ASCAT: Arabic Scientific Benchmark for Advanced Translation Evaluation

Serry Sibaee[1], **Khloud Al Jallad**[2], **Zineb Yousfi**[3], **Israa Elhosiny**[3], **Yousra El-Ghawi**[3], **Batool Balah**[3], **Omer Nacar**[4]
[1]Prince Sultan University, [2]SySSR
[3]NAMAA Community, [4]Tuwaiq Academy
ssibaee@psu.edu.sa, k.jallad.l@gmail.com, yousfi.zineb.yz@gmail.com, israaelhosiny@gmail.com, yousraghawi@gmail.com, batoolnajeh@gmail.com, o.najar@tuwaiq.edu.sa

Abstract

We present ASCAT (Arabic Scientific Corpus for Advanced Translation), a high-quality English-Arabic parallel benchmark corpus designed for scientific translation evaluation constructed through a systematic multi-engine translation and expert post-editing pipeline. Unlike existing Arabic-English corpora that rely on short sentences or single-domain text, ASCAT targets full scientific abstracts averaging 125.3 words (English) and 111.78 words (Arabic), drawn from five scientific domains: physics, mathematics, computer science, quantum mechanics, and artificial intelligence. Each abstract was translated using three complementary architectures: generative AI (Gemini), transformer-based models (Hugging Face `quickmt-en-ar`), and commercial MT API (Google Translate), and subsequently post-edited by domain experts at the lexical, syntactic, and semantic levels. The resulting corpus contains 67,293 English tokens and 60,026 Arabic tokens, with an Arabic vocabulary of 17,604 unique words reflecting the morphological richness of the language. We benchmark three state-of-the-art LLMs on the corpus: GPT-4o-mini (BLEU: 37.07), Gemini-3.0-Flash-Preview (BLEU: 30.44), and Qwen3-235B-A22B (BLEU: 23.68) demonstrating its discriminative power as an evaluation benchmark. ASCAT addresses a critical gap in scientific MT resources for Arabic and is designed to support rigorous evaluation of scientific translation quality and training of domain-specific translation models.

Keywords: Arabic machine translation, parallel corpus, scientific translation, multi-engine translation, neural machine translation, English-Arabic NLP, domain-specific corpus, corpus construction

1. Introduction

The rapid growth of scientific literature has intensified the need for reliable domain-specific translation models, particularly for Arabic spoken by over 400 million people yet grossly underrepresented in scientific discourse. This language gap creates a critical accessibility barrier for Arabic-speaking researchers and professionals. A key bottleneck in improving Arabic scientific machine translation (MT) is the scarcity of high-quality parallel corpora that maintain terminological accuracy and conceptual consistency.

This paper presents the construction and analysis of an English-Arabic scientific translation corpus built through a systematic multi-engine machine translation and expert post-editing process. By combining multiple state-of-the-art translation engines including generative AI (Gemini), transformer-based models (Hugging Face), and commercial MT APIs (Google Translate, DeepL) with rigorous expert post-editing, our approach balances scalability with quality. The corpus covers full scientific abstracts across diverse domains including physics, mathematics, computer science, quantum mechanics, and artificial intelligence, averaging 125.3 words (English) and 111.8 words (Arabic) per abstract far exceeding the complexity of existing datasets. Critically, ASCAT is explicitly designed as an evaluation benchmark rather than a large-scale training dataset, prioritizing depth of validation over dataset size.

2. Background

Existing English-Arabic parallel corpora suffer from several key limitations. General-domain resources such as MultiUN (Eisele and Chen, 2010) and OPUS (Tiedemann, 2012) lack the terminological precision required for scientific translation. Domain-specific datasets like DEAST (Author and Others, 2026) (33,000 thesis title pairs, $\sim$9 words average) and PEACH (Al-Sabbagh, 2024) (51,671 healthcare sentence pairs, $\sim$10–12 words average) are too short to capture the syntactic complexity and discourse-level phenomena of scientific abstracts. While Tarjama-25 (Hennara et al., 2025) targets longer sentences ($\sim$75 words), its 5,000-sentence scale is insufficient for training large MT models. ATHAR (Mohammed and Khalil, 2025) addresses classical Arabic scientific texts, which differ substantially in register from modern scientific writing. Table 1 provides a structured comparison of these corpora against our dataset.

Multi-engine translation approaches have emerged as a promising strategy, exploiting

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 76–80
May 11, 2026.

Table 1: Comparison of English–Arabic Parallel Corpora

Dataset	Domain	Size	Type	Avg. Len.	Val.
DEAST (Author and Others, 2026)	Multi-sci.	33k	Titles	9 w	Expert
PEACH (Al-Sabbagh, 2024)	Health	51.7k	Leaflets	10–12 w	Experts
ATHAR (Mohammed and Khalil, 2025)	Classical	66k	Historical	–	Expert
Tarjama-25 (Hennara et al., 2025)	Multi-dom.	5k	Long sent.	75 w	Prof.
ASCAT (Ours)	Multi-sci.	500	Abstracts	**125 / 112 w**	**Multi-stage**

complementary strengths: statistical MT excels at terminology consistency, neural MT produces fluent output, and large language models handle contextual and idiomatic expressions. Combining their outputs enables comparative analysis and focuses human validation on high-risk segments. Human validation remains essential in scientific translation: full post-editing, selective validation guided by confidence scoring, and inter-annotator agreement metrics each play a role in ensuring quality, yet few existing corpora document their validation protocols transparently.

Our dataset addresses these gaps by targeting full scientific abstracts, employing a multi-engine translation pipeline, and applying multi-stage expert post-editing yielding a resource suitable for both training domain-specific MT models and benchmarking scientific translation quality.

3. Methodology

The dataset construction follows a three-stage pipeline: Data Collection, Multi-Engine Translation, and Expert Post-editing as illustrated in Figure 1.

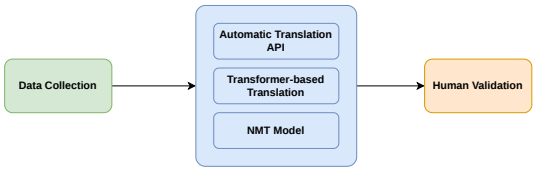

Figure 1: Dataset construction pipeline: Data Collection, Multi-Engine Translation, and Human Validation.

3.1. Data Collection

Scientific abstracts were systematically gathered across multiple domains, including physics, mathematics, computer science, quantum mechanics, and artificial intelligence. To ensure representational diversity and avoid selection bias, abstracts were randomly sampled from papers across these domains, yielding a balanced multi-domain corpus of full-length scientific abstracts.

The abstracts were sourced from arXiv, randomly selected from the gfissore/arxiv-abstracts-2021 dataset, which contains scientific papers from arXiv. arXiv is an open-access repository that allows redistribution of abstracts under its licensing terms, as the content is openly available for research and educational purposes.

3.2. Preprocessing

To ensure consistency and handle scientific content, we applied minimal preprocessing to the English abstracts before translation. LaTeX expressions and mathematical symbols were preserved in their original English form, as they are universally understood in scientific contexts and do not require translation. Scientific symbols (e.g., Greek letters, operators) were similarly kept unchanged. No automated normalization was applied to the text after translation; any adjustments to LaTeX or symbols were made individually by reviewers during the post-editing phase if deemed necessary for clarity or accuracy.

The preprocessing pipeline consisted of:

1. Extraction of plain text from the arXiv abstracts.
2. Random selection of samples for translation.
3. Preservation of LaTeX and scientific symbols.

3.3. Multi-Engine Translation

To maximize linguistic diversity and enable comparative analysis, each abstract was translated using three distinct architectures in the following order: first Gemini, then Hugging Face `quickmt-en-ar` and Google Translate. Samples were randomly selected for translation without additional filtering beyond domain balance.

3.4. Human Validation Process

The final stage involved comprehensive expert review and post-editing of all translated samples. Post-editing was conducted by seven domain experts, each holding at minimum an undergraduate-level degree in either Arabic linguistics or a relevant scientific discipline (physics, mathematics, computer science, or AI).

Validators worked independently using a structured checklist, spending an average of 10-30 minutes per sample. The post-editing process involved

four main types of edits: substitutions (replacing incorrect terms), insertions (adding missing information), deletions (removing redundant or erroneous content), and rewrites (restructuring sentences for better fluency and accuracy).

Since we distributed the abstracts to reviewers and did not conduct formal inter-annotator agreement (IAA), our validation focused on qualifying the quality of the translations and fixing errors rather than measuring inter-annotator consistency. This decision was intentional, as our primary goal was to produce high-quality references for benchmarking (see Table 2).

Table 2: Human Validation Checklist

Level	Criterion	Binary (Y/N)
Lexical	Domain terminology	Y
Lexical	Preservation of NEs	Y
Syntactic	Grammatical correctness	Y
Syntactic	Sentence structure	Y
Semantic	Epistemic hedging	N
Discourse	Sentence consistency	N

Expert post-editors with expertise in both Arabic linguistics and the relevant scientific domains post-edited translations at the lexical, syntactic, and semantic levels, correcting terminology errors, structural inconsistencies, and meaning deviations. Concrete examples of edits include:

- **Substitution:** Changing "machine learning" to (al-taallum al-ālī) for accurate domain terminology.
- **Insertion:** Adding (fī majāl) to specify "in the field of" for clarity.
- **Deletion:** Removing redundant particles that do not contribute to meaning.
- **Rewrite:** Restructuring awkward sentences to improve flow, e.g., converting passive voice to active where appropriate in Arabic.

A detailed analysis of the most frequent translation errors identified during this phase is provided in Section 4.

4. Corpus Statistics

4.1. Domain Coverage

The dataset was strategically curated to ensure broad coverage of modern scientific disciplines. The corpus spans five primary domains: Quantum Mechanics, Artificial Intelligence, Computer Science, Mathematics, Physics. This multi-domain design ensures that the corpus captures the terminological diversity and stylistic variation inherent to different scientific fields, making it suitable for training and evaluating general-purpose scientific MT systems rather than narrow single-domain models. The deliberate balance across domains also mitigates the risk of domain-specific lexical bias, where a model trained predominantly on one field may underperform on others.

4.2. Abstract Length Analysis

To assess the linguistic complexity of the corpus, we analyzed word and character distributions across all abstract pairs. Table 3 summarizes the key statistics for both the English source and Arabic target abstracts.

Table 3: Abstract Length Statistics

Lang	Min	Med	Mean	Max	SD	Mean Ch.
EN	3	113	125.3	297	64.0	822
AR	4	100	111.8	315	58.9	695

The English abstracts average 125.31 words per abstract (median: 113), with a standard deviation of 63.99 words, indicating substantial length variability across domains and paper types. Arabic translations average 111.78 words (median: 100) with a standard deviation of 58.87. The consistently lower word count in Arabic relative to English is linguistically expected, as Arabic's rich morphological system and agglutinative properties allow it to encode more information per word through affixation and cliticization.

The high standard deviation in both languages reflects the natural heterogeneity of scientific abstracts: concise theoretical results may be presented in fewer than 50 words, while comprehensive experimental studies may span close to 300 words. This variability is a desirable property for a training corpus, as it exposes MT models to a wide spectrum of abstract lengths and structural complexities.

4.3. Vocabulary and Lexical Richness

Table 4 presents the vocabulary and token-level statistics for both language sides of the corpus.

Table 4: Vocabulary Statistics

Lang	Tokens	Unique words	TTR	Mean Sent.
EN	67,293	12,685	0.19	7.16
AR	60,026	17,604	0.29	6.99

Several notable observations emerge from the vocabulary analysis. First, despite having fewer total tokens (60,026 vs. 67,293), the Arabic side exhibits a substantially larger vocabulary size (17,604 unique words vs. 12,685 in English). This is reflected in the Type-Token Ratio (TTR) (Richards and Schmidt, 2013), a standard measure of lexical

diversity, where Arabic scores 0.2933 compared to English's 0.1885. The higher Arabic TTR is consistent with the morphological complexity of Arabic, where a single root can generate dozens of surface forms through derivational and inflectional processes, resulting in a much larger effective vocabulary space. This has direct implications for MT model design, as Arabic-side models require larger vocabularies or subword tokenization strategies (e.g., BPE or SentencePiece) to adequately cover the target language's lexical space.

Second, both language sides exhibit comparable mean sentence counts per abstract (7.16 for English and 6.99 for Arabic), confirming that the translation process preserved the discourse segmentation of the original abstracts without undue merging or splitting of sentences. This structural fidelity is important for training models that must learn to produce coherent, multi-sentence scientific discourse rather than isolated sentence translations.

5. Evaluation

The following evaluation demonstrates ASCAT's utility as a discriminative benchmark, a role for which validation quality and domain complexity matter more than corpus size.

To assess the quality of our corpus and benchmark the performance of state-of-the-art translation systems on scientific English-Arabic translation, we evaluated three large language models against our human-validated reference translations. The models evaluated are **GPT-4o-mini** (OpenAI), **Gemini-3.0-Flash-Preview** (Google DeepMind), and **Qwen3-235B-A22B** (Alibaba Cloud). Each model was prompted to translate the English source abstracts, and the outputs were scored against the human-validated Arabic references using two standard automatic evaluation metrics: BLEU and ROUGE.

5.1. Results

Table 5 presents the automatic evaluation results for all three models.

Table 5: Automatic Evaluation on EN–AR Scientific Translation

Model	BLEU ↑	R-1 ↑	R-2 ↑	R-L ↑
GPT-4o-mini	**37.07**	**0.590**	**0.476**	**0.586**
Gem-3.0-Flash	30.44	0.530	0.390	0.522
Qwen3-235B	23.68	0.537	0.410	0.531

5.2. Discussion

GPT-4o-mini achieves the highest BLEU (37.07), indicating strong alignment with expert-validated references(Yan et al., 2024). BLEU scores above 30 signify high-quality translations in Arabic, reflecting robust n-gram overlap in scientific contexts.

Gemini-3.0-Flash-Preview (BLEU: 30.44) shows adequate content coverage but less precise matching, while Qwen3-235B (BLEU: 23.68) performs competitively on ROUGE-1 yet structurally distant.

The 13.4 BLEU gap demonstrates the corpus's discriminative power. Compared to DEAST's 2.15–3.36 BLEU on short titles (Author and Others, 2026), our 23.68–37.07 scores highlight superior quality for benchmarking advanced MT on long-form text.

Error analysis revealed issues in terminology, structure, and semantics, emphasizing scientific MT challenges and expert post-editing value.

Moderate scores across models underscore the difficulty of scientific Arabic translation, justifying domain-specific resources like ASCAT.

6. Limitations and Future Work

Limitations include the dataset's size (500 abstracts, prioritized for quality), uneven domain distribution, and reliance on automatic metrics that miss qualitative aspects like semantic nuance.

Scientific translation challenges include terminological ambiguity, non-standardized terms, epistemic hedging, and acronyms without Arabic equivalents.

Future work: expand corpus for balanced domains, add human evaluation, fine-tune models for improved translation, conduct per-domain evaluation, log edit-level statistics, and analyze lemmatized TTR.

7. Conclusion

We presented ASCAT, a high-quality English-Arabic parallel benchmark corpus for scientific translation, constructed via multi-engine MT and expert post-editing. It addresses gaps in Arabic scientific MT with abstract-level complexity, multi-domain coverage, and transparency.

Analysis shows Arabic's morphological richness (TTR 0.29, vocab 17,604). Benchmarking revealed a 13.4 BLEU gap, proving discriminative power while highlighting translation challenges.

ASCAT advances Arabic scientific communication and supports domain-specific MT research.

8. Acknowledgments

We would like to thank the NAMAA community for their support in this project.

9. Bibliographical References

Rania Al-Sabbagh. 2024. Peach: A sentence-aligned parallel english-arabic corpus for healthcare. *Corpora*, 19(3):395–410.

M. A. Author and Others. 2026. Deast: A dataset for english-arabic scientific translation and vice versa. *Data in Brief*, 64:112381.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

K. Hennara, M. Hreden, M. M. Hamed, Z. Aldallal, S. Chrouf, and S. AlModhayan. 2025. Mutarjim: Advancing bidirectional arabic-english translation with a small language model. *arXiv preprint arXiv:2505.17894*.

Mahmoud S. Mohammed and Mohamed Khalil. 2025. Athar: A high-quality and diverse dataset for classical arabic to english translation. In *Proceedings of the Third Arabic Natural Language Processing Conference*, pages 97–106.

Jack C Richards and Richard W Schmidt. 2013. *Longman dictionary of language teaching and applied linguistics*. Routledge.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of LREC 2012*, pages 2214–2218.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels.