ТЕОРЕТИЧЕСКАЯ, ПРИКЛАДНАЯ И СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНАЯ ЛИНГВИСТИКА/THEORETICAL, APPLIED AND COMPARATIVE LINGUISTICS

DOI: https://doi.org/10.60797/RULB.2026.73.14

LLMS AND THE DOMAIN OF MACHINE TRANSLATION

Review article

Kostinikova O.A.[1, *]

[1] The Russian Presidential Academy of National Economy and Public Administration, Saint-Petersburg, Russian Federation

* Corresponding author (olk2004[at]mail.ru)

Abstract

LLMs have become powerful tools for synthetic data creation, automatic corpus expansion, and structure-aware prompting, all of which have unlocked progress in areas previously constrained by data scarcity. From augmenting speech-to-speech translation resources to generating domain-specific evaluation corpora, LLMs now act as both the objects of study and the instruments that enable continued field advancement. The review of the recent articles published in November 2025 shows that across speech, translation, and multimodal generation, the meta-trend is clear: low-resource, domain-intense, and structurally complex tasks that use synthetic data, linguistically informed representations, and human-aligned evaluation metrics are now the main drivers of progress in the domain of Machine Translation.

Keywords: LLMs, Automatic Speech recognition, Machine Translation.

БЯМИ В СФЕРЕ МАШИННОГО ПЕРЕВОДА

Обзор

Костиникова О.А.[1, *]

[1] Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, Санкт-Петербург, Российская Федерация

* Корреспондирующий автор (olk2004[at]mail.ru)

Аннотация

Большие Языковые Модели стали мощными инструментами для создания синтетических данных, расширения корпусов и структурно-ориентированных запросов. Они открыли путь прогрессу в областях, в которых ранее развитие сдерживалось нехваткой данных. В диапазоне проблем от расширения ресурсов перевода до создания оценочных корпусов, ориентированных на предметную область, БЯМ теперь выступают и как объекты исследования и как инструменты, которые обеспечивают постоянное развитие в этой области. Обзор недавних статей, опубликованных в ноябре 2025 года, показывает, что в речи, переводе и мультимодальной генерации мета-тенденция очевидна: низкозатратные, узко-пердметные и структурно сложные задачи, использующие синтетические данные, лингвистически информированные представления и оценочные метрики, идентичные человеческим, в настоящее время являются основными двигателями прогресса в области машинного перевода.

Ключевые слова: БЯМ, автоматическое распознавание речи, машинный перевод.

Introduction

With the advent of large language models, the field of translation has entered a period of accelerated development, marked by new levels of fluency, robustness, and cross-lingual generalization showing progress in language technology. Yet, the integration of LLMs into translation and speech systems is far from straightforward. Despite their impressive capabilities, LLMs must navigate challenges such as data imbalance, complex morphology, diverse dialects, multimodal constraints, and discourse-level coherence — factors that traditional pipelines often handled through specialized modules and handcrafted linguistic knowledge [1], [5], [8], [10].

In translation, LLMs have demonstrated remarkable gains in general-domain text, but their performance declines in expert-level, culture-heavy, or long-document scenarios. These limitations have motivated the creation of new evaluation frameworks and benchmarks that can capture discourse structure, terminology consistency, and stylistic fidelity — dimensions essential for real-world deployment. At the same time, LLMs have proven invaluable for synthetic parallel data generation, enabling low-resource machine translation (MT) and speech-to-speech translation (S2ST) systems to achieve quality previously impossible with human-curated datasets alone. So the emergence of LLM-centred translation marks a turning point in the field. The integration of generative models with traditional signal-processing and linguistic components is enabling hybrid systems that surpass the existing limitations. As research pushes into more complex, low-resource, and domain-intense scenarios, LLMs are poised not only to enhance translation but to redefine the scientific and engineering principles that underpin them [2], [3], [6], [9]. The idea is clearly presented in some recent research published in November 2025 that is the subject of the review.

Crossing Borders: A Multimodal Challenge for Indian Poetry Translation and Image Generation

This article [6] introduces a multimodal framework designed to translate morphologically rich Indian poetry and generate corresponding visual representations that capture both literal and metaphorical meaning. The authors begin by highlighting that Indian poetry —spanning more than twenty languages and numerous dialects — contains deep cultural context, layered metaphors, and highly complex morphology, making it especially difficult for both humans and machines to interpret. Prior work on poetry translation has focused mainly on structured poetic forms or high-resource languages, leaving Indic poetry

severely understudied in NLP. To address this gap, the paper proposes the Translation and Image Generation (TAI) framework, a two-step system that first translates poems using a Large Language Model guided by an Odds Ratio Preference Alignment algorithm. This alignment step steers the LLM toward translations closer to human poetic interpretations rather than literal, machine-like renderings.

In the second step, TAI uses a semantic graph–based prompt generator to prepare text descriptions for latent diffusion models. These graphs map dependencies, metaphors, and symbolic relations within the poem, enabling the image generation system to produce visuals that reflect conceptual meaning rather than shallow keyword associations. The authors show an example where a Punjabi poem describing a farming scene is translated into English and visualized with culturally accurate elements such as landscape, attire, and mood.

A major contribution of the work is the introduction of MorphoVerse, a new dataset of 1,570 poems across 21 low-resource Indian languages, each with rich morphological variation. This dataset helps address the lack of training resources for Indic literary NLP and supports broader research on cross-cultural poetic understanding. The authors argue that existing LLMs struggle with such texts because morphological richness and metaphorical density exceed the patterns seen in mainstream training corpora. They observe that raw LLM translations often omit cultural nuance or misinterpret figurative expressions, motivating the need for alignment mechanisms.

The paper also emphasizes the importance of multimodal comprehension for global accessibility of literary heritage. By linking translation and visual generation, TAI provides an interpretive bridge for readers unfamiliar with the cultural or linguistic context of Indian poetry. Experimental evaluation — both automated and human—shows that the proposed TAI Diffusion approach outperforms strong baselines in producing accurate, meaningful images aligned with poetic content. The authors conclude that multimodal methods, enriched with structural linguistic knowledge, offer a promising direction for revitalizing interest in low-resource literary traditions and making them accessible to a worldwide audience.

Improving Direct Persian-English Speech-to-Speech Translation with Discrete Units and Synthetic Parallel Data

This article [9] investigates how to improve direct Speech-to-Speech Translation (S2ST) between Persian and English by addressing the core barrier of data scarcity in low-resource languages. The authors begin by noting that direct S2ST avoids the traditional ASR (Automatic Speech recognition) →MT (Machine Translation)→TTS (Text-to-Speech) pipeline, reducing latency and eliminating error propagation, but it requires large amounts of parallel speech data, which Persian lacks. They propose a new architecture consisting of a Conformer-based encoder pretrained on Persian speech, a causal Transformer decoder predicting discrete speech units, and a neural unit-vocoder generating English speech.

A major contribution is the construction of a synthetic Persian–English parallel corpus, created by translating Persian transcripts using GPT-4o and synthesizing English audio with VoiceCraft TTS. This synthetic corpus expands existing parallel data by a factor of six and significantly improves training effectiveness.

Qualitative analyses show that synthetic data increases fluency and reduces omissions, particularly in long or rare constructions. The paper also states that discrete units help stabilize training and disentangle content from prosody. The authors argue that their pipeline is especially suitable for dubbing applications where speaker identity preservation and low latency matter. Their results demonstrate that combining self-supervised pretraining, discrete units, and synthetic corpora is a powerful strategy for low-resource S2ST.

They conclude that this framework can generalize to other under-resourced languages and plan future work involving prosody-aware unit modelling, multilingual training, and cross-lingual pretraining. Overall, this study provides both a new dataset and a strong model that advance the state of direct S2ST for Persian–English translation.

Non-Linear Scoring Models for Translation Quality Evaluation

This article [3] examines why traditional linear scoring models used in translation-quality evaluation fail to reflect real human perception. The authors begin by stating that industry practice typically assumes a direct linear relationship between text length and the number of allowed errors. The authors argue that human tolerance for errors increases with length, but sub-linearly, meaning longer texts allow more errors, but not proportionally more. They introduce empirical data showing that reviewers' acceptable error counts rise quickly at first and then flatten, forming a curve best approximated by a logarithmic function. This aligns with psychophysical principles such as the Weber–Fechner law, which states that human perception of change grows logarithmically. It also aligns with Cognitive Load Theory, because accumulated errors increase mental load nonlinearly, reducing tolerance as text grows.

The authors then present a mathematically simple, two-parameter logarithmic scoring model which captures actual tolerance patterns. Figures in the paper show that logarithmic curves closely follow expert judgments, while linear curves significantly diverge, especially for long texts. Importantly, the authors emphasize that for very small samples (<250 words), no deterministic curve is reliable, and statistical quality-control methods must be used instead.

They argue that implementing this model will bring translation-quality evaluation closer to real-world perception and reduce unfair pass/fail outcomes. They propose integrating this tolerance function into AI-evaluation systems as well, since LLM outputs suffer from similar short- vs. long-text biases, so the idea is that adopting non-linear calibration is essential for fair, human-aligned translation-quality evaluation.

Conclusion

This collection of the research papers reveals a coherent picture of modern computational linguistics: due to LLMs, the field has moved decisively toward low-resource languages, specialized domains, and complex multimodal tasks that go far beyond traditional translation or ASR. Across all papers, a central shared theme is that data scarcity, not modeling limitations, is the primary barrier — and each paper proposes a novel strategy to overcome it.

First, the papers attack low-resource problems through synthetic data generation or major dataset expansion. The Persian–English S2ST work constructs 6× more data using LLM translation and TTS synthesis, proving that synthetic data dramatically

boosts BLEU. The poetry paper introduces MorphoVerse, the first sizeable dataset for 21 Indic languages. Together, these demonstrate that data augmentation using LLMs, TTS, or expert curation consistently yields the largest jumps in performance.

Second, multiple papers demonstrate the importance of representation learning tailored to each linguistic challenge. Discrete speech units simplify S2ST in resource-scarce conditions. Semantic graph prompting captures metaphorical structure in poetry, bridging a major gap in literary translation.

Third, evaluation itself emerges as a central research frontier. The scoring-model paper argues that the entire translation industry misjudges quality due to inappropriate linear scoring models, proposing a perception-aligned logarithmic alternative. Together, these papers show that evaluation is not a solved problem — and future benchmarks must incorporate discourse, domain expertise, and human-like tolerance curves.

Конфликт интересов

Не указан.

Рецензия

Сообщество рецензентов Международного научно-исследовательского журнала
DOI: https://doi.org/10.60797/RULB.2026.73.14.1

Conflict of Interest

None declared.

Review

International Research Journal Reviewers Community
DOI: https://doi.org/10.60797/RULB.2026.73.14.1

Список литературы на английском языке / References in English

1. Berdejo-Espinola V. AI tools can improve equity in science / V.Berdejo-Espinola, T. Amano // Science. — 2023. — Vol. 379 (6636). — P. 991. — DOI: 10.1126/science.adg9714.

2. Dubey P. The Hindi to Dogri machine translation system: grammatical perspective / P.Dubey // International Journal of Information Technology. — 2019. — Vol. 11 (1). — P. 171–182. — DOI: 10.1007/s41870-018-0085-4.

3. Gladkoff S. Non-Linear Scoring Models for Translation Quality Evaluation / S.Gladkoff, H. Lifeng, K. Gasova // arXiv. — 2025. — DOI: 10.48550/arXiv.2511.13467.

4. Gray A. ChatGPT “contamination”: estimating the prevalence of LLMs in the scholarly literature / A.Gray // arXiv. — 2024. — DOI: 10.48550/arXiv.2403.16887.

5. Hu E.J. LoRA: Low-Rank Adaptation of Large Language Models/ E.J. Hu, Y. Shen, P. Wallis [et al.] // arXiv. — 2021. — DOI: 10.48550/arXiv.2106.09685.

6. Jamil S. Crossing Borders: A Multimodal Challenge for Indian Poetry Translation and Image Generation / S.Jamil, K.S. Charan, S. Saha [et al.] // arXiv. — 17 2025. — DOI: 10.48550/arXiv.2511.13689.

7. Litvinova T.A. Writing in the era of large language models: a bibliometric analysis of research field / T.A. Litvinova, G.K. Mikros, O.V. Dekhnich // Research Result. Theoretical and Applied Linguistics. — 2024. — Vol. 10. — № 4. — P. 5–16. — DOI 10.18413/2313-8912-2024-10-4-0-1.

8. Navigli R. Biases in large language models: origins, inventory, and discussion / R. Navigli, S. Simone, B. Ross // ACM Journal of Data and Information Quality. — 2023. — Vol. 15 (2). — P. 1–21. — DOI: 10.1145/3597307.

9. Rashidi S. Improving Direct Persian-English Speech-to-Speech Translation with Discrete Units and Synthetic / S. Rashidi, H. Sameti // arXiv. — 2025. — DOI: 10.48550/arXiv.2511.12690.

10. Wang P. Large language models are not fair evaluators / P. Wang, L. Li, L. Chen [et al.] // arXiv. — 2023. — DOI: 10.48550/arXiv.2305.17926.