ISSN: 3071-1835 Volume 01, Issue 05, June, 2026

Scienta Press Journals

American Scholar
Journal of Interdisciplinary Research and Knowledge

Artificial Intelligence in Language Corpus Construction: A Structured Review of Pipelines, Failure Modes, and Validation Requirements

Sukhrob Avezov
PhD, Associate Professor, Department of Russian Language and Literature
Bukhara State University

Abstract

Background. The construction of language corpora has historically been limited by the cost of human labour: texts had to be located, cleaned, aligned, and annotated by hand. Artificial intelligence (AI) – in particular large language models (LLMs), neural speech recognition systems, and multilingual sentence encoders – has substantially removed that constraint. Corpus builders now routinely delegate acquisition, filtering, alignment, annotation, and even text production itself to automated systems. The methodological consequences of this shift have not yet been consolidated into a coherent account for the corpus-linguistic community.

Objectives. This article had three aims: (i) to map the functions AI currently performs at each stage of corpus construction; (ii) to identify the failure modes that automation introduces into the resulting resources; and (iii) to specify the minimum validation and documentation practices under which an AI-assisted corpus can still serve as linguistic evidence.

Methods. A structured review of approximately sixty primary sources published between 2018 and 2026 was conducted across the ACL Anthology, arXiv, and journals of corpus linguistics, computational linguistics, and research methodology. Sources were coded against a five-stage pipeline model — acquisition, filtering, alignment and modality expansion, annotation, and synthesis and against three cross-cutting

dimensions: quality risk, validation practice, and provenance documentation.

Results. AI performs a distinct and identifiable function at every pipeline stage, and in each case shifts the dominant cost from labour to verification. Three cross-cutting findings emerged. First, a contamination cascade now operates in web-sourced data: machine-translated text dominates lower-resource web content, roughly 35% of newly published websites in mid-2025 were classified as AI-generated or AI-assisted, and recursive training on such material degrades the distributional tails that linguistic research depends upon. Second, validation practice lags well behind automation capability; naive use of model-produced labels yields biased estimates and invalid confidence intervals even at 80-90% surrogate accuracy. Third, provenance documentation has become a legal as well as a scientific requirement, as crawler restrictions expand and text-and-data-mining exceptions tie permissible reuse to institutional status and machine-readable opt-outs.

Conclusions. AI has not solved corpus construction; it has relocated the difficulty from collection to curation and verification. An eight-point protocol is proposed, centring on a probability-sampled human gold standard, explicit machine-provenance tagging at the document level, and reporting of the operational domain actually sampled rather than the domain nominally targeted. The implications are sharpest for low-resource and typologically underserved languages, where automation offers the largest gains and the weakest safeguards simultaneously.

Keywords: corpus construction; large language models; corpus linguistics; data curation; corpus representativeness; annotation validation; low-resource languages; research methodology

This work is Licensed under a Creative Commons Attribution 4.0 International License.

Introduction

For most of its history, corpus linguistics has been a discipline organised around scarcity. The Brown Corpus took years to assemble one million words; the British National Corpus required a consortium and a decade. The constraining resource was never storage or computation but human attention: someone had to select texts, secure permissions, transcribe speech, correct optical character recognition output, align translations sentence by sentence, and apply annotation categories consistently

across thousands of tokens. Every methodological convention of the field – stratified sampling frames, balanced registers, documented selection criteria, inter-annotator agreement statistics – developed as a response to that scarcity. Because data were expensive, they had to be chosen carefully and described precisely.

That constraint has now largely dissolved. A corpus builder in 2026 can obtain billions of words of running text from public web crawls in an afternoon, transcribe thousands of hours of speech overnight with an open-weight recogniser, mine parallel sentences across two hundred languages with a multilingual sentence encoder, and apply a fine-grained pragmatic annotation scheme to a hundred thousand concordance lines for a few tens of dollars in inference costs. The FineWeb dataset alone comprises fifteen trillion tokens derived from ninety-six Common Crawl snapshots. What was once a multi-year institutional undertaking is now, at least in its raw form, a weekend project for an individual researcher.

This is not a marginal improvement in efficiency. It is a change in the nature of the activity. When data are scarce, the central methodological question is what to collect. When data are abundant, the central question becomes what to discard, and on what grounds. The literature on large-scale dataset curation has recognised this: FineWeb-Edu retains roughly one-tenth of its input by applying a quality classifier trained on annotations produced by a large instruction-tuned model. Aggressive discarding is now the principal act of corpus design. Yet corpus linguistics, as a discipline with its own long-standing standards of evidence, has not fully absorbed what this implies for the status of its data.

Three developments have converged to make this lag consequential rather than merely untidy.

First, the raw material has changed composition. Web-crawled multilingual corpora were shown to contain systematic and severe quality problems: an audit of 205 language-specific subsets drawn from five widely used datasets found that at least fifteen contained no usable text in the language they claimed to represent, and that a substantial proportion contained fewer than half acceptable sentences, with mislabelled or non-standard language codes compounding the problem. Subsequent work established that a large share of translated web content and in lower-resource languages, a large share of all web content – is machine-translated output of low quality. Most recently, a representative

sample of archived web pages indicated that roughly 35% of websites newly published in mid-2025 were classified as AI-generated or AI-assisted, up from a baseline of essentially zero before late 2022. A corpus assembled from the contemporary web is therefore not a sample of contemporary language use in the sense the discipline has always intended.

Second, the analytical instrument and the object of analysis have begun to converge. When a linguist uses a language model to annotate constructions in a corpus, and that model was trained on a corpus of the same general provenance, the resulting frequencies are no longer independent observations of usage. This circularity has no precedent in earlier automation: a part-of-speech tagger trained on hand-annotated data introduced error, but not a systematic dependency on the same distribution being measured.

Third, the legal environment governing acquisition has hardened. Within a single year, restrictions expressed through robots.txt rendered a substantial share of the most actively maintained sources in a major training corpus fully off-limits to identified crawlers, while terms-of-service restrictions affected a much larger fraction still. Text-and-data-mining exceptions in European law distinguish sharply between research organisations and other actors, and make general-purpose mining contingent on machine-readable rights reservations. Provenance is thus no longer only a scientific virtue; it determines whether a corpus can be distributed at all.

The present review responds to that lag. It is not a survey of AI techniques as such several exist, and they are written for engineering audiences but an attempt to state what those techniques do to a corpus considered as evidence about language. Three research questions guided the work:

RQ1. What functions does AI currently perform at each stage of language corpus construction, and what does each function replace?

RQ2. What failure modes does automation introduce into the resulting resource, and by what mechanism does each arise?

RQ3. What minimum validation, correction, and documentation practices are required for an AI-assisted corpus to support linguistic claims?

Section 2 sets out the review design and the pipeline model used for coding. Section 3 reports stage-by-stage findings and three cross-cutting results. Section 4 discusses the implications for corpus

representativeness, for the circularity problem, and for low-resource language communities, and proposes a concrete protocol. Section 5 concludes.

Materials and Methods

This study is a structured narrative review with a pre-specified coding framework. It is not a systematic review in the PRISMA sense, and no meta-analytic synthesis was attempted. That choice was deliberate and follows from the object of study: the relevant literature spans corpus linguistics, computational linguistics, machine learning, research methodology, and information law, and these fields do not share indexing conventions, publication venues, terminology, or standards of evidence. An exhaustive protocol keyed to a single database would have produced a systematically distorted sample, over-representing whichever field happened to be best indexed. The review therefore uses purposive sampling across venues with transparent inclusion criteria, and its findings should be read as a structured map rather than as an exhaustive census.

Searches were conducted across four source types: (a) the ACL Anthology, covering the principal computational-linguistics venues and their dataset and benchmark tracks; (b) arXiv preprints in cs.CL, cs.LG, and cs.CY; (c) journals of corpus and applied linguistics, including the International Journal of Corpus Linguistics, Linguistics Vanguard, Applied Corpus Linguistics, and Corpora; and (d) general-science and methodology outlets where dataset-level findings have appeared, including Nature, the Proceedings of the National Academy of Sciences, Transactions of the Association for Computational Linguistics, and Communications of the ACM.

Search terms combined a pipeline term (corpus construction, data curation, web crawling, deduplication, bitext mining, pseudo-labelling, annotation, synthetic data) with a technology term (large language model, LLM, neural machine translation, speech recognition, sentence encoder) and, in a third pass, with an evaluation term (quality, audit, validation, representativeness, provenance, consent). Citation chaining was applied in both directions from the twelve most frequently cited sources. Searches were completed in August 2026.

Sources were included if they satisfied all of the following: they reported an original method, dataset, audit, or measurement rather than an opinion piece; they addressed the production or evaluation of language data

rather than model architecture or downstream task performance alone; they were published between 2018 and 2026; and they were available in full text. Sources were excluded if they concerned image, video, or purely tabular data; if they described a system without reporting evaluation against any external reference; or if they duplicated an earlier report by the same team without adding evidence.

Results

Acquisition was formerly the most labour-intensive stage after annotation. It has been almost entirely automated, but the automation is largely pre-neural: crawling, extraction, and archiving depend on infrastructure that predates the current generation of models. What AI has added at this stage is targeting — the ability to identify candidate documents by language, domain, or register before they are stored, rather than after. Language identification models operating at crawl time determine which portion of a crawl is even seen by downstream processes, and their errors are therefore not correctable later.

The dominant finding at this stage is that acquisition is now governed by consent signals more than by technical capability. An audit of fourteen thousand web domains underlying major training corpora documented a rapid expansion of restrictions: within one year, robots.txt directives rendered roughly 5% of all tokens in a widely used crawl-derived corpus, and more than 28% of tokens from its most actively maintained sources, fully restricted for identified crawlers; terms-of-service restrictions applied to a still larger share (Longpre et al., 2024). The authors characterise the resulting situation as a mismatch between web protocols designed for search indexing and their present use for model training, and note that the restrictions are unevenly distributed across content types — which means their effect is not to shrink the available web uniformly but to bias what remains.

For corpus linguistics this matters in a specific way. The domains most likely to restrict crawling are those with commercial value in their text: news organisations, publishers, and large user-generated-content platforms. These are precisely the registers that balanced corpus designs have historically weighted most heavily. The freely crawlable web is thus drifting away from the register distribution that corpus designs assume, in a direction that no sampling correction applied after the fact can recover.

The legal frame reinforces this. Articles 3 and 4 of the European Union's

Directive on Copyright in the Digital Single Market create two text-and-data-mining exceptions with markedly different scope: Article 3 covers research organisations and cultural heritage institutions carrying out mining for scientific research and cannot be set aside by contract, while Article 4 extends mining to any actor but permits rights holders to reserve their rights through machine-readable means. A university corpus project and a commercial data vendor therefore operate under different rules for the same acquisition act, and the outputs are not equally redistributable — a point with direct consequences for whether an assembled corpus can be shared with the research community that needs to verify it.

The FineWeb-Edu procedure is the reference implementation: a large instruction-tuned model was used to score a sample of documents for educational value, a lightweight classifier was trained on those scores, and that classifier was then applied at corpus scale, retaining approximately one-tenth of the input. Variants have since proliferated – ensembles combining several classifiers, line-level rather than document-level scoring, and lightweight lexical classifiers chosen for inference throughput but the architecture is stable: a large model defines the quality criterion, a small model enforces it at scale.

Three consequences follow, and none is neutral with respect to linguistic research.

First, the criterion is implicit. A heuristic filter can be stated in a methods section in three lines; a classifier trained on model-generated preference scores cannot. When a corpus retains 10% of its input on the basis of a learned quality score, the resulting register composition is an emergent property of the scoring model, not a design decision. It is knowable only by post hoc inspection, and it is rarely inspected.

Second, the criteria in circulation are optimised for language-model training, not for linguistic description. Educational value, instruction-following resemblance, and reasoning density are sensible targets when the corpus is training data. They are actively counterproductive when the corpus is meant to represent language use, because they systematically down-weight informal registers, dialectal and non-standard varieties, code-switched material, phatic and interactional language, and precisely the phenomena that variationist and sociolinguistic corpus work exists to study. Reusing a training-optimised corpus for descriptive linguistics inherits a filter designed to remove the object of inquiry.

Third, deduplication interacts with frequency. Removing near-duplicates improves model training and is standard practice. But formulaic language, legal boilerplate, ritualised openings and closings, and high-frequency constructional patterns are legitimately repetitive in real usage. Aggressive deduplication does not merely remove redundancy; it flattens the frequency profile that corpus-linguistic analysis measures. The review found no dataset paper that reported the effect of its deduplication settings on the frequency distribution of its retained text.

Spoken corpora have historically been the most expensive kind to build, at ratios of transcription time to recording time commonly cited between ten and one and twenty and one. Neural speech recognition has collapsed that cost. The dominant current practice is pseudo-labelling: an existing recogniser transcribes large volumes of untranscribed audio, and the resulting transcripts are treated as training or corpus data after filtering.

The Granary pipeline illustrates both the scale and the caveats. Working across twenty-five European languages, it applied a large multilingual recogniser to roughly 1.06 million hours of audio and retained approximately 643,000 hours after filtration a retention rate near 61%. The project documents the reasons for that attrition explicitly: reduced accuracy in lower-resource languages, a tendency toward hallucinated output, difficulty with noise and non-speech segments requiring separate voice-activity detection, language-identification errors, fixed segment-length requirements, and absence of casing control in the output. Every one of these is a systematic rather than random error source.

The distinction matters. Random transcription error inflates variance and can be handled statistically. Systematic error hallucinated content in low-confidence segments, deletion of disfluencies, normalisation of non-standard forms toward the standard variety, systematic failure on accented or dialectal speech biases exactly the phenomena that spoken-corpus research targets. A pseudo-labelled spoken corpus is a reasonable object for many purposes and a poor one for phonetic, prosodic, disfluency, or variationist analysis unless a human-verified subsample establishes the error profile.

Bitext mining follows a comparable trajectory. Multilingual sentence encoders embed candidate sentences into a shared space and retrieve cross-lingual nearest neighbours, allowing parallel data to be extracted from comparable rather than genuinely parallel sources. The No

Language Left Behind project extended this approach through language-specific student encoders trained against a teacher model, enabling mining for languages that earlier encoders handled poorly.

The project's own results are instructive precisely because they are uneven. For Papiamento, more than seven million mined bitext pairs supported a translation system scoring above 40 BLEU; for Kabuverdianu, mining yielded a system below 5 BLEU. The determining variable, as the authors state, was the amount of available monolingual data mining does not create parallel data where the underlying text does not exist. For several Berber varieties, cross-lingual similarity error rates remained high enough that usable bitext could not be obtained at all.

A more troubling finding concerns what mining retrieves when it does succeed. Analysis of a large mined bitext collection found that multi-way parallel content the same sentence appearing across many languages is disproportionately machine-translated, is of lower quality than two-way parallel content, and is most prevalent precisely in lower-resource languages; the underlying English source material also showed a selection bias toward low-quality content translated at scale. The same study reported that the sentence encoder used for mining exhibits a preference for machine-translation output over human translation. A parallel corpus mined for a low-resource language is therefore liable to be, in substantial part, a record of what machine translation systems previously produced for that language and if it is then used to study translation, or to train a translation system, the circularity is complete.

Table 1. Minimum validation and reporting requirements by pipeline stage

Stage	**Minimum validation requirement**
1. Acquisition	Manual verification of language identification on a random sample per language, with particular attention to lower-resource varieties
2. Filtering	Human review of a random sample of discarded documents to characterise what the filter removes
3. Alignment	Human transcription or verification of a stratified audio subsample; manual adjudication of a bitext subsample
4. Annotation	Expert annotation of a probability sample with known selection probabilities, sized in advance
5. Synthesis	Comparison of generated material against a matched

	attested reference on type–token, lexical, and syntactic diversity measures

Discussion

The single most useful reframing this review supports is that AI has not made corpus construction easier; it has moved the difficulty. Under scarcity, the hard problem was obtaining data, and the discipline built its methodology around principled selection. Under abundance, the hard problem is justifying exclusion, and the discipline has not yet built a corresponding methodology.

This is not a rhetorical distinction. The two problems have different structures. Selection under scarcity is a design act performed by a person who can state their reasons; exclusion under abundance is increasingly performed by a model whose reasons are not statable. The first is documentable by construction; the second is documentable only by measurement.

Contemporary corpus-linguistic theory offers the sharpest available instrument for thinking about this. Representativeness is not a property a corpus possesses in the abstract; it is the extent to which a corpus is adequate for a specified research goal, resting jointly on domain considerations – what the target population of texts is, and how well the corpus covers it and distribution considerations – how much text is required for the linguistic variable in question to be estimated stably.

Conclusion

Artificial intelligence has become a constitutive part of how language corpora are built. It selects what is crawled, decides what is kept, transcribes what is spoken, aligns what is translated, labels what is analysed, and in some projects writes the text itself. This review found that at every one of those stages the technology delivers on capability and shifts the burden to verification and that the verification has largely not been performed.

The central argument is that abundance has changed the discipline's methodological problem without changing its methodological apparatus. Corpus linguistics developed its standards under scarcity, when the hard question was what to collect and the answer could be documented by the person who gave it. Under abundance the hard question is what to exclude, and the answer is increasingly produced by systems whose criteria are not statable in a methods section. The response cannot be to

refuse the tools, which would forfeit the resources that under-served language communities most need. It must be to measure what the tools do through probability-sampled human verification, document-level provenance recording, error reporting broken down by analytic covariate, and honest description of the operational domain actually sampled.

References

1. 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.
2. Dolezal, J., Alam, S., Graham, M., & Bohacek, M. (2026). The impact of AI-generated text on the internet. arXiv preprint arXiv:2604.26965.
3. Egami, N., Hinck, M., Stewart, B., & Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. Advances in Neural Information Processing Systems, 36, 68589-68601.
4. Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.
5. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120.
6. McMillan-Major, A., Bender, E. M., & Friedman, B. (2024). Data statements: From technical concept to community practice. ACM Journal on Responsible Computing, 1(1), 1-17.
7. Mengliev, D., Barakhnin, V., Abdurakhmonova, N., & Eshkulov, M. (2024). Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation. Data in Brief, 54, 110413.
8. Penedo, G., Kydlíček, H., Ben Allal, L., Lozhkov, A., Mitchell, M., Raffel, C., von Werra, L., & Wolf, T. (2024). The FineWeb datasets: Decanting the web for the finest text data at scale. In Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track.
9. 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).

10. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755-759.
11. Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. Applied Corpus Linguistics, 4(1), 100089.
1. Xia, Y., Mukherjee, S., Xie, Z., Wu, J., Li, X., Aponte, R., … McAuley, J. (2025). From selection to generation: A survey of LLM-based active learning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 14552-14569).
12. Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. International Journal of Corpus Linguistics, 29(4), 534-561.