

CRITERIA FOR COMPILING AND ANNOTATING AN UZBEK- ENGLISH CORPUS OF CHEMICAL TERMS

Valiyeva Nilufar Shamsitdinovna

Senior Teacher, Department of English Literature and Translation Studies,

Bukhara State University (Uzbekistan)

E-mail: valiyeva.nilufar2611@gmail.com

Abstract: *This article examines criteria for compiling and annotating an Uzbek-English corpus of chemical terms. On the basis of corpus linguistics, terminology theory, Uzbek terminology studies, and IUPAC chemical terminology, the study proposes principles for source selection, term inclusion, metadata description, bilingual alignment, and annotation quality. The corpus is interpreted as a structured linguistic and terminological resource rather than a simple bilingual word list.*

Keywords: *Uzbek-English terminology, chemical terms, corpus linguistics, annotation, metadata, bilingual corpus.*

Annotatsiya

Maqolada o'zbek-ingliz kimyo terminlari korpusini shakllantirish va annotatsiyalash mezonlari yoritiladi. Korpus lingvistikasi, terminologiya nazariyasi, o'zbek terminologiyasi va IUPAC kimyo terminologiyasi asosida manba tanlash, terminlarni kiritish, metadata berish, ikki tilli moslikni belgilash va annotatsiya sifatini nazorat qilish tamoyillari taklif etiladi. Korpus oddiy ikki tilli so'zlar ro'yxati emas, balki lingvistik va terminologik jihatdan tuzilgan ilmiy resurs sifatida talqin qilinadi.

Kalit soʻzlar: oʻzbek-ingliz terminologiyasi, kimyo terminlari, korpus lingvistikasi, annotatsiya, metadata, ikki tilli korpus.

Аннотация

В статье рассматриваются критерии формирования и аннотирования узбекско-английского корпуса химических терминов. На основе корпусной лингвистики, теории терминологии, исследований узбекской терминологии и химической терминологии IUPAC предлагаются принципы отбора источников, включения терминов, описания метаданных, двуязычного соотнесения и контроля качества аннотации. Корпус понимается не как простой двуязычный список слов, а как структурированный лингвистический и терминологический научный ресурс.

Ключевые слова: узбекско-английская терминология, химические термины, корпусная лингвистика, аннотация, метаданные, двуязычный корпус.

1. Introduction

The development of bilingual scientific terminology requires more than the mechanical collection of isolated equivalents. In chemical terminology, the problem is especially important because terms are connected with highly structured concepts, international standards, specialized communication, and field-specific patterns of use. For this reason, an Uzbek-English corpus of chemical terms should be designed as a research-oriented resource that brings together corpus methodology and terminological analysis.

McEnery and Hardie define corpus linguistics not as the study of one isolated language level, but as a set of methods for studying language through

machine-readable textual data selected for specific research questions [1, pp. 1-2]. This point is central to the present article: a corpus of Uzbek-English chemical terms must be built in relation to a clearly formulated task. It should not merely accumulate words from any available source; it should answer research questions about term structure, equivalence, usage, and standardization.

In Uzbek terminology studies, Dadaboyev emphasizes that terminology occupies a specific position in the structure of the Uzbek literary language and is connected with specialized scientific and technical meanings [4, pp. 3-8]. This view supports the need to treat Uzbek chemical terms as part of a terminological system rather than as ordinary lexical units. At the same time, Cabré's theory of terminology shows that terms are linguistic, cognitive, and communicative units: they have form, meaning, field-related function, and a role in specialized communication [3, pp. 42-48, 80-86]. The article therefore approaches the proposed corpus as a bilingual terminological resource with linguistic, semantic, and communicative layers.

2. Main part

A corpus must be designed according to the purpose for which it is used. Sinclair states that the contents of a corpus should be selected according to external criteria related to communicative function, not according to the researcher's prior expectation about the linguistic forms that may occur in the texts [2, ch. 1]. This principle is important for a chemical terminology corpus because the selection of texts should be based on document type, domain, authorship, and communicative function, not on a desire to find only convenient examples.

McEnery and Hardie also stress that the selected data must be well matched to the research question [1, pp. 2-3]. For the present topic, this means that chemistry textbooks, glossaries, standards, scientific articles, laboratory manuals, and officially recognized terminological materials may serve different functions and therefore should be recorded separately in metadata. The corpus may include both English and Uzbek materials, but their selection must remain transparent and reproducible.

Leech defines corpus annotation as the addition of interpretative linguistic information to a corpus [2, ch. 2]. He identifies several kinds of annotation, including part-of-speech, syntactic, semantic, discourse, stylistic, and lexical annotation. For an Uzbek-English chemical terminology corpus, the most relevant layers are lexical, grammatical, structural, semantic, terminological, and bilingual-equivalence annotation.

Leech also argues that annotation gives added value to the corpus, provided that it is well planned, documented, and consistent [2, ch. 2]. This requirement is directly applicable to chemical terms. If a term is annotated only as a word form, the corpus remains limited. If the same term is linked to its grammatical category, field, source, definition, context, equivalent, and standardization status, the corpus becomes a reusable scientific resource.

Burnard's discussion of metadata is equally important. He describes metadata as information about data and explains that without metadata corpus users cannot determine the provenance, authenticity, or relevance of a text [2, ch. 3]. In the present model, metadata is not an optional addition. It is required for

every source text and for every annotated term record because chemical terms may vary according to subfield, document type, date, and standardization context.

Cabr  defines terminology as a field concerned with specialized terms and emphasizes its interdisciplinary character [3, pp. 1-10, 25-48]. A terminological unit is not only a lexical form; it is a unit connected with a concept in a specialized field. Cabr  notes that terms have formal, semantic, and functional dimensions, and that in discourse they often combine with other terms in phraseological patterns [3, pp. 80-86].

Dadaboyev's description of Uzbek terminology provides the national linguistic basis for this approach. He distinguishes terms from general lexical units and explains that scientific and technical terms function within terminological systems and terminological fields [4, pp. 5-8]. He also discusses term formation, terminological dictionaries, and the role of international term elements in Uzbek terminology [4, pp. 79-121]. Therefore, the Uzbek component of the corpus must be capable of recording both native Uzbek term-formation resources and international elements.

The IUPAC Gold Book serves as the international chemical reference in this article. It is a compendium of chemical terminology and contains standardized entries across branches such as atmospheric chemistry, analytical chemistry, organic chemistry, polymer science, stereochemistry, photochemistry, and related fields [5]. For a bilingual chemical corpus, such a source provides a basis for identifying English chemical terms and for checking whether a term belongs to internationally recognized chemical terminology.

The first criterion is purpose. A corpus is useful only when it is related to a research problem. For an Uzbek-English chemical terminology corpus, the purpose may include describing the structural types of chemical terms, identifying bilingual equivalence, documenting international term elements, supporting technical translation, and preparing a basis for terminological dictionaries. These purposes are consistent with McEnery and Hardie's view of corpus linguistics as a method for investigating language data and with Cabré's view of terminology as both a theoretical and practical field [1; 3].

The second criterion is source selection. Sinclair's distinction between external and internal criteria is essential here [2, ch. 1]. For a chemical terminology corpus, external criteria may include the type of document, the field of chemistry, the publication status, the intended audience, and the language of the text. Internal criteria, such as the researcher's expectation that a source contains many desirable terms, should not be the primary basis for selection. This prevents the corpus from becoming a collection shaped by confirmation bias.

Representativeness and balance should be treated as guiding aims rather than absolute guarantees. Sinclair and McEnery and Hardie both show that corpus representativeness is a complex problem and depends on the research purpose [1, pp. 6-13; 2, ch. 1]. In the present case, the corpus should not claim to represent the whole Uzbek language or the whole English language. It should represent a defined segment: Uzbek and English chemical terminology as reflected in selected specialized sources.

Every source included in the corpus should have metadata. Following Burnard, metadata should allow the researcher to identify the source, evaluate its

relevance, and understand its provenance [2, ch. 3]. For a chemical terminology corpus, the minimum metadata should include source title, author or institution, year, language, genre, chemistry subfield, publication type, page or section reference, and reliability status.

Metadata field	Purpose in the corpus
Source title and bibliographic data	To make every extracted term traceable to a documented source.
Language	To distinguish Uzbek, English, and bilingual or translated materials.
Document type	To separate textbooks, standards, articles, manuals, dictionaries, and glossaries.
Subject field or subfield	To classify terms according to chemistry-related domains.
Date or edition	To record possible diachronic or standardization differences.
Extraction location	To identify the page, section, or entry from which the term was taken.

The corpus should include only units that can be justified as terminological units. Cabré's account of terms as units connected with concepts in specialized fields [3, pp. 80-86] and Dadaboyev's account of terms as units expressing special

scientific and technical meanings [4, pp. 5-8] provide the basis for inclusion. The following types of units may be admitted when they are documented in verified sources.

Unit type	Source-based justification
Simple terms	Cabr� treats terms as linguistic units with formal and semantic properties; Dadaboyev describes terms as special lexical units.
Derived terms	Dadaboyev discusses term formation in Uzbek, including morphological processes; Cabr� describes morphemic structure.
Compound terms	Both general terminology theory and Uzbek terminology studies allow multi-component terminological units.
Terminological phrases	Cabr� notes that terms may occur in phraseological combinations in specialized discourse.
Initialisms and abbreviations	Cabr� treats truncation processes such as initialisms and abbreviations as relevant to terminology.
International term elements	Dadaboyev discusses international prepositive term elements and their role in Uzbek terminology.

A bilingual corpus should also distinguish terms from nomenclature. Dadaboyev discusses the relationship and difference between terminological units and nomenclature [4, pp. 8-10]. This distinction is particularly relevant in chemistry because chemical nomenclature, class names, measurement terms, and general scientific terms may coexist in the same text. The corpus should therefore mark whether a unit functions as a general chemical term, a nomenclature item, a class name, a measurement-related expression, or a phraseological term combination.

Annotation should be layered. A single label cannot capture the linguistic and terminological complexity of chemical terms. Leech's discussion of different annotation types supports a multi-layered model [2, ch. 2]. Cabré's theory supports formal, semantic, functional, and multilingual annotation [3]. Dadaboyev's Uzbek terminology framework supports structural and word-formation annotation [4].

Annotation layer	Fields to be recorded
Formal layer	Term form, spelling, script, abbreviation status, variant form.
Grammatical layer	Part of speech or phrase type where this is relevant and source-supported.
Structural layer	Simple, derived, compound, phrase, initialism, abbreviation, international term element.
Semantic layer	Conceptual note, definition source, subject field,

Annotation layer	Fields to be recorded
	relation to other terms.
Bilingual layer	English form, Uzbek form, equivalent status, source of equivalent, alignment note.
Standardization layer	IUPAC-related status for English chemical terms; Uzbek standardization status when a verified source is available.
Context layer	Documented sentence or context reference, page or entry location.
Editorial layer	Annotator, date, revision status, uncertainty note.

The bilingual layer must be treated with special caution. If no verified Uzbek equivalent is available in the provided sources, the corpus should not invent one. Instead, the term should be marked as requiring verification. This approach is consistent with Cabré's treatment of terminological equivalence as a central issue in multilingual terminology and with the general principle that terminology prepared for translators should include context and conceptual information [3, pp. 47-48].

An Uzbek-English chemical terminology corpus may be parallel, comparable, or mixed, depending on the sources. Azevedo's study, although not devoted to chemistry, is useful as a methodological reference because it treats

corpus annotation as the codification of linguistic knowledge in a parallel corpus and works with alignment units [6, pp. iv, 32-41]. For chemical terminology, alignment should not be based only on surface similarity. It should consider source context, subject field, definition, and standardization evidence.

Equivalence should therefore be annotated by status. A direct equivalent may be recorded only when the source supports it. A partial equivalent may be recorded when the terms overlap conceptually but differ in scope, usage, or standardization. A term without a verified Uzbek or English counterpart should be marked as pending. This status-based approach prevents the corpus from presenting uncertain correspondences as established facts.

The proposed model shows that a bilingual chemical terminology corpus is not identical to a bilingual dictionary. A dictionary normally presents selected entries for users, whereas a corpus preserves evidence of use, source context, metadata, and annotation history. However, Cabré's discussion of terminological records and data banks shows that terminographic practice and computerized terminology are closely connected [3, pp. 152-193]. A corpus built according to the criteria proposed here may later support dictionary compilation, but it should first preserve evidence in a structured and verifiable form.

The model is also compatible with the needs of Uzbek terminology development. Dadaboyev notes that Uzbek terminology includes native and borrowed layers, structural-grammatical patterns, semantic processes, and international term elements [4]. A corpus that records such features can help researchers observe how Uzbek chemical terminology is formed, how international elements are adapted, and how specialized concepts are represented in Uzbek.

The role of IUPAC is not to replace Uzbek terminology, but to provide an internationally recognized reference point for English chemical terminology and standard chemical concepts. The IUPAC Gold Book contains standardized entries and source references for chemical terms [5]. In a bilingual corpus, this can help identify whether an English term has a recognized status in international chemical discourse. The Uzbek equivalent, however, should be entered only when a verified Uzbek source supports it.

The main limitation of the present article is deliberate: it does not provide a fabricated corpus sample. This limitation protects the reliability of the study. Future research may apply the proposed criteria to a pilot dataset, but such a dataset must be compiled from documented Uzbek and English chemical sources and must preserve metadata and annotation history.

3. Conclusion

An Uzbek-English corpus of chemical terms should be compiled as a structured, documented, and annotatable terminological resource. The theoretical sources examined in this article support six main principles. First, the corpus must be designed according to a clear research purpose. Second, sources must be selected through external and transparent criteria. Third, metadata is necessary for source reliability and later verification. Fourth, term inclusion must be based on a terminological understanding of specialized units. Fifth, annotation should include formal, grammatical, structural, semantic, bilingual, standardization, contextual, and editorial layers. Sixth, bilingual equivalence must not be invented; it must be recorded only when supported by verified sources.

The proposed criteria make it possible to move from an ordinary list of Uzbek and English chemical words toward a scientifically grounded corpus resource. Such a corpus can serve terminological research, specialized translation, lexicography, language planning, and the further development of Uzbek scientific terminology.

References

1. McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
2. Wynne, M. (Ed.). (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: AHDS.
3. Cabré, M. T. (1999). *Terminology: Theory, Methods and Applications*. Edited by J. C. Sager. Translated by J. A. DeCesaris. Amsterdam/Philadelphia: John Benjamins.
4. Dadaboyev, H. (2020). *O'zbek terminologiyasi*. Toshkent: Yoshlar nashriyot uyi.
5. International Union of Pure and Applied Chemistry. (2014). *Compendium of Chemical Terminology: Gold Book*. Version 2.3.3.
6. Azevedo, F. (2012). *The Problem of Codifying Linguistic Knowledge in Two Translations of Shakespeare's Sonnets: A Corpus-Based Study*. Doctoral dissertation, Universidade Federal de Santa Catarina.