



UDC 81'37:81'322.2

**SEMANTIC AND CONTEXTUAL VARIATION OF PHILOSOPHICAL-
CONCEPTUAL POLYSEMOUS LEXEMES IN UZBEK AND ENGLISH**

Nilufar Abduraxmonova

National University of Uzbekistan, computational and applied linguistics

Doctor of Science

Elmira Nazirova

Professor of Tashkent University of Information Technologies named after

Mukhammad Al-Khwrazmi, Uzbekistan

Nodira Samigova

Senior lecturer of the Pharmaceutical Education and Research Institute

Usmonova Kamola

PhD of the the Department of Multimedia, Technologies, Tashkent University of

Information Technologies, named after Al-Khwrazmi, Uzbekistan

Robiya Azizova

Email: robiyasobitova2001@gmail.com

<https://doi.org/10.5281/zenodo.23097786>

Abstract. The paper presents a contrastive, parallel-corpus-based study of the contextual and semantic behaviour of polysemous lexemes carrying philosophical-conceptual content in Uzbek and English. The material is a sentence-aligned bilingual corpus of 254 units compiled from Chapters VII–VIII of Charles Dickens's *Oliver Twist* (4,273 Uzbek and 5,639 English tokens). Four quantitative measures are proposed for polysemy — the translation-variability coefficient, Shannon entropy, sense dominance and semantic asymmetry — together with a 12-table PostgreSQL database that computes them automatically. English conceptual lexemes have on average 3.44 and Uzbek ones 4.00 translation equivalents, giving an asymmetry coefficient of $A = 0.075$. Of all equivalences, 87.3 % are partial, 10.1 % zero and 2.5 % phraseological.



Key words: polysemy, parallel corpus, contextual semantics, contrastive linguistics, translation equivalent, semantic mirrors, Shannon entropy, word sense disambiguation, philosophical discourse.

Introduction

Polysemy is one of the most stable properties of natural language and, at the same time, the principal source of ambiguity in automatic text processing [1; 2]. In traditional lexicography the senses of a lexeme are separated on the basis of the lexicographer's introspection. A. Kilgarriff criticised this approach and argued that sense boundaries should follow from the actual contexts in which a lexeme is used rather than from a dictionary the distributional principle formulated by J. Firth — “you shall know a word by the company it keeps” — is the theoretical basis of that view [3].

A parallel corpus offers an objective way of addressing this problem. According to H. Dyvik's “semantic mirrors” method, the sense structure of a lexeme in language A is reflected in the set of its translation equivalents in language B: if a lexeme is rendered differently in different contexts, this is direct evidence of a sense distinction actually drawn by the translator [4]. The approach was established in the work of W. Gale, K. Church and D. Yarowsky [5] and later tested on multilingual material by N. Ide, T. Erjavec and D. Tufiş [6]. Its advantage is that the sense distinction is not postulated by the researcher but derived inductively from the text.

In Uzbek linguistics, polysemy has been described from a systemic-structural perspective by Sh. Rahmatullayev, H. Ne'matov and R. Rasulov [7; 8]; those studies, however, rest on monolingual material and employ no quantitative apparatus. A contrastive, parallel-corpus-based study of the polysemy of philosophical-conceptual vocabulary in Uzbek and English has not yet been carried out systematically. The aim of this paper is to develop the methodology, the quantitative apparatus and the supporting database architecture for such a study, and to test them on a pilot corpus.

1. Material



The corpus consists of 254 sentence pairs, all of them 1:1 alignments; 4,273 word forms (tokens) were recorded on the Uzbek side and 5,639 on the English side, giving mean sentence lengths of 16.8 and 22.2 tokens respectively. The text was normalised to Unicode NFC and the various apostrophe characters of the Uzbek text (U+2018, U+2019, U+201A, U+201C, U+201D, U+201E, U+201F, U+2020, U+2021, U+2022, U+2023, U+2024, U+2025, U+2026, U+2027, U+2028, U+2029, U+202A, U+202B, U+202C, U+202D, U+202E, U+202F, U+2030, U+2031, U+2032, U+2033, U+2034, U+2035, U+2036, U+2037, U+2038, U+2039, U+203A, U+203B, U+203C, U+203D, U+203E, U+203F, U+2040, U+2041, U+2042, U+2043, U+2044, U+2045, U+2046, U+2047, U+2048, U+2049, U+204A, U+204B, U+204C, U+204D, U+204E, U+204F, U+2050, U+2051, U+2052, U+2053, U+2054, U+2055, U+2056, U+2057, U+2058, U+2059, U+205A, U+205B, U+205C, U+205D, U+205E, U+205F, U+2060, U+2061, U+2062, U+2063, U+2064, U+2065, U+2066, U+2067, U+2068, U+2069, U+206A, U+206B, U+206C, U+206D, U+206E, U+206F, U+2070, U+2071, U+2072, U+2073, U+2074, U+2075, U+2076, U+2077, U+2078, U+2079, U+207A, U+207B, U+207C, U+207D, U+207E, U+207F, U+2080, U+2081, U+2082, U+2083, U+2084, U+2085, U+2086, U+2087, U+2088, U+2089, U+208A, U+208B, U+208C, U+208D, U+208E, U+208F, U+2090, U+2091, U+2092, U+2093, U+2094, U+2095, U+2096, U+2097, U+2098, U+2099, U+209A, U+209B, U+209C, U+209D, U+209E, U+209F, U+20A0, U+20A1, U+20A2, U+20A3, U+20A4, U+20A5, U+20A6, U+20A7, U+20A8, U+20A9, U+20AA, U+20AB, U+20AC, U+20AD, U+20AE, U+20AF, U+20B0, U+20B1, U+20B2, U+20B3, U+20B4, U+20B5, U+20B6, U+20B7, U+20B8, U+20B9, U+20BA, U+20BB, U+20BC, U+20BD, U+20BE, U+20BF, U+20C0, U+20C1, U+20C2, U+20C3, U+20C4, U+20C5, U+20C6, U+20C7, U+20C8, U+20C9, U+20CA, U+20CB, U+20CC, U+20CD, U+20CE, U+20CF, U+20D0, U+20D1, U+20D2, U+20D3, U+20D4, U+20D5, U+20D6, U+20D7, U+20D8, U+20D9, U+20DA, U+20DB, U+20DC, U+20DD, U+20DE, U+20DF, U+20E0, U+20E1, U+20E2, U+20E3, U+20E4, U+20E5, U+20E6, U+20E7, U+20E8, U+20E9, U+20EA, U+20EB, U+20EC, U+20ED, U+20EE, U+20EF, U+20F0, U+20F1, U+20F2, U+20F3, U+20F4, U+20F5, U+20F6, U+20F7, U+20F8, U+20F9, U+20FA, U+20FB, U+20FC, U+20FD, U+20FE, U+20FF, U+2010, U+2011, U+2012, U+2013, U+2014, U+2015, U+2016, U+2017, U+2018, U+2019, U+201A, U+201B, U+201C, U+201D, U+201E, U+201F, U+2020, U+2021, U+2022, U+2023, U+2024, U+2025, U+2026, U+2027, U+2028, U+2029, U+202A, U+202B, U+202C, U+202D, U+202E, U+202F, U+2030, U+2031, U+2032, U+2033, U+2034, U+2035, U+2036, U+2037, U+2038, U+2039, U+203A, U+203B, U+203C, U+203D, U+203E, U+203F, U+2040, U+2041, U+2042, U+2043, U+2044, U+2045, U+2046, U+2047, U+2048, U+2049, U+204A, U+204B, U+204C, U+204D, U+204E, U+204F, U+2050, U+2051, U+2052, U+2053, U+2054, U+2055, U+2056, U+2057, U+2058, U+2059, U+205A, U+205B, U+205C, U+205D, U+205E, U+205F, U+2060, U+2061, U+2062, U+2063, U+2064, U+2065, U+2066, U+2067, U+2068, U+2069, U+206A, U+206B, U+206C, U+206D, U+206E, U+206F, U+2070, U+2071, U+2072, U+2073, U+2074, U+2075, U+2076, U+2077, U+2078, U+2079, U+207A, U+207B, U+207C, U+207D, U+207E, U+207F, U+2080, U+2081, U+2082, U+2083, U+2084, U+2085, U+2086, U+2087, U+2088, U+2089, U+208A, U+208B, U+208C, U+208D, U+208E, U+208F, U+2090, U+2091, U+2092, U+2093, U+2094, U+2095, U+2096, U+2097, U+2098, U+2099, U+209A, U+209B, U+209C, U+209D, U+209E, U+209F, U+20A0, U+20A1, U+20A2, U+20A3, U+20A4, U+20A5, U+20A6, U+20A7, U+20A8, U+20A9, U+20AA, U+20AB, U+20AC, U+20AD, U+20AE, U+20AF, U+20B0, U+20B1, U+20B2, U+20B3, U+20B4, U+20B5, U+20B6, U+20B7, U+20B8, U+20B9, U+20BA, U+20BB, U+20BC, U+20BD, U+20BE, U+20BF, U+20C0, U+20C1, U+20C2, U+20C3, U+20C4, U+20C5, U+20C6, U+20C7, U+20C8, U+20C9, U+20CA, U+20CB, U+20CC, U+20CD, U+20CE, U+20CF, U+20D0, U+20D1, U+20D2, U+20D3, U+20D4, U+20D5, U+20D6, U+20D7, U+20D8, U+20D9, U+20DA, U+20DB, U+20DC, U+20DD, U+20DE, U+20DF, U+20E0, U+20E1, U+20E2, U+20E3, U+20E4, U+20E5, U+20E6, U+20E7, U+20E8, U+20E9, U+20EA, U+20EB, U+20EC, U+20ED, U+20EE, U+20EF, U+20F0, U+20F1, U+20F2, U+20F3, U+20F4, U+20F5, U+20F6, U+20F7, U+20F8, U+20F9, U+20FA, U+20FB, U+20FC, U+20FD, U+20FE, U+20FF, U+2010, U+2011, U+2012, U+2013, U+2014, U+2015, U+2016, U+2017, U+2018, U+2019, U+201A, U+201B, U+201C, U+201D, U+201E, U+201F, U+2020, U+2021, U+2022, U+2023, U+2024, U+2025, U+2026, U+2027, U+2028, U+2029, U+202A, U+202B, U+202C, U+202D, U+202E, U+202F, U+2030, U+2031, U+2032, U+2033, U+2034, U+2035, U+2036, U+2037, U+2038, U+2039, U+203A, U+203B, U+203C, U+203D, U+203E, U+203F, U+2040, U+2041, U+2042, U+2043, U+2044, U+2045, U+2046, U+2047, U+2048, U+2049, U+204A, U+204B, U+204C, U+204D, U+204E, U+204F, U+2050, U+2051, U+2052, U+2053, U+2054, U+2055, U+2056, U+2057, U+2058, U+2059, U+205A, U+205B, U+205C, U+205D, U+205E, U+205F, U+2060, U+2061, U+2062, U+2063, U+2064, U+2065, U+2066, U+2067, U+2068, U+2069, U+206A, U+206B, U+206C, U+206D, U+206E, U+206F, U+2070, U+2071, U+2072, U+2073, U+2074, U+2075, U+2076, U+2077, U+2078, U+2079, U+207A, U+207B, U+207C, U+207D, U+207E, U+207F

The analytical procedure was formalised in accordance with GOST 19.701-90 (ISO 5807) (Fig. 1). The algorithm comprises 14 steps, whose content is explained in Table 1.

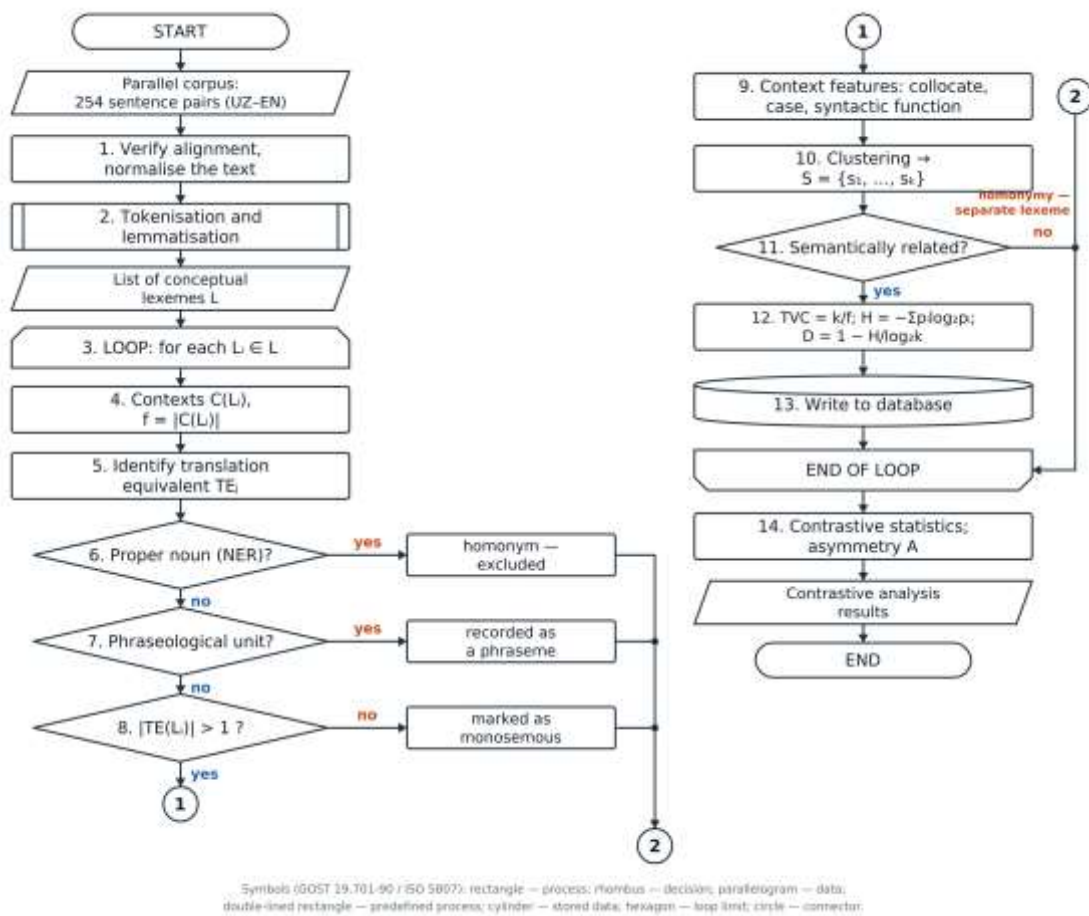


Fig. 1. Algorithm for the contextual-semantic analysis of polysemous lexemes on a parallel corpus

Table 1. Content of the algorithm steps

Step	Content and methodological rationale
1	Sentence-level alignment is verified and the text is normalised. The alignment type (1:1, 1:2, 2:1, zero) is recorded: in 1:2 and 2:1 cases the lexeme boundary shifts and automatic identification of the equivalent becomes unreliable.
2	Tokenisation and lemmatisation. A morphological analyser is required for Uzbek, since in an agglutinative language a single word form decomposes into a stem and several affixes (qalbiga → qalb + POSS3 + DAT); a POS tagger suffices for English.
3	The loop over the list L of philosophical-conceptual lexemes begins. The list covers the ontological (soul, spirit, nature), ethical (good, mercy), epistemic (right, truth) and psychological (heart, feeling) domains.



Step	Content and methodological rationale
4	All contexts $C(L_i)$ of each lexeme L_i in the corpus are extracted; $f = C(L_i) $ is recorded as its frequency.
5	In each context the translation equivalent TE_j is identified on the other side of the aligned sentence — the practical application of the “semantic mirrors” principle [6].
6	Proper-noun check (NER). Units that are formally identical but genetically unrelated are separated as homonyms: in the corpus, the Jon inside “Sent-Jon-Roud” (St. John’s Road) bears no semantic relation to the Uzbek lexeme jon.
7	Phraseological-unit check. Non-compositional units are recorded separately, otherwise they are counted as spurious senses. Example: $oq\ yo‘l \rightarrow$ blessing, where neither oq nor $yo‘l$ separately conveys the sense ‘blessing’.
8	The condition $ TE(L_i) > 1$ is tested: if the lexeme is rendered identically in every context, it is marked as monosemous within this corpus.
9	Context features are extracted: left and right collocate, case affix in Uzbek, preposition in English, syntactic function and thematic domain. This feature set is precisely the differential means that separates the senses.
10	Contexts are clustered by translation equivalent and the candidate senses $S = \{s_1, \dots, s_k\}$ are formed.
11	The semantic relatedness of the senses is assessed: if a common semantic component is present the case is qualified as polysemy, otherwise as homonymy and a separate lexeme is created.
12	The quantitative measures are computed according to formulas (1)–(3).
13	The results are written to the database (sense, token_sense, context_feature, sense_link).
14	When the loop terminates, contrastive statistics for the two languages and the semantic asymmetry coefficient (4) are computed.

Source: compiled by the author

The translation-variability coefficient is defined as the ratio of the number of distinct translation equivalents of a lexeme to its frequency:

$$TVC(L) = k(L) / f(L), \quad 0 < TVC \leq 1 \quad (1)$$



where k is the number of distinct equivalents and f the number of aligned occurrences. $TVC = 1$ means the lexeme was rendered anew in every context, i.e. its sense is wholly context-dependent; $TVC \rightarrow 0$ means the lexeme has a stable equivalent. The evenness of the sense distribution is measured by Shannon entropy [16], and the degree to which one sense predominates by the dominance measure:

$$H(L) = - \sum p_i \cdot \log_2 p_i, \quad p_i = n_i / f \quad (2)$$

$$D(L) = 1 - H(L) / \log_2 k(L) \quad (3)$$

The imbalance of polysemy between the two languages is assessed by the semantic asymmetry coefficient ($A > 0$ indicates higher polysemy in Uzbek):

$$A = (\bar{k}_{uz} - \bar{k}_{en}) / (\bar{k}_{uz} + \bar{k}_{en}), \quad A \in [-1; 1] \quad (4)$$

3. Database architecture

To store the results and generalise them through queries, a four-layer relational database was designed on PostgreSQL: 12 tables and 15 relationships (Fig. 2). The source layer (discourse_type, text_source, align_unit) holds the discourse type and the aligned units; the lexical-semantic layer (lexeme, token, sense, semantic_field) holds the lexeme, its occurrence in the text, its senses and its domain; the annotation layer (annotator, token_sense) holds the sense assigned to each token and the confidence value; the analysis layer (equivalence, context_feature, sense_link) holds the translation equivalents, the context features and the correspondence between the senses of the two languages.

The UNIQUE (token_id, annotator_id) constraint on *token_sense* prevents one annotator from assigning two senses to the same token and makes it possible to compute inter-annotator agreement (Cohen's kappa). Measures (1)–(3) are not stored in a table of their own but derived as the *polysemy_metric* materialized view: this removes data redundancy and rules out any inconsistency between the measures and the annotation when the latter changes. The database was created in a live PostgreSQL 16 environment and loaded with all 254 corpus units and the 79 manually verified equivalences; the



values computed by the view matched independently computed values for all 58 lexemes.

4. Results and discussion

Table 2. Uzbek translation paradigm of the English conceptual lexemes

Lexeme	Uzbek equivalents (frequency)	f	k	TVC	H, bit	D
good	ajoyib 2, xayr 2, Ø 2, yaxshi 1, rahmdil 1, nihoyatda 1	9	6	0.667	2.503	0.032
man	kishi 2, erkak 1, kattalar 1, odam 1, Ø 1	6	5	0.833	2.252	0.030
soul	qalb 2, ruh 1, kimsa (Ø) 1	4	3	0.750	1.500	0.054
heart	yurak 2, qalb 1, jur'at 1	4	3	0.750	1.500	0.054
light	nur 2, yorug' 1, shu'la 1	4	3	0.750	1.500	0.054
right	o'ng 1, to'g'ri 1, joyida 1	3	3	1.000	1.585	0.000
hard	yaqin 1, mahkam 1, og'ir 1	3	3	1.000	1.585	0.000
nature	tus 1, tabiat 1, fe'l 1	3	3	1.000	1.585	0.000
spirit	kayfiyat 2, farosat 1	3	2	0.667	0.918	0.082
Mean	—	4.33	3.44	0.824	1.659	0.038

Table 3. English translation paradigm of the Uzbek lexemes

Lexeme	English equivalents (frequency)	f	k	TVC	H, bit	D
ko'z	eye 14, sight 3, Ø 3	20	3	0.150	1.181	0.255
yo'l	road 4, way 2, blessing 2, footpath 1, Ø 1	10	5	0.500	2.122	0.086
jon	vigour 2, indeed 1, lifeless 1, affectionate 1, John 1	6	5	0.833	2.252	0.030
tus	nature 1, show 1, face 1, character 1	4	4	1.000	2.000	0.000
qalb	soul 2, heart 1, feelings 1	4	3	0.750	1.500	0.054
Mean	—	8.80	4.00	0.647	1.811	0.085



The clearest result is given by right: all three of its occurrences are rendered differently and each expresses a distinct sense — spatial (“He looked to the right and to the left” → “avval o‘ngga qaradi”), epistemic (“the doctor must be right” → “doktor to‘g‘ri aytgan”) and a state reading (“all was right” → “hammasi joyidaligidan”). TVC = 1.000 with D = 0 shows that the sense of this lexeme cannot be determined at all without context.

The pair soul – spirit deserves particular attention. Where both lexemes occur side by side in one sentence (“a artificial soul and spirit”), the Uzbek text renders them as “ruh-u ... kayfiyat”, and in the second sentence (“soul or spirit”) as “Qalb, ruh yoki kayfiyat”: the two-member opposition of English corresponds to a three-member series in Uzbek (qalb – ruh – kayfiyat). The converse confirmation appears in Table 3: qalb in turn aligns with soul (twice) and heart (once), so that the two lexemes stand in a “semantic mirror” relation to each other.

The lexeme jon presents the methodologically most complex case: its six occurrences have at least four different linguistic statuses — free lexeme (jonsiz → lifeless), component of a fixed expression (jon-jahdi-la → with undiminished vigour; Jon-jon derdim → I do, indeed), word-forming element (jonkuyar → affectionate) and a homonymous form inside a proper noun (Sent-Jon-Roud → St. John’s Road); in addition, the corpus contains the diminutive-affectionate suffix -jon (o‘rtoqjon → dear). If steps 6 and 7 of the algorithm are omitted, these five distinct phenomena are counted as “senses” of a single lexeme, k is inflated artificially, and the resulting sense inventory becomes unusable.

By equivalence type, 69 of the 79 alignments (87.3 %) are partial, 8 (10.1 %) zero and 2 (2.5 %) phraseological. The one-in-ten share of zero equivalence is notable: in such cases the lexeme is expressed not by a separate word but through the construction as a whole, or not at all (in “There are a good many of ‘em” the item good is a quantity intensifier and has no independent equivalent in the Uzbek translation). For machine



translation systems this share constitutes the hardest cases, because algorithms based on word-to-word alignment generate a spurious equivalent here.

By the aggregate measures, the Uzbek lexemes have on average 4.00 and the English ones 3.44 equivalents; by formula (4), $A = 0.075$. This value is close to zero and indicates no appreciable asymmetry within the conceptual vocabulary selected. Mean entropy, however, is higher in Uzbek (1.811 against 1.659 bit): the senses of the Uzbek lexemes are distributed relatively evenly, whereas a dominant sense is more common in the English ones.

The practical value of the measures is that for lexemes with high D (ko‘z, $D = 0.255$) the most-frequent-sense strategy yields high accuracy, whereas for lexemes with $D = 0$ (right, hard, nature, tus) that strategy cannot work in principle and a disambiguation module relying on context features is required. D therefore determines in advance for which lexemes the sense-disambiguation module of a machine translation system must be invoked. TVC, by contrast, is sensitive to frequency (ko‘z occurs 20 times with $TVC = 0.150$, right 3 times with $TVC = 1.000$), so entropy is the more reliable measure when comparing lexemes of different frequency.

Limitations of the study: the pilot corpus (254 sentence pairs) is statistically small and the f values range from 3 to 20, which increases the error of the entropy estimate; the annotation was carried out by a single annotator — although the database is ready to accommodate a second annotator and to compute the agreement coefficient, this stage has not yet been performed.

Conclusion

The paper develops a methodology for the contrastive, parallel-corpus-based study of philosophical-conceptual polysemous lexemes in Uzbek and English: a 14-step algorithm four measures for the quantitative assessment of polysemy (TVC, entropy, dominance, semantic asymmetry); and a 12-table PostgreSQL database. Testing on the pilot corpus established that English conceptual lexemes have on average 3.44 and Uzbek ones 4.00 equivalents ($A = 0.075$), and that 87.3 % of the equivalences are partial,



10.1 % zero and 2.5 % phraseological. The theoretical significance of the methodology lies in deriving sense distinctions inductively from translation evidence rather than from the researcher's introspection; its practical significance lies in the possibility of applying the resulting sense inventory and context features directly in the word sense disambiguation module of an Uzbek–English machine translation system. The next stage is to expand the corpus to at least 5,000 sentence pairs and to add a subcorpus of academic philosophical texts.

REFERENCES

1. Kilgarrieff A. I don't believe in word senses // Computers and the Humanities. – 1997. – Vol. 31, No. 2. – P. 91–113.
2. Dyvik H. Translations as semantic mirrors: from parallel corpus to WordNet // Language and Computers. – 2004. – Vol. 49. – P. 311–326.
3. Firth J.R. A synopsis of linguistic theory 1930–1955 // Studies in Linguistic Analysis. – Oxford: Blackwell, 1957. – P. 1–32.
4. Ide N., Erjavec T., Tufiş D. Sense discrimination with parallel corpora // Proceedings of the ACL-02 Workshop on Word Sense Disambiguation. – Philadelphia, 2002. – P. 61–66.
5. Gale W., Church K., Yarowsky D. A method for disambiguating word senses in a large corpus // Computers and the Humanities. – 1992. – Vol. 26. – P. 415–439.
6. Ide N., Erjavec T., Tufiş D. Sense discrimination with parallel corpora // Proceedings of the ACL-02 Workshop on Word Sense Disambiguation. – Philadelphia, 2002. – P. 61–66.
7. Rahmatullayev Sh. Hozirgi adabiy o'zbek tili. – Tashkent: Universitet, 2006. – 464 p. (in Uzbek)
8. Ne'matov H., Rasulov R. O'zbek tili sistem leksikologiyasi asoslari. – Tashkent: O'qituvchi, 1995. – 128 p. (in Uzbek)