Western European Journal of Linguistics and Education
Volume 3, Issue 5, May 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

LINGUISTIC ASPECTS OF AUTOMATIC TEXT ALIGNMENT IN PARALLEL CORPORA

Radjabova Gulnoza Giyosiddinovna
rad.gulnoza@gmail.com
PhD, Associate Professor
Uzbekistan State World Languages University

Abstract
Parallel corpora play a crucial role in multilingual natural language processing, machine translation, and contrastive linguistics. A fundamental task in constructing parallel corpora is automatic text alignment which refers to linking corresponding textual units (sentences or paragraphs) across different languages. This article explores the linguistic aspects influencing alignment accuracy, including syntactic structure, word order, phraseology, and translation strategies. We also examine common alignment techniques and assess their linguistic robustness using case studies from English-Russian corpora. The findings show that integrating linguistic features significantly improves alignment precision, especially in complex or free word order languages.
Keywords: *parallel corpora, text alignment, computational linguistics, syntax, translation studies, corpus linguistics, English-Russian alignment*

Introduction

Parallel corpora which are considered to be collections of texts and their translations in one or more languages, are fundamental resources in corpus linguistics, contrastive studies, and natural language processing (NLP). A critical first step in creating a parallel corpus is automatic text alignment, which involves matching segments of source text with their translations at the sentence, paragraph, or phrase level (Tiedemann, 2011). While various statistical and algorithmic approaches to alignment have been developed (Brown et al., 1991; Gale & Church, 1993), alignment remains a linguistically challenging task, particularly for language pairs with significant syntactic and typological differences. This article investigates linguistic challenges and features that influence alignment accuracy, with a focus on English-Russian parallel texts.

This study addresses the following questions:

- What linguistic factors most significantly affect the accuracy of automatic alignment in parallel corpora?
- How do current alignment methods integrate linguistic information?
- What improvements can be proposed for better alignment of typologically divergent languages?

Methods

The study utilizes data from the OPUS (Open Parallel Corpus) collection, a rich and openly available repository of multilingual parallel texts (Lison & Tiedemann, 2016). Specifically, two subcorpora are selected:

OpenSubtitles2018 corpus comprises movie subtitles that have been translated into multiple languages, including English and Russian. Subtitles provide short, conversational

Western European Journal of Linguistics and Education
Volume 3, Issue 5, May 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

sentence pairs rich in colloquialisms, idiomatic expressions, and rapid turn-taking. These features make them suitable for testing the robustness of alignment algorithms under informal and fragmented linguistic conditions.

Europarl is a collection of transcripts from the European Parliament's multilingual proceedings. It contains well-structured and formalized parallel texts, characterized by complete sentences, consistent syntax, and clearly defined speaker turns. Europarl provides a high-quality benchmark for evaluating alignment tools under formal and institutional language use.

By using these two corpora, the study ensures coverage of both informal and formal registers, enabling a comparative analysis of alignment accuracy across different domains and discourse styles (Radjabova, 2025).

To perform sentence-level alignment between English and Russian texts, the study employs two alignment tools that represent distinct methodological paradigms. First tool is HUNALIGN, a widely-used alignment tool that combines sentence-length heuristics with the use of bilingual dictionaries to identify corresponding sentences. It starts with a statistical estimation of alignment based on sentence lengths (following Gale and Church's model), and then refines the alignment using lexical information when a dictionary is available. Hunalign is particularly effective in texts with relatively consistent translation strategies and sentence boundaries (Varga et al., 2005). The second tool which was used in this study is BLEUALIGN, a semantic-based aligner that utilizes machine translation (MT) and BLEU scoring to determine semantic similarity between sentence pairs (Radjabova, 2024). It translates one side of the parallel corpus (typically the source language) into the target language using MT, and then computes a similarity score with potential matches in the target corpus using the BLEU metric. This approach captures semantic equivalence rather than surface structure similarity, making it more suitable for texts with paraphrasing or free translation (Sennrich & Volk, 2010).

These tools are chosen to contrast two core approaches to alignment, namely, Hunalign emphasizes structural and lexical similarity (statistical-algorithmic), while Bleualign focuses on meaning and translation adequacy (semantic-similarity). This dual-tool strategy allows a deeper exploration of how different linguistic features impact alignment success and error types .

Alignment quality is assessed using both quantitative and qualitative metrics. *Precision and recall* are computed by comparing automatically aligned sentence pairs against a manually aligned gold standard consisting of 1,000 English-Russian sentence pairs sampled evenly from both corpora. Precision measures the proportion of correctly aligned sentences among those identified by the system. Recall measures the proportion of correct alignments found by the system out of all correct alignments present in the gold standard. These metrics offer insight into both the accuracy and coverage of the aligners (Giyosiddinovna, 2022). Linguistic error analysis, to supplement the statistical metrics, is conducted on misaligned pairs. Each incorrect alignment is analyzed to determine underlying linguistic causes, such as:

non-literal translation;
sentence splitting or merging across languages;
idiomatic or culturally specific expressions;
divergences in word order and grammatical structure;
anaphoric references and ellipses not explicitly mirrored in translation;

This qualitative approach helps identify systemic linguistic challenges that alignment tools must overcome, especially in language pairs like English-Russian with different syntactic

Western European Journal of Linguistics and Education
Volume 3, Issue 5, May 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

and morphological characteristics. The study uses segments from the OPUS (Open Parallel Corpus) collection, specifically the OpenSubtitles2018 and Europarl corpora. These include millions of English-Russian aligned sentence pairs from movie subtitles and parliamentary proceedings, respectively.

Results

This section compares two automatic alignment tools, namely, Hunalign and Bleualign which are based on precision and recall when aligning English-Russian sentence pairs from parallel corpora (see Tab. 1).

Table 1. The percentage of Precision and Recall from parallel corpora

Tool	Precision	Recall
Hunalign	85.2%	82.7%
Bleualign	89.6%	87.1%

As it can be seen from the Table 1. precision shows the percentage of sentence alignments produced by the tool that are actually correct. For example, if Hunalign outputs 100 alignments and 85 are correct, precision = 85%. Recall shows the percentage of all correct alignments (according to the gold standard) that the tool successfully identifies and if there are 100 true alignments in the reference set and the tool finds 83 of them, recall is equal to 83%. In contrast, Bleualign shows higher precision (89.6%) and recall (87.1%) than Hunalign. It also performs better in both recognizing correct alignments and avoiding incorrect ones and leverages machine translation and semantic similarity, which allows it to align more contextually similar, yet linguistically diverse, sentence pairs. Both tools are effective, but Bleualign demonstrates superior overall accuracy. The results suggest that semantic-based alignment may offer advantages over purely statistical methods, especially in handling variation in translation.

Table 2. Linguistic reasons of automatic alignment failure

Linguistic Feature	% of Misalignments
Word order variation	24%
Ellipsis/omission	18%
Idiomatic translation	15%
Morphosyntactic divergence	12%
Non-literal translation	11%
Sentence splitting/merging	9%
Others	11%

As it can be seen from Table 2. English and Russian often differ in typical sentence order (SVO vs. SOV/Flexible). Alignment tools relying on surface structure struggle when subject, verb, and object positions are rearranged. For example, English sentence “I gave her the book” can be rendered into Russian as “Книгу я ей дал.” There are also cases of Ellipsis/Omission (18%) where content is deliberately omitted in one language due to stylistic or pragmatic reasons. This results in one-to-zero alignments that many tools are not equipped to handle. In addition, there are cases of idiomatic translation (15%) when idioms often do not translate literally and require semantic equivalence. For example, very common English/Russian idiom “Kick the bucket” – “сыграть в ящик” – alignment tools may not detect

Western European Journal of Linguistics and Education
Volume 3, Issue 5, May 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

them as equivalent. Differences in grammatical structure (e.g., use of cases, aspect, agreement) lead to alignment confusion. Russian's rich morphology makes word-to-word and phrase-to-phrase alignment more complex. In these cases, there might be cases of non-literal translation (11%) and translators often adapt the sentence for clarity, tone, or naturalness, departing from direct correspondence. Tools expecting strict 1-to-1 matches often fail here. One sentence in English may be translated as two or vice versa. Alignment tools not configured for many-to-one or one-to-many alignments will mispair such instances.

Thus, nearly 60% of misalignments are caused by word order variation, ellipsis, and idiomatic expressions. These issues suggest a need for alignment tools that incorporate deeper syntactic parsing and semantic role analysis, especially in languages with different structural typologies like English and Russian. The data emphasize that linguistic complexity significantly affects alignment accuracy. Tools like Bleualign that incorporate semantic similarity show better adaptability, but both tools would benefit from enhancements that account for morphosyntactic and idiomatic variation.

Discussion

The results confirm that linguistic divergence, particularly in syntax, idiomatic usage, and translation strategies, remains a core obstacle to automatic alignment. For instance, free word order in Russian makes surface-level features (e.g., sentence length) unreliable. Moreover, translation strategies like modulation, transposition, and adaptation (Vinay & Darbelnet, 1995) often introduce alignment complexity. Recognizing such shifts requires linguistically enriched models that go beyond statistics.

To improve alignment accuracy, we suggest integrating part-of-speech (POS) tagging, syntactic parsing, semantic similarity measures, phrase alignment models. Accurate alignment is foundational for translation studies, corpus-driven language teaching, machine translation training.

Thus, alignment tools must evolve to linguistically aware systems, not just statistical matchers.

Conclusion

Automatic text alignment in parallel corpora is a linguistically complex task that benefits greatly from incorporating syntactic, semantic, and translational knowledge. Our study reveals that linguistic challenges, particularly those related to divergent word order, idiomatic expressions, and syntactic variation, account for the majority of alignment errors (Radjabova, 2023). These issues are especially pronounced in language pairs with significant typological differences, such as English and Russian, where direct one-to-one correspondence is rare. To address these challenges, future alignment tools must move beyond surface-level string matching and statistical co-occurrence, embracing deeper levels of linguistic analysis. Integrating Natural Language Processing (NLP) techniques such as Part-of-Speech (POS) tagging, syntactic parsing, and semantic similarity models can significantly enhance the accuracy and interpretability of alignments. Moreover, the use of transformer-based architectures, which are capable of capturing long-distance dependencies and nuanced contextual meaning, holds great promise for resolving alignment ambiguities in complex sentence structures. Ultimately, the development of linguistically informed alignment systems is not just a technical upgrade, it is a necessary evolution to support high-quality applications in machine translation, multilingual information retrieval, and corpus-driven language education. As alignment forms the foundation of these fields, improving its accuracy through linguistically rich modeling will lead to more reliable research outcomes, more effective language teaching tools, and more fluent and culturally sensitive machine translation systems.

Western European Journal of Linguistics and Education
Volume 3, Issue 5, May 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

This progression will also facilitate the creation of robust multilingual corpora that are not only quantitatively large but qualitatively precise, enabling deeper linguistic insights and broader cross-linguistic exploration.

References

1. Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7, 597–610.

2. Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 169–176.

3. Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1), 75–102.

4. Giyosiddinovna, Radjabova G. "The Implementation of Spoken Corpora in Creating Teaching Materials." International Journal on Integrated Education, vol. 4, no. 5, 2021, pp. 349-354.

5. Giyosiddinovna, Radjabova G. "Methodological Characteristics of Corpus Technologies in Teaching Foreign Language." *International Journal on Integrated Education*, vol. 5, no. 1, 2022, pp. 157-163, doi:10.31149/ijie.v5i1.2645.

6. Radjabova, G. (2023). Corpus technologies in teaching academic writing. *Foreign Languages in Uzbekistan*, *1*(48), 92-103.

7. Radjabova, G. G. (2024). ADJUSTING THE PERSPECTIVE OF CORPUS LINGUISTICS: BRIDGING RESEARCH AND THE CLASSROOM. *American Journal of Modern World Sciences*, *1*(5), 324–332. Retrieved from https://worldejurnal.ru/index.php/ajmws/article/view/401

8. Раджабова, Г. (2025). Аутентичность как значимая характеристика корпусного подхода DDL для совершенствования письменной компетенции студентов. *Лингвоспектр*, *3*(1), 636–640. извлечено от https://lingvospektr.uz/index.php/lngsp/article/view/591

9. Tiedemann, J. (2011). Bitext alignment. Synthesis Lectures on Human Language Technologies, 4(2), 1–97.

10. Vinay, J.-P., & Darbelnet, J. (1995). Comparative Stylistics of French and English: A Methodology for Translation. John Benjamins.