

METHODOLOGY OF STYLISTIC AND SEMANTIC ANNOTATION IN ENGLISH-UZBEK PARALLEL CORPORA

Vasliddinova Kamola

PhD researcher,

Uzbekistan State World Languages University

Annotatsiya. *Ushbu maqolada tabiiy tilni qayta ishlash (NLP) va amaliy lingvistikaning muhim yo'nalishlaridan biri bo'lgan ingliz-o'zbek parallel korpuslarida uslubiy va semantik annotatsiya qilish metodologiyasi hamda uning nazariy asoslari adabiyotlar tahlili asosida yoritiladi. Kross-lingvistik parallel korpuslar ko'p ma'noli so'zlarni bir qiymatli ajratish (Word Sense Disambiguation – WSD), leksik birliklarning kontekstual va uslubiy muqobilligini aniqlash hamda bilingval leksikonlarni yaratishdagi o'rni tahlil qilingan. Matnlarni jumlar va so'zlar darajasida moslashtirish, semantik ma'no proyeksiyasi hamda uslublararo differentsiallashtirish bosqichlari tizimlashtirilgan.*

Kalit so'zlar: *amaliy lingvistika, kompyuter lingvistikasi, parallel korpus, semantik annotatsiya, uslubiy klassifikatsiya, Word Sense Disambiguation (WSD), kontekstual embedding, bilingval leksikon.*

Аннотация. *В данной статье на основе анализа научной литературы рассматриваются теоретические и методологические основы стилистической и семантической аннотации в англо-узбекских параллельных корпусах как одного из ключевых направлений компьютерной лингвистики и обработки естественного языка (NLP). Анализируется роль параллельных корпусов в устранении лексической многозначности (Word Sense Disambiguation – WSD), определении контекстуальной и стилистической эквивалентности единиц, а также в создании двуязычных лексиконов. Систематизированы этапы выравнивания текстов, семантической проекции и межстилевой дифференциации.*

Ключевые слова: *прикладная лингвистика, компьютерная лингвистика, параллельный корпус, семантическая аннотация, стилистическая классификация, Word Sense Disambiguation (WSD), контекстуальное выравнивание, двуязычный лексикон.*

Abstract. *This article provides a comprehensive theoretical and methodological literature review on the principles of stylistic and semantic annotation in cross-lingual parallel corpora, with specific reference to the English-Uzbek language pair. Within computational linguistics and Natural Language Processing (NLP), parallel corpora serve as essential linguistic knowledge bases for resolving lexical ambiguity (Word Sense Disambiguation – WSD) and capturing register-specific translational variations. The paper synthesizes foundational approaches to bitext alignment,*

semantic sense projection, and multi-dimensional stylistic stratification. By critically analyzing rule-based, distribution-based, and neural contextual paradigms, this study outlines a multi-tier annotation architecture that bridges cross-lingual lexicographical gaps between typologically distinct languages.

Keywords: *computational linguistics, English-Uzbek parallel corpus, semantic annotation, stylistic annotation, Word Sense Disambiguation (WSD), cross-lingual projection, translation equivalence.*

INTRODUCTION

The rapid advancement of digital linguistics and artificial intelligence systems necessitates elevating the automated understanding and processing of natural language resources to a qualitatively new level. One of the most complex phenomena intrinsic to all natural languages is lexical ambiguity, comprising polysemy and homonymy. The computational task of automatically identifying the exact contextual meaning of an ambiguous lexical unit is defined in computational linguistics as Word Sense Disambiguation (WSD) [1. 2009:P.12].

In developing bilingual lexicons and machine translation engines, traditional rule-based or static dictionary-based techniques fail to comprehensively cover the breadth of contextual and stylistic variations. Consequently, contemporary research increasingly recognizes multilingual parallel corpora as one of the most effective empirical foundations for extracting semantic information, discretizing source-language word senses via target-language translation equivalents, and capturing domain-specific registers [2. 2013:P.38].

This paper presents a systematic literature review and methodological framework for stylistic and semantic annotation within English-Uzbek parallel corpora, detailing cross-lingual sense projection mechanisms and theoretical-linguistic modeling strategies.

MAIN PART AND METHODOLOGY

1. Word Sense Disambiguation (WSD) and Cross-Lingual Semantic Potential

In resolving lexical ambiguity, parallel corpora function as a cross-lingual semantic mirror. According to foundational principles of corpus-based lexical semantics, distinct senses of a polysemous word in a source language are frequently lexicalized into distinct lexical units in the target language [3. 2002:P.255].

Compared to monolingual WSD methods relying solely on predefined inventories (such as WordNet), bilingual parallel bitexts exploit human translational choices as an implicit, unsupervised semantic annotation layer. For instance, the semantic spectrum of the English polysemous verb “to adopt” is systematically disambiguated through its Uzbek translation equivalents:

Legal and Official Context: “qabul qilmoq” – when applied to laws, regulations, resolutions, or policies within legislative and administrative discourse;

Social and Familial Context: “farzandlikka olmoq” – when applied to domestic, familial, and social relationships.

Consequently, the deterministic lexical boundary in the target language maps the source polysemous unit into its correct semantic cluster without requiring exhaustive manual annotation [1. 2009:P.28].

2. Multi-Dimensional Stylistic Stratification

Stylistic variations exert a decisive influence on lexical selection and syntactic structuring. In accordance with Douglas Biber’s multi-dimensional analysis, linguistic features demonstrate systematic co-occurrence patterns across communicative registers [4. 2004:P.72]. Neglecting functional register distinctions in parallel corpora leads to unnatural stylistic shifts in translation memories and bilingual lexicons.

Within this framework, parallel bilingual segments are classified across four primary functional styles:

1. Official and Administrative Style (OFF): Distinguished by formulaic phrasing, performative verbs, standardized legal terminology, and nominal passive constructions.

2. Scientific and Academic Style (SCI): Characterized by high nominal density, objective modality, conceptual rigor, and specialized domain terminology.

3. Literary and Artistic Style (LIT): Defined by expressive figurative language, idiomatic phraseology, cultural realia, and contextual metaphors.

4. Publicistic and Media Style (PUB): Relies on evaluative lexicon, persuasive rhetoric, concise informational packaging, and dynamic discourse markers.

3. Alignment Pipelines and Contextual-Lexicographical Modeling

The pipeline for constructing a stylistically and semantically annotated parallel corpus is structured across four sequential linguistic and computational stages:

1. Segmentation and Lexical Normalization: Pre-processing raw bilingual texts, standardizing orthographic conventions in the Uzbek Latin alphabet, performing sentence boundary detection, and identifying multi-word expressions (MWEs).

2. Sentence and Word-Level Alignment: Parallel corpora are aligned at the sentence level using neural sentence encoders (LaBSE, LASER), followed by sub-sentential alignment via statistical models (IBM Models 1–5, HMM) and neural cross-attention representations (awesome-align, fast_align) [5. 2003:P.88].

3. Contextual Semantic and Stylistic Tagging: Pre-trained multilingual contextual models (mBERT, XLM-RoBERTa) analyze syntagmatic relations within a contextual window ($C = \pm 5$ tokens), assigning fine-grained semantic sense identifiers and stylistic register tags.

4. Bilingual Lexicon Integration: The verified bilingual alignments and sense mappings are integrated into an interactive lexical database, transforming static dictionary entries into context-grounded bilingual lookup repositories.

CONCLUSION

The methodology of stylistic and semantic annotation in English-Uzbek parallel corpora provides a rigorous linguistic and computational framework for natural language processing. Leveraging parallel bitexts as cross-lingual semantic anchors significantly improves the accuracy of Word Sense Disambiguation (WSD). Furthermore, incorporating functional register tags ensures register-congruent lexical selection in machine translation engines and enriches domain-specific digital lexicography for the Uzbek language.

REFERENCES

1. Navigli R. Word sense disambiguation: A survey // ACM Computing Surveys (CSUR). – New York: ACM, 2009. – Vol. 41. – No. 2. – P. 1–69.
2. Ide N., Véronis J. Multilingual Corpus Resources and Tools for Word Sense Disambiguation // Word Sense Disambiguation: Algorithms and Applications. – Dordrecht: Springer, 2013. – P. 35–64.
3. Diab M., Resnik P. An Unsupervised Method for Word Sense Tagging using Parallel Corpora // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). – Philadelphia: ACL, 2002. – P. 255–262.
4. Biber D. Register, Genre, and Style // Cambridge Textbooks in Linguistics. – Cambridge: Cambridge University Press, 2004. – P. 45–88.
5. Och F. J., Ney H. A Systematic Comparison of Various Statistical Alignment Models // Computational Linguistics. – Cambridge: MIT Press, 2003. – Vol. 29. – No. 1. – P. 19–51.