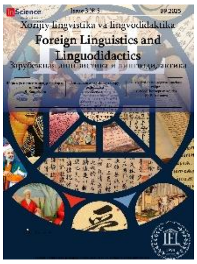
inScience
through time and space

Xorijiy lingvistika va lingvodidaktika – Зарубежная лингвистика и лингводидактика – Foreign Linguistics and Linguodidactics
Journal home page:
https://inscience.uz/index.php/foreign-linguistics



Basic notions and terminology of modern machine translation

Mokhirukh KHOSHIMKHUJAEVA [1]

Termez University of Economics and Service

ARTICLE INFO

Article history:
Received August 2025
Received in revised form 15 August 2025
Accepted 15 September 2025
Available online
05 October 2025

Keywords:
source language,
target language,
structural differences,
morphological differences,
semantics,
pragmatics,
translation model,
lexicon,
monolingual corpora,
multilingual parallel corpora,
natural language processing.

ABSTRACT

This article discusses the fundamental concepts and terminology of modern machine translation. Each idea is analyzed in terms of its function and significance in the translation process. Definitions of terms are considered based on relevant translation examples in English, Japanese, Russian, Turkish, Spanish, Chinese, and French. Thus, the article examines the relationship between the source and target languages based on the primary translation factors, including structural, morphological, and semantic differences. In addition, the importance of concepts such as translation models, language vocabulary, monolingual and multilingual parallel corpora, and natural language processing is substantiated.

2181-3701/© 2025 in Science LLC.
DOI: https://doi.org/10.47689/2181-3701-vol3-iss5-pp74-84

Zamonaviy mashina tarjimasining asosiy tushunchalari va terminologiyasi

Kalit so'zlar:
manba tili,
maqsadli til,
strukturaviy farqlar,
morfologik farqlar,
semantika,
pragmatika,
tarjima modeli,
leksika,

ANNOTATSIYA

Ushbu maqolada zamonaviy mashina tarjimasining asosiy tushunchalari va terminologiyasi muhokama qilinadi. Maqolada har bir tushuncha tarjima jarayonidagi vazifasi va ahamiyati nuqtai nazaridan tahlil qilinadi. Atamalar ta'riflari ingliz, yapon, rus, turk, ispan, xitoy va fransuz tillariga tegishli tarjima misollari asosida ko'rib chiqiladi. Shunday qilib, maqola tarkibiy, morfologik va semantik farqlar kabi asosiy tarjima omillari asosida manba va maqsadli tillar o'rtasidagi munosabatlarni

[1] PhD, Associate Professor, Head of the International Department, Termez University of Economics and Service.
E-mail: moxirux_xoshimxojayeva@tues.uz

bir tilli korpus,
koʻp tilli parallel korpus,
tabiiy tilni qayta ishlash.

oʻrganadi. Tarjima modeli, til lugʻati, bir tilli va koʻp tilli parallel korpus, tabiiy tilni qayta ishlash kabi tushunchalarning ahamiyati asoslanadi.

Основные понятия и терминология современного машинного перевода

Ключевые слова:
исходный язык,
целевой язык,
структурные различия,
морфологические различия,
семантика,
прагматика,
модель перевода,
лексикон,
одноязычный корпус,
многоязычный параллельный корпус,
обработка естественного языка.

АННОТАЦИЯ

В данной статье рассматриваются основные понятия и терминология современного машинного перевода. Каждое понятие анализируется с точки зрения его функции и значимости в процессе перевода. Определения терминов иллюстрируются примерами переводов на английский, японский, русский, турецкий, испанский, китайский и французский языки. Таким образом, статья исследует взаимосвязи между исходным и целевым языками на основе ключевых факторов перевода, таких как структурные, морфологические и семантические различия. Обосновывается важность таких понятий, как модель перевода, языковой словарь, одноязычный и многоязычный параллельный корпус, обработка естественного языка.

INRTODUCTION

In a time of rapid technological progress and growing global connectivity, machine translation (MT) has become a vital tool for overcoming language barriers and promoting cross-cultural communication. Modern MT systems rely on complex linguistic, computational, and statistical models that allow for the automatic translation of texts between languages. Understanding the fundamental concepts and terminology behind these systems is important not only for developers and linguists but also for educators, policymakers, and end-users who want to evaluate and use MT effectively. This article covers key ideas such as source and target languages, translation models, lexical resources, and types of corpora, while also addressing the structural, morphological, semantic, and pragmatic challenges involved in language translation. Special focus is given to translating low-resource languages, which are still underrepresented in current technologies, highlighting the need for fair development of language technology.

MATERIALS AND METHODS

This study adopts a qualitative-descriptive approach to analyze the core concepts and terminology of modern machine translation (MT) within the broader context of computational linguistics. The analysis is based on a comparative review of primary and secondary sources, including scholarly literature, multilingual corpora, and practical examples from existing machine translation systems. Key terms and concepts relevant to machine translation are identified through a comprehensive review of academic publications, textbooks, and conference proceedings in the fields of natural language processing (NLP) and machine translation (MT). Authoritative sources such as Koehn, Vaswani et al., and Papineni et al. (7, 16, 11), along with recent studies on low-resource language processing (3, 13), are selected to ensure both historical depth and contemporary relevance. Terminological definitions are then contextualized using examples from languages with diverse typological structures, including English, Japanese, Turkish, Russian, Spanish, French, and Chinese. To illustrate linguistic phenomena such as structural, morphological, semantic, and pragmatic differences, the study employs contrastive linguistic examples, selecting sample sentences based on their ability to highlight challenges in machine translation such as idiomatic expressions, syntactic reordering, and morphological richness. These examples are either drawn from publicly available machine translation system outputs (e.g., Google Translate, DeepL) or manually constructed to reflect typical translation issues. An overview informs the discussion of monolingual and multilingual corpora of established resources such as the British National Corpus, EUROPARL, and domain-specific corpora referenced in Seraji and Zakharov & Tao (14, 17). Although no new corpora are developed in this study, the structural characteristics, annotation practices, and application domains of these corpora are critically examined. Furthermore, the study outlines different types of MT models, including word-for-word, phrase-based, and neural machine translation (NMT), with descriptions and comparisons based on technical documentation, academic evaluations, and model architecture specifications such as the Transformer model (16). Evaluative metrics such as BLEU score and n-gram models are also reviewed to underscore the methodologies used in assessing MT output quality.

DISCUSSION

Machine translation refers to the automatic conversion of text from one language to another. It is a complex endeavor that relies on the intricate relationship between the source and target languages. The source language (SL), the language in which the original text is written, and the target language (TL), to which the text is being translated, play a crucial role in the success and accuracy of machine translation systems.

For example, when translating a sentence from English to French, English is the SL, and French is the TL. The distinction between SL and TL is critical for building and evaluating MT systems, as it determines the alignment of linguistic structures and meanings between the two languages. The concepts of **source language** and **target language** are foundational in machine translation.

Source Language ---------> Target Language
(Input Text) MT System (Output Text)

This simple scheme illustrates how input text in the SL is processed by MT system to produce output text in the TL. The process is influenced by structural differences, semantic alignment, and morphological variations between the two languages. These differences can manifest in various ways, such as sentence structure, grammatical conventions, and cultural nuances (2). As the translation process attempts to bridge these gaps, it often faces difficulties in preserving the original meaning and intent of the text (15). In the following examples, we can see the various factors that designate these differences.

1. Structural Differences:

Structural differences refer to variations in the syntactic and grammatical rules of languages. For example, English follows a Subject-Verb-Object (SVO) word order: *The cat* (Subject) *eats* (Verb) *the fish* (Object). Japanese, in contrast, follows a Subject-Object-Verb (SOV) word order: *Neko wa sakana o tabemasu* (猫は魚を食べます), which directly translates to "The cat (Subject) the fish (Object) eats (Verb)". These differences create challenges in machine translation, as word order and syntactic alignment must be adjusted to produce grammatically correct target language outputs.

Example:

Source Language: English

The cat eats the fish

Target Language: Japanese

Neko wa sakana o tabemasu

Additionally, languages with rich morphological structures, such as Turkish or Finnish, require careful segmentation and alignment. For example, Turkish: *Evimde* "in my house" combines root and suffixes into a single word, which may require multiple words in English. Such structural and morphological variations necessitate advanced approaches in machine translation to ensure both syntactic and semantic accuracy. Languages may differ in syntax, grammar, and word order (e.g., English follows the Subject-Verb-Object order, while Japanese uses the Subject-Object-Verb order).

2. Morphological differences:

Languages vary in how words are inflected to indicate tense, number, or case. For example, in English, the word *run* can be inflected to *runs* (third-person singular) or *ran* (past simple tense). In contrast, in Turkish, a single word can carry multiple morphological inflections. For instance, the word *evlerimizde* translates to "in our houses", where *ev* means "house", *ler* indicates plurality, *imiz* denotes possession (our), and *de* adds the locative case (in). This complex morphology poses challenges for MT systems, as accurate segmentation and interpretation of inflected words are essential for successful translation.

3. Semantics: The meaning of words may not always align perfectly between languages due to cultural and contextual nuances. For example, the English word *rice* refers to the grain in general, but in Japanese, there are distinct words: *gohan* (ご飯) for cooked rice and *kome* (米) for uncooked rice. Similarly, in Russian, the word *sneg* (снег) means snow, but Inuit languages have multiple terms that describe snow with varying degrees of texture and condition. These cultural and contextual distinctions highlight the challenges of achieving precise semantic equivalence in machine translation.

4. Pragmatics: The way language is used in social and cultural contexts can vary significantly across languages (15). For instance, the conventions of politeness, formality, and conversational norms may differ between languages. This can impact the translation of idiomatic expressions, honorifics, and other pragmatic elements that convey meaning beyond the literal interpretation of words (2). This interplay between source and target languages is a crucial aspect of machine translation, as it determines the level of accuracy and naturalness of the translated output. Addressing the structural, morphological, semantic, and

pragmatic differences between languages is essential for developing effective and reliable machine translation systems (10, 6, 1, 2).

In the translation process, a type of translation model plays a crucial role. A translation model defines how words, phrases, or sentences are mapped from the SL to the TL. There are several types of translation models. They differ from each other according to the method of translation. One and the very first translation model used in machine translation is word-for-word translation. It refers to a word-for-word translation method, where each word in the SL is replaced by its equivalent word in the TL. This method does not consider grammatical structure, context, or idiomatic expressions, which can result in incorrect or unnatural translations. For instance, in the following example, we can see the advantages and drawbacks of this method.

English: *I love dogs*

Russian: *Я люблю собак*

As we can see, word-for-word translation works well for simple sentences with direct word correspondences. But in many cases, it has some limitations. For example:

English: *The car is fast*

Japanese: *Kuruma wa hayai desu*

While this seems correct, a word-for-word approach may miss nuances. For instance, if the context implies *the car runs fast*, the Japanese might use *"Kuruma wa hayaku hashirimasu.*

But the main limitations of word-for-word translation is related to the translation of Idiomatic expressions and complex structures.

English: *It's raining cats and dogs*

Word-for-word French: *Il pleut des chats et des chiens*

Correct idiomatic translation: *Il pleut des cordes* (It’s raining heavily)

T hus, while word-for-word translation is foundational, more advanced models like phrase-based and neural translation are necessary to achieve accurate results (7).

Phrase-based translation is a translation of multi-word phrases. Translation of multi-word phrases involves dividing text into sequences of words (phrases) and translating these units rather than individual words. This approach improves fluency and accuracy compared to a word-for-word translation by considering local context. In English to Spanish phrase-based translation, the sentence *I am going to the store* is divided into 2 phrases:

Phrase 1: *I am going* -> *Voy*

Phrase 2: *to the store* -> *a la tienda*

The final translation will be *Voy a la tienda.* Here, instead of translating word-by-word (which could lead to errors like *Yo soy yendo a la tienda*), the system recognizes phrases and maps them to their correct equivalents in the target language. The key advantages of phrase-based translation are improved context handling (whereby translating phrases, the system avoids errors caused by isolated word meanings) and better idiomatic translation (where the phrases often capture idiomatic expressions more accurately than single words).

The next model is called contextual translation, characterized by capturing contextual meaning. It is a key feature of Neural Machine Translation (NMT). Unlike traditional methods that translate words or phrases independently, NMT systems, such as those using the Transformer model (16), consider the full context of a sentence to generate accurate translations. For example, in English to Chinese contextual translation, the sentence *I went to the bank to deposit money* the word *bank* has multiple meanings (financial institution or riverbank). NMT identifies the context of *deposit money* and correctly translates *bank* as a financial institution and the correct translation in Chinese will be 我去银行存钱 (*Wǎ qù yínháng cún qián*). In the following example, English is an idiomatic sentence. *She broke the ice at the meeting.* NMT recognizes the idiomatic meaning (starting a friendly conversation) and produces the correct translation in French. *Elle a détendu l'atmosphère à la réunion* where the actual (incorrect, word-for-word) translation is Elle *a cassé la glace à la réunion*. By leveraging attention mechanisms and deep learning, NMT systems dynamically focus on the relevant parts of the input sentence, capturing semantics and context to deliver translations that are both accurate and fluent (16, 7).

More types of translation models are implicated in MT. Most of them have a complicated scheme of language processing, and their quality depends on their ability to handle a range of linguistic phenomena and produce fluent, natural-sounding text in the target language.

Another fundamental notion for MT is the Lexicon. It refers to the set of words or terms the MT system knows. It serves as the foundation for understanding and generating translations, as each word or phrase in the lexicon has a corresponding translation in the TL. Lexicons can be categorized into general lexicons and domain-specific lexicons. The general lexicon contains words and phrases used in everyday language. For example:

English: *cat* -> Spanish: *gato*

English: *book* -> French: *livre*

Domain-specific lexicon focuses on vocabulary used in specialized fields such as medicine, law, or technology. For example:

Medical: English: *cardiomyopathy* -> French: *cardiomyopathie*

Legal: English: *plaintiff* -> Spanish: *demandante*

Technical: English: *server* -> French: *serveur*

Consider translating the English sentence *The doctor diagnosed the patient* into French:

General Lexicon: *doctor* -> *docteur, patient* -> *patient*

Domain-Specific Lexicon (Medical): *diagnosed -> a diagnostiqué*

The correct translation would be: *Le docteur a diagnostiqué le patient*

This example highlights how a domain-specific lexicon improves translation accuracy by selecting appropriate terms for specialized contexts. Without it, the system might use incorrect or ambiguous translations.

The lexicon plays a pivotal role in linguistic and computational tasks, particularly in areas such as word sense disambiguation, contextual accuracy, and translation quality. Resolving ambiguity in word meanings is essential, especially when a single term can have multiple interpretations. For example, the word "bank" may refer to a financial institution or a riverbank, depending on the context. In addition, a well-maintained lexicon ensures the use of domain-relevant terms, enhancing precision in specialized fields. In machine translation, a comprehensive lexicon significantly improves fluency and reduces errors, facilitating effective communication across languages (7).

Monolingual corpora, which consist of text data in a single language, are invaluable resources for linguistic and computational linguistics tasks. These structured collections aim to represent the nuances of a language or its specific subsets, providing a foundation for focused research and tool development. A defining feature of monolingual corpora is their exclusive use of one language, allowing for in-depth analysis of linguistic features and patterns. Typically large in size, they aim to capture a broad spectrum of linguistic phenomena to ensure representativeness of the language or a particular domain. They are often categorized by criteria such as genre, source, or author demographics, and many include annotations like part-of-speech tags or syntactic structures to enhance their utility (14).

There are various types of monolingual corpora, each serving specific purposes. General corpora, such as the Brown Corpus for American English and the British National Corpus for British English, represent diverse genres and domains. Specialized corpora focus on specific fields, such as medical or legal texts, and are indispensable for tasks requiring domain-specific language knowledge. Diachronic corpora provide insights into language evolution by featuring texts from different time periods, while synchronic corpora focus on language usage at a specific point in time. Additionally, learner corpora, which contain texts produced by language learners, offer valuable perspectives on language acquisition processes (14).

Monolingual corpora have diverse applications. They are widely used in lexicography to develop dictionaries and thesauruses, as well as in linguistic research to explore grammar, vocabulary, and language usage. In natural language processing (NLP), these corpora serve as essential resources for training and evaluating models for tasks like part-of-speech tagging and machine translation. Furthermore, they play a role in language teaching by aiding the development of learning materials and the assessment of language proficiency, and in stylistics by enabling the analysis of textual style and authorship (17).

In addition to monolingual corpora, multilingual parallel corpora are indispensable for translation studies, contrastive linguistics, and computational linguistics. These corpora align texts in one language with their translations in one or more target languages, supporting a wide range of applications. A key feature of parallel corpora is the alignment of text segments, such as sentences or paragraphs, which ensures their effectiveness for analysis. While bilingual corpora are the most common, multilingual corpora, which include translations in multiple languages, enable comprehensive cross-lingual studies. Like monolingual corpora, parallel corpora benefit from large, diverse datasets and linguistic annotations that enhance their value for research and machine translation development.

Multilingual parallel corpora also come in various forms. Bilingual corpora, such as the Hansard Corpus and EUROPARL, focus on two languages and are widely used for translation studies. Meanwhile, multilingual corpora, which feature texts in multiple languages, are instrumental in developing multilingual translation systems. Specialized parallel corpora are tailored to specific domains, such as medicine or law, providing targeted resources for domain-specific translation and analysis (14).

By offering robust resources for linguistic analysis, lexicography, and computational applications, both monolingual and multilingual corpora significantly advance our understanding and practical use of language

The rest of the terminology used in MT has a more technical character rather than linguistic. They include Model architecture, encoding, decoding, Learning from data, BLEU Score (Bilingual Evaluation Understudy), n-gram language model, Probability, long short-term network, attention mechanism, Recurrent neural network, etc. These metrics help evaluate translation quality, ensuring that machine translations align with human-like output in terms of accuracy, fluency, and relevance (11).

Natural language processing is a crucial component of machine translation, a field that aims to automatically translate text from one language to another. NLP techniques are used to process and understand the input text, which can then be translated into the target language. One of the key aspects of NLP in machine translation is its ability to handle the inherent ambiguity and complexity of natural languages. Unlike programming languages, which have strict syntax and semantics, natural languages are highly flexible and can have multiple meanings for the same phrase or word (6). NLP algorithms must be able to parse the input text, understand its context and meaning, and then generate an appropriate translation. The role of NLP in machine translation has become increasingly important with the growing popularity of the World Wide Web and the need to communicate across language barriers (6). Search engines, for example, rely on NLP to understand user queries and provide relevant results, often across different languages (12).

The realm of natural language processing is a tapestry woven with languages of varying richness and availability of resources. The disparity in the linguistic landscape is often characterized by the dichotomy of "high-resource" and "low-resource" languages, but the reality is far more nuanced (9, 3). For example, some languages are more "equal" than others, with a few dominant languages monopolizing the attention and resources of the NLP community. Recent studies have revealed that languages can be broadly classified into six distinct categories based on the breadth and depth of their computational and data resources (3). At one end of the spectrum lie the languages that enjoy a wealth of annotated data, pre-trained models, and sophisticated natural language processing tools.

These are the "high-resource" languages, which have long been the focus of the NLP research community. In the middle, we find the "medium-resource" languages, which possess a modest but growing set of resources, often lagging behind their high-resource counterparts in terms of technological advancement (13). An example of high-resource languages would be English, Mandarin Chinese, and Spanish, while medium-resource languages could include Hindi, Arabic, and Russian (13, 3). At the other end of the spectrum, the "low-resource" languages face a stark reality. These are the languages that lack the fundamental building blocks for effective NLP, such as large corpora, annotated datasets, and dedicated language models. The absence of these essential resources poses a significant challenge, as researchers and practitioners struggle to develop robust and reliable NLP solutions for these underserved linguistic communities (13). Examples of low-resource languages may include many minority and endangered languages, such as those spoken in parts of Africa, Asia, and Indigenous communities around the world (8, 9).

It's important to note that these assumptions reflect the specific context of the paper, which focuses on neural machine translation for a particular low-resource language pair. The challenges and considerations for other low-resource languages might vary. Consequently, efforts have been made to address this imbalance, including the creation of data sets, benchmarks, and techniques that favor low-resource languages. However, the problem remains far from solved, and the linguistic divide persists (8, 9, 3, 13).

The concept of low-resource languages, as discussed in Kakum et al., highlights significant challenges in machine translation (4). These languages are defined by their limited digital resources, including parallel corpora, monolingual corpora, and other linguistic assets. This scarcity makes it difficult to train robust machine translation models effectively. Classifying a language as low-resource is not straightforward, as the required volume and diversity of data for model training vary. Simply counting parallel sentences does not fully capture the complexity of this issue, emphasizing the need for datasets with diverse sentence structures and linguistic phenomena. The under-representation of low-resource languages, particularly those from Northeast India, such as *Nyishi*, exacerbates the problem. These languages are often neglected in research due to data limitations, creating a cycle of resource scarcity. Moreover, the authors underscore the importance of involving native speakers in data collection and evaluation to ensure the quality and cultural relevance of translations. Although various efforts have aimed to address these challenges by creating datasets, benchmarks, and techniques tailored to low-resource languages, the issue remains unresolved. The linguistic divide persists, necessitating continued focus and innovation to bridge this gap.

RESULTS

The analysis revealed several key insights into the foundational concepts and terminology of modern machine translation (MT), highlighting the complexity and interdisciplinary nature of the field. The study demonstrated that structural, morphological, semantic, and pragmatic differences between languages significantly affect translation quality. For instance, variations in word order, such as SVO in English compared to SOV in Japanese, and morphological richness, as seen in Turkish, require MT systems to implement advanced alignment and segmentation strategies. These disparities underscore the need for context-aware translation models capable of handling linguistic diversity. A comparative analysis of translation models revealed distinct strengths and weaknesses, with word-for-word translation being effective for simple, literal expressions but failing to adequately translate idiomatic or context-sensitive content.

On the other hand, phrase-based models showed improved fluency and accuracy by grouping words into meaningful units, while Neural Machine Translation (NMT), particularly models based on the Transformer architecture, demonstrated the highest contextual sensitivity, producing semantically and pragmatically accurate translations; however, their performance varies depending on the availability of large-scale training data. Results also indicated that the use of general versus domain-specific lexicons substantially impacts translation precision, as general lexicons provide basic vocabulary coverage, while specialized lexicons are essential for translating medical, legal, and technical texts. Incorporating domain-specific terminology ensures not only terminological accuracy but also enhances the relevance and clarity of translations in professional contexts.

Additionally, monolingual and multilingual corpora were found to be vital resources for training, evaluating, and fine-tuning MT systems. General-purpose corpora, such as the British National Corpus, serve as foundational linguistic datasets, while specialized and parallel corpora support domain-specific translation tasks. The study highlighted the ongoing challenges posed by low-resource languages, which lack sufficient annotated corpora, thereby limiting the performance of MT systems in those languages. In summary, the results affirm that successful machine translation hinges on the interplay between linguistic theory, data availability, and model architecture, emphasizing the need for adaptive, context-aware MT solutions that can effectively bridge linguistic and cultural gaps across a diverse array of languages.

CONCLUSION

In conclusion, modern machine translation is a dynamic interdisciplinary field shaped by linguistic theory, computational innovation, and real-world application. By examining the fundamental concepts and terminology-ranging from translation models and lexicons to monolingual and parallel corpora, this article highlights the intricate mechanisms that power MT systems. It also draws attention to the persistent disparities faced by low-resource languages, advocating for more inclusive and culturally aware translation technologies. As machine translation continues to evolve, a clear understanding of its foundational principles is essential for both enhancing its accuracy and expanding its benefits to all linguistic communities.

REFERENCES:

1. Alhaj, A. A. M. (2023). Lexical-Semantic Problems and Constrains Met in Translating Qur'anic Arabic-Specific words "Nafs نفس "into English: A Cross-lingual Perspective. In Technium Social Sciences Journal (Vol. 44, p. 1025). https://doi.org/10.47577/tssj.v44i1.9073

2. Imami, T. R., Mu'in, F., & Nasrullah, N. (2021). Linguistic and Cultural Problems in Translation. In Advances in Social Science, Education and Humanities Research/Advances in social science, education and humanities research. https://doi.org/10.2991/assehr.k.211021.024

3. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. In Multimedia Tools and Applications (Vol. 82, Issue 3, p. 3713). Springer Science+Business Media. https://doi.org/10.1007/s11042-022-13428-4

4. Okur, B. C., TAKCI, H., & Akgül, Y. S. (2013). Rewriting Turkish texts written in the English alphabet using the Turkish alphabet (p. 1). https://doi.org/10.1109/siu.2013.6531394

5. Papineni, K. (2002). Machine Translation Evaluation: N-grams to the Rescue. In Language Resources and Evaluation. Springer Science+Business Media. http://www.lrec-conf.org/proceedings/lrec2002/pdf/347.pdf

6. Seraji, M. (2015). Morphosyntactic Corpora and Tools for Persian. http://www.diva-portal.org/smash/record.jsf?pid=diva2:800998

7. Timalsina, R. (2023). Overcoming Intercultural Obstacles in Translation. In Dristikon A Multidisciplinary Journal (Vol. 13, Issue 1, p. 156). https://doi.org/10.3126/dristikon.v13i1.56096

8. Koehn, P. (2010). Statistical Machine Translation. Cambridge University Press.

9. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems.

10. Zakharov, Victor & Tao, Yuan. (2015). Разработка и использование параллельного корпуса русского и китайского языков. НТИ. Сер. 2. ИНФОРМ. ПРОЦЕССЫ И СИСТЕМЫ.

11. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In arXiv (Cornell University). Cornell University. https://doi.org/10.48550/arxiv.2004.09095

12. Kakum, N., Laskar, S. R., Sambyo, K., & Pakray, P. (2023). Neural machine translation for limited resources English-Nyishi pair. In Sadhana (Vol. 48, Issue 4). Springer Science+Business Media. https://doi.org/10.1007/s12046-023-02308-8

13. Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., … Bashir, A. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. https://doi.org/10.18653/v1/2020.findings-emnlp.195

14. Nigatu, H. H., Tonja, A. L., Rosman, B., Solorio, T., & Choudhury, M. (2024). The Zeno's Paradox of `Low-Resource' Languages. In arXiv (Cornell University). Cornell University. https://doi.org/10.48550/arxiv.2410.20817

15. Ranathunga, S., & Silva, N. de. (2022). Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. In arXiv (Cornell University). Cornell University. https://doi.org/10.48550/arxiv.2210.08523

16. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. In Multimedia Tools and Applications (Vol. 82, Issue 3, p. 3713). Springer Science+Business Media. https://doi.org/10.1007/s11042-022-13428-4

17. Rajput, A. E. (2019). Natural Language Processing, Sentiment Analysis and Clinical Analytics. In arXiv (Cornell University). Cornell University. https://doi.org/10.48550/arxiv.1902.00679