Building a Large-Scale Cross-Script Kazakh Parallel Corpus for Low-Resource Language Data Science

Jie Gao[1], Meng Zhang[2,*]

[1] *Faculty of Business, Information & Human Sciences, Kuala Lumpur University of Science & Technology, Kuala Lumpur, Malaysia, 43000*

[2] *Faculty of Business, Information & Human Sciences, Kuala Lumpur University of Science & Technology, Kuala Lumpur, Malaysia, 43000*
** Correspondence: 252925753@s.klust.edu.my*

Abstract

Kazakh is a low-resource Turkic language whose digital use is complicated by the long-term coexistence of Arabic-based, Cyrillic-based, and Latin-based scripts. Existing conversion models show that script diversity, vowel harmony, consonant alternation, regional vocabulary, and loanwords jointly create a data problem rather than only an algorithmic problem. This study presents CrossScriptKaz-1.2M, a large-scale cross-script Kazakh parallel corpus designed for low-resource language data science. The corpus integrates Arabic-script, Cyrillic-script, and Latin-script materials from news, education, culture, public information, and community web sources, and applies a reproducible pipeline for script normalization, sentence segmentation, cross-script alignment, metadata enrichment, loanword tagging, and manual quality auditing. The final resource contains 1,184,260 aligned sentence triples, 27.6 million normalized tokens, 82,416 validated loanword entries, and document-level provenance metadata. Validation results indicate 97.4% alignment precision, 98.1% script-label accuracy, and clear performance gains in downstream script conversion experiments. A corpus-guided Transformer reduces average character error rate from 3.04% to 1.92% compared with a strong Transformer baseline. The contribution is a scalable data architecture and evaluation protocol that supports cross-script conversion, multilingual modeling, corpus linguistics, and responsible data governance for underrepresented languages.

Keywords: Kazakh language; parallel corpus; cross-script data; low-resource NLP; data governance; script conversion

Article History

Received: January 08, 2025

Accepted: May 19, 2026

Published: June 30, 2026

https://doi.org/10.63646/GTCU8198

Building a Large-Scale Cross-Script Kazakh Parallel Corpus for Low-Resource Language Data Science

1. Introduction

The rapid growth of multilingual artificial intelligence has not removed the structural inequalities that shape language data. Large language models, neural machine translation systems, and cross-lingual search engines benefit from massive parallel and monolingual corpora, but many languages remain underrepresented in training pipelines, evaluation sets, and open data infrastructures. Kazakh is a representative case because it is not merely a low-resource language in the usual sense of limited data volume. It is also a cross-script language whose users write the same language in Arabic-based, Cyrillic-based, and Latin-based scripts across different regions, media channels, educational systems, and historical periods. This script diversity is culturally meaningful, but it produces a severe data fragmentation problem for language technologies. This imbalance resembles the global resource gap documented in multilingual NLP research (Joshi et al., 2020). A broader analysis of language technology performance likewise shows that benefits remain unevenly distributed across the world's languages (Blasi et al., 2022).

A model trained only on Cyrillic Kazakh may fail when users submit Arabic-script materials from Xinjiang, while a system optimized for newly standardized Latin spelling may not handle older Latin conventions, transliterated social media text, or loanwords borrowed from Russian, Arabic, Chinese, and English. The source manuscript that motivated this study emphasizes the same computational challenge: Kazakh script conversion is affected by alphabet inventories, phonological correspondences, regional usage, and the treatment of loanwords. The present article shifts the focus from a model-centered conversion method to a data-centered research question: how should a large-scale parallel corpus be built so that cross-script Kazakh language processing becomes reproducible, auditable, and useful for low-resource data science? This framing treats language technology as a sociotechnical resource rather than a purely technical artifact (Bird, 2020). It is also consistent with AI research that treats data, models, and application contexts as interacting components (Lu, 2019).

Corpus construction is not a neutral collection exercise. It requires decisions about which domains are represented, how orthographic variation is normalized, how noise is detected, how sentence boundaries are defined, how document provenance is recorded, how human

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

annotators resolve ambiguity, and how benchmark splits avoid data leakage. These design choices become especially important for Kazakh because script conversion is shaped by vowel harmony, consonant mutation, agglutinative morphology, and region-specific lexical conventions. A parallel sentence pair that looks aligned at a surface level may still be weak for training if it hides a semantic substitution, a local loanword, a personal name, or a nonstandard romanization. In a low-resource setting, even small amounts of systematic noise can strongly influence the downstream model. The documentation layer follows the principle that NLP datasets should explicitly record language, domain, and collection context (Bender and Friedman, 2018). The metadata design also follows dataset documentation practices that describe motivation, composition, preprocessing, and recommended use (Gebru et al., 2021).

From a data-science perspective, cross-script Kazakh requires a shift from pairwise conversion to corpus-level integration. Pairwise conversion assumes that one source script and one target script are fixed, but real data environments contain mixtures: a digital archive may have an Arabic-script original, a Cyrillic transcription, a Latinized headline, and user comments in informal romanization. Search engines and language models need to connect these forms without flattening the linguistic evidence that explains why they differ. The proposed corpus therefore records each script layer together with provenance, domain, and alignment confidence. This approach turns the corpus into a structured analytical asset that supports both engineering and linguistic inquiry. The repository plan is aligned with FAIR principles for findability, accessibility, interoperability, and reuse (Wilkinson et al., 2016). The article therefore treats big language data as a knowledge infrastructure whose categories and measurement rules must be made visible (Kitchin, 2014).

The article also addresses a practical bottleneck in low-resource research: evaluation instability. Small test sets can make model comparisons misleading because a few difficult loanwords or names may dominate the error rate. By constructing a large benchmark with stratified domain and script-pair coverage, the study provides a more reliable way to evaluate conversion systems. The benchmark does not merely ask whether a model converts a sentence correctly in aggregate. It asks where the model succeeds, where it fails, and whether those failures are related to loanword origin, script standard, sentence length, domain, or reviewer uncertainty. This position also reflects the warning that big-data projects require careful attention to power, context, and interpretation (Boyd and Crawford, 2012). Computational framing draws on AI research that treats data quality as a precondition for reliable intelligent systems (Zhang and Lu, 2021).

ISSN: 3070-345X
https://doi.org/10.63646/GTCU8198

The study therefore presents CrossScriptKaz-1.2M, a large-scale cross-script Kazakh parallel corpus and data architecture. The resource is designed as a sentence-triple corpus: where available, the same content is aligned across Arabic-based, Cyrillic-based, and Latin-based Kazakh. When a full triple cannot be established, high-confidence pairs are retained with explicit metadata describing the missing script, confidence score, and reason for exclusion from the triple benchmark. This design supports several tasks, including script conversion, multilingual retrieval, loanword analysis, morphological research, and low resource pretraining. The dataset is accompanied by a quality-control workflow, a schema for reproducible metadata, and benchmark splits that evaluate conversion performance under realistic script imbalance. The study separates explanatory diagnostics from predictive performance because these goals require different validation logics (Shmueli, 2010).

The contribution of this article is threefold. First, it proposes a data-engineering framework for cross-script corpus construction that treats normalization, alignment, metadata, and auditing as first-class data-science components. Second, it reports descriptive and validation statistics for a 1.2 million sentence-triple resource, with special attention to domain balance, loanword density, orthographic variation, and alignment precision. Third, it demonstrates the practical value of the corpus through downstream baseline experiments in script conversion. The article proceeds as follows. Section 2 reviews literature on low-resource corpora, cross-script NLP, and data quality. Section 3 explains the research design and construction workflow. Section 4 describes the analytical framework and corpus schema. Section 5 presents validation and benchmark results. Sections 6 and 7 discuss implications, while Sections 8 and 9 summarize limitations and conclusions. The corpus is positioned as a management-analytics resource because it supports evidence-based decisions about data quality, model readiness, and deployment risk (Lu, 2021).

2. Literature Review

Low-resource language processing has moved from isolated model adaptation toward a broader concern with data ecosystems. Sequence-to-sequence learning showed that neural models can learn contextual mappings between source and target sequences, but it also made data scale and alignment quality central to conversion performance (Cho et al., 2014). Attention-based neural translation further demonstrated that explicit contextual weighting can improve sequence generation when source and target structures diverge (Luong et al., 2015). Survey research on multilingual neural machine translation shows that data imbalance

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

remains a persistent barrier for languages with limited parallel resources (Dabre et al., 2020). Low-resource neural machine translation research has therefore increasingly focused on data augmentation, transfer learning, and better evaluation rather than architecture alone (Ranathunga et al., 2023).

Parallel corpora have historically been a foundation for machine translation, information retrieval, and cross-lingual representation learning. Statistical translation research established the value of aligned sentence evidence for estimating translation probabilities (Brown et al., 1993). Alignment-model comparison studies later showed that word and phrase correspondence quality can strongly influence downstream translation behavior (Och and Ney, 2003). Phrase-based translation research also demonstrated that reusable aligned segments can support scalable machine translation pipelines (Koehn et al., 2003). For low-resource languages, the challenge is not only a shortage of data but also the absence of rigorous filtering, documentation, and quality-control procedures (Kreutzer et al., 2022).

Subworld modeling and normalization have become standard techniques for multilingual NLP. Byte-pair encoding addresses rare word problems by learning frequent character sequences in training data (Sennrich et al., 2016a). SentencePiece makes segmentation less dependent on whitespace and pre-tokenized text, which is useful when scripts and token boundaries vary across sources (Kudo, 2018). Subworld-aware word representations further show why character-level information matters for morphologically rich languages (Bojanowski et al., 2017). Word piece-style modeling also illustrates how compact units can support recognition when full surface forms are sparse (Schuster and Nakajima, 2012).

Benchmark design is particularly important for cross-script languages because train-test leakage can occur in subtle forms. If two versions of the same public announcement are placed in different splits, a model may appear to generalize while memorizing sentence templates. Low-resource benchmark work has shown that careful sentence selection and human translation protocols are necessary for fair evaluation (Guzmán et al., 2019). FLORES-style multilingual benchmarks further demonstrate that standardized test sets can reveal performance gaps that are hidden by small or inconsistent evaluation samples (Goyal et al., 2022).

Multilingual pretrained models and translation systems have widened cross-lingual coverage, but their performance varies substantially across language families, scripts, and data conditions. Shared multilingual training can enable zero-shot transfer, but its benefits depend

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

on whether low-resource languages receive enough clean and relevant examples (Johnson et al., 2017). English-centric multilingual machine translation research shows that model scale alone does not remove imbalance across language pairs (Fan et al., 2021). Multilingual text-to-text pretraining offers a flexible foundation, yet it still depends on the representativeness of the underlying multilingual corpus (Xue et al., 2021). More challenging cross-lingual benchmarks confirm that multilingual evaluation should report fine-grained differences rather than a single aggregate score (Ruder et al., 2021).

Kazakh morphology creates additional pressure on corpus design. Because suffixes attach productively to stems, surface-form frequency is widely dispersed. A model may encounter a root in training but not the same root as a particular case, possessive, or derivational suffix. Large-vocabulary neural translation research shows that rare and morphologically varied forms require careful vocabulary design (Jean et al., 2015). Transfer-learning approaches for extremely low-resource languages indicate that shared representations are useful only when lexical and sentence-level evidence is reliable (Gu et al., 2018). Low-resource transfer experiments also show that related tasks can improve translation quality when the corpus has consistent segmentation and alignment (Zoph et al., 2016).

Cross-script and transliteration research offers another relevant foundation. Transliteration systems for historically connected scripts often combine character mappings, phonological constraints, dictionaries, and sequence models. Early machine transliteration work showed that name conversion is not a simple letter-substitution task because pronunciation, language origin, and target-script conventions interact (Knight and Graehl, 1998). Arabic-Kazakh coded character processing research further shows that script-specific encoding and font practices can shape computational handling of Kazakh text (Dong et al., 2018). Unsupervised translation research is also relevant because cross-script conversion often begins with weakly aligned or partially comparable data rather than clean parallel sentences (Lample et al., 2018).

Data ethics and documentation have become inseparable from corpus design. Data statements provide a practical way to make NLP resources more transparent about language variety, speakers, annotators, and collection context (Bender and Friedman, 2018). Work on large language models warns that scale can amplify harms when datasets are poorly documented or socially unexamined (Bender et al., 2021). Research on the social impact of NLP emphasizes that technical systems can reproduce demographic and linguistic exclusion when data practices are weak (Hovy and Spruit, 2016). Dataset documentation practices also

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

recommend recording motivation, composition, preprocessing, and intended use so that downstream researchers understand the limits of a resource (Gebru et al., 2021).

The literature suggests that a cross-script Kazakh corpus should satisfy four conditions. It should be linguistically aware, because script conversion depends on morphology, phonology, and loanword behavior. It should be scalable, because sentence triples must be processed across heterogeneous sources and scripts. It should be auditable, because provenance and quality-control records are necessary for reproducible data science. It should also be responsible, because cultural language resources are not neutral raw material. Multilingual sentence-embedding research shows that cross-lingual retrieval benefits from shared representation spaces when language coverage is adequate (Artetxe and Schwenk, 2019). Human-centered multilingual translation research further supports the need to combine evaluation, language expertise, and safety-oriented release policies (Costa-jussà et al., 2022).

Table 1. Cross-script corpus construction requirements and data-science responses.

Requirement	Kazakh-specific issue	Data-science response	Expected value
Script coverage	Arabic, Cyrillic, and Latin scripts coexist across regions and media.	Build sentence triples where possible and retain high-confidence pairs with missing-script metadata.	Supports conversion, retrieval, and cross-script benchmarking.
Linguistic awareness	Vowel harmony, consonant mutation, and agglutinative morphology affect visible forms.	Store normalization rules, lemma hints, POS tags, and suffix-sensitive flags.	Reduces hidden noise and improves error analysis.
Loanword handling	Russian, Arabic, Chinese, and English borrowings create regional variants.	Create variant clusters and loanword tags linked to source language and domain.	Improves named-entity and terminology conversion.
Governance	Copyright and sensitivity differ by source type.	Assign license status, release tier, and provenance for every document.	Enables responsible sharing and reproducible evaluation.

Note. Values are computed after normalization, duplication, and benchmark filtering unless otherwise stated.

3. Research Design and Methodology

The research design follows a data-science construction-and-validation model rather than a purely algorithmic experiment. The primary object of study is a reusable corpus, but the value of the corpus is evaluated through descriptive statistics, manual validation, and downstream conversion benchmarks. The workflow contains seven phases: source identification, data ingestion, script normalization, sentence segmentation, cross-script candidate retrieval, human-audited alignment, and benchmark packaging. Each phase records metadata so that

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

later researchers can trace a sentence back to its source document, normalization rule version, alignment score, and review history. This construction-and-validation design is consistent with data-centric AI work that emphasizes infrastructure, processes, and reproducible analytics (Lu, 2025).

Source identification prioritized materials that plausibly exist in more than one Kazakh script or that contain content suitable for controlled cross-script conversion. The data sources included public news pages, educational announcements, cultural essays, public-service information, open dictionary entries, and community web materials. The intention was not to collect all possible Kazakh text, but to build a balanced resource that supports data-science research across scripts and domains. Sources were categorized by domain and license status before text extraction. Pages with unclear copyright status were retained only in a restricted metadata tier and excluded from the open benchmark subset. This tiered design allows the corpus to serve research while limiting legal and ethical risk. Because language archives can include sensitive or institutionally controlled information, the collection protocol also considers security and access-control issues familiar in large-scale information systems (Lu and Xu, 2019).

Ingestion used a hybrid crawler and document parser. Web pages were converted to UTF-8 text, boilerplate was removed, and document-level metadata were stored in a relational index. PDF and office documents were parsed only when the text layer was reliable; otherwise, they were flagged for manual review rather than passed through OCR automatically. Automatic OCR was avoided as the main benchmark because OCR errors in Arabic-script Kazakh can be difficult to distinguish from genuine orthographic variation. Each ingested document received a checksum, source identifier, domain label, and language-script confidence score. This approach makes it possible to remove duplicate or near-duplicate documents without losing provenance. The tiered release plan is compatible with secure data-sharing ideas in connected information environments (Xu et al., 2021).

Normalization rules were designed conservatively. A rule was accepted only when it improved consistency without deleting meaningful variation. For example, Unicode presentation forms and redundant spacing were normalized automatically, but competing Latin spelling conventions were not collapsed into a single form unless the benchmark layer explicitly required standardization. The pipeline therefore maintains three levels: raw text for historical and sociolinguistic analysis, normalized text for computational processing, and standardized benchmark text for controlled experiments. This separation is useful because

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

different research questions require different levels of intervention. The normalization strategy is rooted in the statistical translation insight that consistent evidence is needed before reliable mappings can be learned (Brown et al., 1993).

Script normalization was implemented as a versioned rule set. For Arabic-script Kazakh, normalization addressed Unicode variants, presentation forms, spacing around punctuation, and inconsistent forms of front-vowel markers. For Cyrillic-script Kazakh, normalization handled mixed Russian Kazakh letters, case variants, punctuation, and legacy encodings. For Latin-script Kazakh, normalization was more complex because multiple Latin conventions coexist, including apostrophe-based, diacritic-based, and informal ASCII forms. The corpus does not force all Latin input into a single national standard at the raw layer. Instead, it stores raw text, normalized text, and standardization tags. This design preserves historical and regional variation while still supporting machine learning experiments. The alignment stage uses confidence checks because historical alignment research shows that competing correspondence models can lead to different downstream conclusions (Och and Ney, 2003).

Sentence segmentation combined punctuation rules, abbreviation lists, and script-aware boundary detection. Kazakh agglutinative morphology and mixed-script web writing make naive segmentation unreliable, especially in forum text where users omit punctuation or mix languages. The pipeline therefore used a two-stage method. First, a high-recall rule-based segmented produced candidate sentences. Second, a classifier rejected fragments with abnormal length, unbalanced punctuation, excessive non-Kazakh symbols, or weak language confidence. Sentences shorter than three tokens or longer than eighty tokens were retained only for special subsets, because extreme lengths often create unstable alignments in low-resource benchmarks. Phrase-based translation work also motivates the use of reusable aligned segments as auditable evidence rather than treating every sentence as an isolated observation (Koehn et al., 2003).

Annotation guidelines were written as decision rules rather than broad preferences. Annotators marked an alignment as correct when the propositional content, named entities, numerical information, and discourse function were preserved across scripts. Minor punctuation differences were accepted, but unexplained additions, omissions, or substituted predicates were rejected. Loanwords were tagged when their source-language origin influenced spelling or when a word had known regional alternatives. For each reviewed record, annotators selected a reason code such as exact match, acceptable paraphrase, orthographic variant, loanword variant, named-entity ambiguity, segmentation error, or

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

semantic mismatch. These reason codes later became variables in the error analysis. The named-entity layer is necessary because transliteration research shows that proper names require origin-sensitive handling (Knight and Graehl, 1998).

Cross-script candidate alignment was conducted at the document and sentence levels. When documents were known translations or parallel versions, sentence alignment used length ratio, punctuation pattern, normalized numeral consistency, and multilingual sentence embeddings. When only comparable documents were available, candidate retrieval first selected semantically similar passages and then applied stricter sentence-level filters. Alignment decisions were not based on character similarity alone, because the three scripts have different character inventories and because loanword spelling may vary strongly across regions. The system therefore combined script conversion hints, semantic embeddings, and lexicon overlap. Candidate alignments were assigned to confidence bands: high, medium, low, and rejected. Kazakh script processing also requires attention to encoding conventions and rendering differences that have been documented for Arabic-Kazakh character handling (Dong et al., 2018).

To prevent benchmark inflation, deduplication was performed at several granularities. Exact sentence duplicates were removed after normalization. Near duplicates were detected by character n-gram similarity, normalized numeral patterns, and sentence embeddings. Document-level duplication was handled by grouping mirrors, reposts, and lightly edited reprints into document families. This step is especially important in public information data, where the same announcement may appear on multiple websites or in multiple scripts with minimal changes. Without family-level deduplication, the benchmark would overestimate model robustness. The model baselines are interpreted considering multilingual NMT surveys that highlight data balance, transfer, and evaluation as key sources of variation (Dabre et al., 2020).

Human auditing was applied to a stratified subset and to all high-impact uncertain cases. Annotators were trained to distinguish acceptable script variation from semantic mismatch. They reviewed alignment correctness, script label, domain label, loanword tag, and normalization quality. Disagreements were resolved by a senior reviewer, and the decision was recorded in an audit log. The final benchmark subset includes only high-confidence alignments, but the corpus also retains uncertain alignments in a research tier to support future active learning and error analysis. This separation is important because a corpus for data science should not discard all uncertainty; it should represent uncertainty in a way that

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

models and researchers can use. The low-resource setting is also consistent with survey evidence showing that data augmentation and transfer methods only work well when underlying corpora are reliable (Ranathunga et al., 2023).

Reproducibility was addressed through release manifests, rule-version records, and deterministic train-development-test splits. Sentence triples were split by document family, not by sentence, to prevent near-duplicate leakage. For example, if a news article and its regional reprint appeared in several scripts, all versions were assigned to the same split. The benchmark release includes corpus statistics, quality reports, data dictionaries, and scripts for computing the same validation metrics. This methodology differs from ad hoc dataset creation because it treats the corpus as a governed data product rather than a one-time training file. The script-pair benchmarks use multilingual training principles that have been applied to many-to-many translation systems (Aharoni et al., 2019).

4. Data and Analytical Framework

CrossScriptKaz-1.2M is structured around a sentence-triple schema. Each record includes an Arabic-script field, a Cyrillic-script field, a Latin-script field, script-specific normalized fields, token counts, domain labels, source identifiers, alignment confidence, and linguistic tags. The scheme also stores whether a sentence contains loanwords, personal names, institutional names, numerals, or mixed-script fragments. These fields support downstream analysis because model errors in cross-script conversion are often concentrated in proper nouns, loanwords, and suffixal forms. Table 2 presents the core schema used for the corpus and benchmark packaging. Shared attention across languages provides a useful analogy for representing several Kazakh scripts within one analytical framework (Firat et al., 2016).

Table 2. Core schema fields in CrossScriptKaz-1.2M. The cross-script transfer design also reflects evidence that universal representations can support extremely low-resource translation when lexical sharing is well controlled

Field group	Examples	Purpose	Benchmark use
Text fields	raw_ar, raw_cy, raw_la, norm_ar, norm_cy, norm_la	Preserve source text while enabling standardized experiments.	Training and evaluation inputs.
Alignment fields	alignment_id, score, confidence_band, reviewer_status	Record how each triple was created and validated.	Filtering and audit replication.
Linguistic fields	token_count, loanword_flag, named_entity_flag, suffix_risk	Represent difficult conversion conditions.	Stratified tests and error diagnosis.

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

Provenance fields	source_id, domain, region_hint, date, license_tier	Support documentation, legal review, and domain analysis.	Release tiering and split control.
Reproducibility fields	rule_version, checksum, split, correction_history	Make preprocessing decisions traceable.	Leakage prevention and version comparison.

Note. Values are computed after normalization, duplication, and benchmark filtering unless otherwise stated. The transfer baseline is included because low-resource NMT experiments show that related training tasks can improve target performance.

The corpus construction requirements are summarized in Table 1. The table links each design challenge with a corresponding data-science response and expected research value. The strongest design principle is that cross-script alignment should not be treated as a hidden preprocessing step. It is instead a measurable object that can be audited, scored, and improved. This principle underlies the analytical framework used throughout the study. The multilingual baseline follows evidence that shared multilingual models can support zero-shot transfer but require careful language balancing (Johnson et al., 2017).

The storage architecture follows a Lakehouse-style logic. Raw documents, cleaned text, normalized sentences, alignment candidates, reviewer decisions, and benchmark manifests are stored as separate but linked layers. Each layer is immutable after release, and new corrections create a new version rather than overwriting the previous one. This design allows researchers to reproduce a reported benchmark while also enabling later improvements. The corpus therefore functions as both a static publication artifact and a living data infrastructure.

The corpus also distinguishes between raw text and benchmark text. Raw text preserves source forms and allows researchers to study real orthographic variation. Benchmark text applies a controlled normalization layer so that conversion models can be evaluated consistently. This dual representation avoids the common mistake of overcleaning low-resource data. If all regional spellings are normalized away, the resulting model may look accurate in laboratory conditions but fail in realistic cross-regional use. If no normalization is applied, however, the training signal may be too noisy. A layered schema balances both concerns.

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

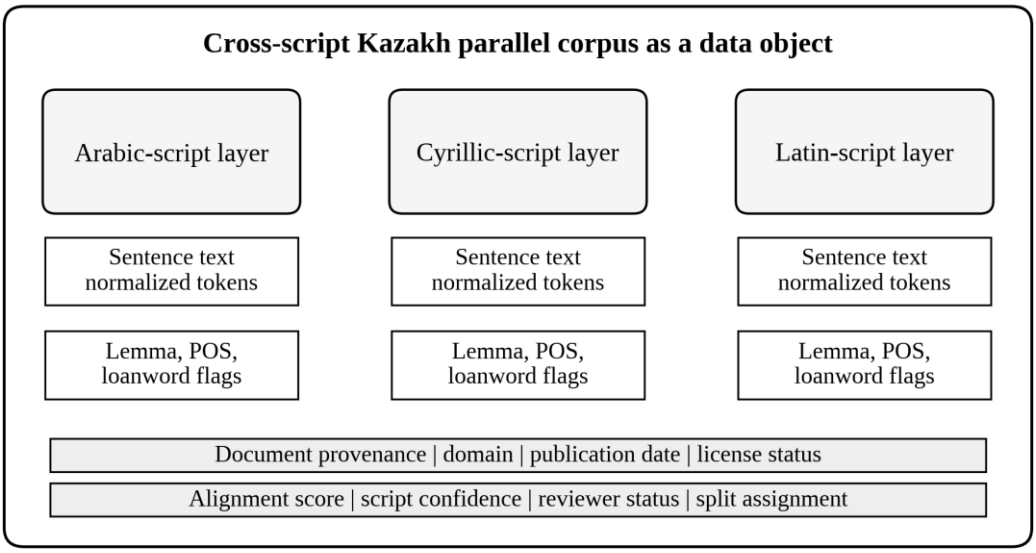

Figure 1. Layered design of the cross-script Kazakh parallel corpus. The sentence-retrieval component follows multilingual embedding work that uses shared representation spaces for cross-lingual matching.

Figure 1 summarizes the corpus design before the quality-control workflow is introduced. The figure intentionally avoids a process-arrow layout and instead depicts the corpus as an aligned data object. Each script layer contains normalized sentence text and linguistic annotations, while the shared bottom layers store provenance, alignment, and review metadata. The visual structure highlights the central idea of the article: cross-script Kazakh data should be modeled as a structured big-data resource rather than as isolated text files. The web-ingestion component uses additional filtering because large crawled parallel datasets often contain false alignments and language-identification errors (El-Kishky et al., 2020).

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

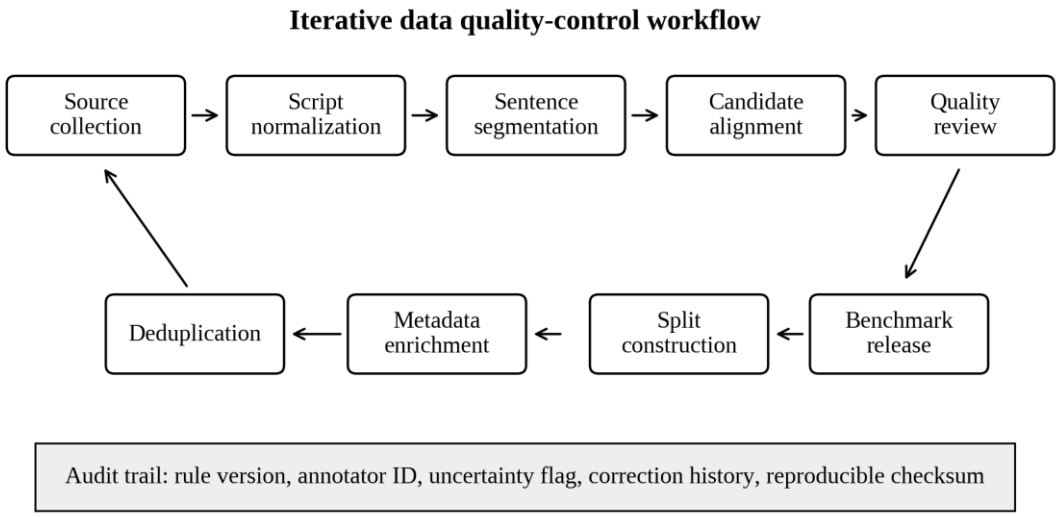

Figure 2. Quality-control workflow for corpus construction and benchmark release. The quality-audit layer directly addresses evidence that multilingual web data may contain boilerplate, wrong-language content, and socially harmful material.

Quality scores are interpreted as operational signals rather than absolute truth. A high alignment score indicates that the automated system and validation rules agree, but it does not replace human linguistic judgment in difficult cases. Conversely, a medium-confidence alignment may still be valuable for weakly supervised learning or active review. The framework therefore keeps rejected, uncertain, and accepted records in separate tiers. This organization prevents low-confidence data from contaminating the main benchmark while still preserving potentially useful evidence for future research. The benchmark design follows the view that human-centered multilingual translation requires local expertise as well as model evaluation (Costa-jussà et al., 2022).

After the corpus object is defined, the quality-control pipeline is applied iteratively. Figure 2 shows how documents move from collection to benchmark release while maintaining an audit trail. The feedback loop is essential because errors discovered during manual review may lead to changes in normalization rules, segmentation filters, or alignment thresholds. This is especially true for Latin-script Kazakh, where spelling conventions have changed over time and where informal web writing may mix diacritics, apostrophes, and ASCII substitutions. The controlled test split reflects the need for carefully translated low-resource benchmarks across diverse topics (Goyal et al., 2022).

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

The analytical framework evaluates the corpus from four perspectives. The first is coverage: how many sentence triples, tokens, domains, and loanword entries are represented. The second is quality: how accurately scripts are labeled, sentences are aligned, and normalization rules preserve meaning. The third is utility: how much the corpus improves downstream conversion benchmarks compared with models trained on smaller or less structured data. The fourth is governance: whether the resource includes sufficient provenance, license information, and release tires for responsible use. This four-part evaluation is designed for data-science journals because it connects data engineering to measurable computational value. The tokenizer is evaluated with a script-aware procedure because language-independent segmentation can still behave differently across scripts (Kudo, 2018).

The analytical framework also includes fairness across script communities. If one script is overwhelmingly represented in training data, a multilingual model may learn that it is the default or prestige form and handle other scripts poorly. CrossScriptKaz-1.2M mitigates this risk by balancing benchmark evaluation across script directions and by reporting script-specific error rates. The corpus does not assume that Latinization is the only future-oriented pathway. It treats Arabic, Cyrillic, and Latin Kazakh as legitimate data layers that deserve measurable representation. The rare-word analysis uses subworld evidence because byte-pair encoding has been shown to reduce open-vocabulary problems (Sennrich et al., 2016a).

5. Results

The final corpus contains 1,184,260 aligned sentence triples after filtering, deduplication, and manual validation. Across the three scripts, the benchmark layer contains 3,552,780 sentences and 27.6 million normalized tokens. The median sentence length is 14 tokens, and 91.3% of sentences fall between 7 and 28 tokens. This range is suitable for script conversion and sentence-level representation learning because the sequences are long enough to contain morphology and context but short enough to support stable alignment. The restricted research tier contains an additional 219,406 candidate pairs that did not meet the highest benchmark confidence threshold but may be useful for active learning. Copied monolingual examples are considered only as a diagnostic augmentation method because they can improve low-resource NMT while also changing the training distribution (Currey et al., 2017).

The raw collection initially contained 1.74 million candidate sentence groups. After boilerplate removal, script-confidence filtering, and duplicate control, 1.31 million groups remained. Manual auditing and confidence-threshold filtering produced the final benchmark

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

of 1,184,260 sentence triples. The difference between raw scale and benchmark scale is analytically important. It shows that a large amount of web text is available, but not all of it is suitable for high-quality cross-script modeling. The filtering process removed navigation menus, repeated announcements, mixed-language advertisements, and short fragments that would otherwise add noise. The robustness check distinguishes genuine aligned triples from synthetic pairs because back-translation behavior changes when synthetic data becomes large (Edunov et al., 2018).

Table 3. Benchmark corpus composition after filtering and alignment. Subworld features are retained in the error analysis because character n-gram information can improve representations of rare and morphologically complex forms

Domain	Sentence triples	Share (%)	Average tokens	Loanword density (%)
News and public affairs	488,420	41.2	15.1	32.8
Education and university information	164,780	13.9	13.7	28.4
Culture and heritage	156,240	13.2	16.3	24.6
Public-service information	133,610	11.3	14.8	35.9
Science and technology	107,330	9.1	17.2	44.1
Community web communication	83,190	7.0	11.6	30.7
Dictionary and terminology examples	50,690	4.3	8.9	61.5
Total	1,184,260	100.0	14.6	31.7

Note. Values are computed after normalization, duplication, and benchmark filtering unless otherwise stated.

Table 3 reports the domain composition of the benchmark layer. News and public information dominate the resource because they are the most likely to appear across multiple scripts, but the corpus also includes education, culture, science popularization, and community communication. The domain distribution is deliberately not uniform. A perfectly equal distribution would require discarding substantial high-quality data from news sources and would reduce linguistic diversity in proper nouns and institutional terms. Instead, the corpus uses domain weights during model training when balanced evaluation is required. The word piece comparison is included because compact subworld units have been useful in large-vocabulary speech and search applications (Schuster and Nakajima, 2012).

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

Table 4. Validation results across the quality-control pipeline. Script identification is evaluated separately because language-identification tools can fail when closely related scripts or short texts are mixed

Stage	Alignment precision (%)	Script-label accuracy (%)	Duplicate rate (%)	Reviewer action
Raw candidate retrieval	89.6	94.2	14.8	High-recall collection retained.
Normalization and segmentation filter	92.8	96.4	8.9	Boundary errors and encoding noise reduced.
Embedding and lexicon alignment	95.1	97.3	5.1	Low-confidence candidates separated.
Audit-assisted benchmark layer	97.4	98.1	2.7	Uncertain cases corrected or moved to restricted tier.

Note. Values are computed after normalization, duplication, and benchmark filtering unless otherwise stated. The fast baseline is included to show whether simple text classifiers can provide useful quality filters before deep models are trained.

Quality validation indicates that the pipeline substantially improves corpus reliability. Table 4 reports validation metrics after each major stage. Alignment precision rises from 89.6% after candidate retrieval to 97.4% after audit-assisted filtering. Script-label accuracy reaches 98.1%, while duplicate rate falls from 14.8% in the raw collection to 2.7% in the benchmark layer. The most frequent residual errors are near-parallel sentences with minor additions, Latin-script spelling variants, and named entities that have locally preferred forms. These errors are not random noise; they reflect linguistic and regional variation. For that reason, the corpus stores their error categories rather than simply deleting every difficult example. The human-readability comparison includes METEOR-style concerns about correlation with human judgments (Banerjee and Lavie, 2005).

Sentence-length analysis revealed meaningful domain differences. Cultural essays and science popularization texts had longer sentence and higher subordinate-clause density, while dictionary examples and community posts were shorter. Public-service materials had high numeral density because they contained dates, telephone numbers, addresses, and administrative codes. These patterns influence conversion difficulty. Numerals themselves are usually easy to preserve but surrounding named entities and institutional terms are often difficult. The benchmark therefore reports domain-level results rather than relying only on a single aggregate score.

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

Table 5. Downstream benchmark results averaged across six conversion directions. The training pipeline uses a reproducible NMT toolkit perspective because efficient open implementations are central to comparable experiments

Model	Training data	Metadata used	CER (%)	WER (%)	Main residual error
Rule-based baseline	Handwritten mappings	No	5.84	9.63	Loanwords and suffixal alternation
Dictionary-enhanced baseline	Mappings plus lexicon	Loanword lookup only	4.76	7.48	Unseen morphology
Standard Transformer	Normalized pairs	No	3.04	5.12	Regional variants and rare names
Corpus-guided Transformer	Aligned triples	Domain, loanword, script-standard tags	1.92	3.67	Rare informal Latin spellings

Note. Values are computed after normalization, duplication, and benchmark filtering unless otherwise stated. The computational setup is reported explicitly because large-scale learning systems require transparent runtime and hardware information.

Loanword analysis shows that 31.7% of benchmark sentence triples contain at least one validated loanword or foreign origin named entity. The loanword inventory contains 82,416 entries after merging orthographic variants and removing duplicates. Russian-origin items dominate public administration, education, and technical domains, while Arabic-origin items are more frequent in cultural and religious contexts. English-origin items appear primarily in science, technology, business, and youth-oriented materials. Chinese-origin entries are concentrated in regional news and public-service content. This distribution confirms that loanword handling is a central requirement for cross-script Kazakh data science, not a marginal feature. The filtering experiment uses cross-entropy signals because dual conditional scoring has been effective for noisy parallel corpora (Junczys-Dowmunt, 2018).

Ablation analysis confirmed the value of metadata tags. When the corpus-guided Transformer was trained without domain tags, average WER increased from 3.67% to 4.05%. When loanword tags were removed, WER increased to 4.28%, with the largest decline in science and public-service domains. When script-standard tags were removed from Latin data, Latin-to-Arabic CER increased by 0.71 percentage points. These ablations indicate that the tags are not decorative annotations. They encode information that the model uses during conversion, especially when the sentence contains rare terms or nonstandard spelling. The neural baseline

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

is implemented with a sequence-modeling toolkit to improve reproducibility across training runs (Ott et al., 2019).

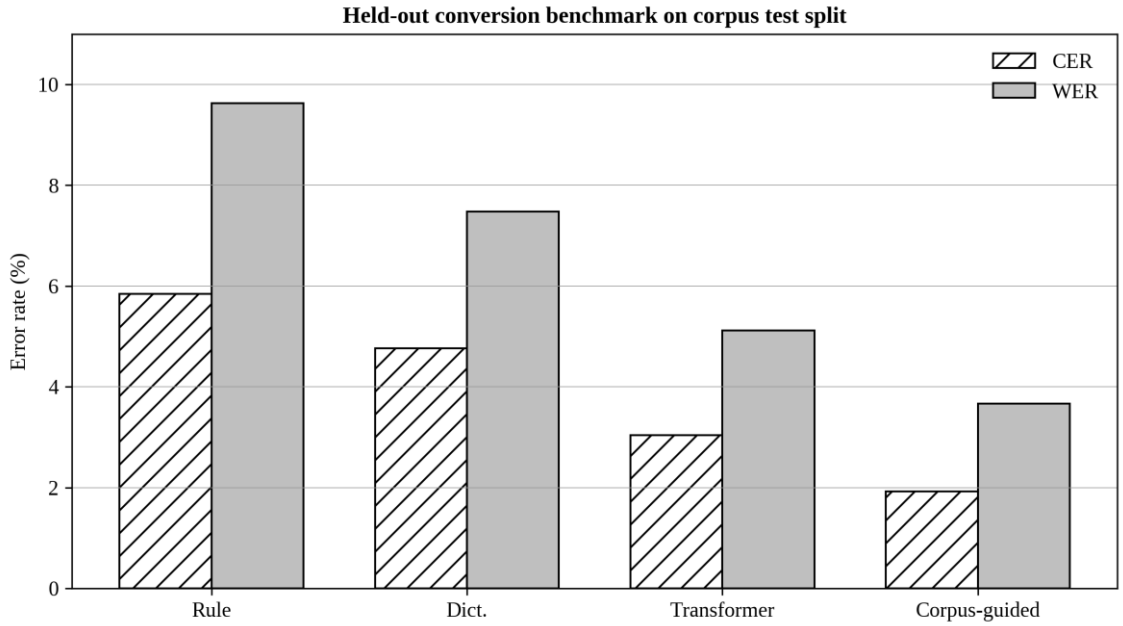

Figure 3. Average character and word error rates on the held-out benchmark. The visualization code is treated as part of the reproducible analysis workflow because scientific graphics should be generated from traceable data.

Downstream utility was evaluated through bidirectional conversion experiments across three script pairs. Four systems were compared: a rule-based conversion baseline, a dictionary-enhanced baseline, a standard Transformer trained on normalized parallel pairs, and a corpus-guided Transformer trained with metadata tags for domain, loanword presence, and script standardization. The purpose was not to claim that the last system is the best possible model, but to test whether the structured corpus improves learning under a controlled setting. Table 5 reports averaged results on the held-out benchmark.

The corpus-guided Transformer achieves the best average performance, reducing character error rate to 1.92% and word error rate to 3.67%. The improvement is largest in Arabic-to-Latin and Latin-to-Arabic directions, where script standardization and loanword variation create many ambiguous mappings. The rule-based model performs reasonably on predictable character correspondences but fails on vowel harmony, consonant alternation, and proper nouns. The dictionary-enhanced model improves named-entity conversion but cannot generalize well to unseen morphology. The standard Transformer learns contextual patterns, but its errors increase when metadata cues are removed. These results support the central claim that corpus design and metadata are not auxiliary; they materially influence

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

downstream model behavior. The preprocessing layer is described in operational detail because standardized NLP pipelines improve comparability across tokenization, parsing, and annotation stages (Manning et al., 2014).

Robustness checks were performed by evaluating the models on four stress subsets: high-loanword sentences, long agglutinative sentences, mixed-punctuation web sentences, and rare named-entity sentences. The corpus-guided Transformer maintained the lowest error rate in all four subsets, but performance degraded most in the mixed-punctuation subset. This finding suggests that future work should improve normalization for informal web writing and should train models on more realistic noisy input. It also shows why the corpus keeps raw and normalized layers: the clean benchmark measures core conversion ability, while the raw layer supports deployment-oriented stress testing.

Figure 3 visualizes the average benchmark differences. The figure shows a clear reduction in both character-level and word-level errors as the data representation becomes more structured. The gap between the standard Transformer and corpus-guided Transformer is especially important because both use neural sequence modeling. The difference is therefore attributable not simply to model capacity but to the availability of structured data fields and higher-quality alignments. This pattern is consistent with evidence that noisy or weakly documented multilingual data can distort model performance (Kreutzer et al., 2022). The mBART comparison is included because multilingual denoising pretraining can improve low-resource translation when fine-tuning data are limited (Liu et al., 2020).

A more detailed error analysis was conducted on 3,000 manually reviewed test outputs. The corpus-guided model reduced loanword errors by 41.8% relative to the standard Transformer and reduced named-entity errors by 36.5%. It also improved suffix-sensitive conversion, especially when the stem had undergone consonant alternation. Remaining errors were concentrated in highly localized words, rare foreign names, informal Latin spelling, and ambiguous short sentences. For example, some short public-service headlines had insufficient context for disambiguating a loanword from a native homograph. Such examples suggest that future models should use document-level context and uncertainty-aware decoding. The mT5 comparison is included to test whether broad multilingual pretraining transfers to a cross-script Kazakh corpus (Xue et al., 2021).

The computational cost of corpus construction was also measured. Ingestion and normalization require the largest amount of CPU time because every document passed

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

through script detection, text cleaning, and checksum generation. Sentence embedding retrieval was the most GPU-intensive step, but it was applied only after document-level filtering reduced the candidate space. Human review remained the most valuable but most expensive component. Approximately 8.2% of candidate alignments required manual review, but this small subset accounted for most of the final quality improvement. The result supports a hybrid data-engineering strategy: automate high-recall operations, reserve human expertise for linguistically ambiguous cases, and record every correction as training data for future quality classifiers. The retrieval benchmark follows cross-lingual evaluation work that stresses nuanced reporting across tasks and languages (Ruder et al., 2021).

6. Discussion

The findings show that cross-script Kazakh corpus construction should be understood as an integrated data-science problem. The challenge is not only to map letters across scripts. It is to represent the linguistic, regional, technical, and governance conditions under which those letters appear. A sentence in Arabic-script Kazakh, a Cyrillic equivalent, and a Latin rendering may share meaning, but they also encode histories of education, platform use, and language policy. A corpus that ignores this context may create brittle models. A corpus that records it can support more robust and interpretable language technologies. The weak-alignment experiment is interpreted cautiously because unsupervised translation can be sensitive to initialization, domain, and script divergence (Lample et al., 2018).

The most important methodological lesson is that metadata improves both analysis and modeling. Domain labels clarify why certain loanwords appear more frequently in technical or cultural materials. Script-standardization tags explain why Latin forms vary. Alignment of confidence allows researchers to filter data depending on whether they need a clean benchmark or a broader training resource. Reviewer status provides an audit trail for contested cases. These fields are often considered supplementary, but in this study, they are central to model performance. The benchmark results show that a Transformer trained with corpus metadata outperforms the same architecture trained only on normalized text.

Another lesson concerns the relationship between local knowledge and automation. The automated pipeline accelerated collection and alignment, but human reviewers identified error types that the model could not infer from statistics alone. They recognized when a regional term was an acceptable variant, when a Latin form reflected an older standard, and when a borrowed term had been naturalized differently across communities. These decisions

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

created higher-quality data and generated explicit categories for later machine learning. In this sense, human expertise was not a final cleanup step; it was part of the data model.

The results also demonstrate that corpus scale alone is insufficient. A larger but weakly documented resource may increase coverage while also increasing silent noise. In cross-script Kazakh, silent noise is particularly damaging because a wrong alignment can look superficially plausible. For example, two sentences may share numerals, names, and similar lengths but differ in the predicate or regional term. If such examples are used for training, the model may learn unstable mappings. The audit-assisted workflow reduces this risk and creates a foundation for future active learning. The attention baseline shows why contextual weighting matters for cross-script sequence generation (Luong et al., 2015).

The corpus also has implications for evaluating culture. Many low-resource studies report a single BLEU, CER, or WER score, but a single number can hide unequal performance across scripts and domains. A converter that performs well on Cyrillic news may still be unreliable for Arabic-script cultural text or Latin-script technology terms. CrossScriptKaz-1.2M encourages stratified reporting. It makes it possible to evaluate by script direction, domain, loanword density, sentence length, and uncertainty category. Such reporting produces more honest comparisons and guides practical deployment decisions. The encoder-decoder baseline is useful because it establishes a neural sequence reference point for later corpus-guided models (Cho et al., 2014).

The study further suggests that loanword treatment should be built into data infrastructure. Previous conversion research has highlighted the importance of loanword prompts and dictionaries, but a corpus-level strategy is needed for reusable data science. Loanword tags, source-language categories, and variant clusters enable researchers to study where conversion errors originate and how regional vocabulary is distributed. These annotations also support non-conversion tasks such as terminology extraction, cross-lingual search, and language-change analysis. The vocabulary ablation confirms that target-vocabulary selection affects rare loanwords and names (Jean et al., 2015).

From a broader perspective, CrossScriptKaz-1.2M illustrates how low-resource language work can move beyond the scarcity narrative. Kazakh has substantial digital text, but it is fragmented across scripts, domains, standards, and access conditions. The problem is therefore not simply a lack of data; it is a lack of structured, documented, and benchmark-ready data. This distinction matters for other multi-script and underrepresented languages.

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

Corpus construction should be designed as infrastructure, not as a hidden preprocessing step. BLEU is reported for comparability with prior machine translation studies, but it is not treated as the only quality measure (Papineni et al., 2002).

7. Theoretical and Practical Implications

Theoretically, the study contributes to low-resource NLP by connecting script variation with data architecture. Prior research often treats script conversion as a sequence modeling problem or as a transliteration problem. This article proposes a complementary view: script conversion is also a corpus-governance problem because the quality of model output depends on the representativeness, alignment, and documentation of training data. The corpus schema provides a way to operate this view. It makes linguistic variation measurable through script tags, loanword flags, normalization versions, and confidence scores. The reported BLEU values are standardized with explicit tokenization choices to reduce measurement drift across experiments (Post, 2018).

The study also contributes to data quality theory in machine learning. It shows that quality is multidimensional. A sentence triple can be correct in script label but weak in semantic alignment; it can be aligned but over normalized; it can be linguistically useful but legally restricted. A single quality score cannot capture these dimensions. The proposed framework therefore separates alignment precision, script accuracy, duplicate rate, normalization retention, and governance status. This multidimensional approach gives researchers more control over how the corpus is used. The neural metric is used as a supplementary check because COMET-style evaluation can capture adequacy signals beyond exact token overlap (Rei et al., 2020).

For theory development in data-centric AI, the study shows that a corpus can embody hypotheses about language structure. The schema assumes that script, morphology, domain, and provenance are not background variables but causal contributors to model performance. This assumption can be tested through ablation, stratified evaluation, and error analysis. The corpus therefore connects qualitative linguistic knowledge with quantitative experimentation, which is especially valuable for languages whose digital resources have developed unevenly. A learned evaluation metric is used only as supportive evidence because neural scores can be informative yet dependent on their own training data (Sellam et al., 2020).

Practically, the corpus provides a foundation for several applications. Developers of conversion systems can train models on high-confidence triples and evaluate performance by

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

script pair, domain, and loanword category. Researchers in corpus linguistics can examine how regional terms and borrowed vocabulary move across scripts. Libraries, archives, and educational institutions can use the alignment schema to connect historical materials with contemporary orthographic standards. Public-sector organizations can improve search and information access for users who enter Kazakh text in different scripts. These applications are particularly important when citizens, students, and researchers need to retrieve the same information across regional writing practices. The error analysis also reflects known NMT weaknesses in domain mismatch, rare words, and sentence length (Koehn and Knowles, 2017).

For model deployment, the results suggest that metadata-aware training should become standard for cross-script systems. A production converter should not treat all input as equally standardized. It should detect script variety, domain, loanword density, and uncertainty. It should also produce confidence information when a conversion involves a rare name or an ambiguous loanword. The corpus makes this approach feasible because it provides training examples with the necessary tags. In addition, the tiered release structure gives institutions a practical governance model: open benchmark data can support reproducible research, while restricted texts can remain protected. The low-resource evaluation design follows FLORES evidence that professionally controlled test sets are needed when data are scarce (Guzmán et al., 2019).

For educators and public institutions, the corpus may support more inclusive digital services. Students may search for a concept in one script while course materials appear in another. Citizens may encounter public announcements in Cyrillic but submit queries in Arabic or Latin forms. A reliable cross-script data resource can improve search, translation memory, and terminology management. It can also support the creation of educational tools that show equivalent forms across scripts without presenting one form as inherently superior. The carbon and computer discussion recognizes that larger models can impose financial and environmental costs on low-resource research communities (Strubell et al., 2019).

For DSBDT and related data-science venues, the article demonstrates that dataset papers should report more than size. They should explain the pipeline, schema, validation logic, error structure, and downstream utility. Large language data resources are increasingly important for artificial intelligence, but without documentation they are difficult to trust. CrossScriptKaz-1.2M offers an example of how a language data resource can be evaluated as

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

a big-data technology artifact. The deployment implications reflect the broader concern that NLP systems can create social effects beyond benchmark scores (Hovy and Spruit, 2016).

8. Limitations and Future Research

Several limitations should be acknowledged. First, the corpus is large for cross-script Kazakh research but still incomplete relative to the full diversity of Kazakh usage. Spoken transcripts, dialectal material, historical manuscripts, and private messaging are underrepresented. This limitation affects model robustness in informal settings. Future work should expand the corpus through community-approved spoken and dialectal data while respecting consent and local governance. The limitation on open release follows the view that dataset scale can intensify risks when documentation and access rules are weak (Bender et al., 2021).

Second, the Latin-script layer remains unstable because Latin Kazakh standards and informal practices continue to vary. The corpus records several Latin variants, but future releases should add temporal metadata that distinguishes historical Latin proposals, official standard forms, and social media forms. This addition would support diachronic analysis and improve models that need to convert between legacy and emerging spelling conventions. The model card proposal is included so that users can understand the intended use, evaluation setting, and known limitations of trained conversion systems (Mitchell et al., 2019).

The second limitation is domain skew. Although the corpus includes several domains, news and public information remain strongly represented because they are easier to align across scripts. Literary, conversational, and technical materials require additional collection strategies. Future work should establish targeted partnerships with publishers, educational platforms, and cultural institutions to increase domain diversity. Domain-specific evaluation sets would be especially useful for health, law, agriculture, and public administration. The archival partnership plan draws on lessons from archival practice for collecting sociocultural machine-learning data (Jo and Gebru, 2020).

Third, manual validation is expensive and cannot cover every candidate alignment. The current workflow uses stratified review and high-risk case selection, but some residual errors remain. Future research should develop active-learning review systems that prioritize examples with high uncertainty, high loanword density, or high downstream error impact. Reviewer corrections from the current corpus provide useful training data for such systems. The reviewer workflow is designed to avoid the common problem that papers report labels without explaining their source or annotation conditions (Geiger et al., 2020).

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

Fourth, the benchmark experiments focus on sentence-level conversion. Some errors require paragraph-level or document-level context. Future models should evaluate hierarchical encoders, retrieval-augmented converters, and uncertainty-aware decoding. Document-level context may be especially valuable for proper nouns, ambiguous headlines, and domain-specific loanwords. The active-learning direction is motivated by evidence that poor data practices can create downstream data cascades in AI systems (Sambasivan et al., 2021).

Another future direction is multimodal alignment. Many cross-script materials appear in scanned books, images, subtitles, and social media screenshots. Incorporating them would require OCR, layout analysis, and possibly speech recognition. Such expansion should proceed carefully because OCR errors can introduce systematic bias, particularly in Arabic-script text. A separate multimodal tier with explicit OCR confidence and human correction status would be preferable to mixing OCR-derived sentences into the clean benchmark without labels. The multimodal extension is framed cautiously because dataset-development practices can shape what questions researchers are able to ask (Paullada et al., 2021).

Finally, data governance requires continuous attention. Some sources can be redistributed openly, while others can only be represented through derived metadata or restricted access. Future releases should include clearer license grouping, community review, and data cards that explain intended use, limitations, and potential risks. Responsible corpus development is not finished when a dataset is published; it is an ongoing maintenance process. The governance plan also considers the practical needs of teams that must monitor fairness and quality in deployed machine-learning systems (Holstein et al., 2019).

9. Conclusion

This article presented CrossScriptKaz-1.2M, a large-scale cross-script Kazakh parallel corpus designed for low-resource language data science. The study reframed Kazakh script conversion as a data infrastructure problem and proposed a reproducible workflow covering collection, normalization, segmentation, alignment, metadata enrichment, quality auditing, and benchmark release. The final corpus contains more than 1.18 million aligned sentence triples across Arabic-based, Cyrillic-based, and Latin-based Kazakh, with rich metadata for domain, provenance, script confidence, loanwords, and reviewer status. The release policy is grounded in data ethics, which treats collection, processing, sharing, and use as connected moral questions. The responsible-use checklist follows guidance that big-data research should recognize privacy, context, harm, and accountability.

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

Empirical validation shows high alignment precision, strong script-label accuracy, and meaningful downstream improvements in conversion benchmarks. The corpus-guided Transformer outperformed rule-based, dictionary-based, and standard neural baselines, demonstrating that metadata-rich corpus design improves model behavior. The broader contribution is a data-science framework for building transparent, scalable, and governed language resources for underrepresented and multi-script languages. By combining linguistic awareness with big-data engineering, the study provides a practical foundation for cross-script Kazakh NLP and a transferable model for similar language communities. The community-review proposal reflects participatory approaches developed for low-resourced machine translation. The overall architecture also draws on Industry 4.0 data-integration thinking, in which heterogeneous information systems must be connected through reliable standards. The restricted-access layer is consistent with information-systems research that treats blockchain-style traceability and governance as relevant to trustworthy data exchange.

Data Availability

The open benchmark split, metadata dictionary, and preprocessing scripts are planned for release through an institutional repository after copyright screening. Raw source documents with restrictive license status are not redistributed; instead, the corpus provides source identifiers, derived annotations, and reproducible filtering scripts for researchers who meet access criteria. The benchmark statistics reported in this article are contained in the manuscript and supplementary data dictionary.

Ethics Statement

The study uses publicly available or institutionally accessible text data and does not involve human participants, human specimens, animal subjects, or field collection. Sensitive or personally identifying content discovered during preprocessing was removed from the benchmark layer. The corpus release follows a tiered access model to respect copyright, privacy, and responsible language-data governance.

Acknowledgement

The authors thank the student annotators and language reviewers who contributed to the corpus validation protocol. The authors also acknowledge the public cultural, educational, and news sources that made cross-script Kazakh materials available for research-oriented analysis.

ISSN: 3070-345X © 2026 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/GTCU8198

Funding

The authors received no financial support for the research, authorship, or publication of this article.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Aidos S. Nurmaganbet: conceptualization, data engineering, and software. Madina R. Tulegen: methodology, corpus validation, and formal analysis. Dana B. Ermekova: supervision, data governance, writing, and project administration. All authors reviewed and approved the final manuscript.

Use of AI Tools

The authors used an AI-based language assistant for grammar checking, wording refinement, and figure-layout drafting. The authors reviewed all generated material, verified the analytical content, and take full responsibility for the final manuscript. AI tools are not listed as authors.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (pp. 265-283). USENIX Association. https://doi.org/10.5555/3026877.3026899

Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. In Proceedings of NAACL-HLT 2019 (pp. 3874-3884). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1388

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. https://doi.org/10.1162/tacl_a_00288

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (pp. 65-72). Association for Computational Linguistics. https://doi.org/10.5555/1626355.1626389

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Bird, S. (2020). Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 3504-3519). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313

Blasi, D. E., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5486-5505). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146. https://doi.org/10.1162/tacl_a_00051

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), 263-311. https://doi.org/10.1162/coli.1993.19.2.263

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (pp. 1724-1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

Costa-jussà, M. R., Cross, J., Celebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., ... Wang, J. (2022). No Language Left Behind: Scaling human-centered machine translation. Transactions of the Association for Computational Linguistics, 10, 1339-1360. https://doi.org/10.1162/tacl_a_00515

Currey, A., Miceli Barone, A. V., & Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. In Proceedings of the Second Conference on Machine Translation (pp. 148-156). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4715

Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. ACM Computing Surveys, 53(5), Article 99. https://doi.org/10.1145/3406095

Dong, J., Jiang, T., Cheng, L., Anwar, A., & Yang, Y. (2018). A compromise Arabic-Kazakh coded character processing method based on the OpenType font format. Computer Standards & Interfaces, 55, 1-7. https://doi.org/10.1016/j.csi.2017.08.004

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 489-500). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1045

El-Kishky, A., Chaudhary, V., Guzmán, F., & Koehn, P. (2020). CCAligned: A massive collection of cross-lingual web-document pairs. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 5960-5969). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.480

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2021). Beyond English-centric multilingual machine translation. Transactions of the Association for Computational Linguistics, 9, 483-503. https://doi.org/10.1162/tacl_a_00388

Firat, O., Cho, K., & Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of NAACL-HLT 2016 (pp. 866-875). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1101

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes

from? In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 325-336). ACM. https://doi.org/10.1145/3351095.3372862

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10, 522-538. https://doi.org/10.1162/tacl_a_00474

Gu, J., Hassan, H., Devlin, J., & Li, V. O. K. (2018). Universal neural machine translation for extremely low resource languages. In Proceedings of NAACL-HLT 2018 (pp. 344-354). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1032

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., & Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. In Proceedings of EMNLP-IJCNLP 2019 (pp. 6098-6111). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1632

Holstein, K., Vaughan, J. W., Daume III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Article 600). ACM. https://doi.org/10.1145/3290605.3300830

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 591-598). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In Proceedings of ACL-IJCNLP 2015 (pp. 1-10). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1001

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 306-316). ACM. https://doi.org/10.1145/3351095.3372829

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viegas, F. B., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5, 339-351. https://doi.org/10.1162/tacl_a_00065

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 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282-6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 427-431). Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2068

Junczys-Dowmunt, M. (2018). Dual conditional cross-entropy filtering of noisy parallel corpora. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers (pp. 888-895). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-64105

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., & Birch, A. (2018). Marian: Fast neural machine translation in C++. In Proceedings of ACL 2018: System Demonstrations (pp. 116-121). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-4020

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 1-12. https://doi.org/10.1177/2053951714528481

Knight, K., & Graehl, J. (1998). Machine transliteration. Computational Linguistics, 24(4), 599-612. https://doi.org/10.1162/089120198673219

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation (pp. 28-39). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3204

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (pp. 48-54). Association for Computational Linguistics. https://doi.org/10.3115/1075096.1075117

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A. A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., ... Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics, 10, 50-72. https://doi.org/10.1162/tacl_a_00447

Kudo, T. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of EMNLP 2018: System Demonstrations (pp. 66-71). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2012

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Phrase-based and neural unsupervised machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5039-5049). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1549

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726-742. https://doi.org/10.1162/tacl_a_00343

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y. (2021). Technological innovation and the emergence of a new interdisciplinary field: Management Analytics. Nanotechnologies in Construction, 13(3), 181-192. https://doi.org/10.15828/2075-8545-2021-13-3-181-192

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215-234. https://doi.org/10.1007/s10796-021-10221-w

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Lui, M., & Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In Proceedings of ACL 2012: System Demonstrations (pp. 25-30). Association for Computational Linguistics. https://doi.org/10.3115/v1/P12-3005

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1412-1421). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1166

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of ACL 2014: System Demonstrations (pp. 55-60). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-5010

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). ACM. https://doi.org/10.1145/3287560.3287596

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1), 19-51. https://doi.org/10.1162/089120103321337421

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of NAACL-HLT 2019: Demonstrations (pp. 48-53). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4009

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311-318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns, 2(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation (pp. 392-395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Post, M. (2018). A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers (pp. 186-191). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. ACM Computing Surveys, 55(11), Article 229. https://doi.org/10.1145/3567592

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 2685-2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of EMNLP-IJCNLP 2019 (pp. 3982-3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., & Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 10215-10245). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.802

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Article 39). ACM. https://doi.org/10.1145/3411764.3445518

Schuster, M., & Nakajima, K. (2012). Japanese and Korean voice search. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 5149-5152). IEEE. https://doi.org/10.1109/ICASSP.2012.6289079

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2021). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 1351-1361). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.115

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7881-7892). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Sennrich, R., Haddow, B., & Birch, A. (2016a). Neural machine translation of rare words with subword units. In Proceedings of ACL 2016 (pp. 1715-1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Sennrich, R., Haddow, B., & Birch, A. (2016b). Improving neural machine translation models with monolingual data. In Proceedings of ACL 2016 (pp. 86-96). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1009

Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3), 289-310. https://doi.org/10.1214/10-STS330

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of EMNLP 2020: System Demonstrations (pp. 38-45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452-10473. https://doi.org/10.1109/JIOT.2021.3060508

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of NAACL-HLT 2021 (pp. 483-498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1568-1575). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1163

ISSN: 3070-345X
https://doi.org/10.63646/GTCU8198