TOKENS AT THE BOUNDARY OF THE TAGSET IN A MANUALLY ANNOTATED UZBEK MICROCORPUS

Abdullayev Ikramjon Khashimdjanovich
PhD researcher, Department of English Language and Literature, Andijan State Institute of Foreign Languages, Andijan, Uzbekistan
ORCID 0009-0007-9353-1036 ·
ikramjona@inbox.ru · tel. (+99891) 605-15-51
Section: Linguistics and Translation Studies

ARTICLE INFO

Received: 04st September 2026
Accepted: 06th September 2026
Online: 08th September 2026

KEYWORDS

part-of-speech annotation, transposition, tagset design, residual category, inter-layer classification, Uzbek corpus

ABSTRACT

Annotation projects normally report what a tagset covers. This paper reports what it does not. In a manually annotated 12,048-token microcorpus of written Uzbek, 575 tokens carry a transposition flag; each was audited against five functional directions of transposition established for Uzbek. Five hundred and twelve tokens (89.0 per cent) matched one direction; 63 (11.0 per cent) were kept as an explicit residual instead of being assigned to the nearest available label. Setting the formal pre-classification, built from tags and morphological features, against the contextual audit shows that formal cues predict function unevenly: the dominant direction lost 6.5 per cent of its members, while the two peripheral directions lost 31.5 and 51.6 per cent. The residual is not confined to one style (31 / 16 / 14 across fiction, academic and journalistic samples) and draws on four source categories. The size and internal composition of a residual class is therefore a measurable property of an annotation scheme, and reporting it openly protects the counts that follow rather than weakening them.

Introduction

Part-of-speech schemes are usually assessed by coverage and by inter-annotator agreement. Both measures reward a tagset that assigns every token to some class. What happens to the tokens that fit no class is rarely stated: in most projects they are resolved silently, either by a default rule or by the annotator's preference, and disappear into the counts. The decision is invisible in the published figures, yet it moves them.

The present study takes the opposite route. In a pilot microcorpus of written Uzbek, tokens that met none of the criteria of the classification were kept apart and counted. This paper describes what that residual contains, how large it is relative to the classified material, and what its composition indicates about the scheme that produced it. The question is not whether the residual could have been eliminated. It is what is learned by refusing to eliminate it.

Materials and methods

The corpus contains 12,048 tokens drawn in roughly equal parts from three functional styles: fiction 4,006, academic prose 4,004 and journalistic writing 4,038. Every token carries

two part-of-speech layers – a national tag reflecting the Uzbek grammatical tradition and a universal tag following the Universal Dependencies inventory [3] – together with lemma, morphological features and a context field holding the full sentence. A third field records the token's categorial status: stable use, boundary status, or transposition. Transposition covers 575 tokens, that is 4.8 per cent of the corpus.

Classification proceeded in two stages. A formal pre-classification first grouped the 575 tokens by their tags, features and annotation comments. Each token was then re-examined in its full sentence against five functional criteria: attributive modification of a following noun (adjectivisation), adverbial specification of a verb or of a whole predication (adverbialisation), independent nominal argument in the position of an elided head (substantivisation), grammatical linkage between two predicative or nominal parts (conjunctionalisation), and the addition of emphasis, restriction or interrogation (particlisation). These five directions follow the classification proposed for Uzbek by G.N.Shodiyeva [1; 2]. A token satisfying none of them was recorded as mismatched, undecidable, or outside the five, and was not forced into the nearest group. An independently executed second pass over all 575 tokens returned the same distribution.

Results and discussion

The two stages diverge, and they diverge unequally. Table 1 sets them side by side.

Table 1. Formal pre-classification and contextual audit of 575 transposition tokens

Functional direction	Formal	Audit	Change
Adjectivisation	490	458	−6.5 %
Conjunctionalisation	54	37	−31.5 %
Adverbialisation	31	15	−51.6 %
Substantivisation	0	2	+2
Particlisation	0	0	0
Residual	0	63	+63
Total	575	575	0

The dominant direction is barely affected by the move from form to context, losing under seven per cent of its members. The peripheral directions lose a third and a half respectively. In our reading this is a property of the material rather than a defect of the pre-classification. Where a function is frequent and structurally regular, its formal signature is reliable. Where it is rare, the same signature carries several readings at once, and the tags alone cannot separate them. The adverbial marker -dek illustrates the point – bound to a verb it specifies manner, bound to a noun it modifies, and the tag is identical in both.

The residual itself is not homogeneous. Sixty-one tokens were mismatched, that is, the transposition flag itself proved too generous; one was undecidable for want of context; and one fell outside the five directions altogether, approaching verbalisation in an analytic verb construction. Their source categories are four: noun 42, verb 17, adjective 3, adverb 1. What unites them is negative – no criterion applied – and this is precisely why a sixth label would have been misleading. Grouping them under a name would have implied a shared function that the evidence does not support.

The largest subgroup repays a closer look. Forty-one nominative nouns had been placed with adjectivisation on formal grounds. In Uzbek an attributive noun requires juxtaposition and a head noun immediately after it; in these forty-one the following token is a finite or defective verb (25 cases), a postposition (8 cases), or a particle, conjunction, determiner pronoun or adverb (8 cases). The nouns therefore stand in the ordinary functions of their own class – subject, object, or complement of a postposition — and no shift has occurred. In *«Illokutiv akt» tushunchasi pragmalingvistikada inson tomonidan nutqiy jarayonda ... bajariladigan harakat sifatida talqin qilinadi* (ID 4858) the noun *inson* is governed by *tomonidan*; in *Fizikadan o‘quv tajribalar tajriba asosida o‘qitishni tashkil qilish* (ID 7069) *tajriba* is governed by *asosida*; in *qay yo‘l bilan bo‘lmasin, planni bajarsa* (ID 273) *yo‘l* is governed by *bilan*. In each of the three, substituting an adjective for the noun destroys the construction instead of shifting its meaning. That substitution test was applied throughout the audit.

Across styles the residual is distributed 31 in fiction, 16 in academic prose and 14 in journalism. The concentration in fiction follows from the material – converb constructions and dialogue produce more borderline readings – but no style is free of it. A residual of this kind is therefore not an artefact of one text type. It marks the places where the scheme, applied honestly, runs out of criteria.

Two consequences follow for annotation practice. Reporting a residual makes the remaining counts stronger, not weaker, because every figure then rests on tokens that met a stated criterion rather than on tokens that were assigned by default. And the proportion of the residual is diagnostic: it locates the directions where form and function come apart, which is where a scheme needs sharper rules or a further layer. On this corpus that place is not the centre of transposition but its periphery.

Conclusion

Of 575 transposition tokens in a manually annotated Uzbek microcorpus, 512 (89.0 per cent) fall into five functional directions and 63 (11.0 per cent) fall outside all of them. Keeping the second group visible costs one line in a table and buys a verifiable basis for every other figure in the study. The comparison of formal and contextual classification shows where that cost is incurred: formal cues serve the dominant direction well and the peripheral ones poorly. We would suggest that the size and composition of a residual class be reported as a standard property of an annotation scheme, alongside coverage and agreement, since it measures something neither of them does.

References:

1. Shodiyeva G.N. O‘zbek tilida so‘z turkumlari transpozitsiyasining o‘ziga xosligi // IQRO. – 2023.–Vol.2, No. 2. – P. 353–356. – URL: https://wordlyknowledge.uz/index.php/iqro/article/view/632
2. Shodiyeva G.N. O‘zbek va ingliz tilida so‘z turkumlari tasnifi va transpozitsiyasi: PhD diss. – Farg‘ona, 2025. – 154 p.
3. de Marneffe M.-C., Manning C.D., Nivre J., Zeman D. Universal Dependencies // Computational Linguistics. – 2021. – Vol. 47, No. 2. – P. 255–308.
4. Akhundjanova A., Talamo L. Universal Dependencies Treebank for Uzbek // Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and

Domains (RESOURCEFUL-2025). – Tallinn: University of Tartu Library, 2025. – P. 1–6. – URL: https://aclanthology.org/2025.resourceful-1.1/
5. Akhundjanova A., Akkurt F., Chontaeva B., Eslami S., Çöltekin Ç. Parallel Universal Dependencies Treebanks for Turkic Languages // Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025). – Ljubljana: Association for Computational Linguistics, 2025. – P. 129–136. – URL: https://aclanthology.org/2025.udw-1.14/
6. Bobojonova L., Akhundjanova A., Ostheimer P.S., Fellenz S. BBPOS: BERT-based Part-of-Speech Tagging for Uzbek // Proceedings of the First Workshop on Language Models for Low-Resource Languages. – Abu Dhabi: Association for Computational Linguistics, 2025. – P. 287–293. – URL: https://aclanthology.org/2025.loreslm-1.23/