RESEARCH ARTICLE

Linguistic Foundations Of Electronic Thesaurus Construction: Structure, Synset-Formation Methodology, And Corpus-Based Evidence From The Uzbek Language

Muhammadjon Najmiddinov
Dean of Philology faculty, DSc, Kokand university.

Abstract

This article presents a comprehensive linguistic and technological analysis of the principles underlying the construction of electronic thesauri. The purpose of the study is threefold: (1) to characterise the structural and functional properties that distinguish an electronic thesaurus from a traditional dictionary; (2) to develop and empirically validate a step-by-step methodology for selecting thesaurus units and forming synsets, integrating corpus statistics, distributional semantics and transformer-based (BERT-type) models with expert verification; and (3) to describe how polysemy, synonymy, homonymy and ontological categorisation are represented within the resulting thesaurus model for four major word classes — object nouns, attributive adjectives, verbs of action/state, and adverbs. Descriptive, componential, distributive, ideographic, ontological and quantitative methods of analysis were applied to material drawn from the Uzbek Educational Corpus and the ARANEUM_UZBEKICUM corpus. As a result, 77 synsets were constructed across five lexical categories, three structural models of thesaurus organisation (hierarchical tree, network graph, and matrix) were formalised, and a six-stage synset-formation methodology was proposed and tested. The findings provide a replicable linguistic foundation for building electronic thesauri for Uzbek and typologically related agglutinative languages, with direct applications in information retrieval, machine translation and natural-language-processing systems.

Keywords: electronic thesaurus, ontology, synset, corpus linguistics, transformer, BERT, semantic network, descriptor, polysemy, synonymy, homonymy, Uzbek language.

1. INTRODUCTION

The accelerating growth of digital information and the demands of the semantic web have made accurate, semantically organised information retrieval one of the central tasks of contemporary linguistics. Traditional, linearly structured dictionaries are no longer sufficient to systematise the semantic content of large text collections (Karaulov, 1976). Consequently, electronic thesauri — digital linguistic resources that systematically represent semantic relations among lexical units — have become a core component of semantic-web and artificial-intelligence technologies (Berners-Lee, Hendler, Lassila, 2001).

The theoretical and methodological foundations of thesaurus construction have been extensively developed within Western and Russian linguistic traditions. The thesaurus model proposed by Yu.N. Karaulov and B.V. Dobrov, the terminological-systems research of S.V.

Grinev and V.M. Leychik, and the WordNet project developed under the direction of G.A. Miller (Miller, 1995; Fellbaum, 1998) constitute the fundamental basis of contemporary thesaurus theory. However, the linguistic foundations of constructing an electronic thesaurus on Uzbek-language material — in particular, the methodology for selecting units and forming synsets — had not previously been the object of dedicated monographic research.

Uzbek presents specific challenges for thesaurus construction as an agglutinative Turkic language with rich derivational morphology, extensive polysemy, and a productive system of synonymic and antonymic relations that has not been systematically modelled in digital form. Existing Uzbek explanatory and synonym dictionaries record lexical meaning but do not formalise the network of semantic relations — synonymy, hyperonymy, hyponymy, meronymy and association — that a thesaurus requires for computational use.

The aim of this article is to characterise the structure and composition of the electronic thesaurus from a linguistic-technological perspective and to develop a step-by-step methodology, grounded in transformer models (in particular BERT), for selecting thesaurus units and forming synsets, and subsequently to demonstrate the application of that methodology to the description of polysemy, synonymy, homonymy, and the verb/adverb/adjective semantic fields of Uzbek (Devlin et al., 2019; Vaswani et al., 2017).

The scientific novelty of the study consists in: a systematic account of the structural properties that distinguish the electronic thesaurus from the traditional dictionary; a stage-by-stage methodology for thesaurus-unit selection combining corpus frequency, semantic relatedness and expert evaluation; a demonstration of how BERT-type contextual embeddings can be used to automatically propose synonymic groupings that are subsequently validated by linguistic experts; and a corpus-grounded description of how polysemy, synonymy, and homonymy are formally represented as distinct or cross-linked synsets in Uzbek.

2. MATERIALS AND METHODS

2.1. Material

The empirical material of the study comprises two corpora: the “Uzbek Educational Corpus” and the ARANEUM_UZBEKICUM corpus, together with the explanatory and synonym dictionaries of the Uzbek language. From these sources, 77 lexical units belonging to five grammatical categories — object nouns, attributive adjectives, action/state verbs, state adjectives, and adverbs — were selected for detailed synset construction (see Table 3 and Figure 3).

2.2. Methods of analysis

The study applied descriptive, componential, ideographic, ontological, onomasiological and semasiological, linguo-statistical, quantitative, distributive and illustrative methods of analysis. Word-sense relations were established through a combination of manual expert analysis and automatic distributional-semantic analysis using word-embedding models; contextual similarity between candidate synonyms was computed as cosine similarity between BERT-type contextual vector representations.

2.3. Six-stage synset-formation methodology

On the basis of the reviewed sources and the corpus material, a six-stage, criterion-based methodology for selecting thesaurus units and constructing synsets was developed (Table 2, Figure 2). The procedure begins with domain analysis and proceeds through frequency-based extraction, semantic-link identification, transformer-based vector analysis (cosine similarity threshold ≥ 0.75), expert verification, and final synset formation. Each stage is governed by an explicit selection criterion, which ensures the reproducibility of the procedure for other lexical fields and, in principle, for other agglutinative languages.

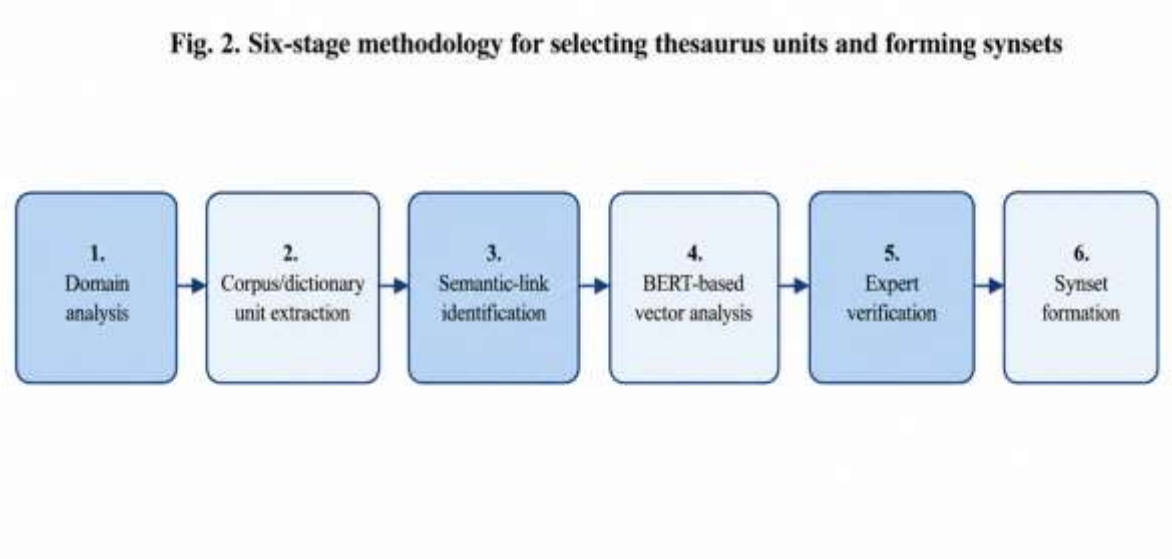

Fig. 2. Six-stage methodology for selecting thesaurus units and forming synsets

2.4. Structural models of thesaurus organisation

Three structural models for organising information within the electronic thesaurus were formalised: (1) a hierarchical tree structure, representing genus–species and part–whole relations; (2) a network (graph) structure, representing hyperonymic–hyponymic relations as a directed semantic network; and (3) a matrix structure, representing meronymic and associative relations as a cross-tabulation of descriptor co-occurrence. These three models correspond to three complementary means of information presentation: hypertext pages, a search-and-filter interface, and cross-referenced links between synsets (Figure 1).

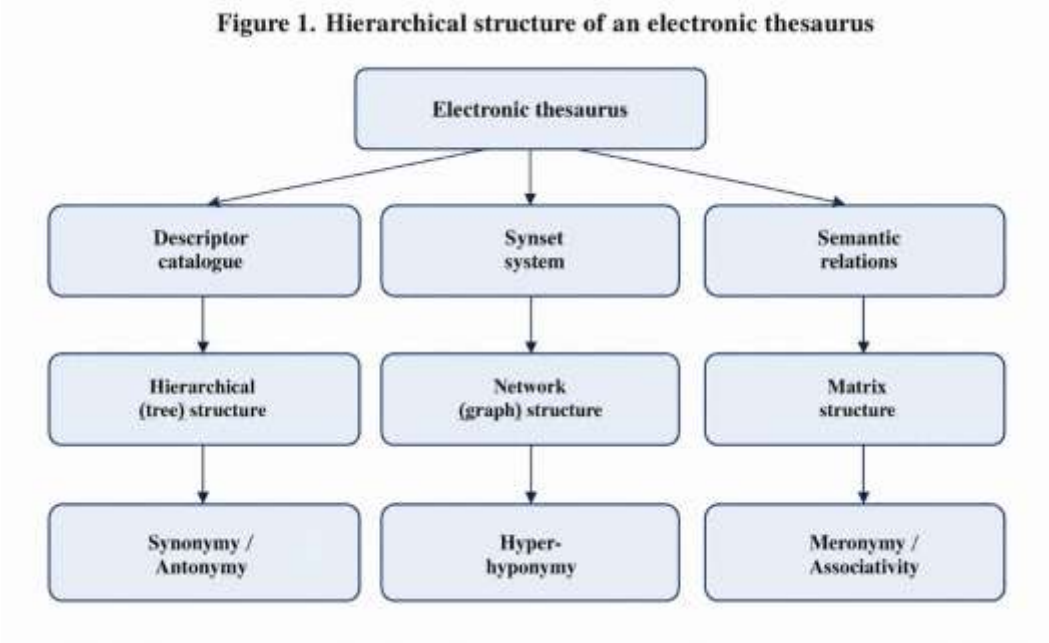

Fig. 1. Hierarchical structure of an electronic thesaurus: descriptor catalogue, synset system and semantic-relation layer

2.5. Ontological basis for descriptor selection

The role of ontology in thesaurus construction was analysed from a linguo-cognitive and computational-linguistics perspective. An ontological structure consisting of classes, properties, instances and relations was adopted as the primary criterion for selecting and systematising thesaurus descriptors and "variable" units. Descriptor selection and placement follow the faceted classification principle of the Kasares–Karaulov system, applying genus–species, meronymic, antonymic and ideographic-field relations to build the descriptor system and its alphabetical index.

2.6. Representation of polysemy, synonymy and homonymy

Uzbek vocabulary is characterised by pronounced polysemy and polyfunctionality. For polysemous lexemes, each distinct sense was assigned to a separate synset; for example, the word quralay ("fawn / young deer") yields one synset in the "animal name" ideographic field and a second, figurative synset in the "eye" field, denoting a tender, deer-like gaze. Homonymous lexemes were treated as structurally distinct entries: for the word bog‘ ("garden" / "binding, tie"), the explanatory dictionary evidence supports splitting it into two independent synsets, cross-linked through an explicit hyperlink marking the homonymic relation (Figure 7). Synonymic rows were formed on the basis of "similar words" analysis in the corpus together with existing dictionary sources, applying the criteria of semantic wholeness, contextual synonymy, cultural conditioning and systemic consistency.

Fig. 7. Splitting a homonymous lexeme into two independent synsets

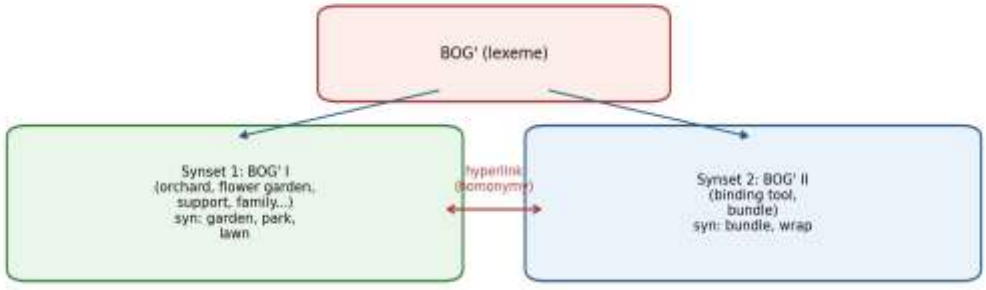

Fig. 7. Splitting a homonymous lexeme (bog') into two independent, cross-linked synsets

3. RESULTS

3.1. Structural comparison with traditional dictionaries

Table 1 summarises the structural and functional features that distinguish the electronic thesaurus from the traditional dictionary. The electronic thesaurus is characterised by a large, expandable data capacity, a hypertextual and networked organisation, an explicitly encoded set of semantic relations (synonymy, hyperonymy, hyponymy, meronymy), interactive search and filtering, integration of multimedia elements, and a combination of manual and AI-assisted automation.

Feature	Traditional dictionary	Electronic thesaurus
Data capacity	Limited	Large-scale, expandable
Structural form	Linear, alphabetical	Hypertextual, networked
Semantic relations	Weakly represented	Synonymy, hyperonymy, hyponymy, meronymy
Interactivity	None	Search, filtering, cross-references
Multimedia elements	None or minimal	Text, audio, image integration
Automation level	Manual	Manual + automatic (AI-based)

Table 1. Structural and functional comparison of the traditional dictionary and the electronic thesaurus.

3.2. Quantitative results of synset construction

Applying the six-stage methodology (Table 2), synsets were constructed for 77 lexical units distributed across five categories, as shown in Table 3 and Figure 3: 25 object nouns (32.5%), 18 attributive adjectives (23.4%), 15 action/state verbs (19.5%), 10 state adjectives (13.0%), and 9 adverbs (11.6%). Object nouns and attributive adjectives together account for more than half of all constructed synsets, reflecting their central role in the ideographic organisation of the general lexicon.

Stage	Content of activity	Selection criterion
1	Domain analysis	Field boundaries, target group

2	Unit extraction from corpus/dictionaries	Frequency index ≥ threshold
3	Identifying semantic links	Contextual co-occurrence
4	BERT-based vector analysis	Cosine similarity ≥ 0.75
5	Expert verification	Linguistic and stylistic consistency
6	Synset formation	Final approved composition

Table 2. Six-stage methodology for thesaurus-unit selection and synset formation.

Lexical category	**Units analysed**	**Share, %**
Object nouns (predmet nomlari)	25	32.5
Attributive adjectives (belgi bildiruvchi)	18	23.4
Action/state verbs	15	19.5
State adjectives	10	13.0
Adverbs	9	11.6
Total	**77**	**100.0**

Table 3. Quantitative distribution of constructed synsets by lexical category.

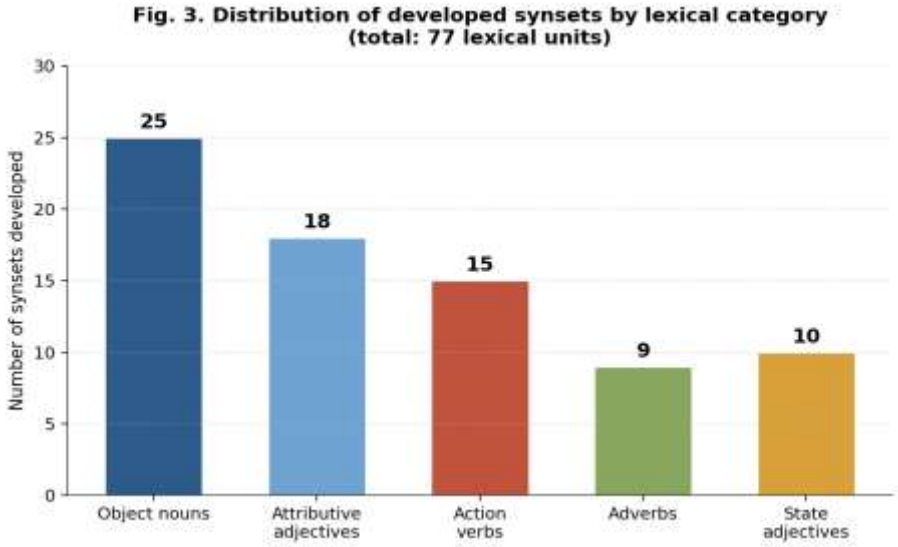

Fig. 3. Distribution of developed synsets by lexical category (total: 77 lexical units)

3.3. Corpus-based evidence for synonymic grouping

To illustrate the distributional-semantic stage of the methodology (Table 2, stage 4), the adjective inja (“delicate, tender”) was analysed in the ARANEUM_UZBEKICUM corpus using a “similar words” function based on contextual word embeddings. Twelve semantically related adjectives were retrieved, with cosine similarity coefficients ranging from 0.445 to 0.537 (Figure 4); the closest candidates — hazin, jilva, she'rxon and yanglig' — were retained as members of the inja synset after expert verification, while lower-similarity candidates were assigned to adjacent but distinct ideographic sub-fields.

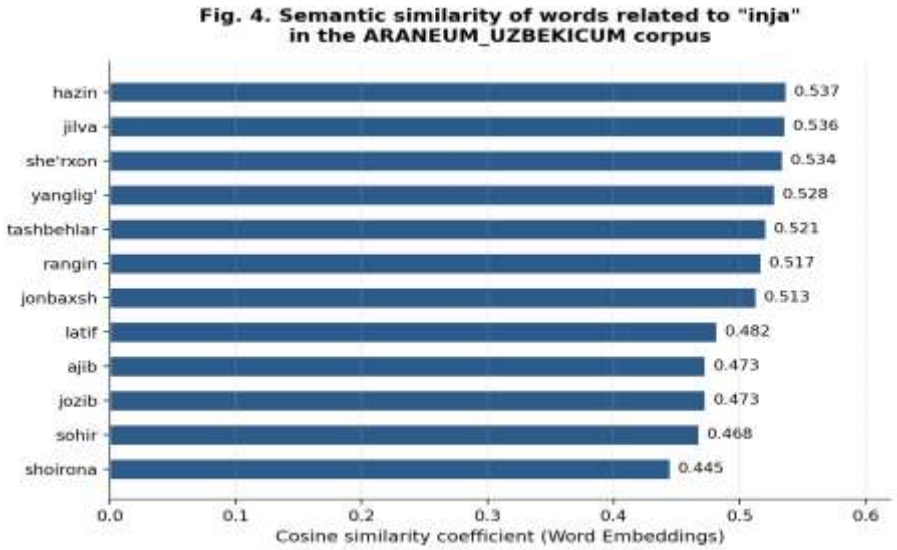

Fig. 4. Cosine-similarity coefficients for words semantically related to "inja" in the ARANEUM_UZBEKICUM corpus

Corpus-frequency analysis of the same synonymic row (Figure 6) shows substantial dispersion, ranging from 132 occurrences (jozib) to 7,152 occurrences (teran) in the Uzbek Educational Corpus. This frequency asymmetry confirms that raw frequency alone is an insufficient criterion for synset membership and must be combined with semantic-similarity scores and expert judgement, as specified in stages 3–5 of the proposed methodology.

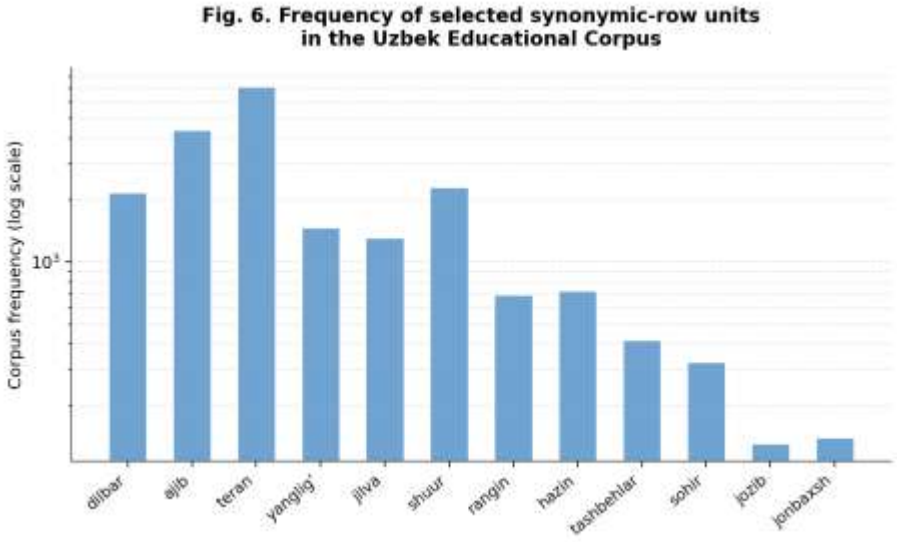

Fig. 6. Corpus frequency of selected synonymic-row units in the Uzbek Educational Corpus (logarithmic scale)

3.4. Dissemination of research results

The results of the dissertation research underlying this article were presented at 11 scientific conferences (6 international, 5 national) and published in 23 scholarly works, comprising 9 articles in journals recommended by the Higher Attestation Commission of Uzbekistan, 2 articles in foreign journals, and 12 conference papers and abstracts (Figure 5).

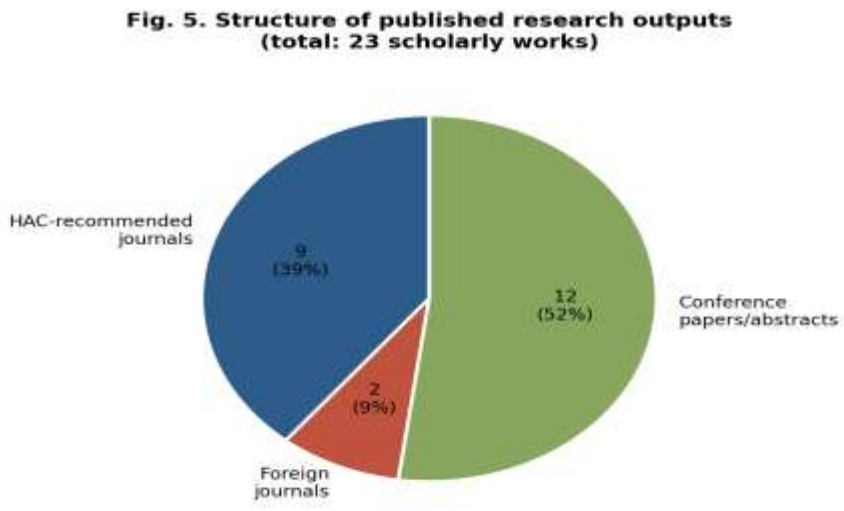

Fig. 5. Structure of published research outputs (total: 23 scholarly works)

4. DISCUSSION

The results confirm that an electronic thesaurus differs from a traditional dictionary not merely in scale but in kind: its defining property is the explicit, machine-tractable encoding of semantic relations, which enables integration with search engines, machine-translation systems and ontology-driven NLP applications (cf. Google, Yandex search integration; Google Translate, DeepL, and GPT-4/Gemini-type systems). This is consistent with the broader shift, documented in the WordNet tradition (Miller, 1995; Fellbaum, 1998), from lexicography as a descriptive practice toward lexicography as a form of applied knowledge engineering.

The six-stage methodology proposed here extends this tradition by formally integrating transformer-based contextual embeddings into the synset-formation workflow for a previously under-resourced language. The use of a similarity threshold (cosine similarity ≥ 0.75) as an explicit, quantifiable criterion, combined with mandatory expert verification (stage 5), addresses a known limitation of purely automatic embedding-based synonym extraction: contextual embeddings capture distributional similarity but do not reliably distinguish true synonymy from mere topical relatedness or antonymy, a distinction that in this study required human linguistic judgement in every case, particularly for polysemous and homonymous items.

The treatment of polysemy and homonymy as distinct synset-splitting operations, illustrated by quralay and bog', demonstrates that the ideographic organisation of the thesaurus — rather than the lexeme as an undifferentiated unit — must serve as the basic unit of description. This finding aligns with the componential and ideographic tradition represented by Karaulov's (1976) general and Russian ideography and echoes the faceted-classification approach of Kasares, adapted here to the genus–species, meronymic and antonymic relational structure of Uzbek.

The observed frequency asymmetry among synonymic-row members (Figure 6) has a practical implication for information-retrieval applications: query-expansion mechanisms built on the resulting thesaurus should not weight synonyms by raw corpus frequency alone, since low-frequency but

semantically central items (e.g., jozib, sohir) would otherwise be under-represented in expanded search results relative to high-frequency but more peripheral items (e.g., teran, ajib).

Certain limitations should be acknowledged. The synset sample (77 units across five categories) is illustrative rather than exhaustive relative to the full Uzbek lexicon, and the 0.75 similarity threshold, while empirically motivated, was calibrated on a limited set of test cases; broader validation against a larger annotated gold-standard set would strengthen confidence in the threshold's generalisability. Moreover, the reliance on a single transformer architecture (BERT-type models) leaves open the question of whether alternative contextual-embedding models would yield materially different candidate groupings for Uzbek, given its agglutinative morphology and comparatively smaller training-corpus size relative to higher-resource languages.

Despite these limitations, the proposed methodology and the resulting conceptual model — in which lexical units, synsets, semantic relations and ontological categories form an interlinked system — provide a transferable linguistic foundation applicable beyond the specific word classes examined here, including for the extension of the thesaurus to other domains such as terminology standardisation, machine translation, and the documentation of endangered or under-resourced language varieties.

5. CONCLUSION

This study has shown that the electronic thesaurus constitutes a qualitatively distinct linguistic resource from the traditional dictionary, defined by its hypertextual organisation, its explicit encoding of synonymic, hyperonymic–hyponymic, meronymic and associative relations, and its capacity for interactive, multimedia-enriched, and AI-assisted use. A six-stage, criterion-based methodology — domain analysis, corpus-based extraction, semantic-link identification, BERT-based vector analysis, expert verification, and synset formation — was developed and empirically applied to 77 Uzbek lexical units across five grammatical categories, yielding a validated set of synsets and demonstrating how polysemy and homonymy can be formally represented through synset splitting and cross-linking.

The results provide both a theoretical contribution to lexicography and semantics and a practical foundation for building Uzbek-language electronic thesauri that can be integrated with search engines, machine-translation pipelines, and other natural-language-processing systems. Future work should extend the methodology to additional lexical-semantic fields, validate the similarity threshold against a larger gold-standard dataset, and explore multilingual alignment of the resulting thesaurus with existing WordNet-type resources for related languages.

REFERENCES

1. Karaulov Yu.N. Общая и русская идеография. – Moscow: Nauka, 1976. – 356 p.

2. Miller G.A. WordNet: A Lexical Database for English // Communications of the ACM. – 1995. – Vol. 38, No. 11. – P. 39–41.

3. Fellbaum C. (ed.) WordNet: An Electronic Lexical Database. – Cambridge, MA: MIT Press, 1998. – 423 p.

4. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.

5. Vaswani A. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

6. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // arXiv:1301.3781. – 2013.

7. Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American. – 2001. – Vol. 284, No. 5. – P. 34–43.

8. Gruber T.R. A Translation Approach to Portable Ontology Specifications // Knowledge Acquisition. – 1993. – Vol. 5, No. 2. – P. 199–220.

9. Апресян Ю.Д. Лексическая семантика: синонимические средства языка. – Moscow: Nauka, 1974. – 367 p.

10. Kasares J. Diccionario ideológico de la lengua española. – Barcelona: Gustavo Gili, 1942.

11. Najmiddinov M.G'. Tezauruslar yaratishning lingvistik asoslari: Filologiya fanlari doktori (DSc) dissertatsiyasi avtoreferati. – Andijon, 2026. – 69 p.

12. Nešpore-Bērzkalne G. et al. ARANEUM corpora project documentation. – 2018.

13. “O'zbek tilining ta'limiy korpusi” [Uzbek Educational Corpus]. Electronic resource. Available at: https://uzbekcorpus.uz

14. Asadov T. Leksik-semantik guruhlarning tasnifi. – Toshkent: Fan, 2015. – 214 p.