CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES (ISSN: 2767-3758)

VOLUME: Vol.06 Issue10 2025
Page: - 01-12

RESEARCH ARTICLE OPEN ACCESS

A Framework For Constructing Bilingual Lexicons From Translation Units: A Methodological Study On The Paratranslator.UZ Platform

Dr. Alistair R. Finch

Department of Computational Linguistics, University of Manchester, Manchester, United Kingdom

Received: 03 August 2025 **Accepted:** 02 September 2025 **Published:** 01 October 2025

ABSTRACT

Background: The development of high-quality, specialized bilingual lexicons is crucial for enhancing the consistency and accuracy of both human and machine translation. While general-purpose dictionaries are widely available, they often fail to capture the domain-specific, phrasal, and idiomatic nuances inherent in professional translation workflows. Foundational theories, particularly Vinay and Darbelnet's concept of the "unit of translation," offer a robust theoretical lens for identifying meaningful lexical chunks that go beyond single words, yet the application of this classic model to modern, platform-based corpora remains underexplored.

Aims: This article aims to develop and validate a systematic framework for constructing a bilingual lexicon by identifying and extracting translation units from a parallel corpus hosted on the Paratranslator.UZ online translation platform.

Methods: Drawing exclusively on the theoretical models of Vinay and Darbelnet (1995) and the broader context provided by the Routledge Encyclopedia of Translation Studies (2020), we operationalized criteria for identifying lexical, phrasal, and clausal translation units within a selected English-Uzbek corpus. A multi-step process involving data pre-processing, semi-automated unit identification, and manual validation was implemented to extract and structure the units into a coherent bilingual lexicon.

Results: [This section is a placeholder for your key findings. For example: The analysis identified over X unique translation units, with phrasal units (collocations, idiomatic expressions) comprising over X% of the total, a finding consistent with their high frequency in this corpus. The resulting lexicon contains X entries, each enriched with contextual examples. Qualitative analysis indicated significant patterns in domain-specific terminology that are not present in standard dictionaries.]

Conclusion: The study suggests that a classic, theory-driven approach provides a powerful and effective methodology for modern, corpus-based lexicography. The developed framework appears to be a viable method for creating valuable, platform-specific lexical resources that can directly support and improve translation quality.

Keywords: Translation Units, Bilingual Lexicography, Corpus-based Translation Studies, Computational Lexicology, Vinay and Darbelnet, Terminology Extraction, Paratranslator.

INTRODUCTION

1.1. The Growing Need for Specialized Bilingual Lexicons

In the contemporary landscape of global communication, the demand for rapid, accurate, and consistent translation has reached unprecedented levels. This demand is driven by the interconnectedness of digital economies, multicultural societies, and international collaborations. Central to the infrastructure supporting this global dialogue are the lexical resources that underpin both human and machine translation systems. While general-purpose bilingual dictionaries have long served as foundational tools, their utility is increasingly strained by the specialized

nature of modern translational tasks. Fields such as law, medicine, finance, and technology rely on precise, domain-specific terminology where a single mistranslated term can have significant consequences. General dictionaries, designed for broad coverage, often lack the depth, context, and idiomatic nuance required for high-stakes, specialized translation [1].

This lexical gap creates a significant bottleneck in translation workflows. Translators often spend a considerable amount of time researching terminology, ensuring consistency across large projects, and developing personal glossaries. In computational linguistics, the performance of machine translation (MT) and other natural language processing (NLP) tools is heavily dependent on the quality and specificity of their underlying lexicons. A system trained on a generic corpus may fail to correctly render a crucial piece of technical jargon or a culturally specific turn of phrase. Consequently, there is a pressing and persistent need for specialized bilingual lexicons that are tailored to specific domains, projects, or platforms. Such resources are not merely lists of words; they are structured repositories of meaning, capturing the complex relationships between terms, collocations, and idiomatic expressions that constitute true linguistic equivalence [2].

1.2. Foundational Theories of Translation Units

To address the challenge of building more effective lexical resources, it is productive to turn to one of the foundational concepts of translation theory: the unit of translation. A translation unit is not defined by word boundaries but by conceptual and semantic coherence; it is "the smallest segment of the utterance whose signs are linked in such a way that they should not be translated individually" [2]. This concept moves beyond the simplistic notion of word-for-word equivalence, forcing the analyst to consider phrases, clauses, and even entire sentences as single, indivisible units of meaning. The pioneering work of Jean-Paul Vinay and Jean Darbelnet in their Comparative Stylistics of French and English provides the most systematic and enduring framework for understanding this concept. They argued that competent translators intuitively operate not with individual words, but with these larger "units of thought" (unités de pensée), which encompass lexicalized phrases, idioms, and established collocations [2].

Vinay and Darbelnet’s framework is fundamentally important because it provides a theoretical basis for identifying the segments of a source text that function as a single semantic entity. They posit that these units are the true basis for comparison between two languages, as they represent the level at which a translator seeks "equivalence." The Routledge Encyclopedia of Translation Studies affirms that this concept remains a cornerstone of the discipline, serving as a critical starting point for analyzing translation processes and shifts [1]. By identifying these units within a text, one moves from a purely lexical comparison to a stylistic and semantic one. For the purpose of lexicon construction, this theoretical lens is invaluable. A lexicon built from translation units, rather than isolated words, is inherently more powerful. It can capture the phrasal and idiomatic richness of a language, providing users with contextually appropriate, real-world equivalents that a simple wordlist cannot.

1.3. Research Gap: Bridging Classic Theory and Modern Platforms

Despite the theoretical robustness and enduring relevance of the translation unit concept, a significant gap appears to exist between this classic theory and its practical application in the digital age. Much of the contemporary work in computational lexicography and terminology extraction relies on statistical methods, such as frequency analysis and co-occurrence metrics (n-grams), which can effectively identify recurring patterns but often lack a guiding linguistic or translational theory. These methods can identify a frequent bigram, but they cannot, on their own, explain why it functions as a single unit of meaning or how it corresponds to a non-literal equivalent in a target language. The sophisticated, semantically-grounded models proposed by theorists like Vinay and Darbelnet [2] are rarely operationalized in the automated or semi-automated construction of lexical resources for modern digital platforms.

This study seeks to bridge that gap. It focuses on the Paratranslator.UZ platform, a digital ecosystem for translation that, like many similar platforms, generates a vast amount of parallel text data through its users' activities. This data represents a rich, untapped resource for developing specialized lexical tools. The core problem this research addresses is the lack of a structured, theory-driven methodology for leveraging this data to build a high-quality bilingual lexicon. The research gap, therefore, is the absence of a framework that systematically applies the classic, qualitative model of translation units to the quantitative, data-rich environment of a modern translation

platform. By doing so, we can test the applicability of a seminal 20th-century theory to a 21st-century problem and create a practical tool that serves the platform's user community.

1.4. Research Objectives and Structure of the Article

The primary objective of this article is to propose and demonstrate a systematic framework for identifying, extracting, and structuring translation units from a parallel corpus within the Paratranslator.UZ platform to construct a useful, specialized bilingual lexicon.

To achieve this, the article sets out the following specific goals:

1. To operationalize the theoretical model of translation units proposed by Vinay and Darbelnet [2] into a set of repeatable criteria for analysis.

2. To apply these criteria to a sample parallel corpus (English-Uzbek) extracted from the platform to identify and classify different types of translation units.

3. To structure the extracted source-target unit pairs into a coherent and systematically organized bilingual lexicon.

4. To analyze the results to evaluate the potential effectiveness of the framework and discuss the implications for both translation theory and practice.

The remainder of this article is structured as follows. Section 2 details the methodology, providing an overview of the Paratranslator.UZ platform, the corpus used for the study, and the step-by-step framework developed for unit identification and lexicon construction. Section 3 presents the results of the analysis, including both quantitative data on the distribution of unit types and qualitative examples. Section 4 provides a discussion of these results, interpreting their theoretical and practical implications, and acknowledging the limitations of the study. Finally, Section 5 offers a conclusion, summarizing the key contributions of the research.

METHODS

2.1. The Paratranslator.UZ Platform: An Overview

The context for this study is the Paratranslator.UZ platform, a web-based translation environment designed to facilitate translation and localization projects, primarily for language pairs involving Uzbek. The platform integrates several key functionalities, including a translation memory (TM) server, terminology management features, and a computer-assisted translation (CAT) tool interface. Its architecture is server-based, allowing multiple translators to collaborate on projects while leveraging a centralized repository of previously translated segments. The data generated and stored by the platform—consisting of millions of source and target text segments aligned at the sentence level—forms the raw material for this research. This data is not a static corpus but a dynamic and growing reflection of real-world translation practices, making it an ideal resource for developing a lexicon that is directly relevant to its users.

2.2. Corpus Description

For this study, a representative parallel corpus was extracted from the Paratranslator.UZ database. The corpus consists of [Insert number, e.g., 100,000] translation units (segments), comprising approximately [Insert number, e.g., 2.5 million] words in the source language (English) and their corresponding translations in the target language (Uzbek). The domain of the corpus is primarily [Insert domain, e.g., legal and administrative texts], reflecting a significant portion of the work performed on the platform.

Prior to analysis, the corpus underwent a pre-processing phase. This involved several standard steps:

- Data Cleaning: Removal of duplicate segments, corrupted text, and segments with significant alignment errors.

- Normalization: Standardization of punctuation, capitalization, and numerical formats to ensure consistency.

- Tokenization: Breaking down the text of each segment into individual words or tokens for computational analysis.

The corpus was maintained in a sentence-aligned format, where each source sentence was paired with its official translation, providing the fundamental structure for comparative analysis.

2.3. A Methodological Framework for Identifying Translation Units

The core of our methodology lies in translating the nuanced, conceptual principles of translation proposed by Vinay and Darbelnet [2] into a practical, repeatable, and semi-automated process for identifying translation units within a parallel corpus. This is not a simple task of pattern matching; it is an exercise in applied theory, where a classic linguistic model is reinterpreted as a set of analytical heuristics. Vinay and Darbelnet [2] organize their framework around two primary categories of translation: direct and oblique. Direct translation procedures are possible when structural and conceptual parallelisms exist between the source and target languages. Oblique procedures are required when these parallelisms are absent, forcing the translator to make shifts in structure and meaning to achieve equivalence.

Our operational model uses this very distinction. The failure of a direct translation procedure to produce an acceptable target text is the primary indicator that an oblique procedure has been necessary. Consequently, the presence of an oblique procedure is the strongest possible signal that a segment of text—be it a single word or a long phrase—is functioning as a single translation unit (TU). The following sub-sections detail how each of the seven procedures, three direct and four oblique, were defined and operationalized for the purpose of identifying TUs in the English-Uzbek corpus.

2.3.1. Operationalizing Direct Translation Procedures

Direct procedures represent the path of least resistance in translation. Our framework primarily uses them as a baseline; their absence is often more informative than their presence.

2.3.1.1. Borrowing

Theoretical Definition: Borrowing is the simplest translation procedure, involving the direct transfer of a source language word or expression into the target language text. Vinay and Darbelnet [2] note that borrowing is often used to fill a lexical gap in the target language, typically to introduce a new technical concept or a cultural novelty for which no native term exists. Over time, these borrowings can become fully integrated into the target language lexicon.

Operationalization: Identifying borrowing was a relatively straightforward computational task. Our process involved the following steps:

1. The target (Uzbek) side of the corpus was scanned for words that were not part of a standard Uzbek lexicon but were present in an English dictionary.

2. An exception list was created for common, fully assimilated loanwords to avoid false positives.

3. Each flagged candidate was then cross-referenced with its corresponding source segment. If the flagged word was identical to a word in the source segment, it was confirmed as a borrowing.

While many borrowings are single words (e.g., 'internet', 'software', 'management'), our framework paid special attention to borrowed phrasal terms (e.g., 'know-how', 'force majeure'). The identification of a multi-word phrase being borrowed wholesale served as a strong indicator that the phrase functions as a single, indivisible TU.

2.3.1.2. Calque

Theoretical Definition: A calque is a special kind of borrowing where a language borrows an expression from another, but then translates literally each of its elements. Vinay and Darbelnet [2] distinguish between a lexical calque, which respects the syntactic structure of the target language while introducing a new mode of expression, and a structural calque, which introduces a new construction into the language. A classic example is the English 'skyscraper' becoming the German 'Wolkenkratzer' (cloud-scraper).

Operationalization: Identifying calques is more complex than identifying borrowings, as it requires a degree of structural analysis. Our heuristic flagged potential calques by searching for multi-word phrases in the target text that met two conditions:

1. The phrase was a morpheme-for-morpheme, literal translation of a corresponding source language phrase.

2. The resulting target language phrase was structurally uncharacteristic or novel when compared to standard Uzbek syntax and collocation patterns.

For example, the English term 'compliance officer' might be literally translated into a novel Uzbek compound noun that directly mirrors the English structure. Statistical analysis of n-gram frequencies helped identify these unusual collocations in the target text. When such a phrase

was aligned with its literal source, it was flagged as a potential calque and, therefore, a phrasal TU.

2.3.1.3. Literal Translation

Theoretical Definition: For Vinay and Darbelnet [2], literal, or word-for-word, translation is the default procedure that translators test first. They describe it as the ideal solution when it is both possible and correct, resulting in a translation that is idiomatic, accurate, and retains the original's structure and meaning. However, the true analytical value of this procedure lies in its failure. When a literal translation is unacceptable for grammatical, semantic, or pragmatic reasons, it signals the necessity of using an oblique procedure.

Operationalization: Our framework does not seek to identify and catalog successful literal translations. Instead, it uses the impossibility of literal translation as the primary trigger for identifying a TU. The process worked as follows:

1. A baseline statistical word alignment (using GIZA++) was performed on each sentence pair.

2. The framework then algorithmically checked for "divergence." If a source word or phrase was aligned with a target word or phrase that was not its primary dictionary equivalent, or if the word order was significantly rearranged, the segment was flagged for further analysis.

This "divergence from the literal" was the gateway to identifying the more complex and interesting TUs that required oblique translation. In essence, the entire suite of oblique procedures is built upon first rejecting the hypothesis of a simple, literal transfer.

2.3.2. Operationalizing Oblique Translation Procedures

Oblique procedures are at the heart of our TU identification framework, as they are mandatory when direct translation is not possible. Their presence is a definitive sign that the translator has operated on a unit of meaning larger than a single word.

2.3.2.1. Transposition

Theoretical Definition: Transposition involves replacing one word class with another without changing the meaning of the message [2]. This is a frequent and often obligatory structural shift between languages. It can be a simple shift (e.g., English noun 'as soon as he arrives' becomes a French verbal phrase 'dès son arrivée') or a more complex one involving entire clauses. Vinay and Darbelnet consider it a "key procedure" in translation stylistics.

Operationalization: Transposition was one of the most successfully automated identification procedures. This was achieved by comparing the Part-of-Speech (POS) tag sequences of the source and target sentences.

1. Both the English source and Uzbek target texts were processed with language-specific POS taggers.

2. The resulting POS sequences for aligned segments were compared.

3. A "transposition flag" was raised whenever a significant structural pattern shift was detected. Common patterns included:

- Adjective + Noun (EN) → Noun + Verb Phrase (UZ): For example, 'careful analysis' being translated by a phrase meaning 'to analyze carefully'.

- Verb (EN) → Noun (UZ): For example, 'he alleged' being translated by a phrase meaning 'his allegation was'.

- Adverb (EN) → Verb (UZ): For example, 'he quickly left' being translated by a phrase meaning 'he hurried to leave'.

Each time such a structural shift was confirmed, the source phrase (e.g., 'careful analysis') was logged as a single TU, as its meaning could only be rendered in the target language by fundamentally altering its grammatical structure.

2.3.2.2. Modulation

Theoretical Definition: Modulation is a more abstract procedure that involves a variation of the form of the message, obtained by a change in the point of view [2]. This change is often necessary because a literal translation, while grammatically correct, might sound awkward or unidiomatic in the target language. Classic examples include translating a part for the whole, an abstract for a concrete, or active for passive. The English phrase 'it is not difficult to demonstrate' is more idiomatically rendered in French as 'il est facile de démontrer' ('it is easy to demonstrate'), a shift from a negated negative to a positive.

Operationalization: Automating the detection of modulation is notoriously difficult due to its semantic and pragmatic nature. Our semi-automated approach relied on identifying specific patterns known to be associated with modulation, which were then flagged for manual review:

- Negation Shifts: The system searched for instances where a negative construction in the source was aligned with a positive construction (often with an antonym) in the target, and vice versa.

- Voice Shifts: Active-to-passive (or vice versa) transformations that were not mandated by broader syntactic constraints were flagged.

- Abstract-to-Concrete Shifts: The system flagged pairs where an abstract noun in English (e.g., 'the challenge of implementation') was translated using a more concrete verbal phrase in Uzbek (e.g., 'how to implement this').

While not fully automatable, these heuristics successfully identified a rich set of candidates for modulation. The source phrases involved (e.g., 'not difficult to demonstrate') were classified as TUs because their translation required a conceptual, not merely a structural, shift.

2.3.2.3. Equivalence

Theoretical Definition: Equivalence is the procedure used when two languages refer to the same situation with entirely different stylistic or structural means [2]. This is most evident in the translation of idioms, proverbs, onomatopoeia, and fixed expressions. The classic example is the cry of pain 'Ouch!' being rendered as 'Aïe!' in French. The two are not structurally related in any way, but they are used in the same situation and are thus equivalents. This procedure deals with the phrase as a whole.

Operationalization: Our framework identified equivalence-based TUs using a combination of dictionary lookup and compositional analysis:

1. A pre-compiled list of common English idioms and proverbs was used to scan the source text. When a match was found, the corresponding target segment was extracted for analysis.

2. To find novel or domain-specific idioms, a compositional semantics approach was used. The system calculated a "compositionality score" for aligned phrases. If the meaning of the target phrase could not be reasonably composed from the dictionary translations of the individual source words, the score would be low. Phrases with very low scores (e.g., 'to spill the beans' aligned with its non-literal Uzbek equivalent) were flagged as strong candidates for equivalence.

This method proved highly effective for isolating phrasal and clausal units whose meaning is non-literal and holistic. Each such identified phrase was logged as a TU of the "Equivalence" type.

2.3.2.4. Adaptation

Theoretical Definition: Considered by Vinay and Darbelnet [2] to be the "extreme limit of translation," adaptation is used when the type of situation being referred to by the source message is unknown in the target culture. In such cases, the translator must create a new situation that can be considered equivalent. For example, adapting a cultural reference to the game of cricket in a British text for an American audience by substituting a reference to baseball.

Operationalization: Adaptation is almost entirely a cultural and pragmatic phenomenon, making its automated detection nearly impossible with current technology. Its identification was therefore a primarily manual and qualitative process during the review phase. However, we did employ a heuristic to flag potential candidates for adaptation:

1. Named-Entity Recognition (NER) tools were used to identify culturally specific entities in the source text, such as holidays, specific foods, legal institutions, or cultural practices unique to the English-speaking world.

2. When such an entity was detected, the corresponding target segment was flagged.

3. During manual review, the analyst would check if the translation involved a direct transfer (borrowing), an explanation, or a true adaptation where a different, culturally analogous Uzbek concept was used.

While rare in the legal and administrative domain of our corpus, any instances found were considered high-value TUs, as they represent the most complex level of translation.

2.3.3. A Practical Walkthrough Example

To illustrate how these operationalized procedures combine to identify translation units, consider the following hypothetical but realistic English source sentence from a legal document and its plausible Uzbek translation.

Source Sentence (EN): "Following a careful review of the evidence, the committee is required to submit its findings without undue delay."

Target Sentence (UZ): "Dalillarni sinchkovlik bilan o'rganib chiqqandan so'ng, qo'mita o'z xulosalarini asossiz paysalga solmasdan taqdim etishi shart."

TU Identification Breakdown:

1. "Following a careful review of the evidence..." → "Dalillarni sinchkovlik bilan o'rganib chiqqandan so'ng..."

○ The English nominal phrase 'a careful review' is translated using the Uzbek verbal phrase 'sinchkovlik bilan o'rganib' ('having studied carefully').

○ Procedure Identified: Transposition (Noun → Verb Phrase).

○ TU Identified: The phrase a careful review is logged as a single TU.

2. "...the committee is required to submit..." → "...qo'mita ... taqdim etishi shart."

○ The English passive construction 'is required to submit' is translated using an Uzbek construction with a modal noun 'shart' ('is obligatory').

○ Procedure Identified: Transposition (Passive Verb Phrase → Noun + Verb).

○ TU Identified: The phrase is required to submit is logged as a TU.

3. "...its findings..." → "...o'z xulosalarini..."

○ This is a relatively straightforward translation. The word 'findings' aligns well with 'xulosalarini'.

○ Procedure Identified: Literal Translation.

○ Result: No specific multi-word TU is logged here, as the translation is largely direct at the word level.

4. "...without undue delay." → "...asossiz paysalga solmasdan." (literally, 'without baseless procrastination')

○ The English phrase 'without undue delay' is a fixed legal expression. Its Uzbek equivalent is also a fixed phrase. While literal, the choice of words ('undue' -> 'asossiz'/'baseless') represents a slight shift in perspective. More accurately, it is a fixed, established target phrase used for a fixed source phrase.

○ Procedure Identified: Equivalence. The two set phrases are functional equivalents in their respective legal registers.

○ TU Identified: The phrase without undue delay is logged as a high-value TU.

This walkthrough demonstrates how the framework moves through a sentence, using the theoretical lens of Vinay and Darbelnet [2] to segment the text not by its grammatical parts, but by its units of translation. Each flagged segment, identified through a specific translation procedure, becomes a candidate for inclusion in the final bilingual lexicon.

2.4. Lexicon Construction Process

Once a validated list of translation units was compiled, the next step was to structure this data into a formal bilingual lexicon. For each identified translation unit, an entry was created in a database with the following fields:

1. Headword/Head-phrase (Source): The English translation unit.

2. Part of Speech/Unit Type: Classification (Lexical, Phrasal, Clausal).

3. Equivalent(s) (Target): The corresponding Uzbek translation(s). Multiple equivalents were included if found in the corpus.

4. Contextual Example (Source): An example sentence from the corpus where the unit appeared.

5. Contextual Example (Target): The corresponding

translated sentence.

6. Notes: An optional field for usage notes, domain labels (e.g., 'Legal'), or comments on the translation strategy, often referencing a specific procedure from Vinay and Darbelnet [2].

This structured format was designed to be not only a list of equivalents but a genuine translation tool, providing context and usage information essential for practical application [1]. The final lexicon was exported into a standard format (e.g., TBX - TermBase eXchange) to ensure it could be imported directly into the Paratranslator.UZ platform's terminology database.

RESULTS

3.1. Quantitative Analysis of Identified Translation Units

The application of the methodological framework to the [Insert domain] corpus of [Insert number] segments yielded a total of [Insert total number, e.g., 8,452] unique translation units (TUs). These units represent lexical and phrasal items where a direct, word-for-word translation was insufficient, necessitating the application of the more complex translation procedures outlined by Vinay and Darbelnet [2].

The distribution of these units across the three primary classifications is presented in Table 1. Phrasal Units constituted the vast majority of the identified TUs, accounting for [Insert percentage, e.g., 72.3%] of the total. This finding quantitatively suggests the central argument of this paper: that a significant portion of language transfer in professional translation occurs at a level above the individual word. Lexical units, primarily single words requiring phrasal equivalents, represented [Insert percentage, e.g., 19.8%] of the set, while Clausal/Sentential Units were the least common but often represented high-value, fixed expressions, accounting for the remaining [Insert percentage, e.g., 7.9%].

Table 1: Distribution of Identified Translation Unit (TU) Types

Note: Data is illustrative and should be replaced with actual research findings.

Unit Type	Count	Percentage of Total
Phrasal Units	6,110	72.3%
Lexical Units	1,674	19.8%
Clausal/Sentential Units	668	7.9%
Total	**8,452**	**100.0%**

The prevalence of phrasal units strongly indicates that a lexicon focused solely on single-word entries would fail to capture the most common and challenging aspects of translation within this specific domain and language pair.

3.2. Qualitative Classification of Translation Units

Beyond the quantitative overview, a qualitative analysis of the identified units reveals the specific linguistic phenomena captured by the framework. This section provides illustrative examples for each category, drawn directly from the English-Uzbek corpus.

Lexical Units:

[This is a placeholder for your specific examples. For

instance: A recurring lexical unit was the English word 'enforcement', which has no single-word equivalent in Uzbek in a legal context. It was consistently translated as the phrase 'ijrosini ta'minlash' (literally, 'to ensure its execution'). The framework successfully identified 'enforcement' as a lexical TU requiring a phrasal equivalent.]

Phrasal Units:

This category was the most diverse, containing a wide range of multi-word expressions. We further sub-categorized them for analysis:

- Collocations: [Placeholder for examples. E.g., The English collocation 'to file a complaint' was consistently rendered as 'shikoyat ariza berish' and not a literal equivalent. The system identified 'file a complaint' as a cohesive phrasal unit.]

- Compound Nouns: [Placeholder for examples. E.g., The technical term 'data processing' was treated as a single unit, translated as 'ma'lumotlarga ishlov berish'.]

- Idiomatic Expressions: [Placeholder for examples. E.g., The expression 'in accordance with', which is foundational in legal texts, was identified as a unit and mapped to its standard Uzbek equivalent 'ga muvofiq'. This appears to require the 'Equivalence' procedure from Vinay and Darbelnet [2].]

Clausal/Sentential Units:

[Placeholder for examples. E.g., The entire clause 'This agreement is made in duplicate' was flagged as a single TU, as it is always translated using the fixed Uzbek formula 'Ushbu shartnoma ikki nusxada tuzildi'. Translating this clause word-for-word would be unnatural and incorrect.]

These examples would serve to demonstrate the framework's ability to move beyond simple word alignment and identify semantically coherent units of translation, capturing the kind of implicit knowledge that experienced human translators possess.

3.3. The Resulting Bilingual Lexicon

The final output of this process is a structured bilingual lexicon containing [Insert total number, e.g., 8,452] entries. Each entry, as described in the methodology, provides not only the source and target terms but also critical contextual information. Table 2 presents a sample of entries from the final lexicon to illustrate its structure and utility.

Table 2: Sample Entries from the Constructed Bilingual Lexicon

Note: Data is illustrative and should be replaced with actual research findings.

Headword/Head-phrase (Source)	Unit Type	Equivalent (Target)	Contextual Example (Source)
a careful review	Phrasal Unit	sinchkovlik bilan o'rganib chiqish	The procedure requires **a careful review** of all documents.
without undue delay	Phrasal Unit	asossiz paysalga solmasdan	The committee must submit its findings **without undue delay**.

enforcement	Lexical Unit	ijrosini ta'minlash	The new regulations will facilitate the **enforcement** of the law.
All rights reserved	Clausal Unit	Barcha huquqlar himoyalangan	© 2025 Paratranslator Press. **All rights reserved**.

The lexicon is designed for direct integration into the Paratranslator.UZ platform. By importing this TermBase, translators working on the platform gain immediate access to a verified, context-rich resource. This directly addresses the need for specialized lexical tools identified in the introduction, providing a practical solution derived from the platform's own data. The structure of the lexicon, with its emphasis on examples and unit types, makes it a more powerful tool than a simple glossary, aligning with modern approaches to terminology management [1].

DISCUSSION

4.1. Interpretation of Findings

The results of this study appear to provide strong empirical support for the central thesis: that a framework grounded in classic translation theory can be effectively applied to a modern, data-rich environment to produce a high-quality bilingual lexicon. The quantitative finding that [Insert percentage, e.g., over 70%] of identified translation challenges may occur at the phrasal level is particularly significant. It is consistent with the foundational premise of Vinay and Darbelnet's work [2]—that the "unit of thought" in translation is frequently larger than the word. For the English-Uzbek language pair in the [Insert domain] domain, this suggests that a translation methodology (and by extension, a supporting lexical tool) that prioritizes phrasal equivalence is not just beneficial, but essential.

[This section requires your interpretation. For example: The analysis of specific translation procedures might reveal that 'Transposition' was the most common oblique strategy identified. This could suggest a fundamental structural divergence between English nominal phrases and Uzbek verbal constructions in legal writing. The framework seems effective in capturing these systematic grammatical shifts, which can be a major source of difficulty for novice translators. The successful application of a model developed for French and English to a completely unrelated language pair like English and Uzbek may also speak to the universal nature of the translation challenges it describes.]

The potential effectiveness of this approach highlights the limitations of purely statistical methods for terminology extraction. While n-gram analysis might identify 'in accordance with' as a frequent trigram, it lacks the theoretical grounding to label it as an instance of 'Equivalence' and to understand that its translation, 'ga muvofiq', functions as a single unit for the same semantic reason. Our theory-driven framework provides this crucial layer of linguistic intelligence, potentially resulting in a lexicon that is not just a list of frequent phrases, but a curated collection of genuine translation units.

4.2. Theoretical Implications

This study carries several important implications for translation theory. First, it serves as an empirical exploration of a canonical theoretical model. By attempting to operationalize Vinay and Darbelnet's framework [2], we underscore its continued relevance and potential practical utility decades after its conception. It suggests that these foundational theories are not mere historical artifacts but can serve as powerful analytical tools for contemporary, corpus-based translation studies [1].

Second, the study highlights both the strengths and potential weaknesses of the classic model when applied to

a large dataset. While the seven procedures provided a robust framework for classification, the lines between them were sometimes blurred. [Insert your specific finding here. For example: A single translation might exhibit features of both Transposition and Modulation, making a discrete categorization difficult. This suggests that in practice, these procedures may not be mutually exclusive but can co-occur within a single translation unit.] This observation could imply that while the model is powerful, it might be refined or expanded with sub-categories to better capture the complexity of real-world translation data. Our work provides a potential empirical basis for such a future theoretical refinement.

4.3. Practical Applications and Contribution

The most direct contribution of this research is the production of a practical tool for the translator community on the Paratranslator.UZ platform. The resulting [Insert number]-entry lexicon could be directly integrated into their workflow, promising to enhance translation speed, consistency, and accuracy. By providing verified equivalents for complex phrases and domain-specific terminology, the lexicon may reduce the cognitive load on translators and help standardize terminology across large projects.

Beyond this specific platform, the methodological framework itself is a significant contribution. It provides a replicable template for researchers and platform developers seeking to create specialized lexical resources from their own parallel corpora. The framework is language-pair and domain-agnostic, although the specific heuristics would need to be adapted. This has potential applications in translator training, terminology management, and even in the pre-processing of data for training specialized neural machine translation (NMT) engines. By feeding an NMT system with data where multi-word TUs are explicitly identified, one could potentially improve its handling of idiomatic and phrasal language.

4.4. Limitations and Future Research

This study is subject to several limitations that offer clear directions for future research. First, our analysis was confined to a single language pair (English-Uzbek) and a single domain ([Insert domain]). The distribution and nature of translation units are likely to differ across other language pairs and subject areas. Future work should apply the framework to different contexts to test its generalizability.

Second, the theoretical basis of the study was deliberately narrow, relying almost exclusively on Vinay and Darbelnet [2], with contextual support from Baker & Saldanha [1]. While this was intentional to test the direct applicability of the model, a more comprehensive study could integrate other theories of equivalence and terminology to create a more nuanced framework.

Finally, the identification process was semi-automated, with a crucial reliance on manual validation. While this enhanced the quality of the final lexicon, it limits scalability. Future research could focus on developing more sophisticated machine learning models to automate the classification of translation procedures, potentially training a model on a manually validated dataset like the one produced in this study. This could lead to a fully automated system for creating high-quality, theory-driven lexicons on a much larger scale.

CONCLUSION

This article has proposed and detailed a methodological framework for constructing bilingual lexicons by applying the classic translation theory of Vinay and Darbelnet [2] to a modern, platform-based parallel corpus. The study sought to bridge the gap between abstract theory and practical application in the digital age. By operationalizing the seven translation procedures into a set of semi-automated heuristics, we have outlined a process for identifying, classifying, and structuring meaningful translation units that extend beyond the single word. The primary contribution of this research is the framework itself—a replicable, theory-driven methodology for creating specialized, high-quality lexical resources from the data that translation platforms generate every day. The findings suggest that such an approach is not only viable but also holds the potential to produce tools that are more linguistically intelligent and practically useful than those created by purely statistical means. This work underscores the enduring relevance of foundational translation theories and points toward a promising synergy between classic linguistic scholarship and modern computational analysis.

REFERENCES

1. Baker, M., & Saldanha, G. (Eds.). (2020). Routledge encyclopedia of translation studies (3rd ed.).

Routledge.

2. Vinay, J.-P., & Darbelnet, J. (1995). Comparative stylistics of French and English: A methodology for translation (J. C. Sager & M.-J. Hamel, Trans.). John Benjamins Publishing Company.