verbs such as the impersonal verbs with expletive subjects like трябва *(have to) and* може *(to be possible to).*

In cases where the modal verb allows for a different subject of the dependent verb, the sentence is considered not simple but complex. Such a verb, as mentioned above, is искам *(want).* Compare Искам ти да дойдеш. *(Want-I you to come, I want you to come.)*

(Penchev, 1993) mentions the control structures of types subject-to-subject (p. 169) and object-to-subject (p. 87, p. 169). For the first type the example is Ти$_1$ забрави pro$_1$ да дойдеш. *(You$_1$ forgot pro$_1$ to come.)* For the second type the example is Принудиха ги$_1$ pro$_1$ да заминат. *(Forced-they them to go, They were forced to go.)*

In (Boyadjiev et al., 1998) (pp. 550-551) Penchev also shows that control is not related to modality only, since some modal verbs behave like content verbs while there are also non-modals that exhibit control characteristics. The author promotes a unifying analysis where both control structures – with modals and non-modals – form a complex sentence.

3.2 Realisations of control structures in the corpus

First, let us look at the heads of the control structures and their frequency. The most frequent one is the modal verb мога *(can)* with 47514 occurrences. In the UD version of Bultreebank modals were treated as full verbs, not as auxiliaries.

In the top 20 lemmas the following types have been observed: other modal verbs ща *(want)*; verbs of phases *(*продължа/продължавам *(continue),* започна/започвам *(start, begin),* спра/спирам и престана *(stop))*; other verbs *(*успея/успявам *(succeed),* опитам се/опитвам се *(try),* пропусна/пропускам *(miss),* отида/отивам *(go somewhere),* откажа/отказвам *(deny)).* Also in the top part come other modal or modal-like verbs like: стремя се *(aim),* възнамерявам *(intend),* умея *(be able),* смея *(dare).*

At the same time some verbs seem to be out of place here because they either express adverbial semantics or allow a non-controlled subject. Such verbs are: изляза/излизам *(go out)*, бързам *(hurry)* with adverbial semantics and thus the expected relation would be *advcl* or призова/призовавам *(call for)*, предложа/предлагам *(suggest)* and thus the expected relation would be *ccomp*. This fact is not surprising because – as mentioned above – such verbs could also share the same subject in some of their realizations.

Let me now turn to the structures with controlling and controlled predicates. I am interested in three questions: a) which are the typical controlling predicates, b) which are the structures that are not really controlling and c) which are the linguistic tests that show the non-controlling usages of the detected verbs in b).

Concerning the modal verbs, the most frequent structure is мога да кажа *(can-1.SG to say-1.SG, I can say).* It has 2230 occurrences. Overall, the perfective verbs are preferred: мога да разбера / приема / дам / направя *(can-1.SG to understand-1.SG / accept-1.SG / give-1.SG / do-1.SG, I can understand/accept/give/do).* One remark should be done here. The third person of the verb can have also an impersonal usage, i.e. mean-

ing that something is possible. Such cases of two possible readings for convenience were annotated in Bultreebank as preferably personal verbs. Thus, many of the examples in the parliamentary corpus also bear this inherited ambiguity.

3.3 'True' control verbs

Here come the 'true' control verbs, or in other words, verbs that would not allow for a different subject of the lower clause verb. Apart from the modal, phase and other verbs, mentioned above, some other verbs are listed below. Please note that some of them are used in their reflexive forms. The semantic classification is made with respect to the lexicographic classes in Princeton WordNet (in contrast to (Henri and Laurens, 2011) where another type of semantic classification is given for Mauritian):

- verbs of cognition: уча се (learn), пропусна/пропускам (miss)
- social verbs: опитам се/опитвам се *(try)*, принудя се/принуждавам се *(force)*, задължа се/задължавам се *(oblige)*, рискувам *(risk)*
- verbs of change: готвя се *(prepare)*
- verbs of communication: откажа/отказвам *(refuse)*

It would be interesting to investigate further the relation between control structures and reflexivity. In general, the reflexive marker се *'se'* ensures the intransitive use (thus – subject-to-subject control) of a transitive verb that provides an object-to-subject control. For example, Учих го да чете *(Taught-1.SG him to read-3.MASC.SG, I taught him to read)* vs. Учих се да чета *(Taught-1.SG REFL to read-1.SG, I taught myself to read)*.

Some insights with respect to the usage and blocking of impersonal and passive se-constructions have been considered in (Penchev, 2001). For example, when a reflexive control verb is used in an impersonal-passive, then either such an usage is semantically blocked (ex. try) or its dependant has to share the same form, and the subject becomes arbitrary (ex. forget): Забравя се да се звъни *(Forget-IMPERS.REFL to REFL ring-IMPERS, Ringing is (being) forgotten)*. It should be noted that such usages are rare.

Another issue that became evident is the role of diathesis. It can be detected in the examples of the verb задължа се *(oblige oneself)*. In all the examples these usages are in reflexive passive. Let us see one: Общината се задължава да извърши проверка *(Municipality-the REFL oblige to perform check, The municipality is obliged to perform the checks)*. Such cases are also considered control structures – not from a lexical but from a syntactic point of view. The role of the reflexive passives is discussed in (Dzhonova and Mihaylova, 2021) where it is mentioned that these forms can have modal meanings when used in a generic way. The reflexive passives can be placed also in the diathesis typology, presented in (Koeva, 2022).

Here it would be also interesting to observe the combinations of a control verb with types of dependant verbs as well as their common subject characteristics.

The modal verb мога *(can)* as the most frequent one has many collocations, thus we will ignore it here. In the group of the phase verb започна/започвам *(start)* the following clusters can be identified: започвам да функционирам *(start functioning)* where the dependant verbs are in active voice and subjects refer to the government, software, assembly, law, portal; започвам да тека *(start to run)* where the dependant verbs are in active voice and subjects refer to mandate, process, deadline, intership; започвам да работя (start to work) where the dependant verbs are in active voice and subjects refer to institution, system, assembly, power. There are also structures where the dependent verb is preferred in se-passive. Here are some examples: започва да се прави компромис/реформа; започва да се гледа бюджет/закон; започва да се говори истина/неистина (start to REFL do compromise/reform; start to REFL look budget/law; start to REFL speak truth/non-truth).

In the group of the phase verbs продължа/продължавам *(continue)* the following clusters can be identified: продължавам да действам *(continue to hold/be in effect)* where the dependant verbs are in active voice and subjects refer to rule, practice, formula, criterion; продължавам да съществувам *(continue to exist)* where the dependant verbs are in active voice and refer to nation, threat, problem, inequality, tension, possibility.

The cognitive verb пропусна/пропускам *(miss)* has a preference to speech-related active dependant verbs like say, note, mention, remind, give an answer.

The social verb принудя се/принуждавам се *(force oneself)* prefers dependant verbs of activities like *'to be forced to come (for a prime-minister); to co-finance (for a municipality); to resort to (for the state)'*.

It turned out that the control verbs other than modal and phase ones are not so frequent in the data.

On the basis of the statistical information about the distribution of these constructions - the combination of the head verbs, the dependent verbs and the grammatical features of the subjects, rules can be formulated to classify the candidate control structures. These are based on grammar characteristics such as shared number and gender where applicable. Then manual evaluation over 3951 examples was performed. From these 3100 were classified as control structures while only 5 cases happened to be misclassified. From the rest there were 651 cases which were classified as structures with non-shared subjects, and 200 that were considered as quasi control structures presented in the next section.

3.4 Quasi control verbs

Some examples were given above with verbs that can take not only the inherited infinitive particle да *(to)*, but also the subordinator за да *(in order to)*. This fact can be used as a test for classifying such verbs as quasi control ones because it allows a structure with different subjects. This group mostly consists of verbs of action. For example, дойда да гласувам *(come to vote),* излизам/отивам да говоря *(go to speak),* чакам да видя *(wait to see),* работя да осигуря *(work to ensure)*, etc.

There is one verb that is ambiguous between a control and quasi control interpretation. This is спра/спирам *(stop)*. In the first meaning – the phase one – it is a verb of control: Спрях да пуша *(Stopped-I to smoke, I quited smoking)*. In the second meaning – the action verb – it is a verb of quasi control: Спрях да купя мляко *(Stopped-I to buy milk, I stopped to buy milk)*. In the parliamentary data only the phase verb has been detected.

There is another group of quasi control verbs that allows for the dependant verb to take a subject in a different number. These verbs belong preferably to the groups of verbs of communication and cognition. For example: предложа/предлагам *(suggest),* ангажирам се *(engage oneself),* апелирам *(apel),* избера/избирам *(choose),* планирам *(plan).* For example, Предлагам да дойдем по-късно*, Suggest-1.SG to come-1.PL later, I would suggest we to come later*.

As a result from these observations, a number of tests were created for the classification of control vs. quasi control usages like the one with the subordinator substitution, and some based on the lexical properties of the verbs like their valency and agreement potential . In addition to using them as features when training parsing models, such tests might be implemented as filters over the search in parsebanks.

4 Conclusions

In this paper some focused observations were shown on the behaviour of Bulgarian structures of raising/control in an automatically parsed UD corpus of parliamentary sessions. The manual checks over the extracted data confirmed the high quality of the UD parser on these data. Thus, it became possible to detect for example the 'true' control structures vs. quasi control structures. The overgeneration seems to be inherited from the Bultreebank model where all cases of shared subjects were marked as coindexed. Due to the distinction between active (*nsubj*) and passive (*nsubjpass*) subjects in the UD schema, it was possible to survey the internal structure of control and observe the preferences of dependant predicates with respect to their control heads to active or passive usages.

One of my goals in this study was also to detect weaknesses in the Bulgarian UD treebank which needs some extensions of the annotation patterns in order to provide better parsed corpora for linguistic research. I think that these analyses of control constructions in the current version of the corpus show the following directions of future work: extension of the treebank coverage with new texts that would demonstrate some of the problematic cases for the parser.

My observations showed that it is difficult to distinguish between similarly presented phenomena in texts, such as control and quasi control structures. These phenomena might be approached by using lexical lists with both types of verbs. However, this is not enough because their contextual realizations also have to be taken into account. In my view the challenge behind the automatic annotation is to find the best balance between lexicon and grammar. If such a balance was achieved, then the parser would be more linguistically informed and would classify

the presented phenomena in a better way.

References

Ash Asudeh. 2005. Control and semantic resource sensitivity. *Linguistics*, 41. Cambridge University Press.

Todor Boyadjiev, Ivan Kutsarov, and Yordan Penchev. 1998. *Contemporary Bulgarian*, name of chapter: Syntax. Petar Beron Publishing House, Sofia.

Marina Dzhonova and Bilyana Mihaylova. 2021. Reflexive passive in Bulgarian and Romanian - forms and uses. *Contrastive Linguistics*, 2-3:25–36. St. Kliment Ohridski University Press.

Fabiola Henri and Frédéric Laurens. 2011. The complementation of raising and control verbs in mauritian. In O. Bonami and P. Cabredo Hofherr, editors, *Empirical Issues in Syntax and Semantics*, 8, pages 195—-219. Http://www.cssp.cnrs.fr/eiss8.

Svetla Koeva. 2022. System of Diatheses in Bulgarian. *Proceedings of the International Annual Conference of the Institute for Bulgarian Language*, pages 80–91. Prof. Marin Drinov Publishing House of Bulgarian Academy of Sciences.

Yordan Penchev. 1993. *Bulgarian Syntax: Government and Binding*. Plovdiv University Press, Plovdiv.

Yordan Penchev. 2001. One component sentences. In Svetla Koeva, editor, *Contemporary Linguistic Theories*, pages 86–93. Plovdiv University Publishing House, Sofia.

Kiril Simov, Alexander Simov, Hristo Ganev, Krasimira Ivanova, and Ilko Grigorov. 2004. The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. *Proceedings of LREC 2004*, pages 235–238.

Kalina Viktorova. 2005. Functional development of da-construction in contemporary bulgarian. In Svetla Koeva, editor, *Argument structure: Problems of the simple and the complex sentence*, pages 185–224. SEMA RSH, Sofia.