Measuring the Semantic Remainder

A corpus-matched, cross-lingual protocol for testing whether untranslatable meaning is a structural floor or a data-quality artifact

Nova Silva, Researcher in Philosophy of Mind and AI

International Academy for Consciousness Studies

Researched and prepared with the Academy. Working draft 1.0

Abstract

The "Language Uncertainty Principle" holds that no finite linguistic expression captures all of a meaning: every symbol system leaves an irreducible semantic remainder. The claim is usually argued philosophically, from cases of untranslatability and from the gap between form and reference. This paper reframes it as an empirical question with a decidable answer. We define a residual quantity R as the meaning that fails to transfer across a language boundary, and we operationalize it through three families of measurement: forward translation against professional references, round-trip (back) translation, and cross-lingual grounding against shared non-linguistic referents. The central methodological move is a matched-resource control. We ask whether R persists when corpus size, domain coverage, tokenizer fairness, and model capacity are equalized across the languages compared. If R is a data-quality artifact, it should shrink toward zero as the poorer side is brought up to parity. If R is structural, it should converge to a positive floor whose height tracks referential grounding gaps rather than data volume. We predict a nonzero floor concentrated in deictic, sensory, socially indexical, and culture-bound vocabulary, and we state an explicit kill condition: if the floor collapses toward measurement noise under matched resources, the irreducible-remainder claim dies. We keep claims about transferable meaning strictly separate from claims about understanding or experience. The protocol is buildable today on existing many-to-many benchmarks.

1. From a philosophical slogan to a measurable quantity

The idea that language always leaves something out is old and, stated loosely, nearly unfalsifiable. Any translation can be called imperfect, and any imperfection can be called a remainder. To make the claim do scientific work, we have to say what the remainder is, how to measure it, and what result would show that it does not exist.

We start from a minimal definition. Take a meaning M carried by an expression e in a source language L1. Render e into a target language L2, producing e'. The semantic remainder R(e, L1, L2) is the portion of M that is recoverable from e in L1 but not recoverable from e' in L2 by a competent user of L2. R is a difference of recoverable information, not a subjective sense of loss. It is directional: R from English to

Yoruba need not equal R from Yoruba to English.

Two things follow immediately. First, R is measurable only against some recovery task. A word list gives one estimate; a truth-conditional entailment battery gives another; a grounded pointing task gives a third. We therefore treat R as a family of task-relative estimators rather than a single scalar, and we report which estimator produced which number. Second, R is confounded with resources. A translation may lose meaning because the target language genuinely lacks a way to close on that meaning, or merely because the model or corpus for that language is thin. The whole empirical interest of the question lives in separating those two causes. That separation is what this paper is built around.

We are not claiming to measure meaning in the mind. We are measuring what a symbol-to-symbol mapping fails to carry across a language boundary, as scored by references and by grounded referents. Section 2 makes that firewall explicit, because it is the point on which loose versions of the thesis usually collapse.

2. Definitions, and a firewall

The literature on this topic runs together three questions that must be kept apart.

The first is a question about transferable content: does the information recoverable from e' match the information recoverable from e? This is a question about symbol systems and their referential coverage. It is what we measure.

The second is a question about understanding: does any agent, human or artificial, grasp what e means? The third is a question about experience: is there something it is like to mean saudade rather than to token the string "saudade"? These are the questions the philosophy of mind cares about, and they are not settled by any translation metric. A perfect translation score would not show that a system understands; a large remainder would not show that it fails to.

FACT: Bender and Koller (2020) argue that a system trained on linguistic form alone has, a priori, no access to the communicative intent that grounds meaning; their octopus thought experiment dramatizes an agent that has seen only the statistics of messages and never their referents. We use their form/meaning distinction, but we use it narrowly. Our R measures a gap in transferred referential content across languages. It is agnostic about whether either the source or target speaker "understands." We flag this because the untranslatability literature often slides from "this word does not transfer" to "this concept cannot be grasped," and that slide is exactly the firewall violation we refuse.

So the object of study is fixed: R is a property of the mapping between two symbol systems relative to a recovery task, measured against professional references and against shared non-linguistic referents. Claims about grasp and about phenomenal character are out of scope, and we will not let a number about transfer masquerade as evidence about either.

3. Three prior signals, and why none of them settles the question

Three existing bodies of work each measure something adjacent to R, and each falls short in a way the protocol is designed to fix.

Untranslatability scholarship. The Dictionary of Untranslatables (Cassin 2014) catalogs terms such as Dasein, pravda, and saudade whose conceptual networks do not superimpose across languages. Crucially, its editors define an untranslatable not as what cannot be translated but as what one never stops (not) translating: the sign of a persistent, structured mismatch between conceptual networks. DISPUTED: whether this mismatch is a fact about languages or a fact about scholarly attention. The catalog is curated by philologists, not sampled from a corpus, so it cannot tell us how much of ordinary language carries a remainder, nor whether the remainder would survive if the poorer language simply had more text. It gives us a hypothesis about where R concentrates, not a measurement of its size.

Round-trip translation. Translating e from L1 to L2 and back to L1, then comparing to e, is an old and tempting proxy for fidelity. FACT: Somers (2005) showed that round-trip translation is a poor predictor of one-way quality, because the round trip exercises two systems at once and errors in the two legs can cancel or compound. This is usually cited as a reason to abandon round-trip translation. We read it differently. The very fact that round-trip error confounds two directions makes it a bad quality metric but a usable lower bound on total transfer loss, provided we decompose the two legs and control each. Somers' warning is a design constraint, not a veto.

Cross-lingual grounding. FACT: Harnad (1990) framed the symbol grounding problem: a formal symbol system interpreted only through other symbols is like learning Chinese from a Chinese-only dictionary, with no exit to the world. Cross-lingual grounding tasks, where a rendering must pick out the same referent in a shared perceptual or structured space, give R an anchor outside language. HYPOTHESIS: the residual that survives grounding, not just the residual measured symbol-to-symbol, is the residual that matters for the structural claim, because grounded recovery cannot be faked by paraphrase.

None of the three, alone, answers our question. Untranslatability tells us where to look. Round-trip translation gives a cheap, confounded bound. Grounding gives an anchor. The protocol combines them under a control none of them applies: matched resources.

4. The protocol

The design has four layers: a matched corpus and model regime, a stratified probe set, three estimators of R, and a scaling sweep.

4.1 Matched-resource regime. The confound to kill is that R might be large only because one language has less and worse data. FACT: FLORES-200 and the No Language Left Behind effort (NLLB Team 2022) provide a many-to-many benchmark with professional translations across 200 languages, which gives us reference-quality targets in both directions. On top of that reference layer we impose parity controls. For each language pair we (a) subsample training corpora to equal token counts after tokenizer-fairness normalization, so that a language whose script inflates token counts is not silently starved; (b) match domain composition across the pair, since a remainder driven by topic mismatch is not a remainder of the language; (c) hold model architecture and parameter count fixed across the directions compared. FACT: Hoffmann et al. (2022) established that model quality is jointly determined by parameters and training tokens at roughly equal scaling rates, which is exactly why capacity and data must be matched together rather than one at a time. Matched resources are the data-quality control: they let us ask whether R is what

remains after the poorer side has been brought to parity.

4.2 Stratified probe set. We do not average R over undifferentiated text, because that hides structure. We stratify expressions into strata predicted to differ in grounding: (i) concrete count nouns with clear physical referents (table, dog); (ii) sensory and affective terms (glossy, saudade, the smell named by a specific verb); (iii) deictic and spatial terms, including absolute versus relative frames of reference; (iv) socially indexical terms (honorifics, kinship systems, politeness registers); (v) grammaticized categories with no lexical counterpart in the other language (evidentiality, aspect, grammatical gender); (vi) abstract logical and quantificational vocabulary. HYPOTHESIS: R will be near floor for stratum (i) and high for strata (ii) through (v), because those are where languages carve the referential and social world differently. Boroditsky's work motivates this stratification.

4.3 Three estimators. We measure R three ways and report all three.

E1, reference-anchored forward loss. Render e into L2, then score how much of the information recoverable from the professional reference for e is recoverable from e'. We use entailment and question-answering probes derived from e, not surface overlap, so that a faithful paraphrase scores as full transfer and a fluent mistranslation does not.

E2, decomposed round-trip loss. Following the caution in Somers (2005), we run L1 to L2 to L1 but score each leg separately against the reference at that leg, and we report the residual only where both legs are individually faithful. This isolates loss that is not attributable to a single weak direction, which is the loss most likely to be structural.

E3, grounded recovery gap. In a shared non-linguistic space (images, structured scenes, coordinate frames, or truth-conditional models), we ask whether a competent L2 user, reading only e', selects the same referent that an L1 user selects from e. E3 is the estimator that speaks to the grounding version of the thesis, because it cannot be satisfied by symbol-internal paraphrase.

4.4 Scaling sweep. For each stratum and each estimator, we compute R at a ladder of matched resource levels: equal token budgets at 10^7, 10^8, 10^9, and beyond, and matched model capacities across a comparable ladder, always raising the poorer side toward the richer. The output is a curve, R as a function of matched resource, per stratum. The shape of that curve, not any single number, is the result.

5. What the estimators can and cannot see

Every estimator has a failure mode, and honesty requires naming them.

E1 can be inflated by probe leakage: if the entailment probes are themselves generated in L1 and translated, their own translation loss contaminates the score. We generate probes independently in both languages from the shared referent, not from each other.

E2 inherits Somers' confound if we skip the per-leg gating; the gating is what converts a discredited metric into a conservative bound. WAGER: I expect E2, properly gated, to track E3 within stratum, and if it does not, E2 should be discarded rather than patched.

E3 depends on the richness of the grounding space. A remainder that shows up in text may vanish in a pointing task simply because the task is too coarse to require the distinction (a politeness register does not change which cup you point at). This cuts against overclaiming: E3 will tend to under-report R for socially indexical strata, so a positive E3 residual there is strong evidence, and a null E3 residual there is weak evidence. We therefore never read a single estimator as decisive; the structural claim requires agreement across E1, E2, and E3 within a stratum.

There is also a ceiling problem. Professional references are themselves finite renderings and carry their own remainder. Measuring R against a reference measures transfer relative to a human translator's best close, not relative to some Platonic complete meaning. This is a feature, not a bug: it keeps the quantity operational. But it means our R is a lower bound on any metaphysical remainder, and we will not dress it up as more.

6. Prediction and kill condition

The whole design exists to make one disagreement decidable. Either the remainder is a structural floor tied to how languages ground and carve the world, or it is an artifact of unequal data that parity erases.

> **Prediction.** Under matched corpus size, matched domain composition, tokenizer-fairness normalization, and matched model capacity, R will not fall to measurement noise. It will converge to a positive floor whose height is ordered by stratum: near zero for concrete count nouns, and significantly positive and stable for sensory/affective, deictic/frame-of-reference, socially indexical, and grammaticized-category strata. Within those strata the floor will be higher for language pairs with larger referential and social grounding gaps, and it will be visible in E3 (grounded recovery) and not only in E1/E2, so that it cannot be explained as paraphrase-level noise. Doubling matched resources will move R by less than it moves out-of-vocabulary or rare-word error, showing that the floor is decoupled from data volume.
>
> **Kill.** If, holding stratum fixed, R declines monotonically toward the noise floor of the estimator as matched corpus size and model capacity rise, so that at parity the high-grounding-gap strata are statistically indistinguishable from the concrete-noun stratum, then the irreducible-remainder claim is false as an empirical matter: the apparent remainder was a data-quality artifact. A single stratum surviving is not enough; if E3 in particular collapses to noise under parity across all strata, the grounding version of the thesis dies even if a symbol-internal E1 residual lingers.

The contrast is the coin versus the weather. A biased coin and a chaotic weather system both look like noise at first, but one has a stable structure that survives more data and the other reduces to sampling error. Matched resources are how we tell which one R is. If R keeps its shape as we pour in matched data, it has structure. If it dissolves, it was poor data wearing the mask of a principle.

7. A Closure reading, held lightly

The Academy's Closure framework offers a compact way to say what R is, and it earns its place only if it adds precision, not decoration. In that framework, an expression is an act of closure: an openness (the space of what could be meant, call it M, with its horizon m) settling into a determinate this, a content c

held in a present act of expression, Cl. Meaning is the closure; what the closure leaves outside itself is the remainder, R.

On this reading the Language Uncertainty Principle is the claim that Cl is always proper: the closure never exhausts the openness it closes on, so R is never empty. That is a strong metaphysical statement, and we do not try to prove it. What we do is measure a shadow of it: the cross-lingual R, the part of one language's closure that a second language's closure cannot re-close. HYPOTHESIS: the stratum ordering we predict is what you would expect if R tracks how much of a meaning is fixed by reference to a shared world (low for count nouns, whose closure is anchored to stable referents both languages share) versus fixed by a language-internal system of contrasts (high for evidentiality, honorifics, absolute frames, where each language's closure carves its own set of thises). The framework thus predicts not just that R is positive but where it should be largest, which is a risk the philosophical slogan alone never takes.

We keep this light on purpose. The Closure vocabulary is a lens for stating the hypothesis crisply; the protocol in Sections 4 and 5 stands or falls on its own, whatever one thinks of the metaphysics.

8. Limits, wager, and conclusion

Limits. The protocol measures transfer against finite human references and finite grounding spaces, so it can only ever return a lower bound on any deeper remainder; it is silent about understanding and about experience, by construction; its grounding estimator under-reports exactly the social and affective strata where the thesis is most interesting, which biases the test toward the kill condition and against the prediction. Matched resources are hard to achieve for genuinely low-resource languages, and a failure to reach parity could be misread either way. DISPUTED: whether tokenizer-fairness normalization can ever be truly neutral across scripts; if it cannot, some residual reflects our measuring instrument rather than the languages. These are reasons the result must be read as a curve with error bars, not a verdict.

Wager. WAGER: I expect the floor to survive, and to survive most visibly in deictic and grammaticized-category strata rather than in the celebrated hard words like saudade. The famous untranslatables are famous because scholars noticed them; I bet the durable remainder is more mundane and more pervasive, living in evidentiality and frame-of-reference and honorification, categories so ordinary that speakers do not experience them as untranslatable at all. If that is right, the interesting remainder is not the poetry we cannot render but the grammar we cannot help imposing. If it is wrong, and parity flattens every stratum to the concrete-noun floor, then the remainder was an accounting error and the principle should be retired.

Conclusion. The Language Uncertainty Principle has lived too long as a slogan that cannot lose. By defining R as task-relative recoverable-information loss, measuring it three ways, and above all controlling it against matched corpus size and model capacity, we convert an unfalsifiable intuition into a curve that can be read and a claim that can be killed. The philosophy told us the remainder is irreducible. The computational-linguistic question is whether it is structural or artifactual, and that question now has a protocol and a kill switch. We keep the firewall intact throughout: a number about what fails to cross a language boundary is evidence about symbol systems and reference, and it is not, on its own, evidence about who or what understands.

References

Bender, E. M., and Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 5185-5198. https://aclanthology.org/2020.acl-main.463/

Boroditsky, L. (2001). Does Language Shape Thought? Mandarin and English Speakers' Conceptions of Time. Cognitive Psychology, 43(1), 1-22.

Cassin, B. (ed.), Apter, E., Lezra, J., and Wood, M. (English trans. eds.) (2014). Dictionary of Untranslatables: A Philosophical Lexicon. Princeton University Press.

Harnad, S. (1990). The Symbol Grounding Problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346.

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. arXiv:2203.15556.

NLLB Team (Costa-jussa, M. R., et al.) (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672.

Somers, H. (2005). Round-Trip Translation: What Is It Good For? Proceedings of the Australasian Language Technology Workshop (ALTW), 127-133.