

CASE MARKING OF THE DIRECT OBJECT IN UZBEK AND ITS REPRESENTATION IN SYNTACTIC ANNOTATION

Khamdamova Dilnoza

Doctoral student of

Karshi state university

DOI: <https://doi.org/10.5281/zenodo.22723369>

Abstract: *The paper examines the annotation of direct objects in Uzbek, focusing on overt and unmarked accusative forms. It analyses their representation within the Universal Dependencies framework and argues that both forms should receive the same obj relation, while accusative marking should be represented at the morphological level. The study also proposes criteria for identifying unmarked objects and compares the Uzbek pattern with English.*

Keywords: *Uzbek, direct object, accusative marking, syntactic annotation, Universal Dependencies, treebanks*

Annotatsiya: *Ushbu maqolada o'zbek tilidagi to'g'ri to'ldiruvchilarning belgilangan (oshkora) va belgisiz tushum kelishigi shakllarida ifodalanishi hamda ularning annotatsiyasi tahlil qilinadi. Tadqiqotda Universal Dependencies (UD) standartlari doirasida har ikki shakl ham sintaktik jihatdan bir xil obj (object) munosabatini olishi kerakligi, tushum kelishigi belgisi esa morfologik satrda (Case=Acc) aks ettirilishi lozimligi asoslab beriladi. Shuningdek, maqolada belgisiz to'ldiruvchilarni aniqlash mezonlari taklif etilib, o'zbek tili strukturasi ingliz tili modeli bilan qiyoslanadi.*

Kalit so'zlar: *O'zbek tili, to'g'ri to'ldiruvchi, tushum kelishigi belgisi, sintaktik annotatsiya, Universal Dependencies, daraxtlar banki (treebank).*

Аннотация: В статье рассматривается аннотирование прямого дополнения в узбекском языке с акцентом на выраженные (оформленные) и невыраженные (неоформленные) формы винительного падежа. Анализируется их репрезентация в рамках модели *Universal Dependencies (UD)*, где аргументируется, что обе формы должны получать одинаковое синтаксическое отношение *obj*, тогда как наличие или отсутствие надежного маркирования должно отражаться на морфологическом уровне (*Case=Acc*). Также предлагаются критерии выявления неоформленных прямых дополнений и проводится сопоставительный анализ узбекской конструкции с английским языком.

Ключевые слова: Узбекский язык, прямое дополнение, показатель винительного падежа, синтаксическая аннотация, *Universal Dependencies*, корпус деревьев (*treebank*)

INTRODUCTION

Syntactically annotated corpora for Uzbek have begun to emerge within the Universal Dependencies framework. Two substantial Uzbek treebanks are currently available. The first, Uzbek-UT, contains 500 sentences and 5,850 tokens drawn from news and fiction [1]. The second, Uzbek-UzUDT, is a manually annotated resource based on literary and educational texts. Its published dataset contains 681 sentences and 7,542 tokens and follows the Universal Dependencies version 2 framework [2]. The availability of these resources makes questions of annotation methodology increasingly important. The use of a particular annotation scheme requires explicit criteria for identifying syntactic relations and assigning the corresponding dependency labels. Such criteria should be sufficiently clear and consistent to allow annotators to apply them to comparable linguistic data.

The direct object presents a relevant case in Uzbek because its syntactic function may be expressed with or without overt accusative marking. The presence or absence of the accusative suffix *-ni* is associated with differences in the interpretation of the object, while the syntactic relation between the object and the predicate remains unchanged. This creates a distinction between morphological marking and syntactic representation that needs to be considered in corpus annotation.

The present paper examines the annotation of direct objects in Uzbek, with particular attention to constructions with overt and unmarked accusative marking. It considers the criteria by which the direct object is identified, the grounds for assigning the *obj* dependency relation, and the extent to which information expressed through case marking is represented in syntactic annotation. The analysis also considers the implications of these issues for the development of consistent annotation guidelines for Uzbek.

The Alternation in Uzbek

The distribution of the accusative suffix *-ni* in Uzbek is conditioned by the properties of the noun phrase and is therefore not entirely free. Pronouns, proper names, demonstrative noun phrases and definite noun phrases generally require overt accusative marking, whereas indefinite noun phrases may occur either with or without the suffix [5]. The range of possibilities is illustrated by the following examples from the same source: (1)

- a. *Ali bu xaritani topdi.*
- b. *Ali bitta xaritani topdi.*
- c. *Ali bitta xarita topdi.*
- d. *Ali xarita topdi.*

In (1a), the demonstrative noun phrase occurs with overt accusative marking. In (1b) and (1c), the indefinite noun phrase occurs with and without the suffix,

respectively. The difference between the two constructions is primarily interpretive rather than syntactic. The marked form favours a specific interpretation, whereas the unmarked form is more readily associated with a non-specific interpretation [5]. In (1d), the bare noun phrase also occurs without overt accusative marking.

The factors underlying this distribution have been examined experimentally. Specificity and animacy, as properties of the object, together with affectedness, which is associated with the event structure of the predicate, have been shown to influence accusative marking in both Turkish and Uzbek. Telicity, by contrast, has not been found to operate as an independent factor [5]. Its contribution to differential object marking is instead related to its interaction with affectedness [5]. For the purposes of syntactic annotation, however, the central point is not the relative contribution of these semantic factors. Rather, the relevant observation is that the noun phrase functions as a direct object in all four constructions. The alternation concerns the morphological realization and interpretation of the object rather than its syntactic status.

This distinction has two consequences for annotation. First, the syntactic function remains constant across the four examples. In each case, the noun phrase is syntactically related to the predicate *topdi* as its direct object. Second, the formal expression of the relation varies. In (1a) and (1b), the object bears overt accusative marking, whereas in (1c) and (1d) accusative marking is not overtly expressed. This contrast is particularly relevant to syntactic annotation because morphological marking cannot serve as a uniform criterion for identifying the object. When overt case marking is present, it provides direct morphological evidence for the analysis. When it is absent, the annotator must rely on other syntactic properties, including the structural position of the noun phrase and the transitivity of the predicate. The Uzbek direct object therefore provides a useful basis for examining how syntactic relations should be identified when morphological evidence is variable.

The Annotation Problem

The Universal Dependencies inventory provides a single dependency relation, *obj*, for direct objects [3]. All four sentences in (1) therefore receive the same syntactic label. This analysis is appropriate because the syntactic function of the noun phrase remains unchanged across the four constructions. The identity of the dependency labels, however, conceals a difference in the evidence available for identifying the relation, and this difference is central to the annotation problem considered here.

When the accusative suffix is overtly expressed, case morphology provides an explicit formal cue for the analysis. The relevant information is directly observable in the word form and can therefore be incorporated into the annotation procedure as morphological evidence. When the suffix is absent, no equivalent morphological cue is available. The annotator must then rely on syntactic evidence, including the position of the noun phrase in relation to the predicate and the transitivity of the verb. These criteria are linguistically motivated, but they require a greater degree of structural interpretation. Uzbek has relatively flexible word order, and although the preverbal position is commonly associated with the direct object, it cannot by itself be treated as an obligatory diagnostic of the relation.

This issue is not unique to the Uzbek object construction. Research on Turkic treebanks has pointed out that the distinction between core and oblique arguments cannot always be established through a simple formal criterion. Since verbal constructions do not necessarily contain an obligatory element and case forms may have a broad range of functions, additional syntactic evidence may be required. One diagnostic proposed in this context is the behaviour of an argument under passivisation or causativisation [4]. Although this observation addresses a broader issue of argument classification, it is relevant to the present discussion because it demonstrates that the

absence of overt morphological marking may require the annotator to use indirect syntactic diagnostics.

These observations have two implications for the formulation of annotation guidelines.

First, the guidelines should specify a hierarchy of criteria for identifying the direct object rather than relying on a single diagnostic. Overt morphological marking should be considered where it is available. When such marking is absent, the analysis should proceed on the basis of syntactic criteria, including the structural position of the noun phrase and the argument structure of the predicate. Stating these criteria explicitly makes the annotation procedure more reproducible because the guidelines specify not only the resulting dependency label but also the evidence supporting the analysis. The reported annotation agreement in the existing Uzbek treebank [2] further demonstrates the importance of formulating criteria that can be applied consistently and evaluated empirically.

Second, the morphological distinction between marked and unmarked objects should be kept separate from the syntactic distinction represented by the dependency relation. Uzbek expresses an important grammatical and semantic contrast through the presence or absence of overt accusative marking, but the UD relation *obj* is designed to represent syntactic function rather than the semantic interpretation associated with that function. Introducing a subtype of *obj* for accusatively marked objects would therefore conflate morphological marking with syntactic dependency. Since the syntactic relation remains the same, the distinction does not provide a basis for a separate syntactic relation.

The distinction is more appropriately represented at the morphological level. Where the accusative suffix is overtly expressed, the corresponding case information can be represented through the morphological feature *Case=Acc*. The dependency

relation remains *obj* in both marked and unmarked constructions. This separation preserves the distinction between morphological form and syntactic function while keeping both types of information available for corpus analysis. Marked objects can consequently be identified through their morphological features without altering the dependency structure. The annotation scheme can thus represent the information at the level to which it properly belongs, while the guidelines specify the criteria used to identify the syntactic relation when overt case marking is absent.

The English Comparison

English provides a useful point of comparison because some of the interpretive distinctions associated with object marking in Uzbek are expressed through different grammatical means. In *Ali found the map* and *Ali found a map*, for example, the noun phrase functions as the direct object in both sentences, while the articles *the* and *a* contribute to its interpretation. Within Universal Dependencies, the articles are analysed as *det* dependents of the noun, whereas the relation between the noun phrase and the verb is represented by *obj* [3].

The two languages therefore differ in the formal realization of information associated with the object. In English, the relevant marker is an independent word that receives its own syntactic relation within the dependency structure. In Uzbek, accusative marking is expressed through the suffix *-ni*, which is represented as morphological information associated with the noun. The distinction is consequently accessible in the annotation of both languages, but it is represented at different levels according to the grammatical structure of each language.

The comparison also highlights a difference in the evidence available for identifying the object relation. In English, word order provides the principal syntactic evidence because the direct object normally follows the verb and full noun phrases do not carry overt accusative case marking. In Uzbek, overt accusative marking provides

an additional morphological cue when it is present, while syntactic evidence becomes particularly important when the object is unmarked. The hierarchy of criteria proposed for Uzbek should therefore be understood as a language-specific response to the formal resources available in the language rather than as a universal procedure for object identification.

This comparison also illustrates a broader principle of multilayer annotation. Information that is not represented by a particular syntactic relation is not necessarily absent from the corpus. Different linguistic properties may be distributed across different annotation layers. In English, information associated with definiteness is represented partly through the dependency structure because articles are independent syntactic units. In Uzbek, information associated with accusative marking is represented morphologically because the marker is an affix. A consistent annotation scheme should therefore preserve the distinction between the information being represented and the structural layer through which it is expressed.

CONCLUSION

The alternation between marked and unmarked direct objects in Uzbek does not require a separate syntactic relation because the syntactic function of the noun phrase remains constant across the two constructions. The alternation nevertheless has important consequences for the formulation of annotation guidelines.

First, the guidelines should establish an explicit hierarchy of criteria for identifying the direct object. When overt accusative marking is present, it provides morphological evidence supporting the analysis. When the suffix is absent, the annotator must rely on syntactic evidence, including the structural position of the noun phrase and the transitivity and argument structure of the predicate. Making these criteria explicit increases the reproducibility of annotation and provides a basis for evaluating the consistency of annotation decisions. This is particularly relevant given the

annotation procedures and agreement assessment reported for the existing Uzbek treebank [2].

Second, the distinction between marked and unmarked objects should be represented at the morphological level rather than through a subtype of the *obj* relation. The latter would treat a difference in morphological realization as though it represented a difference in syntactic function. Maintaining *obj* for the syntactic relation while recording overt accusative marking as morphological information preserves the distinction between these two levels of analysis.

The comparison with English further demonstrates that the location of information within an annotation scheme depends partly on the grammatical form through which that information is expressed. English articles are independent words and can therefore receive syntactic relations such as *det*, whereas the Uzbek accusative marker is an affix and belongs to morphological annotation. The same general principle applies in both cases. Annotation should represent linguistic information at the structural level appropriate to its formal realization without altering the analysis of syntactic relations.

The conclusions proposed here are based on the Uzbek constructions discussed in the literature and on the annotation principles of Universal Dependencies. Their applicability to larger quantities of naturally occurring Uzbek text remains to be evaluated. The existing Uzbek treebanks [1; 2] provide an appropriate basis for such an evaluation, particularly for testing the proposed hierarchy of criteria and examining whether the explicit formulation of these criteria improves annotation consistency.

References

1. Akhundjanova, A., & Talamo, L. (2025). Universal Dependencies Treebank for Uzbek. In Š. A. Holdt, N. Ilinykh, B. Scalvini, M. Bruton, I. N. Debess, & C. M. Tudor (Eds.), *Proceedings of the Third Workshop on Resources and*

Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025) (pp. 1–6). Tallinn: University of Tartu Library.

2. Matlatipov, S., Aripov, M., Bobokandov, M., & Matlatipov, G. (2026). Annotated universal dependencies dataset for literary and educational Uzbek texts. *Data in Brief*, 66, 112857. <https://doi.org/10.1016/j.dib.2026.112857>

3. de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), 255–308. https://doi.org/10.1162/coli_a_00402

4. Tyers, F. M., Washington, J., Çöltekin, Ç., & Makazhanov, A. (2017). An Assessment of Universal Dependency Annotation Guidelines for Turkic Languages. In *Proceedings of the 5th International Conference on Turkic Languages Processing (TurkLang 2017)* (pp. 276–297). Kazan: Tatarstan Academy of Sciences.

5. Kizilkaya, S., Levy-Forsythe, Z., & von Heusinger, K. (2022). Affectedness and Differential Object Marking in Turkish and Uzbek. *Linguistics*, 60(6), 1907–1941. <https://doi.org/10.1515/ling-2020-0216>