SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 5675-5687
Open Access. Online & Print

DOI: 10.5281/zenodo.19264879

ASSESSING STRUCTURAL AMBIGUITY IN AI-ASSISTED ENGLISH-ARABIC TRANSLATION

Yasser Sabtan[1], Abdelhamid Elewa [2], Noha Alowedi[3]

[1] Dhofar University, Salalah, Oman & Al-Azhar University, Cairo, Egypt.
[2] Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, KSA.
[3] Saudi Electronic University, Riyadh, KSA.

Received: 05/02/2026
Accepted: 07/03/2026

Corresponding Author: Abdelhamid Elewa
(aaalewh@imamu.edu.sa)

ABSTRACT

This study investigates how certain linguistic complexities remain unresolved in AI-assisted or AI-generated English–Arabic translation, shedding light on limitations in current machine or AI-assisted translation applications in educational and professional settings. It examines how linguistic ambiguity is left unhandled in AI-assisted English–Arabic translation. Drawing on translations produced by undergraduate translators using AI tools, the paper focuses on structural ambiguity, including pronominal reference, relative clauses, genitive chains, prepositional phrases, verb pattern alternations, and adjective ordering. The data were evaluated through a mixed-methods model that combines translation error analysis with holistic scoring of accuracy, fluency, and adherence to meaning, with substantial inter-rater agreement measured statistically. Quantitative results show particularly high error rates in pronominal, relative pronoun, hyphenation, multiple genitive, and verb pattern ambiguities, while adjective order proves comparatively less problematic. Qualitative analysis reveals that AI systems regularly privilege seemingly fluent structure, overlooking ambiguity that can only be resolved through post-editing by AI tools if translators are aware of the problem. Translation training is still integrally required to help students identify the inadequacy of AI-generated translation. Therefore, AI tools can augment but not replace human intervention, particularly when linguistic structure produces multiple readings.

KEYWORDS: AI-Generated Translation, Linguistic Ambiguity, Ambiguity Resolution, English-Arabic Translation, Post-Editing, Human-AI Collaborative Translation, Translation Quality Assessment, Human Post-Editing.

we only focus on the pervasive types of ambiguity in the context of Arabic and English translation practice. Structural ambiguity, specifically, is often overlooked by students in AI-assisted translation due to its complexity and subtlety. This type of ambiguity arises from the multiple ways a sentence can be parsed or interpreted, leading to potential errors in translation output. In Arabic, most cases of structural ambiguity result from the free word order in sentences that lack diacritics. The grammatical functions in English are mainly based on the word order. For example, the verb is often preceded by a noun and followed by an object. In Arabic, on the other hand, you can identify any grammatical function throughout the sentence by the inflection it carries, while in English sentences may have different readings when a word could possibly have different grammatical functions. For instance, the pronoun "her" in the following example can function as either an accusative or genitive.

*I saw **her** shaking hand.*

The above sentence could be interpreted as follows:

A- I saw her while she was shaking hand [with someone].

B- *I saw her hand shaking [out of fear].*

Attia (2008: 176) identified three sources of structural ambiguity: (1) Alternative c-structure rule interactions that account for word order ambiguity, phrasal modification and coordination boundaries, (2) disjunctions in f-structure descriptions which identify phrases that have multiple feature values and those having multiple grammatical functions, and (3) lexical entries which account for parts of speech that could belong to different categories, different subcategorization frames, different morphological features, and the choice between Multiword expressions and compositional interpretation. Some of these areas of structural ambiguity are listed by Alkhuli (2001: 89-90) in seven categories and the list can be longer or more language-specific.

1- The **and-structure**: e.g. "old men and women".

2- The **preposition-phrase structure** "The man hit the girl with the ball".

3- The **verb voice**: e.g. "The chicken is ready to eat".

4- The **–ing form**: e.g. "Flying planes can be dangerous".

5- The **possessive form**: e.g. "John's painting".

6- The **or-structure**: e.g. "This can be A or B".

7- The **premodification structure**: e.g. "The acquired language output".

To make the above list longer, one can add "verb

1. INTRODUCTION

While some works in cognitive and computational linguistics have explored ambiguity resolution mechanisms and processing (Hirst 1987; Franz 1996; Clifton et al. 1994; Van Gompel et al. 2000; Ravin and Leacock 2000; Gorfein 2001), the focus of this paper lies in understanding the linguistic causes of ambiguity and its implications for translation, particularly between Arabic and English. It explores the challenges posed by linguistic ambiguity in the context of English-Arabic translation. The focus lies on examining instances of structural ambiguity that can lead to misinterpretations and difficulties in conveying meaning accurately between the two languages. By analyzing a corpus of examples extracted from a book on translation practice written by Elewa (2022), the study aims to identify the different cases of ambiguity in Arabic and English and explore strategies for solving these problems in translation. This is an item-based translation quality assessment, that is commonly referred to as "preselected item evaluation" that has been proposed by Kockaert and Segers (2017). This has been popularized in TQA for its practicality and integration of language testing methods into translation quality assessment (Han 2020). The research explores the structural ambiguity, analyzing how differences in sentence structure and grammatical features between the two languages can lead to multiple interpretations and necessitate close attention to syntax and meaning during the translation process. This includes verb pattern ambiguity, pronominal ambiguity, relative pronoun ambiguity, coordination boundaries, ambiguity of prepositional phrases, multiple genitive ambiguity, multiple adjective ambiguity, order of adjectives ambiguity.

By analyzing the complexities of ambiguity in English-Arabic translation and proposing problem solving strategies, this study aims to contribute to a better understanding of cross-linguistic communication and improve the effectiveness of translation practices. The main research question is: What are the main types of structural ambiguity overlooked by students in AI-assisted translation? To answer this question, we categorized the different types of ambiguity before conducting a mixed-methods analysis on 1,800 translation segments produced by 90 translators using AI tools.

1. TYPES OF LINGUISTIC AMBIGUITY

There are many types of ambiguity as far as computation is concerned: orthographic, lexical, structural, semantic, pragmatic, but in this paper,

Models (LLMs).

1.2. Verb pattern Ambiguity

Sometimes phrases/sentences are shortened by dropping some constituents when their meanings are understood through the context. The dropped constituent may cause ambiguity in clauses where verbs have two different patterns, i.e. transitivity and intransitivity. For example, the following sentence has two interpretations:

*The woman I want to **leave**.*

A- The woman I want to leave [her].

B- The woman I want [her] to leave.

The verb "leave" in the embedded clause above could be either transitive or intransitive and both patterns give different readings.

1.3. Pronominal Ambiguity

When pronouns are used after two or more precedents (previously mentioned elements) in a sentence, ambiguity might arise since it would be hard to determine to which antecedent the pronoun refers.

*She threw the **vase** at the **glass door** and **it** broke.*

قذفت المزهرية على الباب الزجاجي فكسرته (فكسرتها).

The pronoun "it" may refer to either "vase" or "glass door". In Arabic this kind of ambiguity could be resolved structurally if the two preceding nouns differ in any grammatical feature because Arabic referential pronouns must agree with their antecedent nouns in number, gender, and person. However, if the referential pronoun shares the same grammatical features of both the nouns, the structure remains ambiguous in translation as in the following example:

*She threw the **vase** at the **window** and **it** broke.*

قذفت المزهرية على النافذة فكسرتها

1.4. Relative pronoun Ambiguity

Like referential pronouns, relative pronouns could create the same sources of ambiguity if preceded by more than one possible antecedent.

رأى الرجل الولد الذي ضربه.

*(1) The man saw the boy **who** hit him.*

*(2) The man saw the boy **whom** he hit.*

The son of the professor whom we met is here.

Whom we met? The son or the professor?

1.5. Coordination Boundaries

When two words or phrases having the same grammatical category are conjoined, we tend to make ellipsis. Here ambiguity may arise when we modify the conjoined constituents altogether as in the following example:

pattern ambiguity", "pronominal ambiguity", "relative pronoun ambiguity", "multiple genitive ambiguity", "multiple adjective ambiguity", "position of adjective ambiguity".

1.1. Ambiguity of Prepositional Phrases

Ambiguity may arise when a prepositional phrase (PP) occurs in a sentence and can either modify a preceding noun or a verb. This is because the PP can be an optional part of NP or VP.

The boy hit the girl with the ball.

In the above example the prepositional phrase could have two meanings:

A- The boy hit the girl that held a ball

B- With a ball, the boy hit the girl.

In fact, prepositional phrases might give different meanings in different positions as follows.

يشاهد معظم الناس في مصر كأس العالم لكرة القدم

ChatGPT/ Gemini / Copilot: Most people in Egypt watch the Football World Cup.

Back translation (BT henceforth):

يشاهد معظم الناس في مصر كأس العالم لكرة القدم.

In the above example we can have the following readings triggered by the position of the prepositional phrase "in Egypt".

1. In Egypt most people watch the Football World Cup. (most Egyptians or people living in Egypt)
2. Most people watch the Football World Cup in Egypt. (most people in the world go to Egypt to watch the Football World Cup)

Now look at how variation in word order brings about different readings that must be considered in translation. Here we should separate the modifying words or phrases from the word they modify and keep the word near the element it modifies. Consider the difference in meaning between the following two sentences:

I saw a bird flying with glasses.

رأيتُ طائرًا يطير وهو يضع نظارات. ChatGPT

BT: *I saw a bird flying while wearing glasses.*

رأيت عصفورًا يطير بنظارات. Gemini

BT: *I saw a bird flying with glasses.*

رأيت طائرًا يطير وهو يرتدي نظارات. Copilot

BT: *I saw a bird flying while wearing glasses.*

With glasses I saw a bird flying.

بالنظارات رأيتُ طائرًا يطير. ChatGPT/ Copilot

BT: *With the glasses, I saw a bird flying.*

بنظارات، رأيت عصفورًا يطير. Gemini

BT: *With glasses, I saw a bird flying.*

The different meanings elicited from the positioning of the PP before or after the head of the sentence are considered in the translations suggested by the three selected Large Language

modified by the word 'new': "head" or "center's board". Therefore, we can use the same technique mentioned above as الرئيس الجديد لمجلس الإدارة. (Lit. the new head for the center's board).

1.7. Multiple adjective Ambiguity

When two adjectives precede a noun, the second adjective can either modify the previous adjective or the following noun. Therefore, a hyphen is used to eliminate ambiguity as in the following examples: *"a valuable-gold detector"* or *"a valuable gold detector"*. Both phrases are grammatically correct, but they give different meanings. The first instrument detects valuable gold, while the second is valuable and detects gold.

A hyphen can solve many ambiguous cases in translation and it can be rendered in different ways in Arabic. For instance, to translate a hyphenated compound from English into Arabic, we can look at the modified head and then use a relative clause.

When translating hyphenated English adjectives like "*American-educated,*" a relative clause is often used in Arabic as in *المدرس الذي تلقى تعليمه في أمريكا* (lit. the-teacher who received education-his in America). For the phrase "*learner-centered strategies", the first version, الاستراتيجيات التي ترتكز على المتعلم, glosses as "the-strategies that center on the learner". The second version, الاستراتيجيات المرتكزة على المتعلم,* glosses ungrammatically as (*the centered strategies on the learner).

Finally, having a look at the difference in meaning between the two phrases below, one can realize that the use of hyphen can solve ambiguity problems that arise from multiple adjectives structures. The intended reading is underlined as follows:

The Israeli-occupied territories.

الأراضي الإسرائيلية المحتلة. *The Israeli lands occupied (i.e., The Israeli lands occupied (by others).*

الأراضي التي تحتلها اسرائيل. *The lands that are occupied by Israel.*

1.8. Order of adjectives ambiguity

If we have more than one adjective in English, we should follow a certain pattern. We often prefer "a nice old bag" to "an old nice bag", "a tall happy man" to "a happy tall man". In Arabic, on the other hand, we can put these adjectives in whatever order we like. حقيبة جميلة قديمة / حقيبة قديمة جميلة

We can realize how important the order of adjectives in translating the following two sentences is.

I bought a Swiss golden watch.

اشتريت ساعة ذهبية (أي من ذهب) سويسرية.

(Lit. I bought a Swiss golden (i.e., made of gold) *Old men and women.*

The adjective "old" may modify "men" separately or "men and women.

1.6. Adjectives of Multiple genitive Structure

In Arabic, genitive forms can function as compounds in English.

غلاف الكتاب "Book cover", غرفة النوم "Bedroom". We can also add more genitives to the head, forming longer structures: تصميم غلاف الكتاب "Book cover design", ستائر غرفة النوم "Bedroom curtains", But if we add an adjective, we may not know to which noun the adjective belongs especially with the lack of diacritics, such as تصميم غلاف الكتاب الجديد "Design of the new book cover" (LLMs), or New design of the book cover". Another common example, ستائر غرفة النوم الجديدة "New bedroom curtains" (LLMs), or "Curtains of the new bedroom".

In these two examples we do not know which noun is modified by the adjective "new":

"new cover", "new book", "new design", or "new bedroom", new curtains". In English, this can be sorted out as follows: "new design of the book cover", "cover design of the new book", or "the new curtains of the bedroom", "the curtains of the new bedroom".

We can also avoid ambiguity in Arabic by using the preposition لـ:

الستائر الجديدة لغرفة النوم (lit. The new curtains for the bedroom)

الستائر لغرفة النوم الجديدة (lit. The curtains for the new bedroom)

التصميم الجديد لغلاف الكتاب (lit. The new design for the book cover)

تصميم الغلاف للكتاب الجديد (lit. design of the cover for the new book)

تصميم الغلاف الجديد للكتاب (lit. The new cover design for the book)

What makes the structure complicated is the absence of case-ending markers and the similarity of gender. If the two nouns are of different genders, we won't have any trouble in identifying the modified noun as follows:

United Nation Secretary General سكرتير الأمم المتحدة العام or السكرتير العام للأمم المتحدة (lit. general secretary (masc.) of the United Nations). This is because the adjectiveالعام is masculine and can only modify the masculine noun "السكرتير"

Ambiguity arises in structures that contain nouns of the same gender or same singular or plural form: "*The new head of the center's board"* رئيس مجلس الإدارة الجديد (lit. head of the new center's board/ new head of the center's board).

The Arabic version does not show which noun is

MT systems, while increasingly sophisticated, often produce output containing errors in meaning, grammar, syntax (including structural ambiguity), style, or terminology (Sabtan et al., 2021). When these translations are intended for casual comprehension ("gisting"), the errors may be acceptable. However, for published or widely circulated materials, accuracy and quality are critical, making error correction essential through post-editing, which is defined by Allen (2003) as the correction of texts that have been translated from a source language (SL) into a target language (TL) by an MT system. This post-editing practice, according to Koponen (2016), can increase the productivity of professional translators compared to manual translation from scratch.

Developing a systematic approach to identify various types of errors for post-editing machine and AI-generated translations could either be holistic or focused. These errors range from superficial issues to deep semantic and pragmatic problems, each demanding a specific kind of intervention.

Pym (1992) notes that his attempt to classify translation errors in a list of fourteen is fundamentally flawed and "wholly unsatisfactory", acknowledging that taxonomies of translation errors impose artificial order on a deeply complex reality. The core problem, Pym explains, is that real-world translation errors resist neat categorization. They often stem from multiple intertwined causes:

[E]rrors may be attributed to numerous causes (lack of comprehension, inappropriateness to readership, misuse of time) and located on numerous levels (language, pragmatics, culture), but also because the terms often employed to describe such errors (overtranslation, under-translation, discursive or semantic inadequacy, etc.) lack commonly agreed distinctions or fixed points of reference: "equivalence" has been used and abused so many times that it is no longer equivalent to anything, and one quickly gets lost following the wanderings of "discourse" and associated concepts

Consequently, rigid systems invariably suffer from having too few categories (oversimplifying complexity) or too many (creating artificial distinctions real errors straddle). One example of a lengthy list of translation errors is the scale proposed by Gouadec (1981, 1989) who proposes a complex scale that contains 675 potential error types (300 lexical, 375 syntactic).

The LISA (Localization Industry Standards Association) QA model was widely adopted in localization and translation industry to establish quality metrics for translated content. It typically

watch)

I bought a golden Swiss watch.

اشتريت ساعة ذهبية (أي لونها) سويسرية.

(lit. I bought a golden (i.e. its color) Swiss watch)

The Arabic for "golden" can denote a material or color in Classical Arabic. The position of "golden" after "Swiss" denotes an adjective of material. But the word ذهبية can denote a material or color. This ambiguity can be worked out if we follow the Classical Arabic style in adjective formation as follows.

اشتريت ساعة ذهبية (أي من ذهب) سويسرية. (lit. I bought a Swiss golden [made] of watch).

Placing the adjective "golden" before "Swiss" makes it an adjective of color.

I bought a golden Swiss watch.

اشتريت ساعة سويسرية ذهبية. LLMs

اشتريت ساعة ذهبية اللون سويسرية. HT

This differentiates between "the watch" being Swiss-made or golden in color.

2. LITERATURE REVIEW

Early attempts to categorize types of ambiguity, such as those by Stageberg (1971), distinguished between lexical ambiguity, syntactic ambiguity (including attachment ambiguity), class ambiguity (part-of-speech ambiguity), and script ambiguity (resolved by intonation). Taha (1983) further explored syntactic ambiguity, highlighting the role of sentence structure and the lack of "formal signals" in creating multiple interpretations.

Oaks (1994) offered a valuable perspective on ambiguity as a deliberate device in fields like advertising and humor, highlighting "class ambiguities" and the linguistic elements that create them.

Giora (2003) provided a broader overview of the psycholinguistic aspects of processing ambiguous language, including jokes and irony. While discussing processing mechanisms is beyond the scope of this research, understanding how ambiguity is perceived and resolved is relevant to the study of its impact on translation.

This literature review demonstrates the multifaceted nature of ambiguity and the need for further investigation into its types, causes and effects in translation, specifically between Arabic and English. This research aims to identify which language is more ambiguous linguistically. It analyzes some challenges posed by ambiguity in translating between these two languages and proposing strategies for achieving accurate and effective communication.

2.1. Categories of the errors to edit

2. Functional vs. Absolute errors: Functional errors violate translation project requirements, while absolute errors breach universal linguistic/cultural norms.

3. Systematic vs. Random errors: Contrasting recurrent flaws with isolated mistakes.

4. Process vs. Product errors: Separating errors in the translation act from those in the final output.

This framework avoids oversimplification or artificial granularity seen in approaches like Gouadec's (1981) 675-category scale, offering instead a flexible method for analyzing errors in context.

2.2. Translation strategies typology

Several attempts have been made to classify the errors or poor quality of translation in general. By identifying common mistakes and areas for improvement, translations can be improved, whether produced manually, CAT-based or fully automated. This can also help developers to fine-tune algorithms and enhance the accuracy of machine translations.

House (2015) proposes two macro translation types, overt translation, leaning to the source text, and covert translation that is target text-oriented. House stresses that both types are valid but serve different purposes: overt translation prioritizes historical and cultural authenticity, while covert translation prioritizes functional equivalence in the target context

Overt translation occurs when the translator makes it explicit that the target text is a translation, preserving the source text's time, place, and cultural setting. It is typically used for texts bound to their original context, such as historical speeches, literary works, or political documents, where the original's author, audience, and socio-cultural backdrop must remain visible. The goal is to reproduce the source text's function in its original culture, allowing readers to look through the translation to the original, even if some elements may feel foreign.

Covert translation, by contrast, aims to produce a target text that functions as if it were originally written in the target language for its audience. The translation is culturally adapted, often through a "cultural filter," so it reads naturally within the target culture's norms and communicative conventions. This approach is suited for texts like advertisements, business communication, or instructions, where the communicative purpose is universal but must be localized for effectiveness.

In his typology of translation solutions, Pym (2016) identifies seven key solution types that categorized errors into types as minor, major, or critical according to the evaluators. The output of this process is to determine whether the translation gets a standard score of 'pass' or 'fail' (Castilho 2018). The strength of such models lies in their ability to offer a common language for quality assessment, enabling consistent evaluation across different projects.

Another example of error typology model is the SAE J2450. It was developed by the American Society for Automotive Engineers in collaboration with General Motors. This is a structured error typology designed as an 'industry-wide metric for the evaluation of translation quality' (https://www.sae.org/standardsdev/j2450p1.htm). The model was specifically designed for the automotive industry, where precision and accuracy in technical documentation are paramount. The primary objective of this metric is to establish a uniform standard for the objective assessment of translation quality within automotive service information. This standard is designed to be language-agnostic, applying equally to variations in both the source and target languages. Furthermore, it aims to be translation method-neutral, accommodating evaluations of both human and machine-generated translation The specificity of SAE J2450 to a particular domain highlights how error typologies can be tailored to meet the unique quality requirements and challenges of different industries.

Pym shifts focus from enumerating error types to proposing a unified definition grounded in translational competence: a translation error reflects a failure to optimally select from multiple viable target-text (TT) options. This redefinition yields his holistic binary/non-binary distinction. A binary error is characterized by a clear right or wrong answer. These errors are often due to a lack of basic linguistic knowledge and can be corrected quickly and authoritatively, similar to how errors are addressed in a language class. In contrast, non-binary errors occur when there are multiple potentially acceptable target text solutions, making the choice more subjective and dependent on context, aesthetics, or intuition. Correcting non-binary errors requires discussion and negotiation rather than simple authoritative decisions.

Further, Melis and Albir (2001) propose that error classification should address four key dimensions rather than exhaustive lists:

1. Source-text vs. Target-text errors: Distinguishing ST-related errors (e.g., wrong sense, omissions) from TT issues (e.g., syntax, cohesion).

match the risk. This seven-type typology moves beyond common binaries (literal vs. free) to offer flexible strategies in practical decision-making. It helps translators to take a wider range of actions, from minimal intervention to creative transformation, by either copying the source language item, expression change, or content change as in Table 1.

translators can draw upon when encountering non-routine problems in "bump mode" (where multiple viable options exist and no single rule dictates the choice). The framework is pedagogical, open-ended, and adaptable to various language pairs, enabling sub-categories for specific contexts. It balances two main factors: translator's effort and credibility/communication risk. The effort should

Table 1: A typology of translation solution types (Pym 2016).

Examples	Subcategory	Category
Copying sounds Copying morphology Copying script...	Copying Words	Copying
Copying prosodic features Copying fixed phrases Copying text structure...	Copying Structure	
Changing sentence focus Changing semantic focus Changing voice	Perspective Change	Expression change
Generalization / Specification Explicitation / Implication Multiple Translation Resegmentation	Density Change	
New level of expression New place in text (notes, paratexts)...	Compensation	
Corresponding idioms Corresponding Culture-Specific Items	Cultural Correspondence	
Correction / censorship / updating Omission of content Addition of content ..	Text Tailoring	Content change

source text element with a target-culture counterpart that serves a similar function. Thirdly, the source content can be tailored and changed to fit the target purpose ideologically, factually, or contextually.

By investigating the challenges of ambiguity in English-Arabic translation and proposing effective strategies for mitigating these challenges, this research aims to contribute to the improvement of translation practices and enhance cross-cultural communication. The findings are relevant to translators, language learners, and anyone involved in intercultural communication.

3. METHODOLOGY

With the first type of translation strategies, elements from the source text (sounds, morphology, or script, syntactic relationships) are retained with minimal alteration. This is similar to Vinay and Darblenet's (1995) two main methods 'direct' and 'oblique'. The second includes presenting the same idea from a different angle in grammar, semantic focus, or register, similar to classical *modulation* or *transposition*. It is also related to redesigning the text size by reduction, expansion, or resegmentation. Loss of meaning can also be compensated here by shifting a solution to another part of the text (paratexts or notes) or to a different expressive level to preserve meaning/effect. Added to the second category is Cultural Correspondence by replacing a

They have been informed about the nature and purpose of the research and they agreed to participate. The test consists of 20 sentences that may cause ambiguity. They are given a list of source texts specifically designed to contain various forms of linguistic ambiguity, including lexical and structural ambiguity. Then they are asked to provide any possible translation of the sentences.

A key procedural control is implemented: participants are instructed to use any AI tool for translation assistance but they are not briefed about the specific ambiguities present in the texts. This is to prevent them from designing specific prompts for disambiguation, thereby ensuring that the data reflects the natural performance of human-AI collaborative translation when faced with inherent ambiguity. The translations they produce form the corpus for the mixed-methods evaluation described above.

3.2. Inter-rater reliability and agreement

To ensure the consistency and objectivity of the evaluations, a critical aspect given the qualitative nature of error identification and holistic scoring, the study implemented statistical measures of inter-rater agreement.

Three independent, qualified raters, all holding PhD degrees in Translation Studies with over five years of experience, were trained on the error typology and holistic rubric. Each rater evaluated the entire corpus of 1,800 translation segments (90 participants × 20 sentences) for both the quantitative (error analysis) and qualitative (holistic scoring) components.

The agreement between raters was calculated using Fleiss' Kappa (κ), a robust statistical measure suitable for three or more raters assessing categorical data. This was applied to:

1. The identification of an error (a binary yes/no for each potential ambiguity in a sentence).
2. The categorization of the error type (e.g., pronominal, genitive, etc.).
3. The holistic score assigned for Accuracy, Fluency, and Adherence to Meaning (scores were first grouped into categories: Poor=1-2, Adequate=3, Good=4-5 for this calculation).

Fleiss' (1971; 2003) Kappa was chosen over Cohen's Kappa because it extends the reliability statistic to more than two raters. The interpretation of the Kappa values follows Landis and Koch's (1977) benchmark:

This research employs a mixed-methods approach for assessing the quality of translation (Chao 2020), specifically as regards ambiguity in English-Arabic translation. The primary data consist of translations produced by human translators and AI tools, which are evaluated using a combination of quantitative and qualitative analytical techniques. The source text to be translated consists of 20 examples (Appendix 1), following Kockaert and Segers' (2017) "preselected item evaluation" mentioned above. The evaluation of the translations is conducted using an integrated scoring model, following the mixed-methods approach established in translation studies (e.g., Waddington, 2001). This model combines two distinct evaluation frameworks:

1. Error Analysis (Quantitative): Translations are systematically analyzed using a detailed error typology to detect, categorize, and quantify instances of error. This analysis specifically focuses on identifying failures to resolve the predefined types of linguistic ambiguity (e.g., structural, pronominal, genitive).
2. Holistic Rubric Scoring (Qualitative): The same translations are evaluated holistically using an analytical rubric. This rubric assesses broader dimensions of translation quality, such as accuracy, fluency, and adherence to the intended meaning, beyond the discrete identification of errors.

The scores from these two components are merged to produce a final unified score for each translation (Waddington, 2001; Amini, 2018). To ensure the reliability of the assessments, the evaluation is conducted by multiple inter-raters. This practice is supported by preliminary empirical evidence suggesting that a mixed-methods approach can yield higher inter-rater reliability than either error analysis or holistic scoring used in isolation (Amini, 2018; Waddington, 2001).

3.1. Participants and procedures

A group of translation students in Imam University in Riyadh, categorized across translation courses offered in the BA translation programs (beginner, intermediate, and advanced) are selected. The meta-data of the test addresses personal (age [17-21 years old] and gender [male]), demographic (Imam University in Riyadh) and study program level (beginner [level 2], intermediate [level 5], advanced [level 7]). The total number of respondents is 90 including 30 beginner students (30. 30%), 30 intermediates (30. 30%) and advanced 30 (30.30%).

Table 2: Landis and Koch's (1977) benchmark scale for measuring the strength of agreement.

K	Interpretation
<0	Poor agreement
0.0-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.0	Almost perfect agreement

meaning. These results highlight that AI-assisted translation tends to produce fluent output that often fails to resolve structurally complex ambiguities, and that advanced training of translators is associated with more effective detection and disambiguation of these issues. The analysis is based on data collected from the translation task (Appendix 1) as performed by a cohort of 90 translators, divided evenly into three proficiency levels: 30 beginners, 30 intermediates, and 30 advanced practitioners. This larger sample size provides a more statistically reliable and generalizable overview of the challenges posed by linguistic ambiguity.

1. Quantitative Analysis: Error Frequency by Ambiguity Type

The translations were first subjected to a quantitative error analysis. Each instance where an ambiguity was not correctly resolved was counted as an error. With 90 participants and 20 sentences, a total of 1,800 translation segments were analyzed. The table below summarizes the total number of errors and the error rate for each type of ambiguity.

This rigorous approach to inter-rater reliability ensures that the subsequent quantitative and qualitative analyses are built upon a foundation of consistent and reliable human judgment.

Through this integrated approach, the study aims to quantitatively identify the most common and challenging types of ambiguity in the English-Arabic pair, as well as qualitatively evaluate the strategies and overall success in resolving them.

4. FINDINGS AND ANALYSIS

This section presents the findings from the mixed-methods evaluation of the translations produced by human translators using AI tools. The 1,800 translation segments produced by 90 participants show that pronominal, relative pronoun, and hyphenation ambiguities yielded the highest error rates, followed closely by multiple genitive, verb pattern, lexical, prepositional phrase, and multiple adjective ambiguities. Coordination boundaries caused a moderate level of difficulty, while order of adjectives was comparatively less problematic. Holistic rubric scores further indicate a clear progression across translators' academic levels, with advanced participants achieving the highest averages in accuracy, fluency, and adherence to

Table 3: Frequency and Error Rate of Ambiguity Types (N=90).

Ambiguity Type	Example Sentence	Total Occurrences	Total Errors	Error Rate
Pronominal	5, 6	180	180	100%
Relative Pronoun	7, 8	180	180	100%
Hyphenated Terms	17, 18	180	180	100%
Multiple Genitive	14, 18	180	141	78.3%
Verb Pattern	3, 4	180	136	75.6%
TOTAL	--	1800	1,339	74.4%
Lexical	15	90	62	69%
Prepositional Phrase	1, 2, 6	270	185	68.5%
Multiple Adjective	16, 17, 19	270	172	63.70%
Coordination Boundaries	9, 10	180	102	56.67%

Order of Adjectives	20	90	26	28.9%

still challenging, the "and-structure" ambiguity is somewhat more manageable for translators to spot and correct. Order of Adjectives maintained the lowest error rate (28.9%). This reinforces the theory that the flexibility of adjective order in Arabic often allows for a passable translation even if the nuanced stylistic preference of English is not perfectly mirrored.

2. Qualitative Analysis: Holistic Rubric Scores

Alongside error counting, translations were evaluated holistically on a scale of 1-5 (1=Poor, 5=Excellent) for Accuracy, Fluency, and Adherence to Intended Meaning. The scores were then averaged across participant groups.

In Table 3, pronominal references (including relative pronouns) and hyphenation ambiguity are the most challenging, with error rates of 100% each. This underscores a critical weakness in AI's ability to handle coreference resolution and hyphenated terms, especially when the context does not provide clear grammatical gender or number distinctions for disambiguation or when they do not understand the function of hyphenated terms. Multiple Genitive, Verb Pattern, Lexical, Prepositional Phrase ambiguities all exhibited high error rates (65-78%), confirming them as major pitfalls in AI-assisted translation. These categories represent a "danger zone" where automated outputs are highly likely to be misleading or incorrect without human intervention. Coordination Boundaries showed a moderate error rate (56.7%), suggesting that while

Table 4: Average Holistic Scores by Translator Level.

Translator Level	Accuracy	Fluency	Adherence to Meaning	Overall Avg.
Advanced (n=30)	4.2	4.4	4.0	4.2
Intermediate (n=30)	3.3	3.7	3.0	3.3
Beginner (n=30)	2.3	2.9	2.0	2.4
All Participants	3.3	3.7	3.0	3.3

advanced translators (a 0.4-point difference). This is a critical insight: AI tools are optimized to produce grammatically smooth and fluent text, often masking underlying semantic inaccuracies. The human skill lies in detecting and correcting these hidden errors.

4.1. Inter-rater Reliability Results

Before presenting the primary findings on ambiguity, the reliability of the raters' assessments was confirmed. The Fleiss' Kappa scores for the different evaluation tasks are presented in Table 5.

A clear positive correlation exists between translator expertise and translation quality. Advanced translators, even when using the same AI tools, produced outputs that were significantly more accurate and faithful to the intended meaning. Fluency scores were consistently higher than Adherence to Meaning scores. This indicates that advanced translators are far more effective at leveraging AI as a tool and overriding its erroneous or ambiguous outputs.

Across all levels, Fluency scores were consistently higher than Adherence to Meaning scores. This gap is most pronounced in the beginner group (a 0.9-point difference) but persists even with

Table 5: Inter-Rater Reliability (Fleiss' Kappa) for Translation Evaluation.

Evaluation Task	Fleiss' Kappa (κ)	Agreement Level
Error Identification (Presence/Absence)	0.78	Substantial
Error Categorization (Type of Ambiguity)	0.72	Substantial
Holistic Rubric Scoring (Accuracy)	0.69	Substantial
Holistic Rubric Scoring (Fluency)	0.75	Substantial
Holistic Rubric Scoring (Adherence to Meaning)	0.71	Substantial

ambiguity in noun phrases. Both failures stem from a misreading of the "modifier-head" relationship, but they differ dramatically in the real-world impact of the error.

The ambiguity lies in what the adjective "heavy" modifies. Is it a *detector* that is heavy, or a detector for *heavy metals*? The hyphen in "metal-detector" explicitly groups those two words, creating a single compound noun. "Heavy" then modifies this entire unit. A possible corrected translation for the hyphenated term can be as follows:

نحتاج إلى كاشف عن المعادن ثقيل الوزن

BT: We need a detector of metals that is heavy.

In the second example, students interpret "Israeli" as an adjective for "territories" (i.e., territories that are Israeli), rather than as the first part of a compound adjective "Israeli-occupied". 100% of students at all levels failed to translate it successfully considering the hyphenation strategy discussed in the paper. A possible translation could be as follows:

الأراضي التي تحتلها إسرائيل هي نقطة خلاف.

BT: The territories that are occupied by Israel are contentious.

Human translators, using AI tools, overlooked these important differences because AI-tools rely heavily on statistical patterns from large datasets that easily pick the wrong, yet statistically common, attachment (e.g., "heavy metal," "Israeli territories").

For the remaining items of the test set, translated by the 90 participants, a clear picture of the human-AI collaborative translation process appears when faced with linguistic ambiguity. The high overall error rate of 69.6% in initial AI-generated outputs underscores that current AI tools function primarily as sophisticated pattern-matching systems, not as understanding entities. They lack the deep contextual and pragmatic knowledge required to reliably resolve ambiguities.

The strong correlation between translator level and final translation quality (from 2.4 for beginners to 4.2 for advanced) highlights that the value of the human translator has shifted, not diminished. In this new paradigm, the core skill is no longer just linguistic transfer, but disambiguation, the ability to critically diagnose AI output, identify potential ambiguities, and apply strategic post-editing.

The consistent gap between Fluency and Adherence to Meaning is particularly consistent. This places a new burden on translator training programs to equip students with advanced analytical and diagnostic skills, moving beyond just evaluating fluency to interrogating semantic

The Kappa values across all evaluation tasks fall within the "Substantial" agreement range (0.61 - 0.80). This indicates a high level of consistency among the three raters, confirming that the error typology and holistic rubric were applied reliably. The slightly higher agreement for Fluency ($\kappa = 0.75$) is expected, as judgments on grammaticality and readability are often more straightforward than judgments on semantic Accuracy ($\kappa = 0.69$) or Adherence to Meaning ($\kappa = 0.71$), which require deeper interpretive analysis. These results validate the dataset and provide confidence in the subsequent analysis.

4.2. Integrated analysis

This section provides a detailed analysis of specific sentences, combining quantitative error data with qualitative observations of the strategies used (or not used) by participants. The data from 90 participants, evaluated with a high degree of inter-rater reliability (Fleiss' Kappa > 0.68 across all measures), show that students rely heavily on AI tools and provide one possible translation of every item as shown in the two examples below (see the translation task in appendix 1).

Example 1: Pronominal Ambiguity (Sentence 5)

ST: Sarah told Amina that she had won the prize.

TT: أخبرت سارة أمينة أنها فازت بالجائزة

BT: Sarah told Amina that she had won the prize.

This translation is ambiguous, as [أنها] "lit. that she" could refer to Sarah or Amina. The inherent ambiguity of the pronoun "she" was not preserved at all of the initial, unedited and edited AI outputs. All students did not disambiguate the structure, accepting the AI-generated output. This can easily be disambiguated, employing grammatical strategies like direct speech قالت سارة لأمينة: "لقد فزت بالجائزة or syntactic restructuring to clarify the referent.

Example 2: Hyphenated and non-hyphenated terms (Sentence 17&18)

Sentence 17

ST: We need a heavy metal-detector.

TT: نحتاج إلى كاشف عن المعادن الثقيلة

Sentence 18

ST: The Israeli occupied territories are a point of contention.

TT: الأراضي الإسرائيلية المحتلة هي نقطة خلاف.

The translations of the two examples above "heavy metal-detector" and "Israeli occupied territories" illustrate a common pitfall in both human and machine translation: syntactic

for students. In this context, linguistic fluency generally outweighs fidelity to meaning, with advanced translators consistently outperforming novice and intermediate ones in detecting and correcting misleading AI suggestions. This indicates that experience not only enhances linguistic accuracy but also improves the ability to read AI output critically, identifying ambiguity where the text appears deceptively polished.

These findings suggest a shift in the translator's role within AI-supported workflows. Rather than merely producing fluent target texts, translators are increasingly required to act as diagnostic and resolution specialists for ambiguity, filtering the output of systems optimized for fluency over fidelity. Consequently, translator training should focus on recurring patterns of ambiguity in English and Arabic, illustrating how and why AI systems mishandle them, and equipping students with practical strategies for disambiguation and subsequent editing. In doing so, translation pedagogy can help ensure that AI becomes a support for, rather than a source of, communication across languages.

precision.

5. CONCLUSION

This study aimed to identify the main categories of ambiguity in translation from English into Arabic and to evaluate their handling in AI-assisted outputs following human post-editing in pedagogical contexts. The analysis showed that structural ambiguity constitutes a significant challenge in AI-assisted preliminary translations. Findings underscore the impact of linguistic ambiguity on translation quality in English-to-Arabic translation, particularly when AI tools are employed at any stage of the process. A systematic error analysis of AI-assisted translations by students revealed that pronoun reference, relative clauses, compound terms, multiple adjuncts, verb pattern alternations, and complex prepositional or adjectival constructions are most prone to misinterpretation. These are precisely the areas where AI-generated output often appears fluent and superficially grammatically sound, while simultaneously introducing serious issues of reference and meaning.

The fluency of AI translations can be misleading

Author Contributions: Conceptualization, Y.S. and A.E.; methodology, Y.S. and E.A; software, A.E., Y.S.; validation, Y.S. and E.A.; formal analysis, Y.S. and E.A.; investigation, Y.S., E.A. and N.A.; resources, A.E., N.A.; data curation, A.E.; writing—original draft preparation, A.E.; writing—review and editing, Y.S. and E.A., N.A.; visualization, Y.S., E.A. and N.A.; supervision, A.E.; project administration, A.E.; Y.S. All authors have read and agreed to the published version of the manuscript.

REFERENCES

Alkhuli, A. (2008) *An Introduction to Semantics*, 2nd edition. Dar Al-Falah for Publishing and Distribution.

Allen, J. (2003) Post-editing. In: *Computers and Translation: A translator's guide*, H. Somers (ed.), John Benjamins Publishing Company: Amsterdam. https://doi.org/10.1075/btl.35.19all

Attia, M. (2008) Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation. *Unpublished PhD Thesis*, the University of Manchester.

Austermühl, F. (2001) *Electronic Tools for Translators*, Manchester: St. Jerome Pub. Co.http://www.computerwoche.de/knowledge_center/software/591945/index.html.

Bar, K. & Dershowitz, N. (2010) Using synonyms for Arabic-to-English example-based translation. *AMTA 2010 - 9th Conference of the Association for Machine Translation in the Americas*, Denver, CO. https://doi.org/10.1075/nlp.9.04bar

Bybee, J. (2003) Cognitive processes in grammaticalization. In: *The New Psychology of Language*, M. Tomasello, Vol. II, 145-167. New Jersey: Lawrence Erlbaum Associates Inc.

Bybee, J. (2003) Mechanisms of change in grammaticization: The role of frequency. In: *The Handbook of Historical Linguistics,* B. D. Janda, 602-623. Oxford: Blackwell.

Castilho, S. (2018) Approaches to human and machine translation quality assessment. In Castilho, S., Doherty, S., Gaspari, F. and Moorkens, J. (eds.) *Translation Quality Assessment: From Principles to Practice.* Springer International Publishing. 9-38.

Chao, H. (2020) Translation quality assessment: A critical methodological review. *The Translator*, DOI: 10.1080/13556509.2020.1834751.

Clifton, C., Frazier, L., & Rayner, K. (eds.). (1994) *Perspectives on Sentence Processing*. New York: Lawrence Erlbaum Associates.

Dries, J. (1995) *Dubbing and Subtitling: Guidelines for Production and Distribution*, European Institute for the

Media.
Elewa, A. (2022) *Levels of Translation*. Alqalam: Cairo.
Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
Fleiss, J. L., Bruce L. and Myunghee C. P. (2003) *Statistical Methods for Rates and Proportions*, Oxford: Wiley.
Franz, A. (1996) *Automatic Ambiguity Resolution in Natural Language Processing*: *An Empirical Approach*. Springer.
Giora, R. (2003) *On our Mind. Salience, Context, and Figurative Language*. Oxford: Oxford University Press
Gouadec, D. (1981) "Paramètres de l'evaluation des traductions." *Meta* 26 (2): 99–116. doi:10.7202/002949ar.
Gouadec, D. (1989) Comprendre, évaluer, prévenir : pratique, enseignement et recherche face à l'erreur et à la faute en traduction. *TTR*, 2(2), 35–54. https://doi.org/10.7202/037045ar.
Han, C. (2020) Translation quality assessment: a critical methodological review. *The Translator*, 26(3), 257–273, DOI: 10.1080/13556509.2020.1834751
Hirst, G. (1987) *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge: Cambridge University Press.
House, J. (2015) *Translation Quality Assessment: Past and Present*. Routledge.
Kockaert, H. J., and Segers, W. (2017) Evaluation of legal translations: PIE method (Preselected items evaluation). *The Journal of Specialised Translation* 27, 148–163.
Koponen, M. (2016) Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25, 131–148.
Landis, J. R., & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
Melis, N. M. & Albir, A. H. (2001), Assessment in Translation Studies: Research Needs, *Meta* 46(2), 272-287.
Oaks, D. D. (1994) Creating structural ambiguities in humor: getting English grammar to cooperate. *Humor: International Journal of Humor Research, 7* (4), 377–401.
Pym, A. (1992) Translation error analysis and the interface with language teaching. In: *The Teaching of Translation,* C. Dollerup & A. Loddegaard (eds.), 279-288. Amsterdam: John Benjamins. https://doi.org/10.1075/z.56.42pym
Pym, A. (2016) *Translation Solutions for Many Languages: Histories of a Flawed Dream*. Bloomsbury Academic.
Ravin, Y., & Leacock, C. (2000) *Polysemy: Theoretical and Computational Approaches*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198238423.001.0001
Sabtan, Y.M.N, Hussein, M.S.M, Ethelb, H & Omar, A. (2021) An evaluation of the accuracy of the machine translation systems of social media language. *International Journal of Advanced Computer Science and Applications*, 12 (7), 406–415. https://doi.org/10.14569/ijacsa.2021.0120746
Stageberg, N. C. (1971) Structural ambiguities in English. In: *The Encyclopedia of Education*, L. C. Deighton (ed.), Macmillan, New York, 356-366.
Taha, A. (1983) Types of syntactic ambiguity in English. *International Review of Applied Linguistics in Language Teaching*, *21* (4), 251–266.
Van Gompel, R.P.G., Pickering, M.J., & Traxler, M.J. (2000) Unrestricted race: A new model of syntactic ambiguity resolution. In: *Reading as a perceptual process*, A. Kennedy, R. Radach, D. Heller, & J. Pynte (eds.), 621-648. North-Holland, Elsevier Science Publishers. https://doi.org/10.1016/B978-008043642-5/50029-2
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.
Waddington, C. (2001) Different methods of evaluating student translations: The question of validity. *Meta*, *46*(2), 311–325. https://doi.org/10.7202/004583ar

Appendix 1:
Translation Task

Please translate the following sentences into Arabic. You are encouraged to use any AI tool of your choice (e.g., Google Translate, ChatGPT, etc.). Please provide any possible translation you consider appropriate.

1.The boy hit the girl with the stick.

2.I saw the man on the hill with the telescope.

3.The chicken is ready to eat.

4.The woman I wanted to leave.

5.Sarah told Amina that she had won the prize.

6.He put the meat on the apple before eating it.

7.The son of the professor who we met is here.

8.I saw the tree with the binoculars that was broken.

9.They are looking for old books and magazines.

10. The project requires skilled data analysts and programmers.

11. Flying planes can be dangerous.

12. Visiting relatives can be boring.

13. The shooting of the hunters was terrible.

14. John's painting was admired.

15. The government is looking for an economical solution.

16. They bought more expensive computer equipment.

17. We need a heavy metal-detector.

18. The Israeli occupied territories are a point of contention.

19. They met the American history teacher.

20. I bought a beautiful green shirt.