NoDaLiDa / Baltic-HLT
RESOURCEFUL-2025
Tallinn, Estonia, 2 March 2025

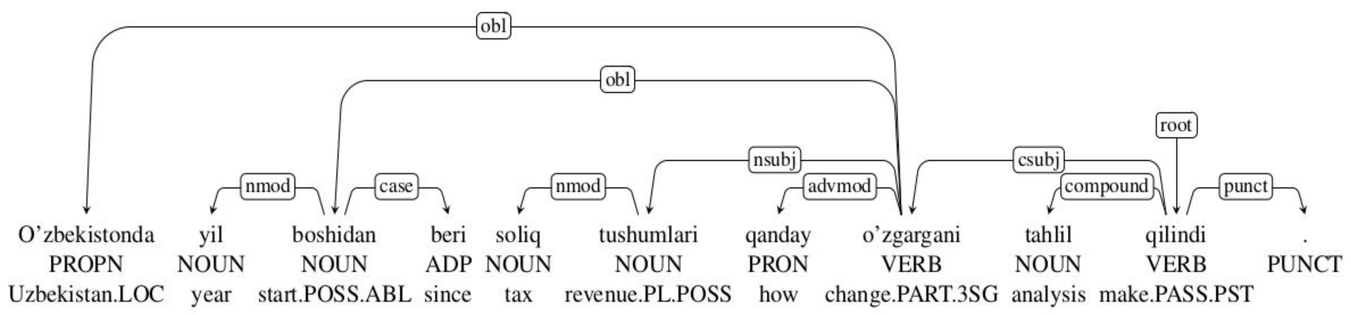
UNIVERSITÄT DES SAARLANDES

Universal Dependencies Treebank for Uzbek

Arofat Akhundjanova, Luigi Talamo
Saarland University

Overview

Uzbek, the second most spoken Turkic language with over 40 million speakers, is a null-subject, agglutinative language with SOV word order and no gender distinctions or articles (Boeschoten, 2021). **Uzbek-UT** includes 500 sentences (5850 tokens) in Latin script: 250 from news articles and 250 from fiction books. The treebank captures diverse domains, levels of formality, and stylistic variation in modern Uzbek.



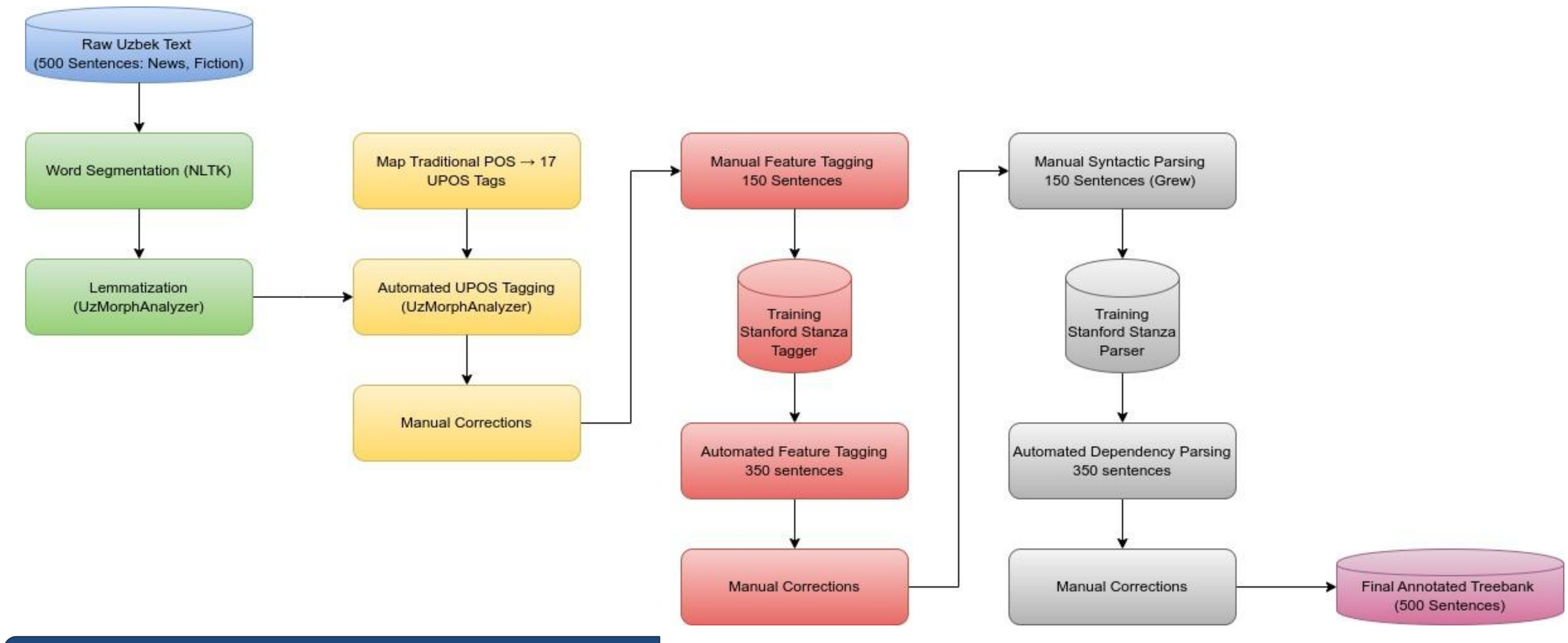
'How tax revenues have changed in Uzbekistan since the beginning of the year was analyzed.'

Figure 1: UD annotation of an Uzbek sentence

	Sentences	Tokens	Unique words	POS tags	Features	Dependencies
No.	500	5850	3523	17	42	32

Table 1: Basic statistics for the UT treebank.

Methodology



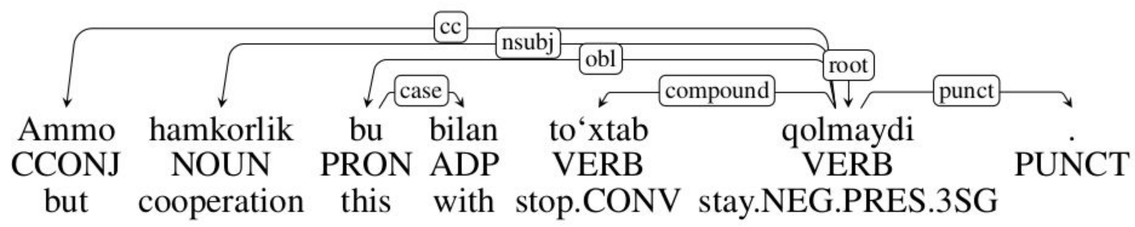
Challenging Constructions

UPOS Tagging Challenge: ***Particle + Verb* MWEs**:

- Particles (e.g., *tashkil* in *tashkil qil* 'establish') lack standalone meaning and POS outside MWEs.
- UD analysis requires separate tokens (PART + VERB), unlike traditional Uzbek grammar.
- **Challenge**: Absence of a comprehensive MWE dictionary; frequent dictionary lookups needed.

Syntactic Relations Challenge: Postverbial Constructions with Auxiliaries:

- Complex verbal phrases formed by a converb and auxiliary verb with strong semantic fusion, e.g. *to'xtab qol* ('lit.: stopping stay) 'end/finish' as in Fig. 2.
- Annotation inconsistencies across Turkic UD treebanks.
- Proposed solution: Introduce **compound:postverb** subtype.
- **Current Approach**: Use **compound** relation, with the postverbial element as the head.



'But the cooperation does not end with this.'

Figure 2: Annotation for the postverbial construction *to'xtab qol*.

Contact information:
arofat.akhundjanova@gmail.com

References:
Hendrik Boeschoten. 2021. Uzbek. In Lars Johanson and Éva Á. Csató, editors, The Turkic languages, pages 388–408. Routledge.
Ulugbek Salaev. 2023. Modeling morphological analysis based on word-ending for Uzbek language.
Shavkat Rahmatullayev. 2006. Hozirgi Adabiy O'zbek Tili [Contemporary Literary Uzbek Language (textbook)]. Universitet.