Faculty of Arts Journal

DOI: 10.36602/faj.2026.n22.08 Faculty of Arts Journal, Issue 22, December 2026

# Evaluating Terminological Consistency in AI-Generated English–Arabic Political Translation: Evidence From ChatGPT and Google Gemini

Osama Ali Musbah Bala*

Department of English, Faculty of Arts, Misurata University

* Osama.m.bala@art.misuratau.edu.ly

**Citation:** Bala, O. (2026). Evaluating terminological consistency in AI-generated English–Arabic political translation: Evidence from ChatGPT and Google Gemini. *Faculty of Arts Journal – Misurata University,* (22), 148–166. https://doi.org/10.36602/faj.2026.n22.08

**Received** 27- 07 - 2026 **Accepted** 18- 08 - 2026 **Published Online** 21- 08 - 2026

## Abstract

Machine translation has become more fluent and contextually accurate with recent advances in artificial intelligence. However, terminological consistency has been underexplored, particularly in political and electoral discourse where lexical repetition and conceptual precision are critical for cohesion and clarity. This study investigates terminological consistency in AI-generated English–Arabic political translations produced by ChatGPT and Google Gemini Advanced. The translations were generated and analyzed between January and June 2026 using the systems' default settings to ensure comparability and avoid potential variations resulting from user-configured parameters. The study employs a mixed-methods corpus-based approach. The study analyzes 30 political and electoral texts with 60 recurring key terms. Quantitative analysis measures the degree of consistency in the form of stability percentages, and qualitative analysis studies lexical variation and its effect on discourse cohesion and clarity. The adequacy and consistency of the translation were checked against a reference translation based on the United Nations Development Programme (UNDP) Arabic Lexicon of Electoral Terminology. The results indicate that ChatGPT achieved higher terminological consistency than Google Gemini. ChatGPT's lexical equivalents for repeated political and electoral terms were more stable than its Gemini counterpart, which showed more lexical variation, especially in context-sensitive terms such as campaign, electoral law, and judicial review. The study concludes that terminological consistency should be considered as a separate dimension of translation quality and emphasizes the importance of terminology control and human post-editing in AI-assisted political translation

**Keywords:** *Terminological consistency; AI translation; Political discourse; English–Arabic translation; Translation quality; Lexical cohesion*

**Citation:** Bala, O. (2026). Evaluating terminological consistency in AI-generated English–Arabic political translation: Evidence from ChatGPT and Google Gemini. *Faculty of Arts Journal – Misurata University,* (22), 148–165. https://doi.org/10.36602/faj.2026.n22.08

## 1. Introduction

Artificial intelligence is growing fast and has revolutionized the translation industry. Systems like ChatGPT and Google Gemini make multilingual communication fast and accessible. Such systems can generate fluid and contextually relevant translations, often achieving human-level performance at the level of the sentence. However, the quality of translation cannot be evaluated at the sentence level alone, especially for specialized domains such as political and institutional discourse.

Repetition of keywords such as governance, legitimacy, political process and transparency is an important element in political speech. These words have institutional meanings and are key to ensuring coherence and clarity throughout a text. An inconsistent translation of such terms may damage lexical cohesion, reduce conceptual clarity and cause subtle shifts of meaning. Hence, terminological consistency is a prerequisite for the translation of political texts.

The fluency of AI-generated translations is high, but the systems work on the basis of probabilistic language generation, which can lead to multiple lexical equivalents of the same source term in the same text. Although such variation is often linguistically acceptable, it might affect the coherence and readability of the translated discourse. This is especially the case in the English–Arabic pair where the differences in linguistic structure and stylistic conventions mean that variations are more probable.

This study investigates the level of terminological consistency in the translation of AI-generated English–Arabic texts by comparing the output of the two AIs, Google Gemini and the chatbot ChatGPT. It uses a controlled corpus of 30 texts with 60 recurring political and electoral terms, to allow systematic analysis of how repetition of terminology is managed across the texts. Stability percentages provide a quantitative measure of terminological consistency, while a qualitative analysis looks into the kind of variation and its effects on cohesion and clarity.

To ensure reliability, the study uses a reference translation from the United Nations Development Programme (UNDP) Arabic Lexicon of Electoral Terminology, which offers standardized equivalents for essential electoral terms. In this research, the quantitative and qualitative methods will be used to measure the terminological stability as a separate quality of translation and analyze the impact of variation on the discourse-level interpretation.

### 1.2 Research Problem

Although the increasing use of AI-based translation systems has improved both fluency and contextual accuracy of translation, terminological consistency has not been sufficiently studied, especially in political and electoral discourse, where repetition of certain terms is essential to maintain cohesion and conceptual accuracy. Existing studies have focused on the translation quality of individual sentences, with little evidence on how AI systems can deal with the repetition of political terms across a full text. Moreover, there is a lack of comparative study on the terminological consistency of English-Arabic political

translation between Google Gemini and ChatGPT. The problem to be addressed, therefore, is the lack of empirical knowledge about the degree to which these AI systems maintain terminological consistency and the effect of terminological variation on the clarity and cohesion of translated political discourse.

### 1.3 Research Questions

1. Which AI system, ChatGPT or Google Gemini, demonstrates a higher level of terminological consistency in translating repeated political and electoral terms from English into Arabic?
2. What kinds of changes in terminology happen in AI-generated translations of political speech?

### 1.4 Research Objectives

1. To assess the terminological consistency of AI translation systems in rendering political discourse from English to Arabic.
2. To examine the effects of terminological inconsistencies on lexical cohesion and conceptual clarity in translated political and electoral texts

### 1.5 Significance of the Study

The present study adds to the knowledge base in AI translation research by looking at terminological consistency in English–Arabic political and electoral translation. It offers a comparative assessment of ChatGPT and Google Gemini and emphasizes the relevance of consistent terminology for maintaining lexical cohesion and conceptual clarity. The results may be useful for researchers, translators and institutions relying on AI-assisted translation in the specialized domains.

## 2.Literature Review

### 2.1 AI Translation & Neural Machine Translation

Neural Machine Translation (NMT) has revolutionized the quality of automatic translation by allowing systems to learn contextual relationships over entire sentences instead of isolated lexical units. Early statistical machine translation systems were heavily dependent on phrase matching and probabilistic alignment, and the translations they produced were often fragmented and structurally unnatural. Neural models, on the other hand, use deep learning architectures that are able to capture contextual dependencies and produce more fluent target-language output (Bahdanau et al., 2015; Koehn, 2020).

More recently, the advent of Large Language Models (LLMs) has further revolutionized machine translation by including advanced contextual reasoning, discourse awareness and multilingual generation capabilities. Systems like ChatGPT and Google Gemini are able to produce very fluent and contextually adaptive translations for many domains and language pairs. LLMs are not traditional machine translation systems designed specifically for translation tasks, but rather general-purpose language generation systems trained on very large multilingual datasets (OpenAI, 2023; Google DeepMind, 2023).

Their probabilistic nature enables LLMs to produce translations that are flexible and context-aware, often improving readability

and stylistic fluency. However, this same flexibility can work against lexical stability where there is repeated terminology throughout a text. Instead of generating identical lexical equivalents many times, LLMs often generate alternative formulations, which are semantically acceptable but terminologically inconsistent. Recent studies have therefore started to challenge the ability of fluency-oriented AI systems to maintain discourse-level consistency in specialized domains that require tight terminological control.

Recent studies on LLM-based translation have revealed that AI systems are more biased toward contextual adaptation than lexical repetition. Wang et al. (2024) found that large language models often generate stylistically natural translations but differ in repeated terms. Hendy et al. (2023) showed that multilingual LLMs are more fluent than many classical machine translation systems, but are still unstable for specialized terminology and domain-specific translation.

These results indicate that the quality of translation in AI systems should not be evaluated solely in terms of sentence-level fluency or adequacy. Instead, discourse-level dimensions like lexical cohesion, consistency and conceptual continuity must also be considered.

## 2.2 Political Discourse and Translation

Political discourse is one of the most sensitive and institutionally limited forms of communication. Political texts are characterized by ideological positioning, conceptual precision and strategic repetition of key terms. In political communication, lexical repetitions are often more than linguistic repetitions: they also have institutional meaning and ideological continuities (Schäffner, 2004).

Hence, political discourse translation involves more than mere semantic equivalence of sentences. Translators have to keep coherence, consistency and conceptual precision throughout the whole of the discourse. Terms like governance, legitimacy, accountability, electoral integrity and political participation have specialized institutional meanings that make for coherence in political communication.

As Schäffner (2004) argues, political translation is closely related to discourse representation and ideological framing. Small lexical variations may affect the interpretation of political concepts by target audiences. Similarly, Venuti (2012) stresses that the decisions made in translation can influence the reception and representation of political realities.

Newmark (1988) also observes that institutional and political translation often demands terminological stability because repeated terminology contributes directly to textual coherence and communicative clarity. In institutional discourse consistency enhances reader understanding and reduces conceptual ambiguity.

This is especially important in the context of English-Arabic translation, as formal lexical repetition is much more common in the traditional use of Arabic political discourse than even in English discourse. Repeated terminology is a rhetorical and organizational device in Arabic political writing that enhances clarity, emphasis and institutional authority.

## 2.3 Terminological Consistency and Lexical Cohesion

Terminological consistency is defined as the systematic use of the same lexical equivalent of a given source language term throughout the entire translated text. Professional translation practice is usually made consistent by using terminology databases, glossaries, translation memories and institutional style guides (House, 2015).

In terms of discourse, terminological consistency is strongly related to lexical cohesion. Halliday and Hasan (1976) define lexical cohesion as the continuity of a text achieved through lexical repetition and semantic relations. Repetition enables readers to make conceptual links between different parts of a discourse and contributes therefore to textual coherence.

According to Baker (2018), repetition of lexical items is particularly important in specialized and institutional texts, since repeated terminology forms lexical chains that help readers understand and stabilize concepts. These lexical chains are broken by inconsistent terminology. This weakens textual cohesion.

Likewise, House (2015) states that the quality of translation should not be evaluated only with respect to grammatical accuracy or semantic equivalence. Instead, quality assessment must be extended to pragmatic functionality, coherence and consistency on the discourse level. This is particularly true for AI-generated translation, where systems can produce fluent individual sentences, but not maintain lexical consistency across larger texts.

Consistency is even more important in the political and electoral debate because often the same terms are used to denote institutional concepts that have stable legal or procedural meanings. Such variation in terminology can permit ambiguity or change in conceptual interpretation.

## 2.4 Challenges of English–Arabic Political Translation

English–Arabic translation presents several linguistic and stylistic challenges that increase the complexity of maintaining terminological consistency. The two languages differ substantially in syntax, rhetorical structure, lexical density, and stylistic conventions.

Habash (2010) explains that Arabic political discourse tends to favor formal lexical repetition and standardized institutional terminology, whereas English discourse allows greater stylistic variation and synonym use. As a result, AI systems trained on multilingual corpora may generate multiple Arabic equivalents for a single English political term.

Many political concepts also possess more than one acceptable Arabic translation. For example, governance may appear as الحوكمة or الإدارة الرشيدة depending on context, while electoral law may be translated as القانون الانتخابي or قانون الانتخابات. Although these alternatives may be semantically acceptable, variation within the same text may weaken lexical cohesion and conceptual continuity.

Arabic morphology also contributes to variation because derivational flexibility allows multiple structurally different forms to represent related concepts. Consequently, AI

systems operating probabilistically may select different lexical realizations depending on contextual predictions.

In addition, political and legal discourse in Arabic often depends on institutional standardization. Organizations such as the United Nations Development Programme (UNDP) and international electoral bodies frequently adopt standardized glossaries to ensure terminological precision and consistency across official translations.

## 2.5 Variance of Terminology in AI Systems

One of the most frequently observed characteristics of AI-generated translation is lexical variability. Neural and LLM-based systems are designed to maximize contextual naturalness and fluency rather than strict lexical repetition. As a result, repeated source-language terms may receive multiple target-language equivalents across the same document.

Toral and Sánchez-Cartagena (2017) found that neural machine translation systems often outperform phrase-based systems in fluency while simultaneously producing more lexical variation. Koehn (2020) similarly explains that probabilistic generation mechanisms encourage the production of alternative but semantically related translations.

Recent research on LLM-based translation suggests that this tendency has become even more pronounced in generative AI systems. Freitas et al. (2024) observe that large language models frequently employ paraphrasing and contextual reformulation strategies during translation, particularly in high-frequency or semantically flexible expressions.

While lexical variation may improve stylistic naturalness, it can become problematic in specialized domains requiring conceptual precision. In political discourse, repeated terminology often functions as a cohesive mechanism linking institutional concepts across the text. Excessive variation may therefore weaken textual cohesion and reduce conceptual clarity.

At the same time, some scholars argue that limited variation is not necessarily undesirable. House (2015) notes that contextual adaptation may improve readability and communicative effectiveness in certain situations. The challenge therefore lies in balancing fluency and lexical stability.

## 2.6 Research Gap

While the fluency and adequacy of neural machine translation and large language models have received extensive attention in recent years, relatively little attention has been paid to terminological consistency as a discourse level quality criterion.

Most current research on AI translation has been on: sentence-level fluency, grammaticality, multilinguality, and vocabulary appropriateness.

Far fewer studies exist on the handling of repeated terminology by AI systems in long stretches of discourse, particularly in the context of sensitive political or institutional constraints.

Moreover, there are few studies that deal with English–Arabic political translation. Specifically, there is a lack of corpus-based

comparative studies on terminological consistency in LLM-generated political translation between Google Gemini and Chat-GPT.

In addition, few studies have integrated quantitative assessment of consistency with qualitative analysis of lexical cohesion and conceptual clarity. Most previous work considers translation quality as a sentence-level phenomenon rather than a discourse-level process.

This study addresses these gaps by: treating terminological consistency as an independent dimension of translation quality; exploring discourse-level lexical cohesion in AI-generated translation; systematically comparing ChatGPT and Google Gemini; focusing specifically on English-Arabic political and electoral discourse; combining quantitative consistency measurement with qualitative discourse analysis.

The study thus advances both translation studies and an expanding research area of AI-created multilingual communication.

## 2.7 Related Studies

Koehn and Knowles (2017), in their study "Six Challenges for Neural Machine Translation," identified terminology translation as one of the major challenges facing neural machine translation systems. The researchers found that NMT systems may produce inconsistent translations of specialized terms, despite improvements in fluency and overall translation quality.

This study is similar to the present research because both examine terminological issues in AI-based translation and emphasize the importance of consistency in translation quality. However, Koehn and Knowles focused on general challenges in neural machine translation, whereas the present study specifically investigates terminological consistency in English–Arabic political and electoral translations generated by ChatGPT and Google Gemini.

According to Juliane House (2015), evaluation of translation quality should take into account not just accuracy but also coherence, cohesion, and pragmatic equivalency. The study highlights how important lexical repetition is to maintaining textual consistency and effective communication. This perspective provides a theoretical basis for seeing terminological stability as an essential component of translation quality, which makes it highly relevant to the present investigation. However, House's work is primarily theoretical, does not provide an actual evaluation of terminological diversity, and does not particularly address AI translation systems. By applying it to AI-generated translations and defining terminological stability as a quantifiable criterion in political discourse, this study builds upon the current paradigm.

According to Philipp Koehn (2020), neural machine translation systems frequently produce many appropriate translations for the same source phrase due to their probabilistic nature. Because it clarifies the causes of the inconsistent language produced by AI systems throughout a document, this insight is particularly relevant to the current investigation. Koehn's research is primarily descriptive and does not look at how such

variance affects textual coherence or discourse-level interpretation. On the other hand, by investigating the impact of terminological variety on coherence and clarity in political discourse and providing a comparative evaluation of two cutting-edge AI systems, ChatGPT and Google Gemini, the current work directly addresses this problem.

## 3.Methodology

### 3.1 Research Design

This study adopts a mixed-methods research design that combines quantitative measurement and qualitative analysis to investigate terminological consistency in AI-generated English–Arabic translation of political and electoral discourse. The quantitative part measures consistency with percentages of stability and the qualitative part analyses the patterns of variation and their effect on lexical cohesion and clarity.

A comparative corpus-based study systematically contrasts the outputs of two AI systems, Google Gemini and ChatGPT, with a dataset created for the repeated use of essential terminology.

### 3.2 Corpus Selection

The corpus is made up of 30 political and electoral texts that contain key terminology. In total 60 key terms were selected based on their frequency, relevance to electoral discourse and, where possible, their alignment with standardized terminology. The texts were selected to have terms repeated in consistent contexts to allow for accurate measurement of terminological consistency.

### 3.3 Data Collection

Data were collected by translating the 30 source texts from English to Arabic using both AI systems. All translations were performed by a standardized prompt with instructions for the systems to produce formal Arabic translations without further elaboration. For consistency, the same prompt was used for all texts and systems.

The translations were generated once per system per text, and the outputs were not regenerated or polished. This allows us to obtain results that mirror the natural translation behavior of the systems without iterative optimization. A reference translation was also prepared using the Arabic Lexicon of Electoral Terminology published by the United Nations Development Programme (UNDP).

### 3.4 Data Analysis Procedure

The analysis was done in two steps, quantitative and qualitative. First, the 60 key terms were identified with their occurrences, and the translations were extracted from the two systems. Lexical equivalents were grouped according to their semantic proximity. Consistency was calculated by the following formula:

Consistency (%) = (Number of consistent uses / Total number of occurrences) * 100

A term was deemed consistent when the same Arabic lexical equivalent was maintained in all the occurrences of a source term in a text. Any change to another lexical equivalent was assigned inconsistency coding, even if the alternative translation was semantically acceptable. Minor grammatical variations such as definiteness, number or

inflectional changes were not taken to be inconsistent unless the lexical form itself was changed. The coding decisions were based on lexical repetition and semantic equivalence with respect to the UNDP reference terminology.

Second, a qualitative analysis was conducted to find types of variation including synonym substitution, lexical expansion, reduction and paraphrasing. The analysis also considered the effect of these variations on lexical cohesion and clarity.

### 3.5 Reliability Considerations

Reliability was guaranteed by applying consistent criteria in identifying and classifying terminological variation. Lexical equivalents were grouped by semantic similarity using a coding protocol. Counting was of real lexical differences, not grammatical or stylistic variation.

The researcher performed all classifications based on pre-defined rules. Ambiguous cases were reviewed carefully and cross-checked against the UNDP lexicon where applicable, however, no second annotator was involved. This is a limitation and future work may involve multiple evaluators to assess inter-rater reliability.

### 3.6 Limitations

This study is limited to the analysis of 60 political and electoral terms and the translation outputs of two AI systems, ChatGPT and Google Gemini. Therefore, the findings cannot be generalized to all types of political discourse, translation contexts, or AI translation systems. In addition, the analysis was conducted by a single evaluator, which may introduce a degree of subjectivity in identifying and classifying terminological variation.

Finally, the use of the UNDP Arabic Lexicon of Electoral Terminology as a reference provides a standardized basis for evaluation; however, it may not account for all acceptable contextual or stylistic variations that could occur in political translation.

## 4. Results& Discussion

This chapter presents and discusses the quantitative and qualitative findings on terminological consistency in English–Arabic political and electoral translations generated by ChatGPT and Google Gemini. The analysis covers 60 recurring terms with 820 occurrences across 30 source texts. It addresses the two research questions by comparing the systems' overall consistency and identifying the main forms and discourse-level effects of terminological variation.

A term was coded as consistent when the same Arabic lexical equivalent was maintained across its repeated occurrences. An alternative rendering was coded as inconsistent even when it was semantically acceptable. This criterion distinguishes terminological stability from general accuracy: a sentence may be correctly translated while the document remains inconsistent if the same concept is repeatedly given different Arabic labels.

### 4.1 Quantitative Results

Table 1 presents the frequency of the 60 selected terms, their consistency percentages, and the numbers of consistent and inconsistent uses produced by each system. Table 2 summarizes the systems' overall term-level averages. The tables are followed by an interpretation of the main patterns.

**Table 1. Term-Level Consistency Results for ChatGPT and Gemini**

| Term | Category | Occ | ChatGPT_pct | Gemini_pct | ChatGPT_cons | ChatGPT_incons | Gemini_cons | Gemini_incons |
|---|---|---|---|---|---|---|---|---|
| electoral process | Institutional/electoral | 64 | 91.8 | 95.2 | 59 | 5 | 61 | 3 |
| electoral law | Legal | 21 | 51.7 | 73.9 | 11 | 10 | 16 | 5 |
| electoral commission | Administrative/procedural | 35 | 100 | 100 | 35 | 0 | 35 | 0 |
| election | Institutional/electoral | 48 | 95.8 | 92.1 | 46 | 2 | 44 | 4 |
| electoral system | Institutional/electoral | 17 | 100 | 88.9 | 17 | 0 | 15 | 2 |
| electoral framework | Institutional/electoral | 6 | 83.3 | 66.7 | 5 | 1 | 4 | 2 |
| election cycle | Institutional/electoral | 8 | 100 | 87.5 | 8 | 0 | 7 | 1 |
| electoral reform | Institutional/electoral | 11 | 90.9 | 72.7 | 10 | 1 | 8 | 3 |
| voter | Campaign/participation | 28 | 96.4 | 92.8 | 27 | 1 | 26 | 2 |
| vote | Campaign/participation | 54 | 88.9 | 83.3 | 48 | 6 | 45 | 9 |
| ballot | Administrative/procedural | 21 | 100 | 95.2 | 21 | 0 | 20 | 1 |
| ballot box | Administrative/procedural | 6 | 100 | 83.3 | 6 | 0 | 5 | 1 |
| vote counting | Administrative/procedural | 23 | 100 | 86.4 | 23 | 0 | 20 | 3 |
| recount | Administrative/procedural | 8 | 100 | 87.5 | 8 | 0 | 7 | 1 |
| valid vote | Administrative/procedural | 7 | 100 | 85.7 | 7 | 0 | 6 | 1 |
| invalid ballot | Administrative/procedural | 7 | 100 | 85.7 | 7 | 0 | 6 | 1 |
| majority | Governance/rights/integrity | 14 | 92.8 | 78.6 | 13 | 1 | 11 | 3 |
| simple majority | Governance/rights/integrity | 6 | 100 | 83.3 | 6 | 0 | 5 | 1 |
| absolute majority | Governance/rights/integrity | 6 | 100 | 83.3 | 6 | 0 | 5 | 1 |
| representation | Governance/rights/integrity | 19 | 89.5 | 73.9 | 17 | 2 | 14 | 5 |
| proportional representation | Governance/rights/integrity | 6 | 100 | 83.3 | 6 | 0 | 5 | 1 |
| electoral threshold | Governance/rights/integrity | 9 | 100 | 88.9 | 9 | 0 | 8 | 1 |

| Term | Category | Occ | ChatGPT_pct | Gemini_pct | ChatGPT_cons | ChatGPT_incons | Gemini_cons | Gemini_incons |
|---|---|---|---|---|---|---|---|---|
| parliament | Institutional/electoral | 11 | 100 | 100 | 11 | 0 | 11 | 0 |
| candidate | Campaign/participation | 18 | 83.3 | 72.2 | 15 | 3 | 13 | 5 |
| political party | Campaign/participation | 17 | 82.4 | 70.6 | 14 | 3 | 12 | 5 |
| coalition | Campaign/participation | 9 | 77.8 | 66.7 | 7 | 2 | 6 | 3 |
| campaign | Campaign/participation | 33 | 60.6 | 41 | 20 | 13 | 14 | 19 |
| campaign strategy | Campaign/participation | 12 | 83.3 | 66.7 | 10 | 2 | 8 | 4 |
| campaign finance | Campaign/participation | 9 | 100 | 87.5 | 9 | 0 | 8 | 1 |
| media coverage | Campaign/participation | 13 | 92.3 | 76.9 | 12 | 1 | 10 | 3 |
| opinion polls | Campaign/participation | 10 | 90 | 80 | 9 | 1 | 8 | 2 |
| voter participation | Campaign/participation | 15 | 93.3 | 80 | 14 | 1 | 12 | 3 |
| turnout | Campaign/participation | 19 | 100 | 80 | 19 | 0 | 15 | 4 |
| political participation | Campaign/participation | 14 | 92.8 | 78.6 | 13 | 1 | 11 | 3 |
| public trust | Governance/rights/integrity | 12 | 100 | 83.3 | 12 | 0 | 10 | 2 |
| governance | Governance/rights/integrity | 12 | 100 | 92.3 | 12 | 0 | 11 | 1 |
| transparency | Governance/rights/integrity | 16 | 100 | 90 | 16 | 0 | 14 | 2 |
| accountability | Governance/rights/integrity | 19 | 100 | 100 | 19 | 0 | 19 | 0 |
| legitimacy | Governance/rights/integrity | 14 | 100 | 92.8 | 14 | 0 | 13 | 1 |
| democracy | Governance/rights/integrity | 10 | 100 | 90 | 10 | 0 | 9 | 1 |
| human rights | Governance/rights/integrity | 3 | 100 | 100 | 3 | 0 | 3 | 0 |
| political rights | Governance/rights/integrity | 3 | 100 | 100 | 3 | 0 | 3 | 0 |
| civil society | Governance/rights/integrity | 8 | 100 | 87.5 | 8 | 0 | 7 | 1 |
| electoral integrity | Institutional/electoral | 11 | 100 | 90.9 | 11 | 0 | 10 | 1 |
| fraud | Governance/rights/integrity | 9 | 100 | 100 | 9 | 0 | 9 | 0 |

| Term | Category | Occ | ChatGPT_pct | Gemini_pct | ChatGPT_cons | ChatGPT_incons | Gemini_cons | Gemini_incons |
|---|---|---|---|---|---|---|---|---|
| vote buying | Governance/rights/integrity | 4 | 100 | 100 | 4 | 0 | 4 | 0 |
| electoral violence | Governance/rights/integrity | 9 | 100 | 100 | 9 | 0 | 9 | 0 |
| electoral dispute | Governance/rights/integrity | 12 | 83.3 | 66.7 | 10 | 2 | 8 | 4 |
| judicial review | Legal | 5 | 80 | 57.1 | 4 | 1 | 3 | 2 |
| legal framework | Legal | 6 | 83.3 | 66.7 | 5 | 1 | 4 | 2 |
| referendum | Institutional/electoral | 6 | 100 | 100 | 6 | 0 | 6 | 0 |
| suffrage | Legal | 4 | 100 | 75 | 4 | 0 | 3 | 1 |
| polling station | Administrative/procedural | 8 | 100 | 87.5 | 8 | 0 | 7 | 1 |
| polling day | Administrative/procedural | 7 | 100 | 85.7 | 7 | 0 | 6 | 1 |
| voter registration | Administrative/procedural | 9 | 100 | 100 | 9 | 0 | 9 | 0 |
| electoral observer | Administrative/procedural | 7 | 100 | 85.7 | 7 | 0 | 6 | 1 |
| monitoring | Administrative/procedural | 6 | 100 | 83.3 | 6 | 0 | 5 | 1 |
| public policy | Governance/rights/integrity | 5 | 100 | 80 | 5 | 0 | 4 | 1 |
| gender roles | Governance/rights/integrity | 5 | 100 | 80 | 5 | 0 | 4 | 1 |
| political actors | Campaign/participation | 6 | 83.3 | 66.7 | 5 | 1 | 4 | 2 |

Table 2: The Systems' Overall Term-level Averages

| System | Average Consistency (%) | Avg. Consistent Uses | Avg. Inconsistent Uses |
|---|---|---|---|
| ChatGPT | **94.4%** | High | Low |
| Gemini | **83.8%** | Moderate | Higher |

The results show a clear overall advantage for ChatGPT. Its average term-level consistency rate was 94.4%, compared with 83.8% for Gemini, a difference of 10.6 percentage points. ChatGPT produced 759 consistent uses out of 820 occurrences, whereas Gemini produced 691. Based on these occurrence counts, the weighted rates were approximately 92.6% for ChatGPT and 84.3% for Gemini. Although term-level and occurrence-weighted averages assign different importance to frequent terms, both

measures indicate that ChatGPT maintained Arabic equivalents more consistently.

ChatGPT scored higher for 58 of the 60 terms. Gemini performed better only for electoral process (95.2% compared with 91.8%) and electoral law (73.9% compared with 51.7%). The result for electoral law is particularly notable because it was ChatGPT's least stable item. Thus, ChatGPT's stronger overall performance should not be interpreted as uniform superiority for every term; individual results remained sensitive to lexical meaning and context.

Both systems reached 100% consistency for electoral commission, parliament, accountability, human rights, political rights, fraud, vote buying, electoral violence, referendum, and voter registration. These expressions denote relatively fixed institutional, procedural, or rights-related concepts and generally have widely recognized Arabic equivalents. Their stability suggests that established institutional usage limits the range of alternatives selected by the systems.

The weakest results were concentrated among semantically flexible terms. Campaign scored 60.6% in ChatGPT and 41.0% in Gemini, while judicial review scored 80.0% and 57.1%, respectively. Other unstable items included coalition, political party, candidate, campaign strategy, electoral framework, legal framework, electoral dispute, and political actors. These terms permit several plausible Arabic formulations, making contextual reformulation more likely.

Frequency alone did not guarantee consistency. Vote occurred 54 times but achieved 88.9% consistency in ChatGPT and 83.3% in Gemini. Campaign appeared 33 times and remained one of the least stable terms in both systems. By contrast, electoral commission occurred 35 times and was fully consistent in both outputs. Stability therefore appears to depend more on the availability of a conventional institutional equivalent than on repetition alone.

### 4.2 Results by Terminological Category

Category-level analysis further clarifies the difference between the systems. ChatGPT achieved 100% average consistency for administrative and procedural terminology, compared with 88.8% for Gemini. This category includes ballot, vote counting, recount, polling station, polling day, voter registration, and electoral observer. The high scores reflect the procedural specificity of these concepts and the availability of conventional Arabic forms.

Governance, rights, and integrity terminology was also highly stable. ChatGPT achieved 98.4%, while Gemini achieved 88.3%. Institutional and electoral terminology produced averages of 95.8% and 88.2%, respectively. Terms such as accountability, transparency, legitimacy, parliament, referendum, fraud, and electoral violence were especially stable because they are common in official political and institutional discourse.

The lowest averages occurred in legal terminology and campaign or participation terminology. Legal terms averaged 78.8% in

ChatGPT and 68.2% in Gemini, while campaign and participation terms averaged 87.5% and 74.5%, respectively. The legal category contains only four items and should therefore be interpreted cautiously. Nevertheless, the low scores for electoral law and judicial review show that legal expressions with competing Arabic equivalents are particularly vulnerable to inconsistency. Campaign-related language was similarly unstable because terms such as campaign, coalition, candidate, and political participation can be reformulated according to context.

## 4.3 Types of Terminological Variation

The qualitative analysis identified five recurring patterns: synonym substitution, lexical expansion or specification, reduction or generalization, paraphrasing, and mixed usage. These patterns appeared in both systems but were more frequent in Gemini's output.

Synonym substitution was the most common type. Judicial review, for example, was rendered as القضائية المراجعة, القضائية الرقابة, and القضائي الطعن. These expressions overlap semantically but may emphasize different procedures: review, oversight, or legal challenge. Electoral law likewise appeared as أحكام القانون, القانون, القانوني القانون, or الانتخابات قانون. The alternatives may be acceptable in isolation, but alternating among them reduces terminological precision and can alter the scope of the concept.

Lexical expansion occurred when the translation added a descriptive element. Campaign could appear as الحملة and later as الانتخابية الحملة, while representation could alternate between التمثيل and التمثيل البرلماني. The expanded form may clarify a particular sentence, but inconsistent expansion may suggest that the general and specific forms refer to different concepts.

Reduction or generalization produced the opposite effect. A specific expression such as electoral law was sometimes shortened to القانون, or electoral process was referred to without repeating the complete term. The sentence may remain understandable, but the broader wording removes information and weakens the lexical connection between repeated mentions.

Paraphrasing involved expressing a repeated concept through a different lexical or syntactic structure instead of maintaining one fixed term. It was most visible in flexible items such as campaign, vote, and political participation. Paraphrase can improve local naturalness, but it makes a concept more difficult to trace across a document and complicates comparison with glossaries and reference translations.

Mixed usage occurred when a system alternated between several valid equivalents within the same text. This pattern was more prominent in Gemini. It suggests that Gemini placed greater emphasis on immediate contextual adaptation, whereas ChatGPT showed greater persistence in maintaining an earlier lexical choice.

## 4.4 Discussion

The findings answer the first research question by showing that ChatGPT demonstrated greater terminological

consistency than Gemini in the examined corpus. Its higher average, larger number of consistent occurrences, and stronger performance for 58 terms indicate a systematic advantage in lexical stability. However, Gemini's better results for electoral process and electoral law show that system-level rankings must be supported by term-level examination. Neither system was uniformly reliable.

The second research question concerned the forms of terminological change. The results show that inconsistency did not usually arise from obvious mistranslation. Instead, it resulted mainly from alternation among plausible synonyms, expanded or reduced expressions, paraphrases, and context-dependent formulations. This distinction is important because fluent output may conceal discourse-level instability. A translation can be grammatically accurate and natural at the sentence level while failing to maintain a stable terminology system across the complete text.

The results support the connection between lexical repetition and cohesion described by Halliday and Hasan (1976) and Baker (2018). Repeated terminology helps readers recognize that the same institution, procedure, right, or political concept remains under discussion. When the Arabic equivalent changes, the lexical chain becomes less visible and the reader must decide whether the forms are co-referential or signal a conceptual difference.

This effect is especially important in legal and institutional discourse. Related Arabic expressions may differ in procedural meaning. For instance, alternation among the equivalents of judicial review may shift emphasis from review to oversight or appeal, while reducing electoral law to القانون removes its specific electoral reference. Terminological variation may therefore affect conceptual clarity even when the general meaning of the sentence is retained.

The findings are also consistent with House's (2015) view that translation quality extends beyond grammatical accuracy and semantic equivalence to include coherence and pragmatic function. ChatGPT's higher stability makes it more suitable than Gemini for tasks requiring strict terminology control, but its low performance for electoral law and campaign confirms the need for human review. Gemini's greater lexical flexibility may improve the naturalness of isolated sentences, yet it can weaken consistency across longer documents.

In practice, AI-assisted political translation should be supported by an approved bilingual glossary, explicit instructions to preserve designated equivalents, and a post-translation consistency check. Reviewers should prioritize legal and campaign-related terms because they showed the greatest instability. Human oversight remains necessary whenever competing Arabic expressions may carry different legal, procedural, or institutional implications.

These findings should be interpreted within the study's limits. The corpus consisted of 30 texts and 60 terms, each system produced one unedited translation per text,

and the classifications were completed by a single evaluator. The results describe the behavior observed under this controlled design rather than permanent properties of either platform. Model versions, prompt wording, context length, and repeated generation may produce different outcomes. Future research should therefore use larger corpora, multiple outputs, clearly documented model versions, and more than one evaluator.

## 5. Conclusion and Recommendations

### 5.1 Conclusion

The results of the analysis reveal clear differences between ChatGPT and Google Gemini in terms of terminological consistency. Overall, ChatGPT demonstrates a higher level of stability, with an average consistency rate of 94.4%, compared to 83.8 % for Gemini. This indicates that ChatGPT is generally more consistent in maintaining a single lexical equivalent for repeated political and electoral terms across the corpus.

Both systems show high consistency when translating standardized institutional terminology, such as *electoral commission*, *accountability*, *fraud*, and *electoral violence*. These terms tend to have fixed and widely accepted equivalents in Arabic, which reduces the likelihood of variation. However, differences become more apparent in terms that allow multiple valid translations. Gemini frequently introduces lexical variation in such cases, as seen with terms like *judicial review*, *electoral law*, and *campaign*, where multiple Arabic equivalents are used within the same text. While these variations are often semantically accurate, they reduce terminological stability and may weaken lexical cohesion.

The analysis also shows that high-frequency terms are not necessarily more stable. In some cases, frequently repeated terms such as *campaign* and *vote* exhibit lower consistency due to stylistic reformulation or synonym use. This suggests that AI systems prioritize fluency and naturalness over strict repetition, particularly in contexts where variation is linguistically acceptable.

In conclusion, the findings confirm that terminological consistency remains a key challenge in AI-generated translation. ChatGPT tends to favor consistency, making it more suitable for contexts that require strict terminological control, such as political and institutional texts. Gemini, on the other hand, demonstrates greater lexical flexibility, which may enhance readability but can reduce coherence at the discourse level. These results highlight the importance of considering terminological stability as a distinct and measurable dimension of translation quality in AI systems.

### 5.2 Recommendations

Future studies in this area should consider the issue of terminological consistency in larger and more varied corpora to look at whether the patterns found in this study are present across different types of political and institutional discourse.

Additional research could include comparisons of other AI translation systems and newer iterations of large language models, providing a wider scope of terminological performance.

Researchers are invited to explore the problem of terminological consistency in other specialized fields such as legal, diplomatic, medical and technical translation where conceptual precision is equally important.

Future research may also examine the effect of prompt design, terminology constraints and custom glossaries in improving consistency in AI-generated translations.

Finally, it would be interesting to study the impact of consistent versus inconsistent terminology on lexical cohesion and overall translation quality in order to better understand the impact of terminology variation on reader understanding and discourse coherence.

## Conflict of Interest

The author declares no conflict of interest

## Declaration of AI Use

The author declares that ChatGPT and Google Gemini were used to generate translation outputs examined in this study as the study aims to compare the terminological consistency in translation generated by these two platforms. AI tools were also used to assist in preparing comparison tables, language editing, proofreading, and improving the clarity of the manuscript. All research design, evaluation criteria, analysis, interpretation of results, discussion, and conclusions were conducted and verified by the author.

## References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/1409.0473

Baker, M. (2018). *In other words: A coursebook on translation* (3rd ed.). Routledge.

Freitas, A., Guerberof-Arenas, A., & Moorkens, J. (2024). Large language models and translation variation: Challenges for terminology consistency in specialized domains. *Machine Translation*, 38(1), 55–78.

Google DeepMind. (2023). *Gemini: A family of highly capable multimodal models.* https://arxiv.org/abs/2312.11805

Habash, N. (2010). *Introduction to Arabic natural language processing.* Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201008HLT010

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* Longman.

Hendy, A., Abdelrehim, M., Sharaf, A., et al. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint.* https://arxiv.org/abs/2302.09210

House, J. (2015). *Translation quality assessment: Past and present.* Routledge.

Koehn, P. (2020). *Neural machine translation.* Cambridge University Press. https://doi.org/10.1017/9781108608485

Newmark, P. (1988). *A textbook of translation.* Prentice Hall.

OpenAI. (2023). *GPT-4 technical report.* https://arxiv.org/abs/2303.08774

Schäffner, C. (2004). Political discourse analysis from the point of view of translation studies. *Journal of Language and Politics, 3*(1), 117–150. https://doi.org/10.1075/jlp.3.1.07sch

Toral, A., & Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation. *Proceedings of EACL*, 1063–1073.

Venuti, L. (2012). *The translation studies reader* (3rd ed.). Routledge.

Wang, Z., Li, Y., & Chen, X. (2024). Terminology consistency in large language model translation: A comparative evaluation of GPT-based systems. *Journal of Artificial Intelligence and Language Processing, 8*(2), 77–96.

**الاقتباس:** بالة، أسامة علي. (2026). الاتساق الاصطلاحي في الترجمة السياسية المولَّدة بالذكاء الاصطناعي من الإنجليزية إلى العربية: دراسة مقارنة بين ChatGPT وGoogle Gemini. *مجلة كلية الآداب جامعة مصراتة (Faculty of Arts Journal)*. (22)، 148–166. https://doi.org/10.36602/faj.2026.n22.08

# الاتساق الاصطلاحي في الترجمة السياسية المُولَّدة بالذكاء الاصطناعي من الإنجليزية إلى العربية: دراسة مقارنة بين ChatGPT وGoogle Gemini

أسامة علي مصباح باله *

قسم اللغة الانجليزية، كلية الآداب، جامعة مصراتة

*Osama.m.bala@art.misuratau.edu.ly

تاريخ التقديم: 2026-07-27 تاريخ القبول: 18 -2026-08 نشر إلكترونيا في: 2026-08-21

**ملخص البحث:**

أصبحت الترجمة الآلية أكثر طلاقة ودقة سياقية بفضل التطورات الأخيرة في مجال الذكاء الاصطناعي. ومع ذلك، لا يزال الاتساق المصطلحي مجالًا لم ينل حقه الكافي من البحث والدراسة، لا سيما في الخطاب السياسي والانتخابي حيث يُعد التكرار اللفظي والدقة المفاهيمية أمرين حاسمين لتحقيق التماسك والوضوح النصي. تبحث هذه الدراسة في الاتساق المصطلحي في الترجمات السياسية من الإنجليزية إلى العربية المولَّدة بواسطة تقنيات الذكاء الاصطناعي، وتحديدًا عبر نظامي "تشات جي بي تي" و"جوجل جيميناي". واجريت المقارنات في الترجمات والتحليلات في الفترة الممتدة بين يناير ويونيو 2026 باستخدام الإعدادات الافتراضية للنظامين، وذلك لضمان القابلية للمقارنة وتجنب أي تباينات محتملة قد تنجم عن المعلمات المتغيرة التي يحددها المستخدم. تعتمد الدراسة على منهج وصفي تحليلي قائم على تحليل (30) نصًا سياسيًا وانتخابيًا يتضمن (60) مصطلحًا رئيسيًا متكررًا. ويقيس التحليل الكمي درجة الاتساق في شكل نسب مئوية للاستقرار المصطلحي، بينما يدرس التحليل الكيفي التباين اللفظي وأثره على تماسك الخطاب ووضوحه. كما تم التحقق من مدى كفاية الترجمة واتساقها مقارنةً بترجمة مرجعية مستندة إلى "معجم المصطلحات الانتخابية" الصادر عن برنامج الأمم المتحدة الإنمائي (UNDP). وتشير النتائج إلى أن "تشات جي بي تي" حقق اتساقًا مصطلحيًا أعلى مقارنة بنظام "جوجل جيميناي"؛ إذ كانت المقابلات اللفظية التي قدمها "تشات جي بي تي" للمصطلحات السياسية والانتخابية المتكررة أكثر استقرارًا وثباتًا من نظيرتها في "جوجل جيميناي"، والتي أظهرت تباينًا لفظيًا أكبر، لا سيما في المصطلحات الحساسة للسياق مثل: (campaign – الحملة)، و(electoral law – القانون الانتخابي)، و( judicial review – المراجعة القضائية / الطعن القضائي). وتخلص الدراسة إلى ضرورة اعتبار الاتساق المصطلحي بُعدًا مستقلًا من أبعاد جودة الترجمة، كما تؤكد على أهمية ضبط المصطلحات والتحرير البشري اللاحق في الترجمة السياسية المدعومة بالذكاء الاصطناعي.

**الكلمات المفتاحية:** *الاتساق المصطلحي؛ الترجمة بالذكاء الاصطناعي؛ الخطاب السياسي؛ الترجمة من الإنجليزية إلى العربية؛ جودة الترجمة؛ التماسك اللفظي.*