

Lost in Translation: Structured-Span Corruption in Machine-Translated Multilingual Web Corpora

Ruiyang Zhang

Flock.io & Ryonix Labs Inc.

zhangruiyang36@gmail.com

Abstract

Machine translation of English web text is a default recipe for building multilingual and low-resource corpora. We show that this recipe systematically damages the *structured spans* embedded in web text—URLs, email addresses, numbers and identifiers, code, markup and quotation delimiters—and that fluency metrics do not reveal the damage. Translating 56,163 span-bearing segments from a CC-MAIN-2026-30 Common Crawl sample into nine typologically diverse languages with NLLB-200, we find that 12.6% of 73,881 structured spans do not survive, ranging from 91.1% for numbers to 50.6% for quotation delimiters. Damage is strongly language-dependent (18.7 points), but in the *opposite* direction to the obvious prediction: the highest-resource targets are damaged most, an inversion that survives controlling for the systems’ tendency to copy low-resource input through untranslated. Round-trip fluency barely tracks any of this ($\rho = 0.48$ across languages, $\rho = 0.19$; $\rho = 0.25$ per segment). We pre-registered the study, report the refuted hypothesis alongside the supported ones, and release the annotated parallel-span corpus with an independent re-derivation of every number.

1 Introduction

For most of the world’s languages there is not enough native web text to build a large corpus, so a common substitute is to machine-translate English web text into the target language. The resulting corpora are then filtered and validated with corpus-level quality metrics: BLEU (Papineni et al., 2002), chrF (Popović, 2015), or learned estimators such as COMET (Rei et al., 2020) and its reference-free variant (Rei et al., 2022).

Those metrics score a translation by comparing surface forms in bulk—against a reference, or, when none exists, against the source via back-translation. Web text, however, is not only language. It is dense with *structured spans* that carry

meaning through their exact surface form: a URL, an email address, a product code, a version string, an HTML entity, a quoted phrase. A translation can score well and still have silently rewritten a URL, converted digits to another numeral system, or dropped a quotation mark. Aggregate surface metrics are, by construction, nearly blind to this: a handful of altered characters barely moves chrF, but it is the difference between a resolvable URL and a dead one.

We quantify that gap. From a CC-MAIN-2026-30 Common Crawl WET sample we draw 5,000 English documents, extract 73,881 structured spans of six frozen types with documented detectors, and translate the 56,163 span-bearing segments into nine languages with NLLB-200 (NLLB Team et al., 2024), plus a second, LLM-based system as a robustness check.

Three hypotheses were registered before any translation was run. **H1** (span damage): translation does not preserve embedded structured spans, and the loss rate depends on the span type. **H2** (language dependence): preservation varies across target languages, with lower-resourced targets losing more. **H3** (blindness): corpus-level quality scores do not predict span damage. Each was given a falsification criterion in advance. H1 and H3 are supported; **H2 is refuted, and inverted**—the highest-resource targets are damaged most. We report the refutation rather than reframe it (§4).

Our contributions are:

- A **span-typed, language-dependent corruption measurement** on real web text: preservation rates by span type and target language.
- A **corruption taxonomy** that names how spans break—deletion, script transliteration, digit alteration, truncation, casing, delimiter substitution—and its distribution across languages.
- The **blindness result**: the relationship (or lack

of one) between round-trip chrF2—source against back-translation—and span preservation, reported as measured.

- A **released annotated parallel-span resource**—source span, translated segment, preservation verdict and corruption label—together with the frozen detectors, the scoring code, and an independent verifier that re-derives every number in this paper from the raw translations.

The study was pre-registered before any translation was run; the pre-registration, a deviations log, and all code are released with the resource.¹

2 Related Work

Web corpora and their quality. Building corpora from the web is the founding concern of this workshop’s community (Kilgariff and Grefenstette, 2003; Baroni et al., 2009). Common Crawl now underpins C4 (Raffel et al., 2020), CCNet (Wenzek et al., 2020), OSCAR (Ortiz Suárez et al., 2020) and FineWeb (Penedo et al., 2024). Quality auditing of the multilingual portions of these resources has found serious problems: Kreutzer et al. (2022) audit web-crawled multilingual datasets and report substantial fractions of mislabelled or non-linguistic content, and MADLAD-400 (Kudugunta et al., 2023) documents extensive manual auditing. That line of work asks whether the *text* is good; we ask whether the *non-linguistic material inside* otherwise good text survives.

MT quality and its failure modes. Low-resource MT quality is uneven even in strong multilingual systems (NLLB Team et al., 2024), and hallucination and omission are well documented (Guerreiro et al., 2023). Round-trip translation (Sennrich et al., 2016) is a common cheap proxy for adequacy. Our concern is orthogonal to overall adequacy: a segment can be adequately translated as language while its embedded spans are destroyed.

Data documentation. Data statements (Bender and Friedman, 2018) and datasheets (Gebu et al., 2021) argue that resources should ship with explicit accounts of provenance and limitation. We propose span preservation as a concrete, cheap statistic that MT-built corpora should report alongside fluency.

¹Corpus, detectors, scoring code, pre-registration, deviations log and the independent verifier: <https://github.com/ryonzhang/wac13-span-corruption>

3 Method

Corpus. We sample the CC-MAIN-2026-30 Common Crawl WET release (extracted plain text), using 100 WET files chosen by seeded random selection from the snapshot manifest, and 5,000 documents drawn by seeded reservoir sampling within those files. We keep records whose first WARC-Identified-Content-Language is eng, of 500–5000 characters, containing at least one detected span, deduplicated on normalised content hash.

We use WET rather than WARC deliberately. WET is what MT-based corpus pipelines actually consume, it is reproducible from a snapshot id plus a seed with no HTML-extraction confound, and code and markup still occur in it as literal text in tutorials, documentation and forum posts. The cost is that WET discards page structure, so “markup” here means markup-as-content rather than page scaffolding (§5).

Span detectors. We detect six types—URL, EMAIL, NUMBER, CODE, MARKUP, QUOTE—with deterministic regular expressions plus validators. Spans are non-overlapping, resolved by a fixed priority order, so `<a href="http://x.com">` is a single MARKUP span rather than a MARKUP/URL/QUOTE pile-up. CODE covers source-code fragments and code-like identifiers (snake_case, dotted paths, call forms); NUMBER requires at least three characters so that trivially preserved single digits do not dominate.

Detectors were developed on a 400-document development split drawn from WET files disjoint from the evaluation files. We manually adjudicated a stratified sample of 204 detected spans and 25 exhaustively hand-marked windows, made a single revision pass, then froze the detectors. Measured precision is 98.5% and recall 91.4% overall (Table 2).

Unit of analysis. The unit is the *span-bearing segment*. Documents are split at newlines and sentence ends—never inside a detected span—and hard-capped at 350 characters at a span-safe boundary. Because NLLB is a sentence-level system, segments are translated independently of their neighbours, so restricting translation to span-bearing segments produces output for those segments that is identical to translating the whole corpus, at roughly one eighth of the compute. Since our splitter never

cuts a span, any breakage we observe is attributable to translation rather than to preprocessing.

Systems. The primary system is NLLB-200-distilled-600M with deterministic decoding (beam 4, no sampling), translating all 43,981 unique segments into nine languages: French, German, Chinese, Arabic, Hindi, Swahili, Icelandic, Maltese and Yoruba. As a robustness check against system-specific behaviour we additionally translate a 10% seeded subsample with an instruction-tuned LLM (Llama-3.3-70B-Instruct, temperature 0).

Matching. Both sides are NFKC-normalised, stripped of zero-width characters and whitespace-collapsed; URL and EMAIL are compared case-insensitively. Every normalisation choice is deliberately lenient—it can only make a span easier to count as preserved—so the damage we report is a lower bound. For the five verbatim types, a span occurrence is preserved if and only if its normalised form occurs in the normalised translation, counted as $\min(\text{count}_{\text{src}}, \text{count}_{\text{tgt}})$.

QUOTE is scored differently, on **delimiter integrity**: the text inside a quotation is natural language and is supposed to change, so requiring verbatim preservation would be wrong. A QUOTE span counts as preserved if and only if a matched delimiter pair of the same class survives.

Metrics. We report (i) preservation by span type $\times$ language with Wilson intervals, plus a script-normalised variant that folds non-ASCII digits; (ii) a corruption taxonomy over non-preserved spans, assigned by deterministic rules in fixed priority order; (iii) fluency as chrF2 between the source segment and its NLLB round-trip back-translation, restricted to segments of at least eight words, together with a span-masked variant; and (iv) a parse proxy measuring whether a span is still recoverable and syntactically valid after translation.

4 Results

Span damage is large and strongly type-dependent. Across 73,881 spans and nine languages, 12.6% of span occurrences do not survive translation. The rate depends far more on span type than any fluency score would suggest: NUMBER survives at 91.1% and EMAIL at 84.1%, but CODE at only 60.0% and QUOTE at 50.6%. H1 is supported.

QUOTE is the flattest and the worst: preservation sits near 50.6% in *every* target language (Fig-

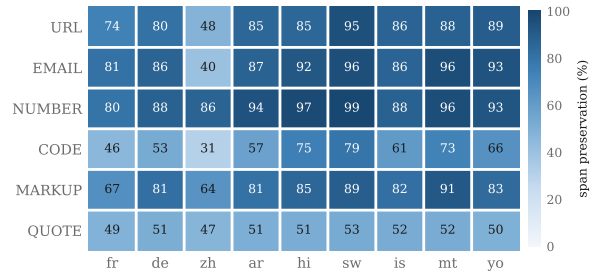


Figure 1: Span preservation (%) by span type and target language, NLLB-200-distilled-600M, 73,881 spans. Darker is better.

ure 1). Quotation-mark integrity is essentially independent of the target language, which suggests a decoder-level rather than a language-specific cause.

Language dependence is large, but the direction is the opposite of what we predicted. Preservation ranges 18.7 points, from 76.7% (fr) to 95.4% (sw). The ordering, however, *inverts* H2: the high-resource targets are damaged most (fr 76.7%, zh 79.6%, de 84.3%) and the lower-resource ones least (sw 95.4%, hi 93.4%, mt 92.7%). **H2 as pre-registered is refuted, with the sign reversed.**

The obvious confound is non-translation: if the model simply copies English through for low-resource targets, their spans survive trivially. Copy-through is indeed about twice as common there (32.5% for sw against 15.0% for fr and 7.0% for zh). Restricting to units the model actually rewrote lowers every language by 2–5 points and leaves the ordering intact (fr 73.0%, zh 78.0%, sw 93.4%), so the inversion is a property of translation behaviour, not of its absence. Chinese is the limiting case: it has the lowest copy-through (7.0%) — it rewrites most aggressively — and destroys URLs, emails and code worst of all. Span damage tracks how much the system rewrites.

Failures are dominated by deletion, not transliteration. Of 81,512 failures, 46.8% are outright *deletion* and 18.0% truncation or splitting. Script transliteration — the failure mode a digit-system argument would predict, and the one most visible in prior controlled studies — accounts for just 0.3% overall (5.0% for Hindi, 0.3% for Arabic). Spans are overwhelmingly not converted; they are lost.

A further 6.6% are delimiter substitutions, many of which are locale-*correct* typography ("..." to «...»). The same caveat applies to numbers: French NUMBER scores 79.6% under the strict matcher but 90.4% once separators and numeral

systems are ignored, because 14,000 becomes 14 000. This is good translation and unusable data, and we report both brackets rather than choosing one.

Round-trip similarity does not predict span damage (H3). Round-trip chrF2 averages 65.7 and spans 20.8 points across languages, but it barely tracks preservation. Across the nine language means, Spearman $\rho = 0.48$ ($p = 0.19$, $n = 9$): not significant, and below the $|\rho| \geq 0.6$ threshold we fixed for treating fluency as a usable predictor. Per unit, over 111,573 scored segments, $\rho = 0.25$ — certain in sign given the sample size, but explaining roughly 6% of variance. Masking spans before scoring, which removes the mechanical coupling whereby damaged spans depress round-trip chrF2, lowers it further to $\rho = 0.21$. **H3 is supported:** fluency is detectably but negligibly informative about span damage. Figure 2 shows the dispersion: Arabic and French have near-identical chrF2 yet differ by five points of preservation.

Downstream cost. Of the spans that were syntactically valid before translation, 87.0% remain recoverable and valid afterwards. For code the figure is 59.8% and for quotations 48.3%. A pipeline that re-extracts URLs, addresses or code from an MT-built corpus should expect to lose roughly one span in eight, and far more for the structured types.

Second system: the effect replicates, the ranking does not. On the 10% subsample, Llama-3.3-70B preserves 82.6% of spans overall — the same order of damage as NLLB, with much lower copy-through. The per-language profile, however, differs sharply: it is markedly better on Chinese (88.4% against 79.6%) and markedly worse on Hindi (68.1% against 93.4%). Structured-span damage is therefore a property of MT-based corpus construction rather than of one model, but *which* languages suffer is system-specific and cannot be read off a published table.

5 Conclusions and recommendations

Report span preservation next to the quality score. Span preservation is cheap to compute, needs no reference translation and no model, and measures what aggregate surface metrics cannot see. Any release of an MT-built corpus should report a preservation rate per span type and target language beside its quality figure; our detectors and scoring code are released for that purpose.

	NLLB-200			Llama
	Preserv.	Transl. only	chrF2	Preserv.
fr	76.7	73.0	68.1	73.5
de	84.3	79.8	68.7	80.6
zh	79.6	78.0	53.8	88.4
ar	89.9	87.9	61.6	81.3
hi	93.4	91.7	70.8	68.1
sw	95.4	93.4	74.5	89.0
is	84.9	80.6	64.2	83.7
mt	92.7	89.5	66.7	91.2
yo	89.7	84.9	62.7	87.3

Table 1: Per-language span preservation (%) for both systems, against round-trip fluency. “Transl. only” excludes units the model returned unchanged. Higher-resource targets are damaged *more*, and fluency does not order the columns.

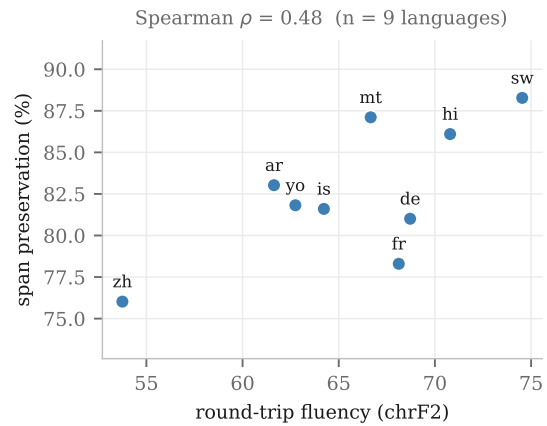


Figure 2: Round-trip fluency does not track span preservation across target languages.

Protect high-value spans rather than repair them. Because our segmenter never splits a span, the damage happens inside the translation model, not around it, so the remedy is not post-hoc repair but the placeholder discipline standard in commercial localisation and largely absent from corpus building: mask detected spans, translate, restore. It also converts an unbounded problem into a bounded one—detector coverage—which can be measured (Appendix A).

State per-language caveats, especially for low-resource targets. A single corpus-level quality claim hides the variation we observe across target languages—and the languages with the least native web text are exactly those most likely to be built by translation, so the caveat matters most where it is least often given.

Is the damage necessarily harmful? A reviewer raised the reasonable counter-hypothesis that losing URLs and email addresses is a side benefit,

scrubbing personal information from training data. The question separates by span type. For EMAIL the point has force, though corruption is a poor substitute for a filter that can be audited. For the types carrying most of the mass it does not: NUMBER is 62,157 of our 73,881 spans, and a silently altered quantity, version string or identifier is not anonymised but *wrong*—undetectably so, because the segment still reads well. The useful distinction is removal, which is visible and can be a policy choice, against alteration, which is neither: 46.8% of failures are deletions, but 3.8% are digit alterations and 22.7% other alterations.

Distinguish reformatting from destruction.

Not every non-preserved span is an error: locale-correct rewriting—14,000 to 14 000, "... " to «...»—is good translation and bad exact-match data. Reporting the strict rate with a lenient bracket (digit-sequence recovery) separates the two and keeps the headline number honest.

Conclusion. Translating English web text to build a corpus in another language damages 12.6% of the structured spans it contains, and the corpus-level scores used to validate such corpora do not reveal it. The damage is not uniform: it concentrates by span type, from 91.1% preservation for numbers down to 50.6% for quotation delimiters, and it varies by target language in the direction opposite to the one we pre-registered. None of this argues against building corpora by translation; it argues that span preservation belongs next to the quality score when such a corpus is released, and that masking spans before translation is the cheap fix. We release the annotated corpus, the detectors and a verifier so both can be checked.

Limitations

Systems and settings. We test one dedicated NMT model and one instruction-tuned LLM, at fixed deterministic decoding settings. Larger NLLB variants, other architectures, and sampling-based decoding may behave differently. The second system runs on a 10% subsample and is a robustness check, not a co-headline.

English source only. All source text is English. Translation out of other source languages, and pivot translation, are untested; the span inventory of non-English web text also differs.

Detector coverage. Our detectors are frozen, documented and measured (precision 98.5%, recall 91.4% on a held-out development split), but they are regular expressions, not an oracle. Straight single-quoted strings, numeric tokens under three characters, and digit runs followed by letters are excluded by definition; version strings and hyphenated package filenames are known recall gaps. Missed spans do not enter the denominator, so coverage limits what we can see rather than biasing the rates we report.

WET extraction. WET is plain text, so HTML structure is gone before we see it. “Markup” here therefore means markup appearing as content — entities and literal tags in tutorials and documentation — and MARKUP is much the rarest of our six types (359 spans), with correspondingly wide intervals. Measuring page scaffolding would require WARC and an extraction step that would itself confound the comparison.

Segment-level translation. Units are segments, not documents. This matches how sentence-level NMT is actually applied and makes our numbers identical to full-corpus translation for those segments, but it means we cannot observe document-context effects.

Parse proxy is syntactic. We test whether a span is recoverable and well-formed after translation, not whether a URL still resolves or an address still receives mail. Live resolution would be neither reproducible nor kind to the hosts involved.

Ethics Statement

The source text is derived from Common Crawl and may contain personal data such as the email addresses our EMAIL detector is designed to find. We release span annotations and verdicts together with the document URLs and WET file identifiers needed to reconstruct the sample, and we do not construct or distribute any aggregated contact list. A random web sample also contains adult and spam content; we apply no topical filtering, since filtering on content would bias the span inventory we set out to measure.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. [The WaCky wide web: a](#)

- collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
- Emily M. Bender and Batya Friedman. 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.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Communications of the ACM*, 64(12):86–92.
- Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. [Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.
- Adam Kilgarriff and Gregory Grefenstette. 2003. [Introduction to the special issue on the web as corpus](#). *American Journal of Computational Linguistics*, 29(3):333–348.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmunkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. [Quality at a glance: An audit of web-crawled multilingual datasets](#). *Transactions of the Association for Computational Linguistics*, 10:50–72.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [MADLAD-400: A multilingual and document-level large audited dataset](#). In *Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track*.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. [A monolingual approach to contextualized word embeddings for mid-resource languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 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*.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. [CCNet: Extracting high quality monolingual datasets from web crawl data](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

A Detector quality

Span type	Precision	Recall
URL	95.0	100.0
EMAIL	100.0	100.0
NUMBER	100.0	92.0
CODE	97.5	71.4
MARKUP	100.0	–
QUOTE	100.0	100.0
Overall	98.5	91.4

Table 2: Detector quality (%), hand-adjudicated on the development split (204 sampled spans; 96 detected and 9 missed spans in 25 exhaustively marked windows).