WXRKS RESEARCH · WORKING PAPER · WP-2026-01

The Edit Distance Report

Post-editing effort across 27 million translated segments, 2024–2026

wxrks Research September 2026 Second measurement; first ran 2022–2023 research.wxrks.com

ABSTRACT

Our first study — *Findings on Context Sensitive Efficiencies*, published 2024 and revised Q1 2025 — measured 4.3 million segments from June 2022 to December 2023: 20 language pairs, translation edit rate (TER), Microsoft neural machine translation (NMT) as the baseline. Its headline: humans changed 18.23% of NMT output and 11.17% of context-sensitive output — 39% less editing. This paper repeats the measurement at production scale: 27.0 million anonymized segment measurements, recorded August 2024 to September 2026. Three results. First, the gap widened: reviewers now change 8.1% of context-sensitive suggestions against 15.5% of NMT output — **47% less editing than NMT**, and 84% of suggestions ship with zero edits. Second, the improvement compounds: an account's edit distance falls from 41.7% in its first month to about 16% by month ten, because every correction feeds the account's knowledge base, never a model. Third, half of the measured segments — 51.6% — ship untouched. The study reads abstracted workflow measurements only: every figure is a group aggregate, no client is identifiable, and no source or target text was ever accessible to it.

1 Headline result: 2024 vs 2026

Same metric, same question. The first study's window closed in December 2023; this one opens in August 2024 — the two do not overlap. Editing fell on both sides. It fell faster where context is in the loop.

MEASURE (TER)	2024 STUDY	2026 PRODUCTION	CHANGE
Editing on context-sensitive suggestions	11.17%	**8.14%**	**−27%**
Editing on NMT	18.23%	**15.49%**	**−15%**
Advantage over NMT	39%	**47%**	**+8 pts**

Table 1. Editing effort, 2024 controlled study (4.3M segments, Microsoft neural MT baseline) vs 2026 production (7.7M origin-labeled segments, ≥5 accounts per bucket). Two years of better NMT models cut NMT editing by 15%. Context cut its own side by 27% and widened the advantage from 39% to 47%. The two editions average differently; under the matched construction the gap widens further, 22% to 47% — the note in section 14 has the detail.

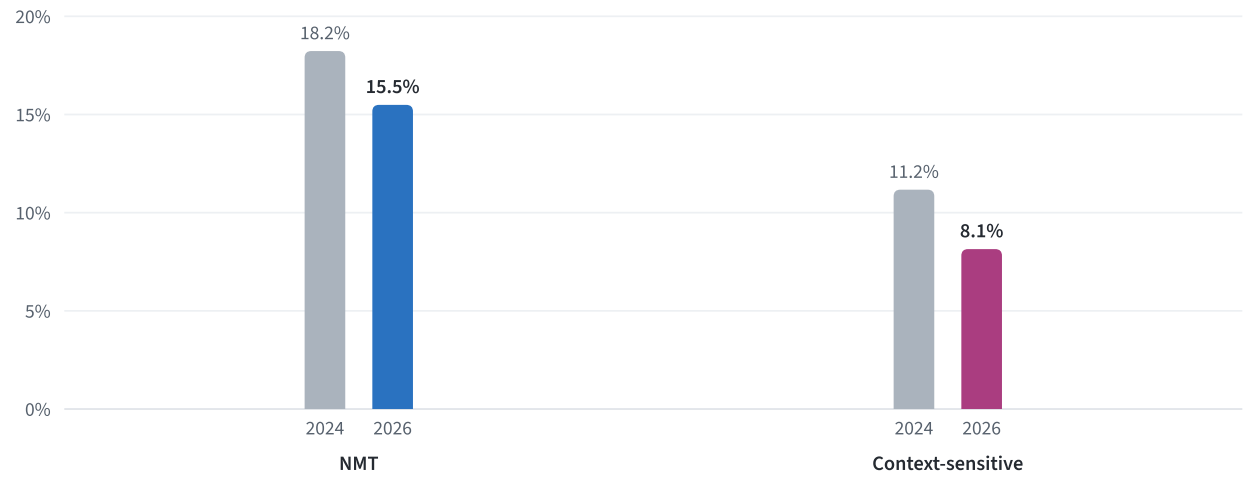

Figure 1. Average TER, the 2024 study beside 2026 production. Lower is less editing. Both eras use the same metric; the 2024 bars are the controlled study, the 2026 bars are live production.

The first study closed with open hypotheses. This edition answers three of them with production data: whether editing effort moves with working speed (its hypothesis 3 — section 8), what continuous learning earns an account over time (section 4), and how the spread across languages behaves at scale (section 6).

2 Data

The corpus: 27.0 million anonymized segment measurements, August 2024 through September 2026. Each record is an abstracted workflow metric — an edit-distance value with coarse context such as origin and language pair — held apart from customer content, which the study never touches. *Edit distance*: how much of a suggested translation the human changed before confirming — 0% is accepted as written, 100% is fully rewritten. *TER* counts those changes the same way the 2024 study did, so the two measurements compare.

27.0M	**51.6%**	**17.1%**	**6.3×**
segments measured	shipped with zero edits	average edit distance	the size of the 2024 study

3 Editing by origin

Since January 2026 every segment records where its starting text came from: a **translation memory** match (TM — a sentence your team already approved), a **context-sensitive suggestion** (built from your memories, glossaries, and preferences), **neural machine translation** (NMT — with none of that context), or **human translation** (HT).

Reviewers change 8.1% of context-sensitive output. They change 15.5% of NMT output — almost twice as much. And **84% of context-sensitive suggestions are confirmed without a single edit**; NMT manages 67%.

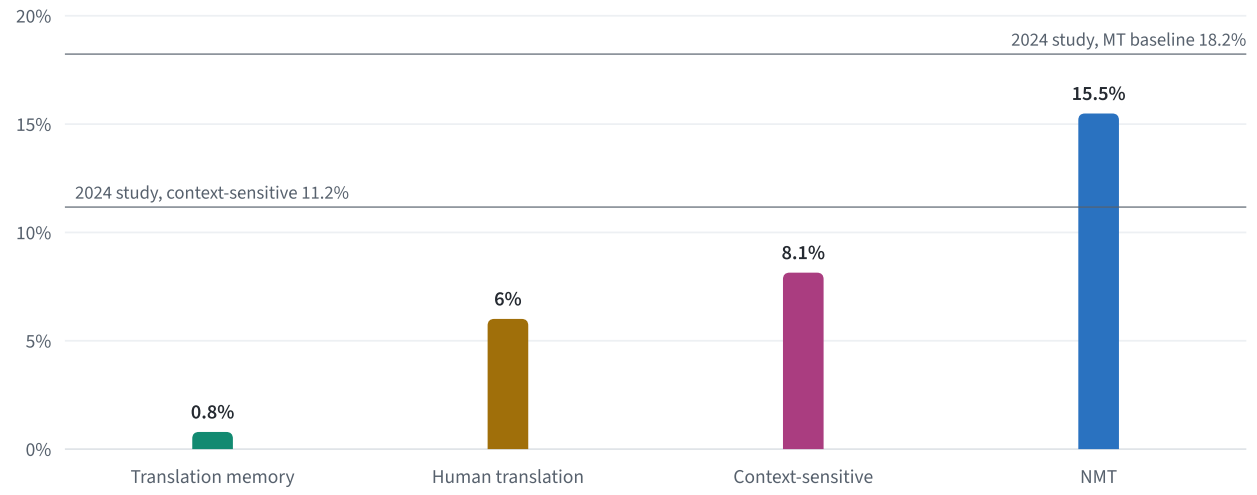

Figure 2. Average TER by segment origin, 2026 production. Reference lines mark the 2024 study: its MT baseline (18.2%) and its context-sensitive result (11.2%). Production sits below both.

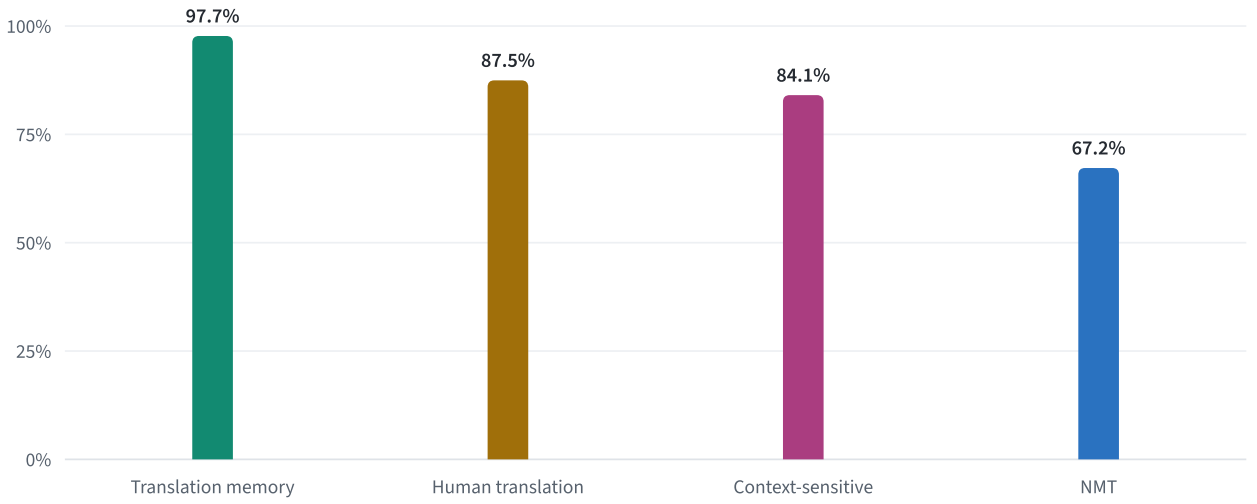

Figure 3. Share of segments confirmed without a single edit, by origin, 2026 production.

4 The learning curve

Every confirmation and correction teaches the account's knowledge base — never a model. If that is true, accounts must edit less the longer they work. They do.

Accounts start at a **41.7% average edit distance in month zero**. By month ten they sit near **16%** — about 60% less correcting — and hold there through month seventeen. Each account counts once per month, so no single large customer moves the line. Months 0–2 run high by design: a new account's knowledge base is empty.

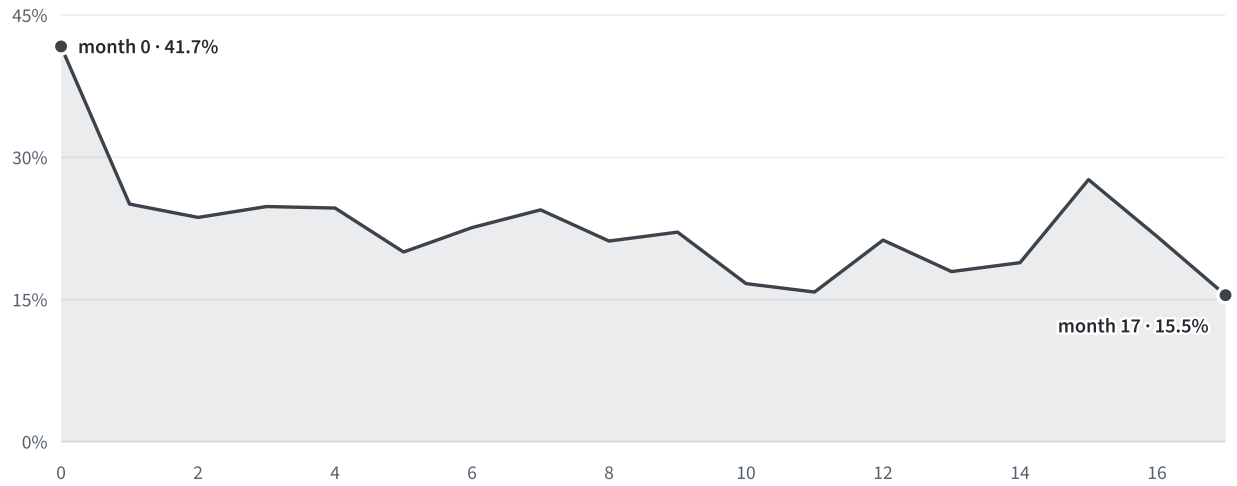

Figure 4. Average edit distance by account age in months. Per-account means averaged with equal weight; 4,344,758 segments; every bucket spans at least 10 accounts; cohort limited to accounts first seen November 2024 or later.

A trained model would keep this learning inside the vendor's weights. Here it lives in the customer's knowledge base, and this curve is what it earns.

The same curve, read from the other side: the share of segments shipped without a single edit climbs from about 40% in month zero to 64% by month seventeen.

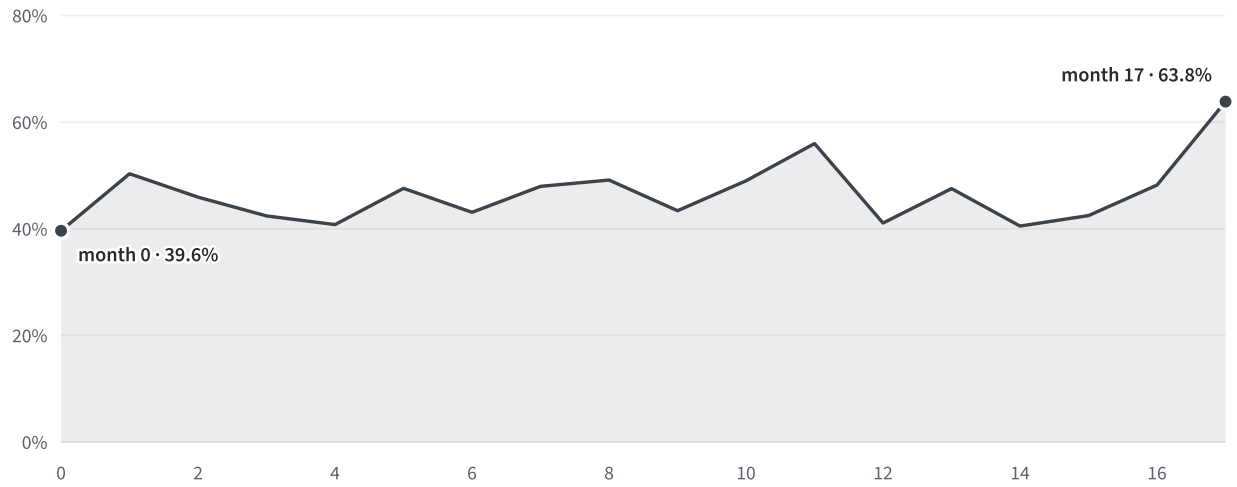

Figure 5. Share of segments confirmed without a single edit, by account age in months. Same cohort, weighting, and suppression as the previous figure.

5 Half untouched, a thin middle

The first study found translators changing as little as 1% and as much as 77% of a text. Production sharpens that picture into two poles: **half of all segments ship untouched**, and one in seven is fully rewritten. The middle — the patch-it-up editing that MTPE was named for — is the thinnest part of the distribution.

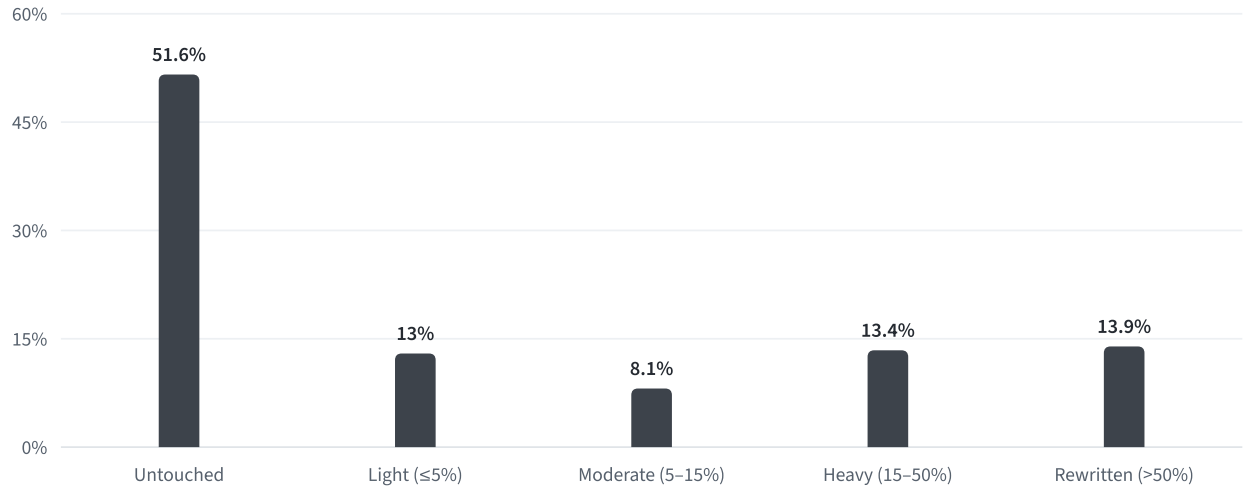

Figure 6. Share of segments by edit-distance band, whole corpus. Bands are the fraction of the suggested text changed before confirmation.

6 Where the editing concentrates

Fourteen busiest language pairs, ten or more accounts each. English into Spanish (US) gets 9% editing. English into Italian gets 26.3% — almost three times as much. The spread maps where human judgment spends its time, and where a glossary or style preference pays back first.

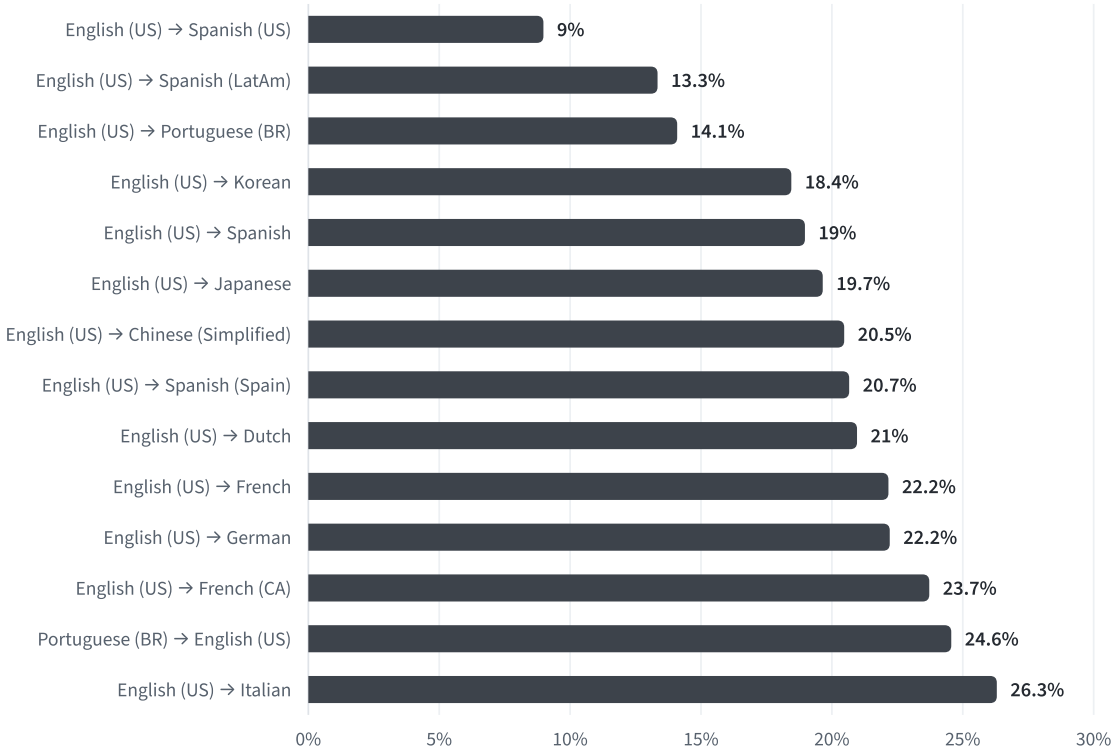

Figure 7. Average edit distance by language pair: the fourteen busiest pairs, each spanning at least 10 accounts, measured within one hosting region for consistent locale coding, sorted lightest-touch first.

7 The second pass confirms, it does not correct

Six human pipeline stages span at least five accounts each. The first translation pass carries nearly all the change; the passes after it mostly confirm — review changes 0.5% of what it reads, in-country review 0.2%. The one stage that still edits heavily is the legacy **full post-editing** workflow, at 9.2%: the old MTPE pattern, visible in the data exactly where teams still run it.

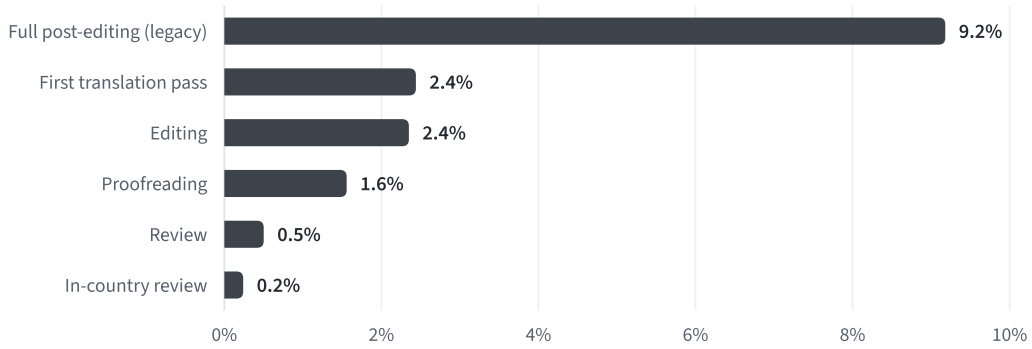

Figure 8. Average edit distance by pipeline stage, heaviest first. Human stages spanning at least 5 accounts; machine-only steps excluded.

8 Review speed

Seconds per source word, over segments with a recorded working time: memory matches review fastest at 1.12. Context-sensitive suggestions review at human-translation speed — 1.35 against 1.38. NMT is slowest at 1.45. Modest gaps, shown from zero.

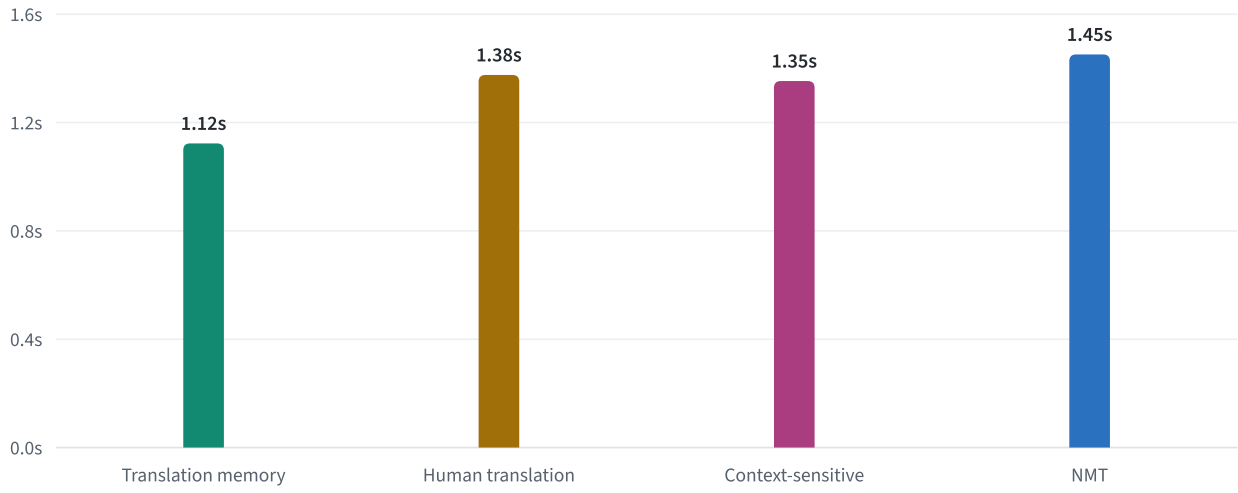

Figure 9. Working time in seconds per source word, by origin, across timed segments. The axis starts at zero.

9 Volume, and one number that does not move

The corpus sits on forty-one months of production history. We chart it as an index rather than absolute volume — data abstraction applies to our own numbers too. Words processed per month grew more than fivefold from mid-2023 to 2026. Words routed through NMT are metered separately and stay a small share throughout.

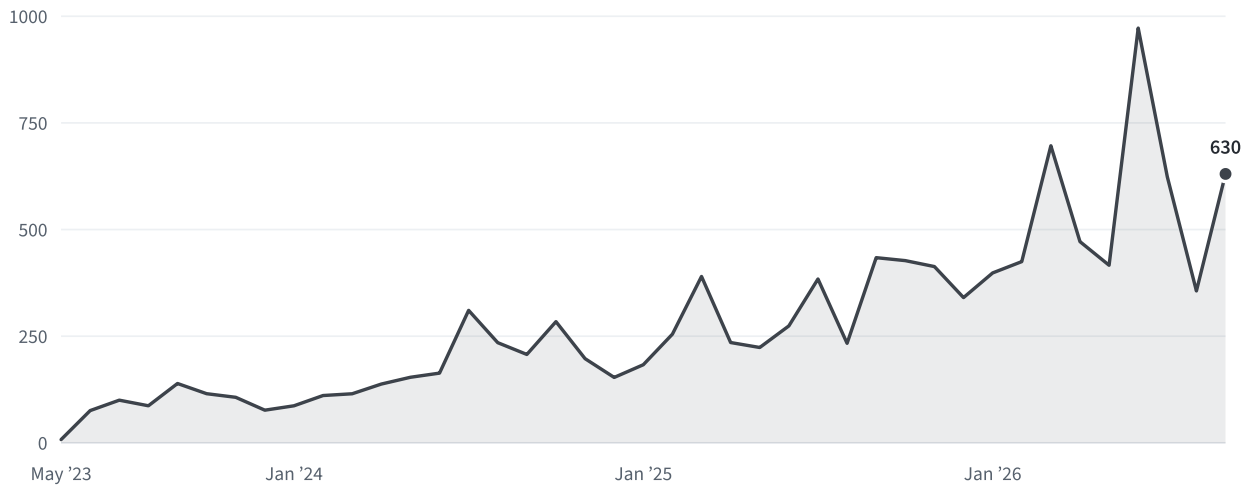

Figure 10. Words processed per month, indexed to July 2023 = 100. Absolute volumes are withheld by design; the shape carries the finding.

The blended edit distance across the whole corpus holds at 15–20% every quarter — flat for two years. New accounts enter every month at month-zero editing rates and hold the blend up. The improvement is real; it shows per cohort (section 4), not in the blend. We chart it anyway.

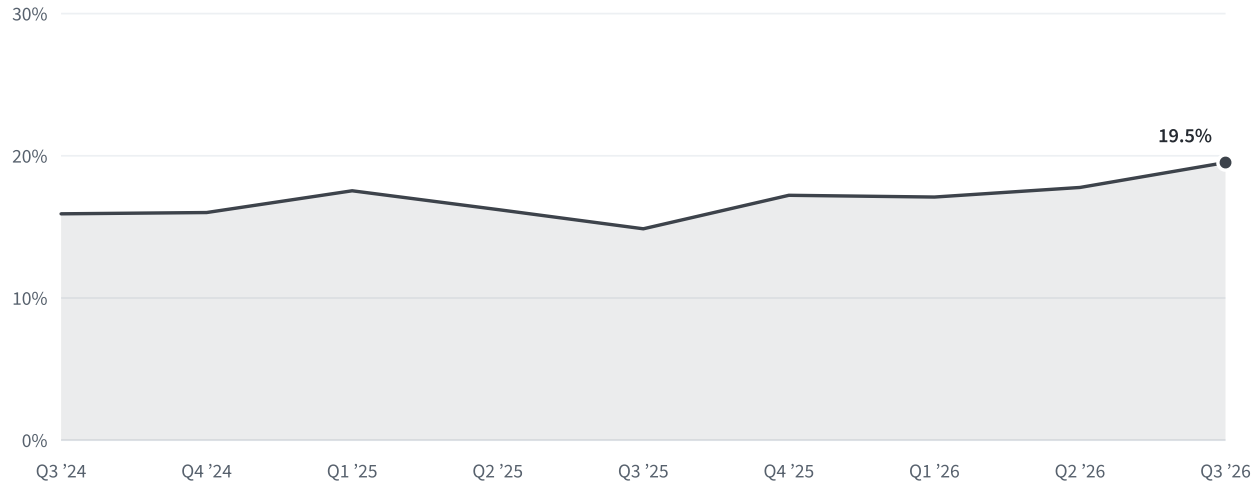

Figure 11. Average edit distance across the whole measured corpus per quarter. Flat by construction: new accounts keep entering the blend at month-zero rates.

10 The first study's hypotheses, revisited

The first study closed with eight hypotheses. Here is where each stands now. Section 13 adds this edition's own, numbered on from these.

#	HYPOTHESIS, AS POSED	Q1 2025 STATUS	THIS EDITION
1	Language variance comes from tokenization of character-based languages	Confirmed — the revision standardized it	**Settled**
2	Translator methodology swings TER by ±50%	Confirmed — edits ranged from 1% to 77% of a text	**Extended** — production is bimodal (section 5)
3	TER correlates with speed	Not yet tested	**Measured** — lighter-edited origins review faster; the per-word gap is modest (section 8)
4	Domain variance of ±20%	Not yet tested	Open — the abstracted metrics layer carries no domain field
5	Learned terms improve TER by at least 10%	Not yet tested	Open
6	TER shifts ±10% across NMT engines	Not yet tested	Open
7	Translator well-being improves	More data required	Open — telemetry cannot answer it; a survey can
8	Translation quality improves	Not yet tested	Open — zero-edit share measures acceptance, not quality

11 From findings to mechanism

The numbers in this paper describe the pipeline as it ran. Three product layers now sit on top of it, and they turn these findings into a loop clients operate themselves.

Smells, the runtime semantic verifier, reads the whole document and flags potential meaning deviations — omissions, mistranslations, inconsistencies — as they happen. With half of all segments already shipping untouched (section 5), review can change shape: approve the document, and analyze what Smells raised.

Sous-Chef applies audience, tone, and formality at scale — defined once, applied to an account, a project, or a language. The pre-translation arrives already tuned to each client and context instead of as one deterministic output, so fewer of the remaining edits are about tone at all.

The Confidence Score totals what Smells found and tracks it over time. It gives a team the number to watch: where it holds, approve; where it moves, look.

Together they close the loop this paper measured: corrections teach the knowledge base, Sous-Chef carries preferences into the next suggestion, Smells inspects the output, and the Confidence Score keeps the running tally — constant improvement of the pre-translation, run by the client. Their effect on edit distance is deliberately not claimed here: it is not yet measured, and it belongs to the next edition alongside the open hypotheses.

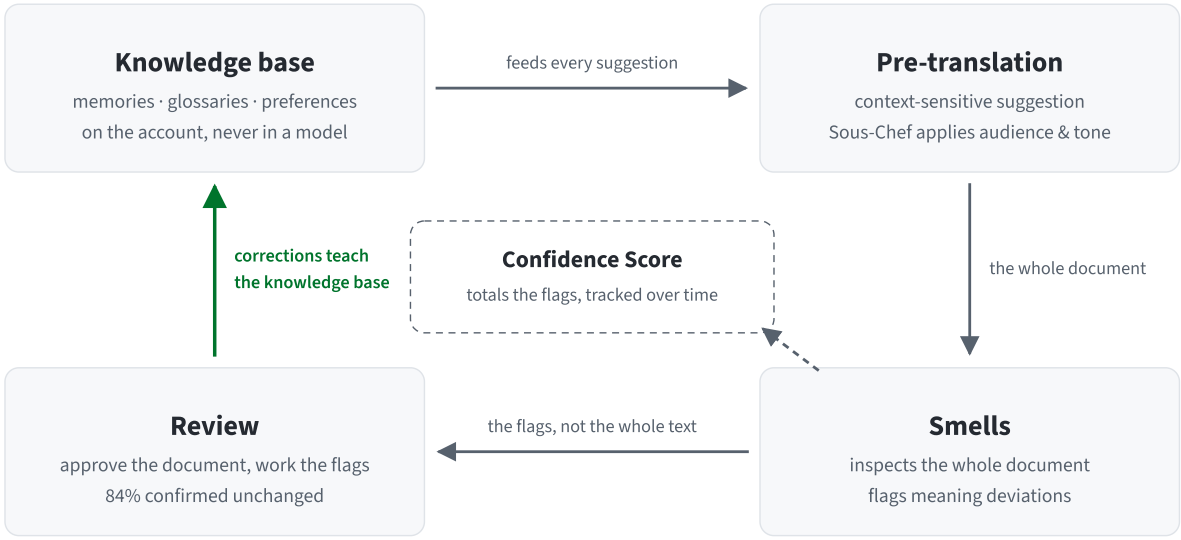

Figure 12. The improvement loop around the pre-translation. The green edge is the finding of section 4, measured; the three product layers make the rest of the loop operable. Their effect on edit distance is the next edition's subject.

12 Conclusions

1. **The context advantage is real and widening.** Under the matched segment-weighted construction it moved from 22% to 47% in two years, and 84% of context-sensitive suggestions ship exactly as suggested.
2. **The learning compounds — in the knowledge base, never a model.** A first-year account goes from editing 41.7% of its text to about 16%, and its untouched share climbs from 40% to 64%.
3. **Classic MTPE's middle ground is disappearing.** Half the corpus ships untouched and one segment in seven is rewritten; wholesale correction survives mainly in the legacy full-post-editing stage.

4. **Downstream passes confirm rather than correct.** Review changes 0.5% of what it reads, in-country review 0.2%. Quality is settled at the first pass.
5. **The findings become a mechanism.** Smells inspects whole documents and flags meaning deviations, Sous-Chef tunes the pre-translation to audience and tone, and the Confidence Score tracks the result — the improvement loop, run by the client. Measuring its effect is the next edition's work.
6. **Five hypotheses stay open** — domains, learned terms, engines, well-being, and quality. Section 13 adds six more from this edition's data. That is the agenda for the next edition.

13 Reflections, and what is worth testing next

The first study ended by thinking out loud. It said what was worth testing and what was worth noticing, and those questions shaped this edition. We end this one the same way.

Nothing in this section is a finding. Each item is something the data made us notice, the hypothesis it suggests, and how the next edition could test it. We number them on from the first study's eight, so one table can track every open question.

9. **A seeded knowledge base shortens the learning curve.** It is worth noticing where the curve starts. A new account edits 41.7% of its text in month zero, because its knowledge base is empty (section 4). Many teams arrive with years of approved translations and glossaries. If the curve is the knowledge base filling up, loading that history on day one should start an account lower and bring it to the 16% floor sooner.

   How to test it. Compare the first year of accounts that imported their memories and glossaries at the start with accounts that began empty.

10. **The 16% floor is judgment, not error.** The curve flattens near 16% from month ten to month seventeen (section 4). It is worth asking what is left in that 16%. Our guess is that it is no longer fixing mistakes. It is choices about tone, audience, and preference that no memory holds yet, which is the work Sous-Chef is built to carry (section 11). If that is right, accounts that define audience and tone should settle on a lower floor.

    How to test it. Compare the floor after month ten for accounts that use Sous-Chef with accounts that do not.

11. **A full rewrite is a different job, not a failed suggestion.** One segment in seven is fully rewritten (section 5). It is easy to read that as the suggestion failing. Another reading is worth testing. Some content is meant to be rewritten. Slogans, ad copy, and transcreation (rewriting a message so it works in a new market) ask a writer to start over, whatever the suggestion says. The first study already set copywriting and transcreation apart in its domain hypothesis (number 4).

    How to test it. Once the metrics layer carries a content-type signal, check whether rewrites cluster by content type and project or spread evenly.

12. **Review can work by exception.** Review changes 0.5% of what it reads, and in-country review changes 0.2% (section 7). It is worth noticing what that means for a reviewer's day. Almost all of it goes to reading text they will confirm as written. If Smells flags the segments where meaning drifted, a reviewer who reads only the flags should catch nearly everything a full read catches (section 11).

    How to test it. On projects that run Smells, compare what reviewers change in flagged segments with what they change in unflagged ones. If the unflagged side stays near zero, review by exception is safe.

13. **High acceptance means good suggestions, not quick trust.** 84% of context-sensitive suggestions ship without an edit (section 3). Acceptance is not quality, which is why the first study's hypothesis 8 is still open. The hard question is worth asking. Does a high zero-edit share mean the suggestions are good, or that people trust them too fast? The low change rate in later review passes (section 7) points to the first answer. It does not prove it.

    How to test it. Run a blind quality review of zero-edit segments from every origin, scored by reviewers who do not know where the text came from. It needs customer consent, because it means reading text this study never touched.

14. **The language spread follows the knowledge base, not the language.** In section 6, English into Italian gets almost three times the editing of English into Spanish (US). The first study traced part of its spread to how text in character-based languages such as Chinese is split into words for counting (number 1). It also saw a hint of a second cause. Spanish set for Spain or for Latin America needed less than half the editing of generic Spanish. Context Sensitivity gives specific information more weight when it ranks what it finds. Pairs with deep, specific memories and glossaries may edit less, whatever the language.

    How to test it. Compare edit distance per pair with the size of the knowledge base behind it, reported only as groups.

One more thing is worth noticing, and it is not a hypothesis. The blended average holds flat at 15 to 20% (section 9) while accounts edit less the longer they work (section 4). A single blended number hides the learning. The next edition should lead with cohort numbers.

That makes eleven open questions, five from the first study and six from this one. Some need data we do not collect yet, such as a content-type signal, a quality sample, and a survey. We name them now so the next edition can start collecting it.

14 Method and privacy

1. **Data abstraction.** The study reads an anonymized metrics layer held apart from customer content: edit-distance measurements (Aug 2024 – Sep 2026) and monthly volume metering (May 2023 – Sep 2026). No source or target text, translation memory, or glossary was accessible to the study at any point — the layer carries measurements, not content.

2. **Group-only reporting.** No account names or identifiers anywhere in the pipeline. Origin figures use only quarters with at least 5 accounts; tenure buckets require at least 10; language pairs at least 10; pipeline stages at least 5, with machine-only steps excluded. Suppressed buckets are dropped, not merged.
3. **Equal weighting.** The learning curve averages per-account means: a million-word account and a ten-thousand-word account count the same, and no client is recoverable from the line.
4. **Origin field.** Populated at production scale from January 2026. The origin comparison is a 2026 snapshot, not a two-year trend.
5. **First study (the baseline).** *Findings on Context Sensitive Efficiencies*, published 2024, revised Q1 2025. Window June 2022 – December 2023: about 2.3M segments through neural MTPE (Microsoft NMT the main engine) and 2M through Context Sensitive; 20 language pairs, English source; fourteen domains. An edit is a word replaced, added, deleted, or moved, divided by reference length; scores above 1 are capped at 1; translation-memory leverage was normalized between the two arms. The Q1 2025 revision standardized tokenization for character-based languages. This report uses the same 0–1 scale.
6. **Tenure.** Month 0 is an account's first measured activity. Only accounts first seen at least three months after the window opens are cohorted, so month 0 is a real start. Survivorship applies: accounts that stay may differ from accounts that leave.
7. **Speed.** Segments with a recorded working time between 1 ms and 1 hour; seconds per source word, summed then divided, per origin.

Note on Table 1 and Figure 1. The 2024 figures come from a controlled study against one NMT baseline; the 2026 figures are observational production data — accounts choose their own mix of work. The averaging also differs: the study's headline (18.23% and 11.17%) weighs its 20 language pairs equally, while its segment-weighted result was TER .162 against .128 — about a 22% advantage. The 2026 figures are segment-weighted, so the matched comparison is weighted against weighted: the advantage moved from 22% to 47%. We compare the editions because they use the same metric on the same platform, and we label every figure so you can weigh it.

---

wxrks research · WP-2026-01 Built from anonymized aggregates · revised 2026-10-14

research.wxrks.com/edit-distance-report