Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2026”

June 24–26, 2026

Building Chukchi Corpora: A Chukchi–Russian Parallel Corpus and a Monolingual Chukchi Corpus for Low-Resource NLP

Sergey Gofman
National Research University
Higher School of Economics
Moscow, Russia
gofmansergeyy@mail.ru

Egor Lebedev
National Research University
Higher School of Economics
Moscow, Russia
lebegorr@gmail.com

Anton Perfilev
National Research University
Higher School of Economics
Moscow, Russia
kekys778@gmail.com

Elizaveta Plohuta-Plakutina
National Research University
Higher School of Economics
Moscow, Russia
e.plakutina@yandex.ru

Liudmila Shlyakhtina
National Research University
Higher School of Economics
Moscow, Russia
lucya135@gmail.com

Abstract

Many endangered and under-resourced languages lack publicly available corpora suitable for modern NLP, limiting research and downstream applications such as language documentation, morphological modeling, search, and alignment-based text study. We present a new Chukchi–Russian parallel corpus and a complementary monolingual Chukchi corpus compiled from heterogeneous online and printed resources. The parallel corpus was assembled via structured extraction, segment-level alignment, cleaning, and semantic diagnostics, yielding 70,508 Chukchi–Russian aligned segments (translation units); the majority of these units are lexicographic word/phrase glosses rather than full sentences. The monolingual corpus contains 15,610 Chukchi segments gathered from untranslated and unaligned materials. We describe data sources, normalization rules for Chukchi orthography, automatic and manual quality control procedures, and corpus statistics. The resulting corpora provide a practical foundation for Chukchi computational research and language preservation efforts.

Keywords: Chukchi, low-resource NLP, parallel corpus, corpus construction, segment alignment
DOI: 10.29003/2075-7182-2026-24-116-124

Создание чукотских корпусов: чукотско-русский параллельный корпус и одноязычный чукотский корпус для NLP в условиях низкоресурсности

Сергей Гофман
Национальный исследовательский университет
Высшая школа экономики
Москва, Россия
gofmansergeyy@mail.ru

Егор Лебедев
Национальный исследовательский университет
Высшая школа экономики
Москва, Россия
lebegorr@gmail.com

Антон Перфильев
Национальный исследовательский университет
Высшая школа экономики
Москва, Россия
kekys778@gmail.com

Елизавета Плохута-Плакутина
Национальный исследовательский университет
Высшая школа экономики
Москва, Россия
e.plakutina@yandex.ru

Людмила Шляхтина
Национальный исследовательский университет
Высшая школа экономики
Москва, Россия
lucya135@gmail.com

Аннотация

Многие исчезающие и малоресурсные языки не имеют публично доступных корпусов, пригодных для современных задач NLP, что ограничивает исследования и прикладные направления, включая языковую документацию, морфологическое моделирование, поиск и выравнивание текстов. Мы представляем новый чукотско-русский параллельный корпус и дополняющий его моноязычный чукотский корпус, собранные из разнородных онлайн- и печатных источников. Параллельный корпус получен с помощью структурированного извлечения, выравнивания на уровне сегментов, очистки и семантической диагностики; он содержит 70 508 выровненных чукотско-русских сегментов (единиц перевода). Моноязычный корпус включает 15 610 чукотских сегментов из непереведённых и невыровненных материалов. Мы описываем источники данных, правила нормализации чукотской орфографии, автоматические и ручные процедуры контроля качества, а также статистику корпусов. Полученные ресурсы создают практическую основу для вычислительных исследований чукотского языка и задач его сохранения.

Ключевые слова: чукотский язык, малоресурсный NLP, параллельный корпус, создание корпусов, выравнивание сегментов

1 Introduction

Corpora are a prerequisite for most computational studies of language, yet for many minority languages the available digital text is sparse, fragmented, and noisy. This is especially acute for endangered languages: limited online presence results in small datasets, inconsistent orthography, and a lack of benchmarks.

Chukchi is an endangered language of northeastern Siberia (Russian Far East). While the umbrella term “Paleo-Asiatic” is sometimes used as a geographic label in the literature, Chukchi is genealogically part of the Chukotko–Kamchatkan language family, with some accounts proposing a narrower Chukchi–Koryak subgroup. The 2010 census (as summarized in a 2024 secondary source) reports an ethnic Chukchi population of around 15,908, while only about 5,095 report being able to speak Chukchi (Institute of Linguistics, Russian Academy of Sciences, 2024); the 2020/2021 census reports about 16,200 ethnic Chukchi. However, contemporary language use may be lower and uneven across age groups and settlements. Chukchi has a very limited web presence. It does not appear as a separately reported language in W3Techs content-language statistics, which mainly track languages with measurable presence among websites covered by their methodology (W3Techs – Web Technology Surveys, 2024). We therefore use this observation only as indirect evidence of limited web visibility, not as a comprehensive estimate of online Chukchi content.

From a corpus-linguistic perspective, Chukchi is also important because it is an incorporating language with rich morphology. This typological profile is underrepresented in many NLP resources and creates challenges for both segmentation and alignment. A single Chukchi word may correspond to a multi-word Russian expression, while Russian function words, inflectional morphology, or syntactic relations may be expressed differently in Chukchi morphology. Consequently, word-level correspondence is often misleading, and even sentence-level alignment may obscure substantial differences in internal structure. For

Table 1: Sources of the parallel corpus (post-filtering totals).

Source	Pairs	Type / genre
Charles Weinstein dictionary (RU–FR–EN–Chukchi)	48,383	Dictionary
RAS Russian–Chukchi dictionary	13,501	Dictionary
Krayniy Sever newspaper	3,881	Regional media
The Little Prince	1,491	Fiction
Folktales	1,616	Folklore
UD Chukchi treebank	644	Annotated sentences
Russian–Chukchi phrasebook	611	Everyday expressions
Tatoeba	275	Short curated sentences
Proverbs and sayings	106	Folklore / sayings
Total after filtering	70,508	

this reason, our resource is organized around translation units and source metadata rather than assuming one-to-one word or sentence correspondence across the two languages.

This paper focuses on corpus creation rather than downstream model training. We contribute: (i) a Chukchi–Russian parallel corpus of 70,508 aligned segments; (ii) a monolingual Chukchi corpus of 15,610 segments; (iii) a documented pipeline for extraction, normalization, alignment, and quality control; and (iv) a release package with metadata and scripts. We use *corpus* in a broad resource-construction sense: the data are segmented and aligned, but not morphologically or syntactically annotated, and the parallel part mixes dictionary entries, short phrases, and full sentences.

2 Related Work

Publicly available resources for low-resource languages often originate from (i) community translation platforms, (ii) linguistic treebanks, and (iii) digitized books and regional media. Tatoeba provides manually curated multilingual sentence pairs (Tiedemann, 2020). Universal Dependencies (UD) treebanks supply high-quality annotated sentences, occasionally with parallel or comparable alignments (Tyers and Mishchenkova, 2020). Large web-crawled collections such as OPUS (Tiedemann, 2012) have expanded coverage for many languages, but minority-language subsets are frequently small or noisy, requiring careful filtering.

For alignment and corpus cleaning, embedding-based semantic similarity models have become standard tools. Sentence-embedding models such as LaBSE (Feng et al., 2022) allow robust matching and filtering across languages, and practical alignment pipelines (e.g., Lingtrain Aligner) support document-level alignment through embedding similarity (Averkij, 2024). Our work follows this direction, adapting alignment and filtering to the specific orthographic and data-quality challenges of Chukchi corpora.

3 Data Sources

We compiled two datasets: (1) a parallel Chukchi–Russian corpus, and (2) a monolingual Chukchi corpus. The datasets are available on Hugging Face.[1]

3.1 Parallel sources

Source overview

The parallel corpus combines lexicographic, curated, and text-based sources. Most pairs come from two dictionaries: the RAS Russian–Chukchi dictionary (Russian Academy of Sciences, 2025) and Charles

[1] Parallel corpus: `https://huggingface.co/datasets/HSE-Chukchi-NLP/russian-chukchi-parallel-corpora`; monolingual corpus: `https://huggingface.co/datasets/HSE-Chukchi-NLP/chukchi-mono-corpora`

Weinstein's Russian–French–English–Chukchi dictionary (Weinstein, 2025). Smaller sentence- and phrase-level subsets come from Tatoeba (Tiedemann, 2020), the UD Chukchi HSE treebank (Tyers and Mishchenkova, 2020), and the Russian–Chukchi phrasebook (Keulkut, 1958). For UD, we include only Chukchi sentences and available Russian translations as aligned text segments; the original morphological and syntactic annotation remains in the UD treebank.

We also collected bilingual materials from *Krayniy Sever* (Krayniy Sever, 2024), including the *Murgin nutənut* section, and from fiction and folklore. Folklore is included because it is one of the few sources of extended connected Chukchi text, but it is not treated as representative of contemporary everyday usage; it is marked as a separate source type in the metadata. *The Little Prince* contributes 1,491 aligned sentence pairs (Tokyo University of Foreign Studies, 2009).

3.2 Monolingual sources

The monolingual Chukchi corpus was constructed primarily from:

- **Untranslated materials** from *Krayniy Sever* (Krayniy Sever, 2024) that appeared only in Chukchi;
- **Unaligned materials** where a Russian counterpart existed but could not be aligned reliably at segment level;
- additional Chukchi text resources collected for low-resource speech and text processing in prior work (Safonova et al., 2022).

In total, we collected 15,610 monolingual Chukchi segments.

4 Corpus Construction Pipeline

4.1 Implementation, deduplication, and access

The pipeline was implemented by the authors as Python scripts for source-specific extraction, normalization, duplicate checking, alignment, and export to UTF-8 CSV / plain-text formats. Deduplication was performed after normalization: exact repeated Chukchi–Russian pairs were removed when they were clearly mechanical duplicates. We did not globally collapse repeated Chukchi or Russian strings, since identical forms may correspond to different dictionary senses, formulaic folklore expressions, or genre-specific translations.

The datasets are distributed on Hugging Face as files rather than through a dedicated corpus manager. The parallel corpus is provided as UTF-8 CSV with text and metadata fields; the monolingual corpus is provided as Chukchi segments with source metadata.

4.2 Extraction and segmentation

We used three extraction pathways depending on source structure:

- **Structured parsing of HTML trees** for well-marked bilingual pages;
- **Manual segmentation and alignment** for short and clearly corresponding texts (e.g., children's folklore);
- **Automatic mining and alignment** for longer documents, supported by sentence embeddings and dynamic programming alignment.

To separate languages into columns during early aggregation, we used GlotLID (a fastText-based language identification model) (Kargaran et al., 2023). While effective for coarse grouping, it required human supervision in edge cases such as Chukchi segments containing Russian proper names.

4.3 Chukchi–Russian sentence embedding model (LaBSE) for alignment

To support mining, filtering, and alignment, we fine-tuned the Language-agnostic BERT Sentence Embedding (LaBSE) model (Feng et al., 2022) on naturally occurring Chukchi–Russian aligned segments, yielding an encoder adapted for Chukchi–Russian. We used *Multiple Negatives Ranking Loss* to pull true pairs together in embedding space and push non-pairs apart.

For document alignment, we used Lingtrain Aligner (Averkij, 2024), which compares sentence embeddings across comparable documents and applies dynamic programming to compute an alignment path. The fine-tuned encoder improved stability for low-resource Chukchi materials and was used to:

- mine candidate parallel segments from comparable documents;
- filter noisy segment pairs prior to inclusion;
- detect segments that fail alignment and should be treated as monolingual.

4.4 Human validation and expert review

Selected difficult segments were checked by native speakers and language experts, especially folklore, OCR-derived materials, and low-scoring alignment candidates. The review was not exhaustive, but it helped identify cases requiring re-segmentation, correction, or exclusion. Most corrections concerned folklore, where versions are often non-literal and not structurally parallel.

5 Text Preprocessing

5.1 Orthographic normalization and cleanup

Since a large portion of Chukchi text originates from OCR or heterogeneous web sources, normalization was essential. The most critical steps were:

- unifying multiple apostrophe-like characters (ʼ, ’, ', ‘, etc.) into a single chosen form;
- normalizing letter-plus-apostrophe sequences to canonical Chukchi characters (e.g., к’/К’ → ҝ/Ҝ; н’/Н’ → ң/Ң; л’/Л’ → ӆ/Ӆ);
- replacing all dash/hyphen variants with a single standard hyphen (-);
- removing duplicated spaces and obvious artifacts (HTML tags, email addresses, phone numbers);
- fixing OCR-induced Latin/Cyrillic confusions by mapping Latin letters to their Cyrillic counterparts where appropriate.

All processing was implemented with regular expressions (Python `re`) and auxiliary regex tooling.

5.2 Segmentation and tokenization conventions

We use *segment* rather than *sentence* because units differ by source: Tatoeba and UD preserve original sentence boundaries, dictionaries and phrasebooks contain words or phrases, and longer texts are segmented by markup or, when needed, conservatively by punctuation and paragraph structure. Token counts are descriptive only: we use whitespace- and punctuation-based splitting after normalization, not a linguistically motivated Chukchi tokenizer. UD-derived segments keep source identifiers, so they can be compared with the original UD annotation.

5.3 Semantic diagnostics with LaBSE

We used the fine-tuned LaBSE encoder as a diagnostic tool for finding anomalous alignments. Cosine similarity was not treated as an independent evaluation metric or a hard threshold, since many valid dictionary-style pairs are short and context-poor. Low-scoring pairs were inspected manually; fewer than 0.01% of aligned segments were removed at this stage. Filtering was applied to aligned segment pairs, not to whole documents. For continuous texts, document identifiers are preserved, while dictionary and phrasebook entries are treated separately because textual completeness is not directly applicable to lexicographic resources.

LaBSE was fine-tuned for one epoch with batch size 8, learning rate $2{\cdot}10^{-5}$, warmup cosine scheduling, and Multiple Negatives Ranking Loss. The full list of normalization rules and scripts is provided in the release repository.

(1) Russian: Старик; Chukchi: Ыӈпыначга иничгыту нинэӆгықин тъэвыӆьатгыргын ынин ранмыӈъёнэн; cosine similarity score: 0.3322; source: Charles Weinstein

After filtering, the parallel dataset contains 70,508 Chukchi–Russian aligned segments. Figure 2 shows length distributions for both sides.

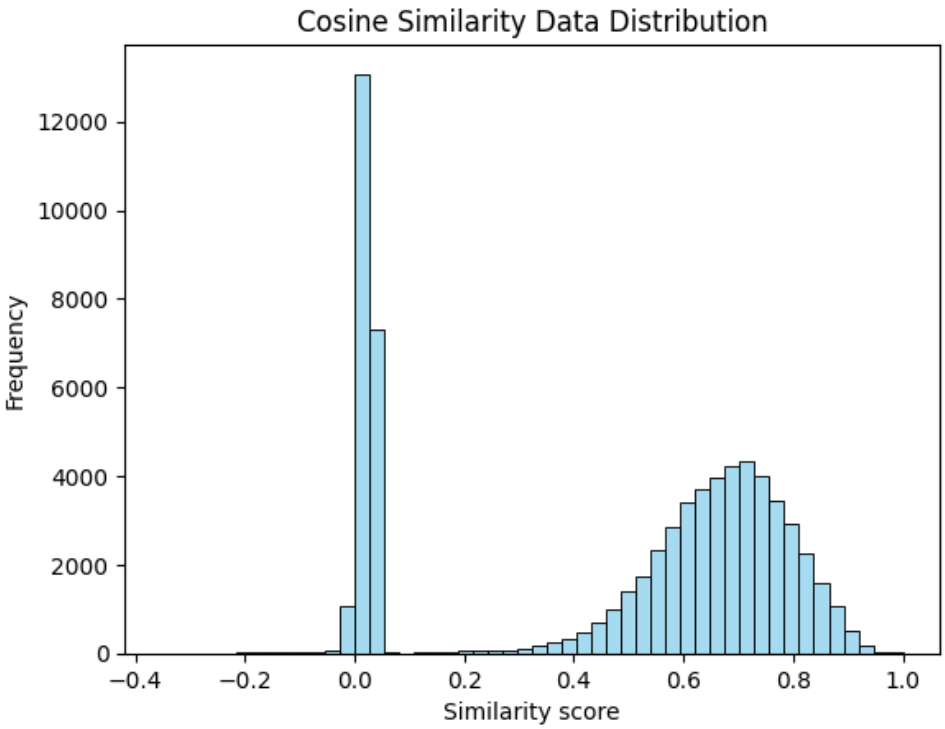

Figure 1: Distribution of cosine similarity scores across Chukchi–Russian aligned segments.

6 Corpus Statistics and Format

6.1 Parallel corpus

The final parallel corpus contains 70,508 aligned segments (translation units). It covers a mixture of domains, including phrasebook expressions, folklore, journalistic texts, cultural descriptions, and fiction. The distribution of sources across the corpus is shown in Figure 3.

The dataset is stored as a UTF-8-encoded CSV file with multiple columns. In addition to the two primary language columns (Chukchi and Russian), the file contains several metadata columns. We preserve original sentence boundaries whenever they are explicitly marked in the source; otherwise, we apply conservative segmentation aligned to punctuation and document structure.

For every parallel fragment, the source of the text is explicitly specified. Where available, we also provide more information about the source, such as article or story identifiers, publication details, or internal numbering (e.g., article number rather than only the general source title). In some cases we include an indication of whether the translation is literal or literary.

6.2 Monolingual corpus

The monolingual corpus contains 15,610 Chukchi segments. It is stored as one segment per line in UTF-8. Source information is also provided for the monolingual subset. These data are intended for linguistic analysis, vocabulary coverage estimation, and as additional Chukchi text for tasks that do not require parallel alignment.

7 Limitations

The corpus has several limitations. First, it follows a mixed access model. Some sources can be redistributed as full text, whereas copyrighted materials such as parts of regional media and *The Little Prince*

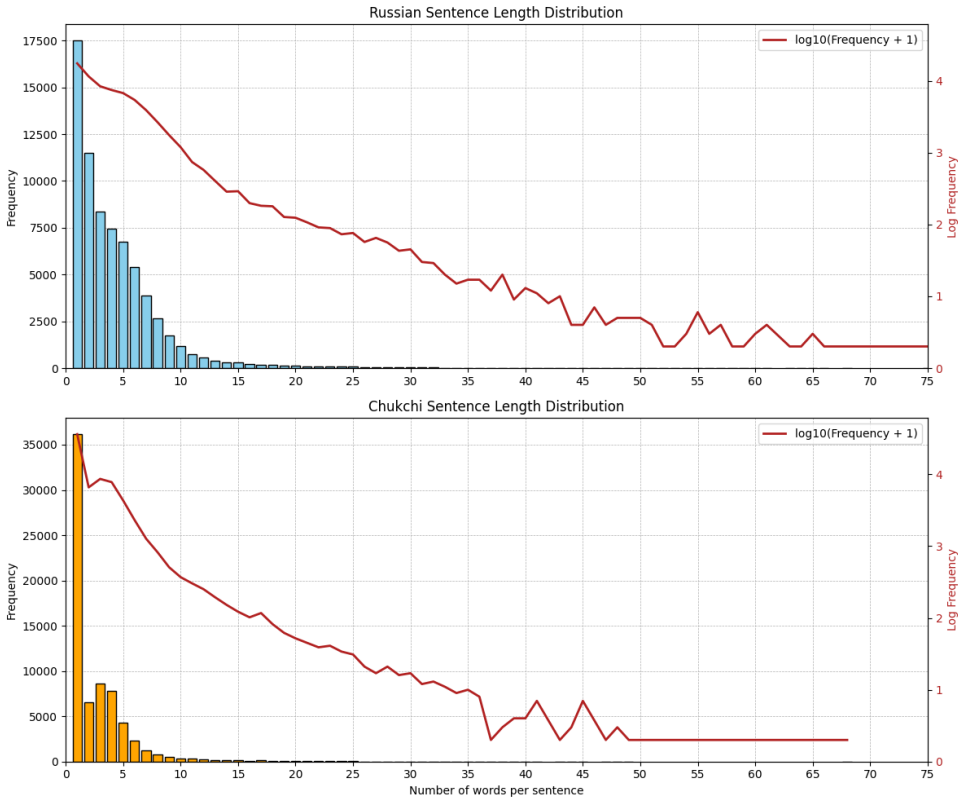

Figure 2: Length distributions for the Russian and Chukchi sides of the parallel corpus.

are represented through metadata, source identifiers, and reconstruction scripts. We preserve provenance and maintain a takedown policy for rights holders.

Second, the resource is not balanced. It is large for Chukchi but small compared to corpora for high-resource languages, and it is skewed toward dictionaries, folklore, phrasebook expressions, and regional media. Most aligned pairs are lexicographic word or short-phrase glosses, roughly 53,000 of 70,508, rather than full sentences. Users should therefore treat the dataset as a mixed-granularity parallel resource and filter it by source type or segment length for sentence-level experiments.

Third, the current release does not include full linguistic annotation such as lemmatization, morphological analysis, POS tags, syntactic dependencies, or glossing. Residual noise is also possible despite normalization and manual checks, especially in OCR-derived texts, folklore variants, and automatically aligned segments. Finally, we did not conduct downstream evaluation such as MT training with BLEU or chrF; this is left for future work because standard sentence-level MT evaluation would not adequately reflect the lexicographic structure of much of the dataset.

8 Conclusions and Future Work

We introduced a curated Chukchi–Russian parallel corpus of 70,508 aligned segments and a monolingual Chukchi corpus of 15,610 segments, together with a pipeline for extraction, normalization, alignment, and quality control. The resources support Chukchi computational research, linguistic analysis, and documentation.

Future work includes expanding contemporary and conversational materials, improving source-aware normalization, conducting broader expert review, adding linguistic annotation, documenting duplicates and source completeness, and exploring additional Chukchi–English and Chukchi–French pairs from the Weinstein dictionary.

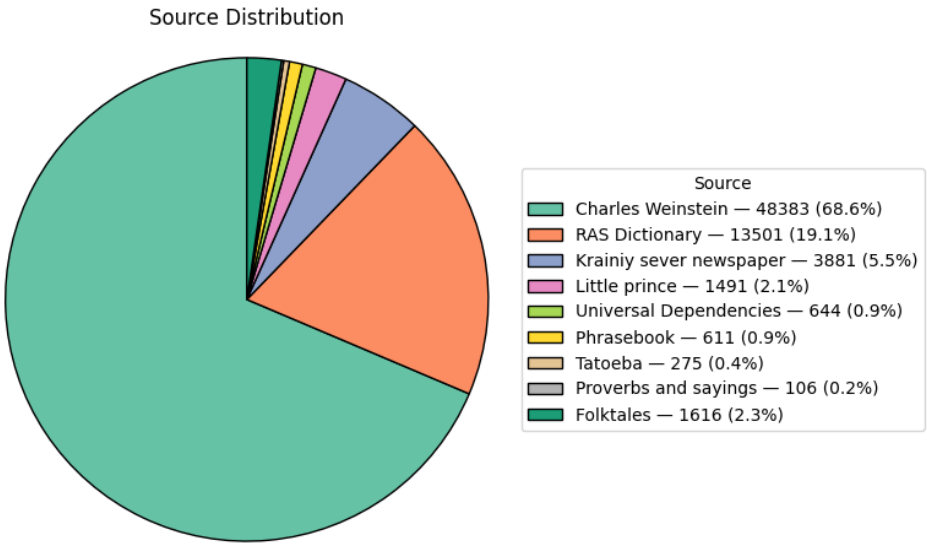

Figure 3: Source distribution for the parallel corpus.

Acknowledgements

We thank Alexander Antonov for consultation and Victoria Kavry for assistance in organizing sentence verification.

References

Averkij. 2024. Lingtrain aligner – ml powered library for the accurate texts alignment. GitHub, `https://github.com/averkij/lingtrain-aligner`.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 878–891, Dublin, Ireland. Association for Computational Linguistics.

Institute of Linguistics, Russian Academy of Sciences. 2024. Minority languages of russia: Chukchi. `https://minlang.iling-ran.ru/en/node/115`. Accessed: 2024-12-16.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. arXiv. arXiv:2310.16248.

V.G. Keulkut. 1958. *Russian-Chukchi Phrasebook*. Magadan Book Publishing House, Magadan.

Krayniy Sever. 2024. Krayniy sever. `https://www.ks87.ru/`. Accessed: 2024-12-16.

Russian Academy of Sciences. 2025. Russian–chukchi dictionary. `https://chukdict.com/`. Accessed: 2025-05-20.

Alla Safonova, Tatiana Yudina, Erkin Nadimanov, and Cory Davenport. 2022. Automatic speech recognition of low-resource languages based on chukchi. *arXiv preprint*. arXiv:2210.05726.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. // *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, P 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt. // *Proceedings of the Fifth Conference on Machine Translation*, P 1174–1182. Association for Computational Linguistics.

Tokyo University of Foreign Studies. 2009. *Маленький принц. Параллельный чукотско-русский текст*. Tokyo University of Foreign Studies, Tokyo, Japan. Retyping: 2024, 152 p. Translator: Alla P. K'erginto (from the Russian translation of Nora Gal'). ISBN 978-4-86337-163-7.

F. M. Tyers and K. Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. // *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, P 195–204.

W3Techs – Web Technology Surveys. 2024. Usage statistics of content languages for websites. `https://w3techs.com/`. Accessed: 2024-12-16.

Charles Weinstein. 2025. Russian–french–english–chukchi dictionary. `http://charles.weinstein.free.fr/chukches/index_ru.html`. Accessed: 2025-05-25.