187 Languages on Paper: An Audit of a Machine-Translated Multilingual Sensitive-Word Lexicon

Daozhendong Yang
School of Intelligent Science and Engineering,
Yunnan Technology and Business University, Kunming, China
hi@yxpil.com
https://github.com/yxpil/multilingual-sensitive-words

Abstract

A keyword list is the cheapest fallback for content moderation when a classifier is unavailable, and a machine-translated keyword list is the cheapest way to cover many languages. This paper audits one such artifact: a public lexicon that advertises 187 languages, five categories, and 2,951,608 entries. Every advertised language directory exists, every one of the five category files is present, and the JSON parses without error. That is where the simple story ends. Of the 2,951,608 entries, 2,685,296 are copies of the same 14,594-term URL list, passed through unchanged in every language, and 218,960 are semantic translations of 1,190 Chinese source terms in the 184 non-Chinese targets; the remainder belongs to the three Chinese variants. Only 70.9% of the semantic target entries differ from their source. The non-Chinese targets split into two sharply separated groups: 113 languages with 0.71% identity on Han-containing source entries, and 71 languages with 72.2%. The partial group is not a set of 71 different failures. Its members leave almost exactly the same source entries untranslated—the pairwise Jaccard similarity to the modal failure set is 99.98% for political terms and 99.62% for porn terms—and only 18 distinct failure signatures appear across the 71 languages. The pattern is consistent with a pipeline that stopped on specific inputs, not with 71 languages that are individually hard to translate. An expected-script check separates real script coverage from superficially non-empty output: 13 languages exceed 98.5%, while Japanese kana reach 64.6% (an all-kanji lower bound) and Serbian reaches 27.9%. A duplicate analysis finds 5,933 excess duplicate occurrences, and 26,681 translated entries (17.2%) are shared verbatim across at least two languages, rising to 46.2% for political terms. A 40-pair inspection sample of benign advertising terms finds 27 accurate, 7 acceptable, 4 wrong-sense, and 2 other outcomes. The paper does not print the sensitive source strings; it reports the audit method, the coverage arithmetic, and what a maintainer should fix first.

Artifact. The lexicon, the audit script (`scripts/audit.py`), the table generator (`scripts/make_tables.py`), the figure generator, the generated `numbers.tex`, and the raw `audit.json` are public at https://github.com/yxpil/multilingual-sensitive-words. The audited commit is `1a8de85` (2026-10-18). The lexicon is Apache-2.0; the Chinese source lists are derived from an upstream open-source list, as the repository states. The audit uses only the Python standard library and performs no network access. This paper intentionally does not reproduce the moderated terms; the only strings shown are benign advertising vocabulary in Table 5.

1 Introduction

Content moderation systems often combine a classifier with a keyword fallback: if the classifier is uncertain or unavailable, a hit in a sensitive-word list rejects the content, and a miss approves it. The fallback is attractive because it is cheap, auditable, and explainable. It is also only as good as its coverage: a list that covers Chinese but not the languages a site actually serves silently approves content in those languages.

The repository audited here addresses that gap by machine-translating a Chinese sensitive-word list into 187 languages. It is a reasonable strategy for a small project, and the repository is candid that the translations are machine-generated and that URLs are passed through unchanged. The question this paper asks is not whether the strategy is sound, but what the artifact actually contains after translation. The answer matters to anyone who would deploy it: "187 languages" can mean 187 directories, 187 effective lists, or something in between.

We audit the artifact at one commit with four independent signals: *identity* (the target string equals the source string after normalisation), *expected script* (for languages with a distinctive orthography, the output contains that script), *vocabulary collapse* (duplicates within a language and verbatim sharing across languages), and a small *human inspection sample* of benign terms. Identity alone is not a perfect untranslated detector—a proper noun or a borrowed word can legitimately be identical—so the paper treats it as a signal with a known false-positive rate, and the sample estimates that rate on a benign subset. Script match is independent of identity and catches output that is non-identical but in the wrong language. Duplicate analysis catches a different failure: a translator that returns the same target for several distinct source terms, which

shrinks the effective vocabulary even when every entry is non-identical.

Contributions.

1. *A complete inventory of a 187-language resource.* All 187 claimed directories exist with all five files, no empty files, and no parse errors; the artifact contains 2,951,608 entries (Section 4.1).
2. *Entry-level coverage arithmetic.* 92.5% of the entries per language are the same URL list, passed through unchanged; semantic translation covers 70.9% of the 218,960 semantic target entries. A coverage figure that counts entries without separating URLs overstates translation by an order of magnitude (Section 4.2).
3. *A bimodal population with one shared failure set.* 113 languages have 0.71% identity on Han entries; 71 have 72.2%. The partial group's failures are nearly identical across languages (Jaccard 99.62%–99.98% to the modal set; 18 signatures), which points to a fixable pipeline defect rather than a language-specific difficulty (Section 4.3).
4. *Script and collapse diagnostics.* The expected-script check separates 13 clean languages from Japanese (64.6%, lower bound) and Serbian (27.9%); duplicate and cross-language collision counts quantify vocabulary collapse (Sections 4.4 and 4.5).
5. *A small, honest quality sample.* On 40 benign advertising terms in seven high-resource Latin-script languages, 27 are accurate, 7 acceptable, 4 wrong-sense, and 2 other; the sample is too small to estimate a rate but large enough to show that non-identity does not imply correctness (Section 4.7).
6. *Maintenance recommendations.* The audit identifies the entries that should be re-translated first, and it argues that coverage claims should be reported per entry and per script rather than per directory (Section 5).

2 The artifact

The repository has five Chinese source files and a directory per target language. The source categories are porn, political, advertising, weapons, and URLs. The first four are translated; the URL list is copied verbatim by design. Each language directory contains five JSON arrays whose indices align with the source arrays. The source has 15,784 entries in total, of which 1,190 are semantic (the four translated categories) and 14,594 are URLs. The source is not perfectly deduplicated: it contains 1,161 distinct semantic strings, so some source terms repeat.

The advertised list contains 187 language codes, mixing two-letter and three-letter ISO 639 codes (130 and 57 respectively) and three Chinese variants: `zh` for the source, `cht` for traditional Chinese, and `yue` for Cantonese. The audit excludes the three Chinese variants from translation-quality rates, because identity is expected when the target shares the source script; it reports them separately.

Ethics and dual use. Sensitive-word lists are used for censorship as well as for spam and abuse filtering, and the categories include political terms. This paper does not evaluate whether any term should be moderated, does not reproduce the terms, and does not rank languages by political content. It audits translation coverage, a property that is independent of the merits of the list. The inspection sample is restricted to benign advertising vocabulary so that the paper's examples are publishable.

3 Method

3.1 Inventory and alignment

The audit reads `languages.txt` and the directory tree, compares the two sets, and checks that every language has all five files. It parses every JSON array and records parse errors and empty files. It aligns each target array with its source array by index and records whether the lengths match; a language whose lengths differ is excluded from rate calculations for that category, and the exclusion is reported rather than silently dropped.

3.2 Identity

For each aligned pair, the target is normalised to NFC and stripped of outer whitespace, and the pair is *identical* if the result equals the similarly normalised source. Rates are computed over the 184 non-Chinese target languages. Two views are reported: all entries, and the subset whose source contains at least one Han character. The second view is the one used for the main result, because non-Han source entries (product names, Latin acronyms, numbers) can legitimately remain unchanged.

Identity is a conservative signal. It cannot detect a translation that is wrong but non-identical, and it counts a legitimately unchanged proper noun as a failure. We quantify the second error on a benign sample (Section 4.7) and treat the first as a limitation.

3.3 Expected script

For fifteen languages with a distinctive orthography (Japanese kana, Cyrillic for Russian and Serbian, Arabic, Hangul, Thai, Devanagari, Hebrew, Greek, Georgian, Armenian, Bengali, Tamil, Myanmar, Ethiopic), the audit checks whether a Han-containing source entry's target contains at least one character from the expected script. For Japanese the check is deliberately conservative: an all-kanji translation contains no kana and is counted as a miss, so the Japanese rate is a lower bound. For Serbian the check detects Latin-script output in a language that is also written in Cyrillic; the rate is a coverage indicator, not an error count, because Latin-script Serbian is a legitimate orthographic choice. The script check is independent of identity: it can pass on a non-identical but correct translation and fail on a non-identical translation in the wrong language.

3.4 Duplicates and cross-language sharing

Within each language and category, the audit counts duplicate target strings after normalisation, subtracts the duplicates already present in the source, and reports the excess. Cross-language sharing counts a translated entry when the same target string appears for at least two languages at the same source index. This is a lower bound on duplication: a translator can return the same word with different diacritics, and the audit does not cluster those.

3.5 Human inspection sample

A fixed vocabulary of 40 benign advertising source terms (recruitment, shipping, download, customer service, and similar) is sampled without replacement across seven high-resource Latin-script languages (`en`, `de`, `fr`, `es`, `pt`, `it`, `nl`), giving 280 possible pairs; 40 are drawn with seed 20261020. The author labelled each pair as *accurate*, *acceptable* (literal but usable), *wrong sense*, or *other* (proper noun, transliteration, unreviewed). The labels are subjective and single-annotator; they are not used in any quantitative metric and are reported as a sanity check, not as an evaluation.

4 Results

4.1 Inventory: all directories, unequal content

Table 1 gives the inventory. The claimed and actual directory counts agree exactly (187 and 187), with 0 missing and 0 unlisted directories. Every language has all five files (187), there are 0 empty files, 0 missing categories, and 0 parse errors. The artifact is structurally complete. It contains 2,951,608 entries. Across the 184 non-Chinese targets, 218,960 are semantic and 2,685,296 are URL copies; the remaining entries belong to the three Chinese variants. The URL share is not a minor accounting detail: it is 92.5% of the entries in each non-Chinese language.

The per-category view shows the asymmetry. The URL category has 14,594 source entries and 2,685,296 target entries, with 100.0% identity—the documented pass-through. The four semantic categories have 1,190 source entries and 218,960 target entries; 155,163 of the latter differ from the source, a rate of 70.9%. Identity over all semantic entries is 29.1%; restricted to Han-containing source entries it is 28.3%. In other words, roughly seven of ten semantic entries have some translation, and slightly less than three of ten are byte-identical to the Chinese source.

4.2 Coverage is bimodal, not uniform

A per-language mean identity rate is bimodal (Figure 1). 113 languages have a mean identity on Han entries at or below 50%; their aggregate identity is 0.71% and their per-category rates are 0.05%, 2.33%, 0.02%, and 0.18%. 71 languages have a mean above 50%; their aggregate identity is 72.2% and their per-category rates are 43.4%, 90.3%, 18.1%, and 93.4%.

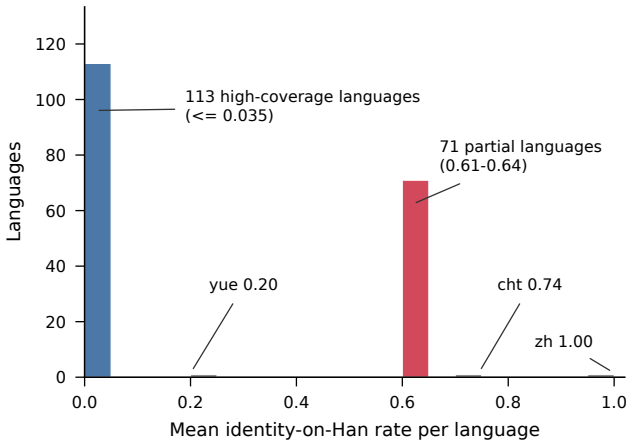


Figure 1: Distribution of the per-language mean identity-on-Han rate. The population is bimodal: a high-coverage group near zero and a partial group near 0.6–0.7, with little mass in between.

The two groups partition the 184 non-Chinese targets. Table 2 gives the group breakdown, including the three Chinese variants, which are shown for completeness but excluded from the partition.

The partial group is not a marginal tail. Its members include languages with substantial speaker populations and Latin-script targets that are easy to inspect: the maximum Latin-script identity is 72.1% for `no`, and 22 of the 24 Latin-script targets are below 1% identity on Han entries. A Latin-script language with a high identity rate is therefore not an artifact of script sharing.

Figure 2 shows the per-language, per-category identity matrix sorted by group. The partial group is dark across all four semantic categories, with political and weapons the darkest. The high-coverage group is near white, with a few category-specific exceptions.

4.3 The partial group shares one failure set

The most useful result is not the size of the partial group but its structure. For each partial language, the audit computes the set of source indices that were left untranslated. Table 3 reports per-entry language coverage; Figure 3 shows the ECDF of how many of the 184 targets translate each source entry. The distribution is not smooth: a large mass of entries is translated by almost all targets, and a small mass is translated by few. That shape is the signature of entry-specific failures.

The overlap analysis makes the point directly. For porn, the modal untranslated set has 132 entries, or 43.4% of the category; the mean Jaccard similarity of a partial language's failure set to that mode is 99.62%, with a minimum of 89.2%. For political the modal set is 295 entries (90.5%), with mean Jaccard 99.98% and minimum 99.3%. For advertising the corresponding values are 27 (22.0%), 99.31%, and 81.8%; for weapons they are 408 (93.4%), 99.97%, and 99.0%. Across all four categories, 52 of the 71 partial languages have exactly the modal

Table 1: Inventory and surface translation signals by category. Target columns cover the 184 non-Chinese target directories (the source `zh` and the two Chinese variants `cht` and `yue` are excluded from translation rates). *Identity (all)* counts every entry; *Identity (Han)* is restricted to source entries containing at least one Han character. *Median langs* is the median number of the 184 targets whose string differs from the source; *<50%* counts Han-containing source entries translated in fewer than 92 targets. URLs are passed through unchanged by design.

Category	Source	Distinct	Target	Ident. all	Ident. Han	Median langs	<50%
Porn	304	304	55,936	17.1%	16.8%	183	0
Political	326	303	59,984	37.6%	36.3%	113	7
Ads	123	120	22,632	11.5%	7.0%	184	0
Weapons	437	434	80,408	36.1%	36.1%	113	0
URLs	14,594	14,594	2,685,296	100.0%	–	0	14,594
All semantic	1,190	1,161	218,960	29.1%	28.3%	–	7

Table 2: Language groups and identity-on-Han rates. The 71-language partial group and the 113-language high-coverage group partition the 184 non-Chinese targets. The two Chinese variants and the source language are shown for completeness but are not counted in either group.

Group	n	Porn	Political	Ads	Weapons	Overall
Source (`zh`)	1	100.00%	100.00%	100.00%	100.00%	100.00%
Variant (`cht`)	1	69.31%	93.73%	41.03%	93.59%	82.14%
Variant (`yue`)	1	27.72%	26.02%	23.08%	3.20%	17.69%
Partial group	71	43.41%	90.30%	18.13%	93.39%	72.19%
High-coverage group	113	0.05%	2.33%	0.02%	0.18%	0.71%
All non-Chinese targets	184	16.78%	36.28%	7.00%	36.14%	28.29%

Table 3: Per-entry language coverage. For each Han-containing source term we count how many of the 184 non-Chinese targets return a string different from the source. The URL row counts non-Han terms and is included as a reference for the documented pass-through.

Category	Terms	Median	Mean	Min	All 184	<50%
Porn	303	183	153.1	106	137	0
Political	319	113	117.2	0	24	7
Ads	117	184	171.1	112	82	0
Weapons	437	113	117.5	111	22	0
URLs (by design)	14,594	0	0	0	0	14,594

failure set, and the 71 languages produce only 18 distinct failure signatures.

This matters for maintenance. If the failures were language-specific, fixing coverage would require translating all four categories into 71 languages, a task of roughly 71 times the source vocabulary. If the failures are entry-specific, the maintainer can re-translate the modal failure sets—on the order of 132 porn terms, 295 political terms, 27 advertising terms, and 408 weapons terms—and recover most of the missing coverage. The audit cannot prove the cause, but the concentration is strong evidence for the second hypothesis. The repository's own description—machine translation through a web interface with fallbacks—is consistent with per-request failures that were written back as the source string.

Table 4: Expected-script match for 15 languages with a distinctive orthography. Each cell is the fraction of Han-containing source entries whose returned string contains at least one character of the expected script. `ja` counts kana, so all-kanji entries are false negatives and its rate is a lower bound; `sr` belongs to the partial group and is included as a contrast.

Code	Script	Overall	Porn	Political	Ads	Weapons
`ko`	Hangul	99.7%	100.0%	99.1%	100.0%	100.0%
`th`	Thai	99.7%	100.0%	99.4%	99.1%	100.0%
`ar`	Arabic	99.5%	100.0%	98.1%	100.0%	100.0%
`ru`	Cyrillic	99.2%	99.3%	98.7%	97.4%	100.0%
`hi`	Devanagari	99.1%	99.7%	98.4%	99.1%	99.3%
`ta`	Tamil	99.1%	100.0%	97.2%	99.1%	100.0%
`he`	Hebrew	99.1%	99.7%	97.2%	100.0%	99.8%
`bn`	Bengali	99.1%	99.7%	97.2%	100.0%	99.8%
`ka`	Georgian	99.0%	100.0%	97.8%	95.7%	100.0%
`el`	Greek	98.9%	99.7%	97.2%	97.4%	100.0%
`am`	Ethiopic	98.8%	99.7%	96.6%	98.3%	100.0%
`my`	Myanmar	98.7%	99.0%	97.5%	97.4%	99.8%
`hy`	Armenian	98.5%	99.3%	97.2%	94.9%	99.8%
`ja`	Kana	64.6%	66.0%	33.5%	64.1%	86.5%
`sr`	Cyrillic	27.9%	56.8%	9.7%	82.1%	6.6%

4.4 Script match

Table 4 reports the expected-script check. Thirteen of the fifteen checked languages exceed 98.5%; Korean and Thai reach 99.7% and 99.7%, Arabic 99.5%, Russian 99.2%, and the remaining Indic and other scripts range from 98.5% to 99.1%. Japanese kana reach 64.6%, but this is a lower bound: all-kanji entries are counted as misses, and the category variation (from 33.5% for political to 86.5% for weapons) reflects how often the translation uses kana rather than kanji. Serbian reaches 27.9%, consistent with its partial status; Serbian is written in both Cyrillic and Latin, so the check undercounts correct Latin-script output but still separates it from the other partial languages.

The script check is the audit's best defense against a coverage claim based on non-empty files. A target array can be non-empty, aligned, and non-identical to the source while still being in the wrong language—for example, an English string in a Japanese array, or a transliteration that contains no expected script. Identity would not catch it, and a directory-count metric

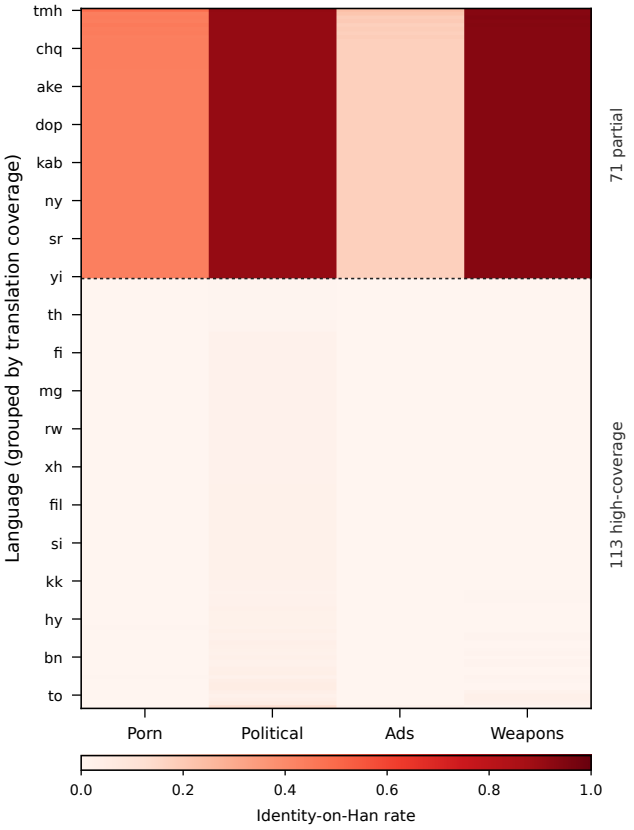


Figure 2: Identity-on-Han rate by language and category, languages grouped by translation coverage. All 184 target languages appear as rows; every tenth row carries a label. The partial group fails across categories; the high-coverage group has small category-specific residues.

would not either. The check is not a translation-quality metric: it says the output is in the right script, not that it means the right thing.

4.5 Duplicates and cross-language collisions

Within-language duplicates are modest in most categories but not in all. The audit subtracts the source's own duplicate occurrences and reports the excess: 5,933 excess occurrences in total, with 2,612 for porn, 1,832 for political, 299 for advertising, and 1,190 for weapons. The duplicate rates are 4.7%, 10.1%, 3.8%, and 2.2%. A duplicate in a keyword list is not merely redundant: it changes the effective vocabulary and can signal that several distinct source terms collapsed to one translation.

Cross-language sharing is larger. Of 26,681 translated semantic entries, 17.2% have the same target string in at least two languages. By category the rate is 10.6% for porn, 46.2% for political, 14.5% for advertising, and 3.1% for weapons. Political terms are the extreme: nearly half of the translated political entries are shared verbatim across languages. Some of this is legitimate (names, acronyms, borrowed terms), and some is

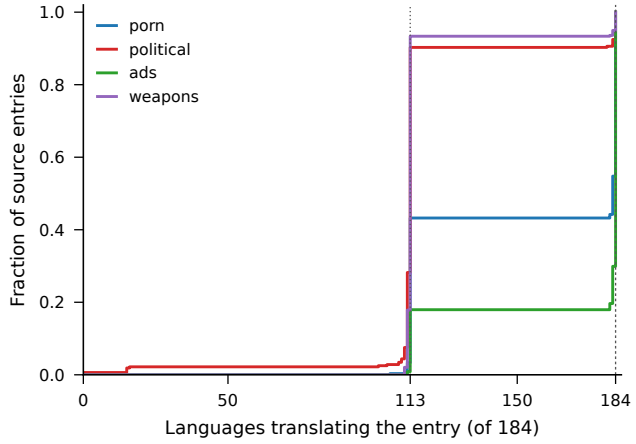


Figure 3: ECDF of the number of target languages that translate each source entry. The two mass points—almost all targets and almost none—are the entry-level counterpart of the bimodal language population.

likely a translator returning a common English string instead of translating. The audit does not distinguish the two; it reports the magnitude and leaves the interpretation to a sample.

The Chinese variants behave as expected for a source-script target: traditional Chinese has 82.1% identity on Han entries, and Cantonese has 17.7%. Cantonese is the least complete variant, with 3.2% identity on weapons terms, which is consistent with a partial translation rather than a different script.

4.6 Anomalies

The audit counts zero-width characters and length outliers. It finds 21 entries that consist only of zero-width characters (51 zero-width spaces, 719 zero-width non-joiners, and 262 zero-width joiners across the corpus), concentrated in Japanese (17), Hindi (2), Arabic (1), and Korean (1). These entries are non-empty and non-identical, so an identity metric would count them as translated, but they carry no moderation signal. The audit also finds 190 entries longer than 100 characters, 187 of which are URLs; the maximum is 763 characters in `ur` in the Porn category. Long entries are not necessarily errors, but they are worth a manual look because a keyword list is most useful when entries are short enough to match.

4.7 Human sample

Table 5 lists the 40 sampled pairs, their English concept glosses, the returned string, and the label. Of the 40, 27 are accurate, 7 acceptable, 4 wrong-sense, and 2 other (proper noun unchanged or transliteration). The wrong-sense cases are informative: a request to click or like becomes a request to leave a rating in German, and a six-digit QQ number becomes the literal "six QQs" in German and Italian. These are not identity failures—the output is correctly non-Chinese—and a coverage metric would count them as successes. The sample is

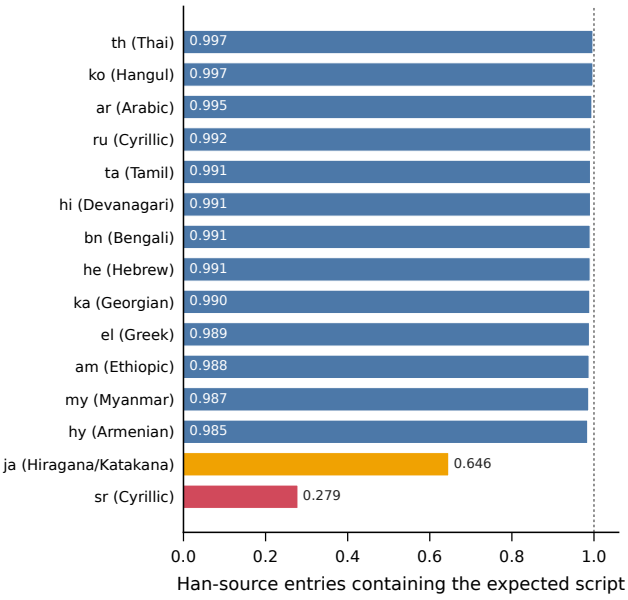


Figure 4: Expected-script match for the fifteen checked languages. Japanese (kana) is a lower bound; Serbian is a partial-group contrast.

too small and too restricted (seven Latin-script languages, advertising terms only) to estimate a corpus-wide accuracy, but it establishes that the gap between "translated" and "correct" is real and measurable.

5 Discussion

Report coverage per entry, not per directory. The artifact's headline number is 187 languages, and all 187 directories are real. But 92.5% of the entries per language are URL pass-throughs, and 71 languages translate only about two-thirds of the Han entries. A maintainer who reports "187 languages" is reporting the directory count, not the coverage. A more useful report has three numbers: languages with all files, languages whose Han identity is below a threshold, and the fraction of source terms translated in each language.

Fix the shared failure set first. The partial group's failures are nearly identical across languages. Re-translating the modal failure sets is a small, bounded task compared with re-translating 71 languages. The audit's per-entry coverage table is the work list; the maximal entries are the ones that fail everywhere.

Separate URLs from semantic coverage. URL pass-through is documented and correct, but mixing it into an entry count inflates coverage by an order of magnitude. The audit reports the two separately, and the paper recommends that the artifact's own coverage statement do the same.

Add a script check to the pipeline. Identity cannot distinguish a wrong-language translation from a correct one. A script check is cheap, needs no model, and catches the most common failure mode for languages with a distinctive orthography. It should be run at translation time, not after publication.

Detect collapse. Cross-language sharing and within-language duplicates are early warnings that a translator is returning a small set of common strings. A maintainer can flag a category when its cross-language sharing rate exceeds a threshold; political terms at 46.2% would have been flagged immediately.

Do not over-read the sample. The 40-pair sample is a sanity check, not an evaluation. It shows that wrong-sense translations exist; it does not estimate their prevalence. A proper evaluation would sample across categories and scripts, use at least two annotators, and report agreement.

6 Limitations

One commit, one artifact. The audit describes commit 1a8de85 (2026-10-18). The repository may change; the JSON and the scripts are the reproducible record.

Identity is a signal, not a label. A legitimately unchanged proper noun is counted as untranslated, and a wrong but non-identical translation is counted as translated. The script check and the sample reduce the first error but do not eliminate it; the second is only visible in the sample.

Alignment assumes index correspondence. The audit aligns target arrays to source arrays by index. If a translator reordered or dropped entries without changing the length, the audit would mis-pair them. The repository's files are generated in a single pass, and the source and target lengths match in every audited language and category, which makes index alignment plausible but not proven.

No semantic evaluation at scale. The audit does not use a translation-quality model, a reference translation, or a bilingual speaker for every language. It cannot say whether a non-identical entry is correct; it can only say whether it is identical and in the expected script.

The sample is small and narrow. Forty pairs, seven Latin-script languages, advertising terms only, one annotator. It is not a prevalence estimate and does not cover the partial group.

Language codes are not validated. The audit takes the directory names as given and does not check that a code denotes the language its files actually contain. Some of the three-letter codes are rare or ambiguous; a speaker-based audit would be needed to confirm the mapping. The codes are compared

Table 5: Author inspection sample: 40 (language, entry) pairs drawn with seed `20261020` from 40 manually vetted benign ads source terms and 7 high-resource Latin-script languages (280 possible pairs). Strings are reproduced verbatim after LaTeX escaping. Assessment labels are the single author's subjective judgement, not a formal evaluation; they are not used in any quantitative metric.

#	Lang	Concept (English gloss)	Returned string	Assessment
1	en	agent	agent	accurate
2	en	please click to enter	Please click to enter	accurate
3	en	free ordering hotline	Free Ordering Hotline	accurate
4	en	franchise chain	Franchising	accurate
5	en	high-definition online	HD Online	accurate
6	en	stream online	Stream online	accurate
7	de	recruitment	Personalrekrutierung	acceptable (literal compound)
8	de	our shop	Unser Geschäft	accurate
9	de	purchasing agent	Kaufvermittler	acceptable
10	de	request to click / like	Bitte geben Sie eine Bewertung ab.	wrong sense (rating instead of click)
11	de	download speed	Downloadgeschwindigkeit	accurate
12	de	six-digit QQ number	Sechs QQs	wrong sense (literal six QQs)
13	de	capital turnover	Kapitalumlauf	acceptable
14	fr	recruitment	Recrutement	accurate
15	fr	purchasing agent	Acheteur mandataire	acceptable
16	fr	payment before shipping	Expédition après réception du paiement	accurate
17	fr	franchise chain	Franchisage	accurate
18	fr	obtain for free	Obtenir gratuitement	accurate
19	fr	buy a small car	Acheter une petite voiture	accurate
20	fr	download speed	Vitesse de téléchargement	accurate
21	fr	stream online	Diffusion en ligne	accurate
22	es	our shop	Nuestra tienda	accurate
23	es	professional agency	Agente profesional	accurate
24	es	details inside	Para más información, acceda aquí	accurate
25	es	funding shortage	Escasez de fondos	accurate
26	pt	part-time work	Trabalho a tempo parcial	accurate
27	pt	QQ (messaging handle)	QQ	proper noun, unchanged
28	pt	interested applicants	Interessados	accurate
29	pt	goods arrived	Recebido	acceptable
30	it	Koukou (QQ transliteration)	Koukou	transliteration
31	it	request to click / like	Per favore, dai un like!	acceptable
32	it	free to use	Uso gratuito	accurate
33	it	text-file download	Scarica il file .txt	accurate
34	it	six-digit QQ number	Sei QQ	wrong sense (literal six QQ)
35	it	capital turnover	Rotazione dei fondi	acceptable
36	nl	our shop	Onze winkel	accurate
37	nl	full-time work	Voltime	wrong form (not a Dutch word)
38	nl	please click to enter	Klik hier om door te gaan	accurate
39	nl	details inside	Raadpleeg de pagina voor meer informatie.	accurate
40	nl	download speed	Downloadsnelheid	accurate

only against the directory listing, not against the ISO 639 registry International Organization for Standardization (2007).

Ethics. The artifact contains political terms and terms for illegal content. The paper does not reproduce them, does not evaluate their appropriateness, and does not provide a deployment recipe beyond the audit. A keyword list can be used to suppress speech; the audit's purpose is to make its coverage claims honest, not to endorse its use. The content harms that deploying such a filter can create are catalogued in the risk taxonomy for language models Weidinger et al. (2022); this paper measures coverage, not harm.

7 Related work

Multilingual data audits are an established response to the gap between a dataset's headline coverage and its actual content. Kreutzer et al. audit web-crawled multilingual datasets and find that language labels, sentence counts, and domain coverage routinely disagree Kreutzer et al. (2022). The same gap appears inside a single released corpus, where documenting what a web-scale collection actually contains is itself a research task Dodge et al. (2021), and the size and provenance of a language resource have been argued to be a risk in their own right Bender et al. (2021). Joshi et al. document the unequal representation of languages in NLP resources Joshi et al. (2020), and Caswell et al. show that language identification in the wild is itself unreliable Caswell et al. (2020). The present audit is narrower—one lexicon, one commit—but it uses the same principle: measure the artifact, not its description. Documentation frameworks such as datasheets Gebru et al. (2021), data statements Bender and Friedman (2018), and model cards Mitchell et al. (2019) make a related argument for recording provenance and intended use; this paper's contribution is a concrete measurement for a resource that has neither.

Translation quality is usually measured with reference-based metrics such as BLEU Papineni et al. (2002), chrF Popović (2015), or learned metrics such as COMET Rei et al. (2020). Those metrics require references or a model; the

audit here has neither, and it therefore uses identity, script, and duplication as no-reference diagnostics. The trade-off is explicit: the audit can say that an entry was left in the source language or written in the wrong script, but it cannot say that a non-identical entry is a good translation. System-level behaviour that no single sentence score captures is a known limit of reference-based evaluation Koehn and Knowles (2017), and the metrics community has warned that aggregate scores can mislead when the systems behind them differ Post (2018).

Content moderation with keyword lists is widely used as a fallback and widely criticized for context blindness. The governance literature discusses the political and technical limits of automated moderation Gorwa et al. (2020); Gillespie (2020), and the platform-side account of moderation describes the organisational context that a purely lexical filter cannot see Gillespie (2018); this paper does not take a position on those debates. It focuses on a narrower question—whether the multilingual list does what its coverage claim says—and leaves the policy question to the deployer.

8 Conclusion

The audited lexicon is structurally complete and substantively uneven. All 187 language directories exist with all five files and no parse errors, but 92.5% of the entries are the same URL list, only 70.9% of semantic entries differ from the Chinese source, and 71 of the 184 non-Chinese languages translate only about two-thirds of Han entries. The partial group is not 71 independent failures: its members share almost the same untranslated entries, with Jaccard similarity above 89.2% in every semantic category and only 18 distinct failure signatures. That is good news for maintenance—the fix is a bounded re-translation of roughly one thousand entries—and bad news for the headline claim, which counts directories rather than coverage. A script check, a duplicate/collision check, and a small human sample add three independent views that identity alone cannot provide. The broader lesson is that a machine-translated lexicon should report coverage per entry, per script, and per category, and that the cheapest audit is often the one that asks what the artifact contains rather than how many languages it lists.

References

E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

E. M. Bender, T. Gebru, A. McMillan-More, and M. Mitchell. On the dangers of Stochastic Parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623. Association for Computing Machinery, 2021.

I. Caswell, T. Breiner, D. van Esch, and A. Bapna. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. arXiv preprint arXiv:2010.14571, 2020.

J. Dodge, M. Sap, A. Marasović, et al. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2021.

T. Gebru, J. Morgenstern, B. Vecchione, et al. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

T. Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven, 2018.

T. Gillespie. Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 2020. doi: 10.1177/2053951720942541.

R. Gorwa, R. Binns, and C. Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7 (1), 2020. doi: 10.1177/2053951720908109.

International Organization for Standardization. ISO 639-3: Codes for the representation of names of languages — part 3: Alpha-3 code for comprehensive coverage of languages. International standard, maintained by SIL International, 2007.

P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. 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 (ACL)*. Association for Computational Linguistics, 2020.

P. Koehn and R. Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics, 2017.

J. Kreutzer, I. Caswell, L. Wang, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:112–130, 2022.

M. Mitchell, S. Wu, A. Zaldivar, et al. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pages 220–229. Association for Computing Machinery, 2019.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002. doi: 10.3115/1073083.1073153.

M. Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics, 2015.

M. Post. The call for system papers for the third conference on machine translation (WMT 2018): Evaluated metrics can lie and evaluation must change. Conference call for system papers, circulated 2018, 2018.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics, 2020.

L. Weidinger, J. Uesato, M. Rauh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 2022.