Towards transferring Bulgarian sentences with elliptical elements to Universal Dependencies: issues and strategies

Petya Osenova
LMaKP
IICT-BAS
Sofia, Bulgaria
petya@bultreebank.org

Kiril Simov
LMaKP
IICT-BAS
Sofia, Bulgaria
kivs@bultreebank.org

Abstract

The paper considers the problems in transferring the sentences with elliptical elements from the original BulTreeBank into the Universal Dependencies style. The similarities and differences between the original constituency annotation scheme and the target dependency one are outlined to show that the current UD scheme needs elaboration to capture more complex cases.

1 Introduction

BulTreeBank (BTB) — an HPSG-based treebank of Bulgarian (Simov et al., 2005) — encodes both constituent and head-dependant structure in each phrase.[1] This facilitates the conversion from a constituent to a dependency representation. It should be noted, however, that BTB is not a typical HPSG treebank per se. It reflects the main principles of this theory and makes use of the structure sharing mechanism (especially for encoding phenomena like control) but it does not represent complex feature structures. The detailed features and their interaction are encoded within the node labels like VPS for verbal head-subject phrase and the structure of the constituent tree.

The current conversion of the treebank into the Universal Dependencies (UD) annotation scheme does not include the sentences with elliptical elements. These sentences amount to 1007 altogether. In the recent release editions most of the enhanced dependencies have already been added. The enhanced dependencies include the following phenomena: null nodes for elided predicates; propagation of conjuncts; additional subject relations for control and raising constructions; arguments of passives (and other valency-changing constructions); coreference in relative clause constructions and modifier labels that contain the preposition or other case-marking information.

However, the sentences with ellipsis (including the null nodes for elided predicates in the list of enhanced dependencies above) are still not present in the resource. These sentences constitute about 7% of the treebank. Needless to say, they are very important because they illustrate frequent structures that are typical for the organization of Bulgarian sentences.

In this paper the annotation typology of ellipsis in BulTreeBank is presented together with discussion on strategies for transferring these annotations into the UD framework. We also discuss the complexity of the envisaged transfer with respect to the specific types.

Our submission is expected to contribute to the discussion on the proper handling of elliptical phenomena, especially when transferring them from a constituency to a dependency-based treebank. Despite the fact that it focuses on Bulgarian language only, we believe that our considerations would be useful also for modeling ellipsis in treebanks of other languages.

The structure of the paper is as follows: in the next section related work is outlined briefly. Section 3 focuses on modeling ellipsis in the original BulTreeBank. Section 4 compares the annotation of elliptical phenomena in the original treebank with the strategies within UD. Section 5 concludes the paper.

[1] Coordination is an exception. It is considered as non-headed phrase.

2 Related Work

There is extensive literature on the ellipsis and its treatment in one or more languages, on grammatical and cognitive levels, etc. Here, however, only few findings will be mentioned that mainly focus on annotations within dependency-based frameworks.

(Mikulova, 2014) presents the typology of ellipsis in Czech in the dependency theory of Functional Generative Description. Since this is a multistratal theory and thus distinguishes between surface and deep levels, ellipsis is mainly modeled on deep (tectogrammatical) level. Within the scope of the surface ellipses the author includes the so-called'structural ellipses' (type 1). Here belong the following subtypes: a) ellipsis of the governing verb (I like coffee, but you [like] tea) and b) ellipsis of governing noun (Central [Europe] and Eastern Europe). Within the scope of the deep syntactic ellipses the author includes the so-called 'valency ellipses' (type 2). These include phenomena like textual ellipsis, general argument, control and reciprocity. While the former type (1) is analyzed with the insertion of an empty node, most cases that belong to the latter type (2) are marked with coreference arrows. In BTB the treatment of ellipsis is only on one level. Also, 'structural ellipses' group includes functional words/dependants (auxiliaries, modals, prepositions, etc.). 'Valency ellipses' are treated either with an insertion of a node (textual ellipsis), or with coreference (control), or not approached at all (general argument, reciprocity) i.e. a mixed strategy has been followed.

Another approach on the same language – Czech – is adopted by (Jelinek et al., 2015). The authors propose a constituent-based analysis for handling ellipsis, because it includes more information than the dependency-based one and also restores the syntactic structures. The constituent structures are output from the conversion of the dependency tree parses.

In a more theoretical paradigm is the survey of (Osborne and Liang, 2015). The authors used the dependency-based notion of catena to prove that in spite of the differing types of ellipsis in English and Chinese, the mechanism of analysis is equally suitable. Thus, the preferences of a language to certain types are made visible.

(Schuster et al., 2017) give arguments in favor of introducing distinct nodes for gapping constructions in the enhanced representation of UD guidelines version 2, instead of the previously used relations *remnant* and *orphan*. The similarity with the BTB approach is that they applied a recovery procedure to verbal ellipses. The difference is that in their case one node substituted the head verb and all its missing dependants, whereas in our case various recovery nodes are provided for the head and the dependants.

(Droganova and Zeman, 2017) discuss the varieties in the annotation of ellipsis within the UD treebanks. Their focus is on the dependent promotion when a head is elided. In the statistics survey of ellipsis in 41 treebanks (Table 1, p. 51) Bulgarian is given with a nearly zero representation of orphans (3/2) which is true given the fact that no sentences with ellipsis have been added in the releases. An example sentence is commented on Fig. 10 where the appearance of orphan is considered an error. In this particular case of introducing *orphan* instead of conjoining the two definite adjectives to the noun head, an error occurs because of the colloquial nuance (see the gloss: peaceful-the, political-the efforts). However, this nominal structure typically introduces a new entity. For example, in 'Bulgarian-the and Greek-the government' the governments are actually two referents.

There is also a line of work that describes the steps and challenges when transferring the linguistic information from the Lexical Functional Grammar (LFG) to UD — (Przepiorkowski and Patejuk, 2019). The authors chose to use the f structure for forming dependencies. Their transformation processes include the following stages: moving from LFG to LFG-like dependencies and then – rearranging dependencies. All these processes are not trivial. As it was mentioned before, our treebank is HPSG-based and HPSG-inspired. Thus our conversion started from the constituent structure, enriched with dependencies. Even from such a starting point our transformation suffered from lost information similarly to the one reported in (Przepiorkowski and Patejuk, 2019). This is for example, pro-dropness, control relations, etc.

3 Modeling of Ellipsis in the original BulTreeBank

In the original treebank that is constituency-based, the ellipsis is viewed as an expression that lacks an overt element. This element, however, is presupposed and thus recoverable or easily predicted by the context. Context might refer to many other phenomena: mostly grammar-based ellipsis (as pro-drop), optional arguments (as missing complements of transitive verbs), coreference and anaphora (especially in long texts and wider — including extralinguistic — discourse). Context can also be viewed as local and global.

The complexity of modeling ellipsis is also due to its close relatedness to and interference with linguistic phenomena like coordination and substantivization. The first one often licenses the insertion of recovery nodes in the position of the missing element, while the second licences the strategy of promoting dependants into heads when some head is missing.

As it was mentioned above, in BTB recovery markers for ellipsis were consistently added explicitly for all the modeled elliptical phrases. Thus, the idea is to preserve full syntactic structures.

Ellipsis was introduced through a mechanism of adding a special artificial node at the place of ellipsis, and connecting it with an index to the overt corresponding part (if there is such a part) or connecting it at the sentence level only (if the ellipsis is recoverable in a broader context or from world knowledge). Ellipsis was indicated on two levels: a) syntactic (V-Elip, N-Elip, A-Elip, PP-Elip, Prep-Elip) and b) discourse (VD-Elip, ND-Elip, PrepD-Elip). Please note that verbal ellipsis was briefly discussed in (Osenova and Simov, 2018) in relation to handling enhanced dependencies.

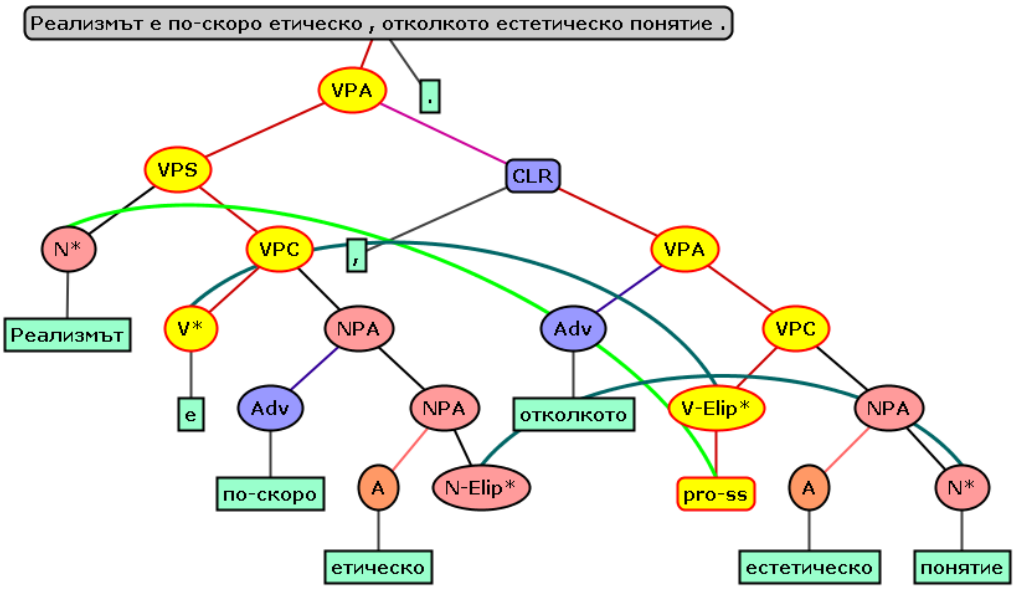

Figure 1: An example of ellipsis as encoded in the original HPSG treebank.

Fig. 1 depicts the representation of the sentence Реализмът е по-скоро етическо, отколкото естетическо понятие. "Realizmyt e po-skoro etichesko, otkolkoto estetichesko ponyatie." (Realism is rather an ethical than an aesthetic concept.). The sentence contains the two main types of ellipses in BTB — verbal and nominal ones. In both cases an empty node is inserted where the elided element has to be present. Thus, in the example there is an ellipsis of the copula and an ellipsis of the noun "concept". In addition to the verbal ellipsis node there is an additional node `pro-ss` representing the unexpressed subject which is coreferent with the subject within the first clause.

In Table 1 it can be seen that the most frequent type in BulTreeBank is the syntactic nominal

ellipsis (N-Elip), immediately followed by the syntactic verbal ellipsis (V-Elip). It is interesting to note that the frequencies of the syntactic and the discourse verbal types are quite similar. The prepositional and adjectival types are rare.

However, it should be kept in mind that since ellipsis is related to other phenomena like substantivization and coordination, in some cases an alternative strategy might have been preferred. These interfering cases are discussed in more detail below.

In the annotation guidelines the coordination has been modeled at the following two levels: lexical and clausal. In both cases the following requirement is posited: if some heads are coordinated, they have to select the same arguments; similarly, if some dependants are coordinated, they have to be selected by the same head.

Substantivization being a promoter of dependants to heads, is used in very limited cases, such as: the missing head refers to general referents (the sick [people]; the three [men], etc.) or in systematic cases (one [student] of the students).

Type of Ellipsis	**Occurrences**
N-Elip	327
V-Elip	262
VD-Elip	255
ND-Elip	70
PP-Elip	12
Prep-Elip	3
PrepD-Elip	2
A-Elip	1

Table 1: Ellipsis Types in the original BTB and the number of their occurrences.

First, let us give some examples for each of the four most frequent types and briefly discuss them – N-Elip, V-Elip, VD-Elip and ND-Elip (examples 1–4). Then more complex cases are considered as well (examples 5–8).

In example (1) the nominal ellipsis is related to nominal phrasal coordination since the definite article in both adjectives (as indicated in the gloss) shows that the entities are different. Thus adjectival coordination is blocked here.

(1) Също така е развита химическата и техническата индустрия .
Also such is developed [chemical-the **N-Elip**] and [technical-the **industry**] .
'Also, the chemical and technical industries have been developed.'

In example (2) verbal ellipsis occurs in the second conjunct of a clausal coordination. The negative particle had not been promoted into a verbal head. Instead, a null node was inserted.

(2) Иван отиде в градината , но Петър не .
Ivan **[went in garden-the]** , but [Peter not **V-Elip**] .
'Ivan entered the garden but Peter did not.'

Example (3) shows a case where the verb 'to be' in present tense is missing. This characteristics is typical for the titles in newspapers. This ellipsis is considered as a kind of discourse ellipsis of subtype 'exist' instead of being analyzed as a truncated construction. The annotation repeats the one for the non-auxiliary verbs, namely inserting an artificial node as a place holder of the missing element.

(3) Социалните центрове пред стачка .
Social centers **VD-Elip [are]** before strike .
'Social centers about to strike.'

In example (4) the discourse nominal ellipsis can be recovered only on the level of the whole article as well as from the local social knowledge. The elided element is the word for Bulgarian currency 'levs'. Again, instead of promoting the numeral as a head, the nominal phrase structure is preserved through the addition of the node ND-Elip.

(4) Дори и с 5000 **ND-Elip [levs]** не бих се чувствала богата .
Even and with 5000 **ND-Elip [levs]** not would se.REFL felt rich .
'Even with 5000 I would not feel rich enough.'

Let us now consider some more complex cases. In example (5) two things are interesting in the clausal coordination structure както..., така и...(as..., in a such a way...), 'Similarly to X, Y did something'. First, the nominal coordination of the nominal subjects — Vulgaris and prince Mihaylo — is not eligible, because in this case the verb (crave for) agrees only with prince Mihaylo. Even if the verb was in plural agreement, adding more dependants around the heads – Vulgaris and prince Mihaylo – would prevent the coordinating of subjects only. Thus, a clausal coordination with restored ellipsis of the whole verbal phrase 'was craving for Bulgarian land' in the first clause is needed.

(5) Както Вулгарис **VD-Elip [craved for Bulgarian lands]** , така и княз Михайло
As Vulgaris **VD-Elip [craved for Bulgarian lands]** , such and prince Mihaylo
ламтеше за български земи .
craved for Bulgarian lands .
'Similarly to Vulgaris, prince Mihaylo was craving for Bulgarian land.'

In example (6) the problem is that the elided verb (there was) is a lexical opposite to the overt one (there was not). Thus, it is not enough just to copy the structural node, but is necessary to also indicate its opposite meaning. In BulTreeBank this was managed by providing subtypes of ellipses: identity in morphology and meaning; difference only in the morphological form; and change of the meaning into its opposite.

(6) Там нямаше заплаха, а **V-Elip [was]** само радост .
There **was**-not threat, but **V-Elip [was]** only joy .
'There was not any threat, but only joy.'

In example (7) a valency-based ellipsis is introduced. The form *да отиде (to go-he)* is missing in the frame 'I order someone to do something'. The overt past form of 'went' is not morphologically identical to the elided structure 'to go'.

(7) Наскоро замина където му бяха наредили **V-Elip [to go]** .
Recently **went-he** where him were ordered **V-Elip [to go]** .
'Recently he went to where he was ordered to go.'

One case that might be reconsidered in favor of UD strategy, is presented in example (8). Here the structure of the NP has been preserved by inserting an N-Elip node. However, such cases might be extended into the application of substantivization strategy:

(8) Който се грижи за хорските работи, хората се грижат за неговите
Who se.REFL cares about people's **things**, people se.REFL care about his
.
N-Elip [things] .
'Who takes care of people's business, people take care about his [business].'

From all the above illustrations it becomes clear that in the original BTB the goal was to maximally restore the clausal structure. Concerning the competition with coordination, the cases were solved with predefined structures that can coordinate only if they have the same selectional restrictions (from both points of view - being heads or being dependants). Concerning substantivization, it might be extended beyond the initially defined cases. There are also cases where the elided material is not related only to phenomena like coordination or substantivization, but also to phenomena like complementation (example 7) or free relatives (example 8).

4 Considering the sentences with ellipsis in UD framework

UD proposes the following strategies for handling ellipsis: a) a surface-based one (in which a special *orphan* relation is used) and b) a recovery-based one (in which null elements for the elided material are used – as in the enhanced dependencies) or promotion from the elided head to its dependants (when present) is introduced. The former relation adheres to surface syntax and thus – to truncated phrases where, in the absence of the head, non-typical relations connect their dependents (ex. in the sentence *John drinks water and Maria wine*, *Maria* and *wine* would be connected by the orphan relation). The latter relations are closer to the strategy that was adopted in the original BTB.

As it became clear already, in BTB the ellipsis has been always recovered, i.e. in this respect it followed somewhat a non-surface-like analysis. A full syntactic analysis of the structures was aimed, thus not considering the idea of having truncated phrases unless they form a typical constituent phrase (eg. [NP Good job!]).

The first type of the UD enhanced dependencies, called 'Null nodes for elided predicates', involves the addition of special null nodes in clauses with an elided predicate. An illustration of this idea is the following sentence: *I go to Varna, and you [V-Elip - go] to Sofia.* In Bulgarian V-Elip differs from the overt element by the category of person only. With this ellipsis recovery, the grammatical relations are maintained also in clauses without an explicit predicate. In BTB such predicates are introduced as V-Elip nodes in an appropriate place in the structure. Thus, this label can be mapped directly into the so-called null nodes. There are two cases of usage of V-Elip — representation of elided single verbal form; and representation of elided phrase — VP-ellipsis. The first case is the more straightforward one. In the second case in UD we need to introduce several null nodes in order to represent the whole VP. In addition to the null nodes in BTB also some variation of the grammatical features are encoded such as change in the number, tense, etc. For the moment it is not clear how to represent this in UD — see example 6. In such cases we encode the modified grammatical features as such for the null nodes.

Also, in UD each verb in a verbal complex is marked with a null node, while in BTB there is only one such substitute node for all the elided material. The principle is that the introduced recovery node refers to the maximal material that is elided.

In contrast to V-Elip, the null nodes annotated with VD-Elip label in BTB provide discourse information that is difficult to identify by type (let alone the form) of the missing element(s). These difficult cases can be processed only manually. Usually the additional information could be recovered within the whole text or on the basis of general knowledge. In this case within UD we could use *orphan* relation, but then the encoded information would be lost. In order to preserve this information, we modify the *orphan* relation in order to specify the value the discourse information. For example, *orphan:cop* is used to represent the case of an elided copula licensed by discourse.

If we put the comparison on a broader scale of approaches and thus — beyond enhanced dependencies, then the following observations are to be made. First of all, the idea of using null elements instead of verbs or verbal groups does not cover all other cases with elided elements in UD. For example, in the nominal ellipsis the elided head is substituted by its dependent (if such a dependant is present). This means that a process of substantivization is performed. A similar promotion strategy holds for auxiliaries. In BTB in such cases null elements have been used (recall examples 2 and 4).

In the case of BTB, the process of substantivization is restricted to: a) adjectives promoted to nouns; b) numerals in the structure 'one of them; three of them', etc. In general, the annotation scheme puts a preference to constituent coordination instead of introducing ellipsis. Constituent coordination, however, applies in cases when the coordinated heads have the same selectional criteria or when dependants are selected in the same way by the same head.

There are two special cases of ellipses in BTB which require more attention. The first case is illustrated in Fig. 1 — the `pro-ss` element could be encoded in two ways in UD: (1) via a new

null node for the subject; or (2) express it via enhanced dependencies. The second is illustrated in Fig. 2 where the second clause contains an explicit marker for the place of the ellipsis (a dash). In UD this dash could be used functionally instead of introducing a null node.

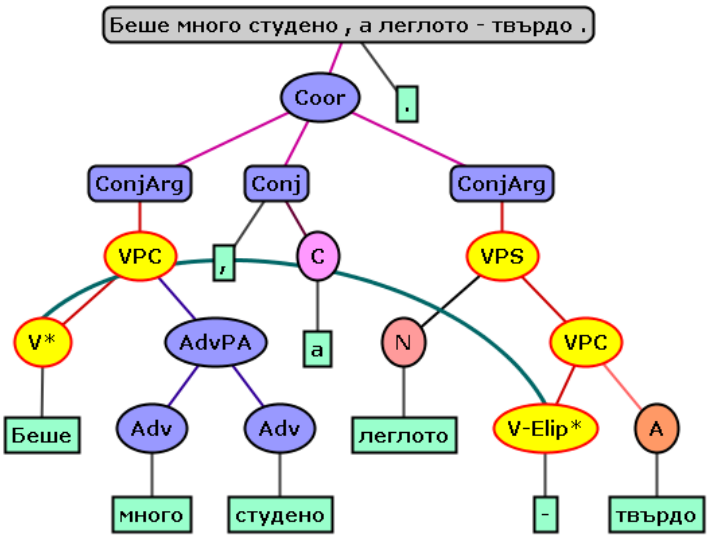

Figure 2: An example of ellipsis marked as a dash in the sentence. Беше много студено, а леглото - твърдо. "Beshe mnogo studeno, a legloto - tvyrdo." (It was very cold, and the bed - firm.)

5 Conclusions

The current general principles behind UD for handling ellipsis are as follows: a) elided element with no dependents is not processed at all; b) if it has dependants, then they are promoted as heads and c) the promoted element uses the relation *orphan* when other functional elements are attached to it. In BTB, besides the systematically applied null-node-insertion-strategy, ellipsis subtypes were added as a specification relation. Substantivation was kept for the lexicalized in the dictionary dependants.

One possible direction of the UD development would be to extend the null node introduction. Another one is to continue with the mixed strategy of treating ellipses in the basic and enhanced dependencies as it is now.

From our point of view, in both cases it would be useful to add more information on the ellipsis type and characteristics, and also to consider language specific features as it was done for other phenomena. For example, in Bulgarian it is not typical to elide the main predicate and to leave the auxiliary/modal to be promoted as it is in English (e.g. *She wants to go there, but he does not* or *She wanted to go there, but he did not want to*).

As discussed in the cited literature here (and beyond it), the proper treatment of ellipsis in an explicit way is important for the mono- and cross-lingual as well as for reasonable typological surveys across languages.

6 Acknowledgements

This research was partially funded by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-164/28.08.2018 and the Bulgarian National Science Fund, Grant number 02/12/2016 — *Deep Models of Semantic Knowledge (DemoSem)*. We would like to thank all reviewers for their valuable feedback. All errors remain our own.

References

Kira Droganova and Daniel Zeman. 2017. *Gapping Constructions in Universal Dependencies v2*, Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pp. 48–57. Gothenburg, Sweden.

Tomáš Jelínek, Vladimír Petkevic, Alexandr Rosen, Hana Skoumalová and Premysl Vítovec. 2015. *Taking Care of Orphans: Ellipsis in Dependency and Constituency-Based Treebanks*, Proceedings of the TLT14, pp. 119–133. Warsaw, Polland.

Marie Mikulova. 2014. *Semantic Representation of Ellipsis in the Prague Dependency Treebanks*, Proceedings of ROCLING 2014, pp. 125—137. Prague, Czech Republic.

Timothy Osborne and Junying Liang. 2014. *A Survey of Ellipsis in Chinese*,Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pp. 271—280. Uppsala, Sweden.

Petya Osenova and Kiril Simov. 2017. *Recent Developments within BulTreeBank*, Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, pp. 129—137. Prague, Czech Republic.

Adam Przepiorkowski and Agnieszka Patejuk. 2019. *From Lexical Functional Grammar to enhanced Universal Dependencies*, Languages Resources and Evaluation, published online: 04 February 2019. Springer.

Sebastian Schuster, Matthew Lamm and Christopher D. Manning. 2017. *Gapping Constructions in Universal Dependencies v2*, Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pp. 123–132. Gothenburg, Sweden.

Kiril Simov, Petya Osenova, Alexander Simov and Milen Kouylekov. 2005. *Design and Implementation of the Bulgarian HPSG-based Treebank*, ournal of Research on Language and Computation. Special Issue, pp. 495–522. Kluwer Academic Publisher.