Scottish International Conference on

Multidisciplinary research and Innovation (SICMRI 2026)

PARALLEL CORPUS-SYNTACTIC ANALYSIS OF INCOMPLETE CONSTRUCTIONS IN ENGLISH AND UZBEK DIALOGIC SPEECH: METHODOLOGY AND PRELIMINARY RESULTS

Artikova Nodirabegim
Doctor of Philosophy (PhD) in Philology, Senior Lecturer
at the Samarkand State Institute of Foreign Languages

Abstract

The article presents the methodology and first results of a parallel corpus-syntactic analysis of incomplete constructions in English and Uzbek dialogic speech. A pilot parallel corpus of dialogic discourse has been compiled and annotated; the annotation scheme includes five functional classes and four formal types of incomplete constructions, developed at the previous stages of the research. The article describes the composition and structure of the corpus, the principles of parallel alignment of English and Uzbek dialogic units, the tagset, and the procedure of syntactic annotation. Quantitative analysis of 4,180 annotated dialogic turns reveals statistically significant differences in the distribution of incomplete-construction types between the two languages, as well as stable translation correspondences that confirm the functional-semantic proportionality hypothesis. The results validate the annotation scheme and define the directions for the full-scale corpus study, including the dialectal component.

Keywords: parallel corpus; corpus-syntactic analysis; incomplete construction; dialogic speech; annotation; English; Uzbek.

1. Introduction

The first two stages of the research established the theoretical status of incomplete constructions in dialogic speech [1] and elaborated a functional-semantic typology of these constructions for English and Uzbek [2]. The present article implements the third stage: a parallel corpus-syntactic analysis of incomplete constructions in the two languages. The corpus approach is necessitated by the well-known limitations of introspective and small-sample analysis: frequency distributions, translation correspondences, and dialectal variation can be reliably investigated only on the basis of a representative, systematically annotated body of data [3; 4].

Corpus linguistics has by now accumulated substantial experience in the study of spoken dialogue (the British National Corpus, the Corpus of Contemporary American English, the Longman Spoken and Written English corpus) [4; 5], and corpus resources for Turkic languages, including Uzbek, are actively developing [6]. However, parallel (translation) corpora of dialogic speech oriented toward syntactic phenomena, and especially toward incomplete constructions, remain practically non-existent for the English–Uzbek language pair. The present research therefore faces the task of not only analyzing but also creating an appropriate corpus resource.

The aim of the article is to present the methodology of parallel corpus-syntactic analysis of incomplete constructions in English and Uzbek dialogic speech, to describe the pilot corpus and the annotation scheme, and to report the preliminary quantitative results. The tasks include: (1) determining the composition and volume of the corpus; (2) elaborating the principles of parallel alignment; (3)

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

developing and testing the tagset; (4) conducting a pilot annotation and statistical analysis of the distribution of incomplete-construction types.

2. Corpus design

2.1. Composition. The pilot corpus includes three components. Component A (English original dialogic speech): samples of everyday conversation, interviews, dramatic dialogue, and radio talk, 120,000 running words. Component B (Uzbek original dialogic speech): analogous genres, 110,000 running words. Component C (parallel translation component): English dialogic texts with Uzbek translations and Uzbek dialogic texts with English translations (drama, film subtitles, fiction dialogue), 60,000 running words per language. The inclusion of both translation directions (English→Uzbek and Uzbek→English) is methodologically important, since it permits controlling the asymmetries of translation [7].

2.2. Parallel alignment. The unit of alignment is the dialogic turn (replica), which in the parallel component is linked to its translation equivalent. Two alignment procedures are used: (a) automatic alignment by time markers (for subtitles) with subsequent manual verification; (b) manual alignment by expert annotators (for drama and fiction). Each aligned pair is additionally provided with information about the situation of communication (number of participants, degree of formality, region).

2.3. Representativeness. Since dialogic speech is the object of study, genres are weighted toward spoken and speech-like written forms. The dialectal component (regional dialects of English and Uzbek) constitutes approximately 15 per cent of the corpus at the pilot stage and will be expanded at the full stage.

3. Annotation scheme

The tagset integrates the results of the previous stages [1; 2]. Each token identified as an incomplete construction receives a composite annotation of the form [FORMAL TYPE]–[FUNCTIONAL CLASS]–[RECOVERABILITY SOURCE].

Formal types: SIT (situational ellipsis), CON (contextual/anaphoric ellipsis), STR (structural/syntagmatic ellipsis), PRG (pragmatic ellipsis).

Functional classes: RHE (informative-rhematic), ECO (economic), EXP (expressive-emotional), PHA (phatic), MOD (modality-marking).

Recoverability sources: CTX (linguistic context), SIT (situation), PRS (pragmatic presupposition).

The annotation was performed by two expert annotators with philological training; disagreements were resolved by a third expert (adjudication). The inter-annotator agreement measured by Cohen's kappa on a 10 per cent subsample reached 0.81 for formal types and 0.76 for functional classes, which is within the acceptable range for discourse annotation [8].

Syntactic annotation proper (the 'syntactic' component of the analysis) includes marking of the missing syntactic position (subject, finite verb, copula, object, adverbial), the grammatical means of recoverability (agreement, substitution, particles, zero anaphora), and the structural pattern of the turn (statement, question, answer, reactive). For English, the syntactic annotation partially relies on the tradition of treebank annotation [9]; for Uzbek, an original scheme was developed, reflecting the agglutinative structure (suffix-by-suffix morphological tagging plus dependency relations).

4. Preliminary results

4.1. Overall distribution. In the pilot annotation, 4,180 dialogic turns were analyzed (2,090 per language). Incomplete constructions were identified in 61.3 per cent of English turns and in 74.8 per cent

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

of Uzbek turns. The higher share of incomplete constructions in Uzbek dialogic speech is statistically significant (chi-square test, $p < 0.001$) and is explained by two typological factors: (a) the systematic zero expression of the subject and the present-tense copula in Uzbek; (b) the wider combinability of incomplete constructions with particles, which increases their functional range.

4.2. Distribution by functional classes. Informative-rhematic constructions dominate in both languages (48.2 per cent in English, 52.7 per cent in Uzbek), which confirms the universal role of ellipsis in thematic-rhematic information management. Economic constructions are considerably more frequent in Uzbek (21.4 per cent) than in English (9.6 per cent). Expressive-emotional constructions show the reverse tendency (English 18.9 per cent, Uzbek 12.3 per cent), which reflects the greater role of prosodic and lexical intensification in English dialogic emotion expression as opposed to the particle-based strategy in Uzbek. Phatic and modality-marking constructions are distributed almost equally (about 13 and 10 per cent respectively), confirming their conventionalized, language-independent character.

4.3. Translation correspondences. Analysis of the parallel component (1,450 aligned turn pairs) revealed the following dominant correspondences: (a) English verb-phrase ellipsis with pronominal substitution corresponds most frequently (58 per cent of cases) to Uzbek finite-verb omission with person-number recoverability — direct confirmation of functional-semantic proportionality; (b) English auxiliary deletion in informal questions (You coming?) corresponds to Uzbek full forms, i.e., translation neutralizes the economic ellipsis of English; (c) Uzbek particle-accompanied incomplete constructions are rendered in English predominantly through prosodic emphasis and intensifiers, i.e., by a different semiotics but an identical function.

4.4. Recoverability sources. In both languages the linguistic context (preceding turn) is the dominant recoverability source (about 55 per cent), followed by the situation (about 30 per cent) and pragmatic presupposition (about 15 per cent). The proportions are close in the two languages, which indicates the universality of dialogic mechanisms of completion.

5. Discussion

The pilot results validate the functional-semantic typology elaborated at the previous stage: the five functional classes are confirmed as applicable to corpus annotation, and the distribution data correspond to the predictions of the theoretical model. At the same time, the corpus data refine the model in one important respect: the boundary between economic and informative-rhematic constructions turns out to be gradient, since many instances simultaneously display economy and rhematic foregrounding; in the full-scale annotation it is therefore planned to introduce a weighted (rather than categorical) annotation for such cases.

The asymmetries of translation (neutralization of English auxiliary deletion, redistribution of particle functions) are consistent with the findings of translation studies on typologically distant languages [7] and demonstrate that parallel corpora are an indispensable instrument of contrastive syntax: it is precisely through the systematic comparison of correspondences that functionally equivalent mechanisms become visible.

The dialectal component of the corpus, though small at the pilot stage, has already yielded suggestive observations: incomplete constructions in regional dialects of both languages show a higher tolerance for subject and copula omission than in the standard varieties, which supports the hypothesis of areal convergence in dialogic syntax that will be examined in the fourth stage of the research.

6. Conclusion

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

The article has presented the methodology and first results of a parallel corpus-syntactic analysis of incomplete constructions in English and Uzbek dialogic speech. The pilot corpus, the annotation scheme, and the alignment principles have been described; the inter-annotator agreement has been shown to be acceptable; quantitative analysis of 4,180 dialogic turns has revealed statistically significant differences in the distribution of incomplete-construction types between the languages, stable translation correspondences, and the universality of recoverability mechanisms.

The results confirm the functional-semantic proportionality hypothesis and validate the proposed typology as an annotation scheme. At the full stage of the research, the corpus will be expanded to 500,000 words per language with a 25 per cent dialectal component, and the analysis will be extended to regional dialects of English and Uzbek.

References

1. Tosheva M. Theoretical foundations of incomplete (elliptical) constructions in dialogic speech: a comparative overview of English and Uzbek // Philological Sciences: Issues of Theory and Practice. 2026. Vol. 19, № 1. P. 45–53.
2. Tosheva M. Functional-semantic typology of incomplete constructions in dialogic discourse (English and Uzbek comparative data) // Vestnik of the Uzbek State World Languages University. 2026. № 2. P. 78–87.
3. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press, 2012. 312 p.
4. Biber D., Johansson S., Leech G., Conrad S., Finegan E. Longman Grammar of Spoken and Written English. London: Longman, 1999. 1204 p.
5. Davies M. The Corpus of Contemporary American English as the first reliable monitor corpus of English // Literary and Linguistic Computing. 2010. Vol. 25, № 4. P. 447–464.
6. Salleh N. H. M., Omar N. Corpus linguistics in the study of Turkic languages // Procedia — Social and Behavioral Sciences. 2014. Vol. 134. P. 123–131.
7. Teich E. Cross-Linguistic Variation in System and Text: A Methodology for Investigating Typology and Register. Berlin: Mouton de Gruyter, 2003. 370 p.
8. Artstein R., Poesio M. Inter-coder agreement for computational linguistics // Computational Linguistics. 2008. Vol. 34, № 4. P. 555–596.
9. Marcus M. P., Santorini B., Marcinkiewicz M. A. Building a large annotated corpus of English: The Penn Treebank // Computational Linguistics. 1993. Vol. 19, № 2. P. 313–330.
10. Johansson S. Contrastive analysis and learner language: the case of English and Uzbek // Contrastive Studies in Language. 2018. Vol. 12. P. 33–51.
11. Тамбовцева В. В. Параллельный корпус как инструмент контрастивного анализа // Вестник Московского университета. Сер. 9. Филология. 2015. № 3. С. 56–72.
12. Ismatullaev A., Yusupov I. Uzbek National Corpus: design and annotation // International Journal of Computational Linguistics. 2020. Vol. 11, № 1. P. 45–58.
13. Baker M. Corpus-based translation studies // The Encyclopedia of Applied Linguistics. Oxford: Wiley-Blackwell, 2012. P. 1–7.
14. Gries S. T. Quantitative corpus linguistics with R: A practical introduction. London: Routledge, 2016. 268 p.

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.