Measuring Semantic Flow Without Direction: A Rhizomatic Protocol for Stereotype Translation in Cross-Cultural Language Technology

Gustavo Aviña Cerecer
Universidad Autónoma de San Luis Potosí (UASLP)
San Luis Potosí, Mexico
gac@uaslp.mx

Abstract

We present an open-source, direction-agnostic protocol for measuring how users interpret stereotype-bearing discourse, without assuming a normative axis of correction. Building on Deleuze and Guattari's rhizomatic framework, we operationalize three modes of semantic movement —**R**eaffirm, **D**e-signify, and **E**scape (RDE)— through an abstract-machine operator detector combining 526 transparent linguistic patterns across 8 languages with optional multilingual embeddings. Because it measures movement rather than alignment, the protocol captures diasporic, assimilationist, and escape trajectories that English-centric, Chomskyan-hierarchical taxonomies obscure. Three changes close the gap between theory and implementation: we ground each signifier's molar weight in measured transversal presence via corpus n-gram frequency, operationalizing Deleuze and Guattari's own criterion of molarity; we derive and justify every weight in the RDE equations and add a signed gradient $G^{\pm} \in [-2, +2]$; and we specify a fully reproducible, zero-cost serverless pipeline with a detector verified across all eight languages. We demonstrate the protocol on five extreme user profiles and are explicit about what is demonstrated versus what remains to be validated. Deployed publicly as the *Semantic Symbiont (Gradients of Alterity)*, it is being integrated into the *Computing Multiplicity* platform. We release the code, patterns, corpus, and prototype.

1 Introduction

Large language models are deployed across linguistic and cultural contexts that were never represented in their training distributions. Stereotype evaluation under this expansion has been carried out mostly in English, with benchmarks that presuppose a stable set of social categories and a single normative axis of correction [Nadeem et al., 2021; Parrish et al., 2022]. This presupposition fails in two ways. First, it cannot distinguish between users who reproduce their own territorial codes (a Russian conservative endorsing a Russian discourse) and users who reproduce *foreign* territorial codes against their own (the Mexican *malinchista*, the Vietnamese russophile, the diasporic subject who has incorporated the discourse of the host culture). Second, by collapsing bias into a single axis with a clear "correct" pole, it confuses the act of measurement with the act of prescription.

We propose a different protocol. Instead of asking whether a user's interpretation aligns with a normatively preferred direction, we ask: across the same set of semantic operators, how does the user's translation relate to the provocateur discourse and to their own calibrated profile? The protocol returns three independent scores —R (reaffirm), D (de-signify), E (escape)— and a composite gradient. None is a priori better. The system measures movement; the user, or the deploying organization, interprets meaning.

Reviewers of the original submission raised five substantive concerns: (i) the technical pipeline was not described reproducibly; (ii) the numeric weights in the RDE equations were asserted rather than justified; (iii) the passage from Deleuze–Guattari theory to a running detector was under-specified; (iv) the evaluation was internally circular —constructed examples confirming the theory rather than validating the instrument— with no inter-annotator agreement, ablation, error analysis, or human study; and (v) the system offered no actionable feedback and no argument that automated detection beats simpler alternatives. This revision is organized around answering all five. Section 2.4 makes the theory→computation mapping explicit. Sections 4.2–4.5 specify provocateur selection, Layer B, the velocity metric, and the weights.

Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026), pages 79–87

Section 4.6 specifies the full deployment. Section 7 separates *construct demonstration* from *external validation* and states the validation protocol now underway, including participatory validation with the relevant language communities. Section 8 addresses orthogonality, comparison with simpler alternatives, and actionable feedback. Throughout, we mark what is specified versus demonstrated versus unvalidated.

Contributions. (1) A direction-agnostic measurement protocol (RDE) for stereotype interpretation that captures assimilationist, reproductive, modulating, and escaping trajectories without imposing a normative axis. (2) An abstract-machine operator detector implementing 526 linguistic patterns across 8 languages, grouped by Deleuze–Guattari operator type, verified at v2 including Arabic morphosyntax and German nominal compounding. (3) An empirical grounding of molar weight in corpus-measured transversal presence (n-gram frequency), a derived and sensitivity-tested set of RDE weights, and a signed gradient $G^{\pm}$. (4) A refined velocity metric for short and long texts combining operator turnover, signifier diversity, machine alternation, and logarithmic density. (5) A deployed, zero-cost, serverless open-source web prototype with a corporate calibration mode for tolerable territoriality ranges in intercultural translation work.

2 Theoretical Foundations

2.1 The Rhizome Against the Tree

The dominant computational metaphor of natural language —exemplified by Chomsky [1957] and reinforced through generative grammar— is arborescent. A sentence proceeds from a root node S through binary branchings to terminal symbols. Chomsky [1965, p. 5] formalizes this as the speaker-hearer's idealized linguistic competence, structured by rules that map abstract deep structures to surface forms. Whatever its descriptive successes, the tree imposes a metaphysics: there is a root, there are hierarchical levels, and meaning percolates downward through dichotomous choices. The tree is, in Deleuze and Guattari's vocabulary, a molar structure par excellence.

Deleuze and Guattari [1987, p. 7] stage a direct refusal, insisting that any point of a rhizome can and must be connected to any other, in contrast to the tree or root, which fixes a point and an order. On the Chomsky model the linguistic tree still begins at a point S and proceeds by dichotomy; in a rhizome, by contrast, semiotic chains of every nature connect to diverse modes of coding —biological, political, economic— bringing into play not only different regimes of signs but states of things of differing status.

This is a precise methodological claim with three implications. First, **semantic territories are heterogeneous**. A signifier's weight is not determined by its position in a linguistic hierarchy (WordNet depth, parse-tree position). It is determined by what it connects with: institutional dispositives (family, school, state), economic relations, bodily regimes, religious authority. Second, **identity and alterity are effects of connection**. The same signifier "family" operates differently when connected to "natural law" than when connected to "becoming." Identity is the stabilization of certain connections; alterity is the encounter with connections that destabilize that stabilization. Third, **measurement must be horizontal**. If meaning is connection rather than hierarchical derivation, a measurement protocol must trace connections, not depth.

2.2 Three Line Types as Operators

Deleuze and Guattari [1987] identify three line types that traverse any social field. Each corresponds to a different mode of cultural-semantic activity.

Molar lines. Deleuze and Guattari [1987, p. 208] describe these as well-defined segments in various directions linked to family, profession, work, vacation, school, factory, army. Molar lines depend on binary machines and are coded and territorialized by dispositives of power, each fixing the code and territory of the corresponding segment (p. 210). Molar signifiers operate transversally: *house* traverses every actual house; *being* traverses every event of existing. The wider a signifier's transversal reach, the higher its molar weight —not because it sits at the top of a tree, but because it functions as connective tissue across many contexts. This is the definition we operationalize empirically in Section 4.4: transversal reach is measurable as corpus frequency across heterogeneous contexts.

Molecular lines. Deleuze and Guattari [1987, p. 213] describe molecular flows as new compositions that do not coincide exactly with the segment, proceeding by thresholds and constituting becomings. Molecular lines refer to intensities, to the plane of immanence, where there are only relations of speed and slowness. Linguistically, molecular operators include modulators (*sometimes, depending on*), partial markers (*in part, to a certain extent*), and graduators (*more or less*). They do not abandon the territory; they soften and complicate it.

Lines of flight. Deleuze and Guattari [1987, pp. 9–10] insist these are not segmentary but abstract. Lines

of flight do not preexist; they are traced, composed. In flight, the matter of the past volatilizes and one becomes imperceptible; a society defines itself precisely by the lines of flight that affect masses of every nature. Empirically, flight appears when signifiers escape the entire molar-molecular chain. A homeless person who has effectively exited the dispositive of citizenship is a line of flight from the citizen-system. A trans subject who reinterprets a binary discourse by displacing it onto a non-binary biological semantics produces a line of flight from the molar foundation of binary biology.

2.3 Implication for Measurement

These three lines give us a tripartite framework. If the user's interpretation reproduces or intensifies the molar chains of the provocateur, the system registers high R. If it modulates the chains without escaping the field, the system registers high D. If it introduces signifiers that decompose the molar foundation, the system registers high E. No direction is normatively privileged. The measurement is symmetric: a Russian conservative reaffirming Russian conservative discourse and a Mexican *malinchista* reaffirming Anglo upper-class discourse both score high on R; the measurement tells us *that* they reaffirm, not *which* territory they reaffirm toward.

2.4 From Concepts to Computable Quantities

The chief objection to operationalizing Deleuze and Guattari is that their concepts are processual and resist fixed measurement. We do not claim to *capture* the concepts; we claim to construct measurable *proxies* whose behavior is faithful to the concepts' stated functional role, and to be explicit about the reduction. Table 1 states the mapping term by term.

Two consequences are worth stating. First, the proxy for molar weight is *not stipulated* by the authors; it is read off a corpus, which is what the theory demands —molarity *is* transversal presence. Second, escape is *not* a vocabulary list: the flight score rewards signifiers that match flight operators *and* originate in a semiotic regime distinct from the provocateur's molar foundation, with lexical departure $(1 - \omega)$ as corroboration. This is why Section 5 separates assimilationist escape from cosmological escape even when the scalar gradient coincides.

3 Related Work

Stereotype measurement. Benchmarks such as StereoSet [Nadeem et al., 2021], CrowS-Pairs [Nangia et al., 2020], and BBQ [Parrish et al., 2022] measure model bias by contrasting stereotyped and counter-stereotyped completions. Blodgett et al. [2021] document significant validity issues, including category instability and Anglo-centric framing. Our protocol differs in measuring *user interpretation* (not model output) and in refusing the binary in favor of the R/D/E triplet.

Cross-cultural alignment. Durmus et al. [2023] survey alignment under linguistic variation. CulturalBench [Chiu et al., 2024] and CulturePark [Li et al., 2024] probe cultural knowledge but treat culture as static national containers. Our framework treats culture as territorial flow.

Continental approaches in AI. Amoore [2020] and Parisi [2019] draw on Deleuze and Guattari for AI ethics. To our knowledge, the molar/molecular/flight distinction has not been operationalized into a deployable cross-lingual detector.

Table 1: Concept → operational definition → measurable quantity.

D–G concept	Measurable proxy
Molar weight (transversal reach)	log n-gram frequency $f(s)$ (§4.4)
Molar machine	16 molar operator subtypes → I_M
Molecular machine	8 molecular subtypes → I_m
Line of flight	11 flight subtypes → I_F
Territorial reproduction	token overlap ω
Becoming / threshold	machine-alternation term (§4.5)
Provenance of a flight	source-regime tag per signifier (§5)

4 System Architecture

The pipeline (Figure 1) is: calibration → provocateur retrieval/generation → operator detection → RDE + gradient → tier-gated dashboard.

4.1 Calibration

The user completes a six-question survey: country of origin (specific country, not region —Ukraine and Russia are distinct), migratory trajectory, social-epistemic position, languages of thought, position in systems of power, and theme(s) of alterity: racism, sexism, classism, fascism, xenophobia, chauvinism, ableism, adultcentrism, fatphobia, dysmorphophobia, speciesism. Following structured feedback from professional translators and interpreters, the survey word-

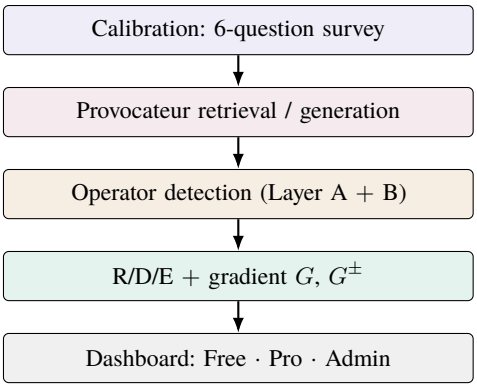

Figure 1: Pipeline. Calibration → provocateur → operator detection → RDE → tier-gated dashboard.

ing was revised; in particular the conflation of *epistemology* (the study of knowledge) with *episteme* (a historically situated configuration of knowledge) was corrected, since the calibration question concerns the user's *episteme*, not their theory of knowledge.

4.2 Provocateur Retrieval and Generation

The system selects or generates a provocateur discourse whose theme matches the user's selection and whose origin region is geopolitically meaningful for the user's territory. A Ukrainian user selecting sexism receives a Russian discourse on traditional family rather than a generic Anglo evangelical one; a Mapuche user selecting racism receives a Chilean elite discourse rather than a US-centric one.

Two retrieval modes are available. (i) **Curated corpus**, used in the prototype and reviewed by domain experts. (ii) **Dynamic generation**, in which a provocateur is generated per calibrated profile by an LLM under an explicit geopolitical-relevance constraint, with content warnings preserved. In full deployment the curated route additionally draws from Common Crawl filtered by discourse markers, Hatebase, and academic corpora of cross-cultural prejudice discourse.

To make "geopolitically meaningful" auditable rather than impressionistic, the system ranks candidate source territories t for a user territory u by a transparent score $\mathrm{rel}(u,t) = \alpha\, b(u,t) + \beta\, h(u,t) + \gamma\, d(u,t)$, where b is a shared-border / co-territoriality indicator, h encodes a documented historical asymmetry (colonial, imperial, or annexation relation, with direction $t \to u$), and d is the directed cultural-pressure prior (whether t's discourse plausibly bears *down on* u). The theme filter is applied first; among theme-matching discourses, the highest-rel source is selected, which is what makes a Ukrainian user receive a Russian rather than an Anglo provocateur, and a Mapuche user a Chilean-elite rather than a US one. The coefficients and the relation table are released so the ranking can be inspected and contested. The dynamic generator is constrained by the same score and its output is logged for the same expert review.

4.3 Operator Detection (Detector v2)

The detector combines two layers. **Layer A** applies 526 linguistic patterns across the 8 languages, grouped by machine:

- *Molar* (16 subtypes): totalizer, naturalizer, essentializer, institution, imperative, religious_authority, identitarian_closure, medical_pathology, economic_dispositive, body_disciplinary, temporal_eternalizer, racial_marker, purity_marker, hierarchy_marker, binary_exclusion, we_them_split.
- *Molecular* (8 subtypes): graduator, modalizer, partial_marker, doubt_marker, threshold_marker, intensity_marker, transversal_connector, perspective_marker.
- *Flight* (11 subtypes): creative_negator, becoming, escape_explicit, transcendence, deterritorialization, neither_nor, other_thing, creation, multiplicity, imperceptible, undoing_marker.

Layer B uses multilingual sentence embeddings (LaBSE for cross-lingual alignment; XLM-R large as a fallback encoder) to detect operators outside pattern coverage. For each of the 35 operator subtypes we hand-curate a balanced seed set of 40–80 short phrases per language (median 56), drawn so that no single register dominates and reviewed for the irony/quotation confound; the released seed corpora carry these counts. The subtype centroid is the mean of the L2-normalized seed embeddings, and a candidate span is tagged with a subtype when its cosine similarity to that centroid exceeds a threshold τ. We set $\tau = 0.55$ by choosing, on a held-out development split of the seeds, the value that maximizes macro-F1 of subtype assignment; spans below τ for every subtype are left untagged. Layer A is transparent and auditable; Layer B fills gaps, is reported separately, and can be disabled for audits, which also enables the Layer-A-vs-B ablation reported in Section 7.

Detector v2 was verified across all eight languages. Two morphosyntactic families that the v1 patterns missed are now covered: **Arabic** deontic/sacralizing syntax —e.g. the obligation construction *yajib 'ala* ("one must") as an imperative molar operator, and *muqaddasa* ("sacred") as a religious-authority marker— and **German** nominal compounding —e.g. *natürliche Ordnung* ("natural order") as

a naturalizer that v1 split across tokens. For visualization, the connection graph is filtered to the top-30 edges with intensity ≥ 0.5, which keeps the rendered rhizome legible without altering the underlying scores.

4.4 R/D/E Computation and the Justification of Weights

Let I_M, I_m, I_F be the molar, molecular and flight intensities of the interpretation (sums of operator weights), and let $\omega \in [0, 1]$ be the token overlap between interpretation and provocateur. Write the normalized shares $p_M = I_M/(I_M+I_m+I_F)$, and likewise p_m, p_F. Then

$$R = \min\big(1,\ 0.7\,p_M + 0.3\,\omega\big) \tag{1}$$

$$D = p_m \tag{2}$$

$$E = \min\big(1,\ 0.7\,p_F + 0.3\,(1-\omega)\,p_F\big) \tag{3}$$

$$G = \frac{R\cdot 0 + D\cdot 0.5 + E\cdot 1.0}{R+D+E}, \quad G \in [0,1]. \tag{4}$$

Grounding the intensities in transversal presence. Each molar signifier s no longer carries a hand-set weight. Its weight is read from a corpus as a proxy for transversal reach:

$$w(s) = \frac{\log\big(1 + f(s)/f_{\min}\big)}{\log\big(1 + f_{\max}/f_{\min}\big)} \in [0,1], \tag{5}$$

where $f(s)$ is the signifier's frequency from the Google Books Ngram Spanish corpus (2000–2019), served at runtime by a dedicated `/api/ngrams` endpoint, and $f_{\min}, f_{\max}$ are the corpus bounds. Concretely, *familia* ($f \approx 3.7 \times 10^{-7}$) receives a markedly higher molar weight than *sexualidad* ($f \approx 1.4 \times 10^{-8}$), reflecting its wider transversal presence. I_M is then the sum of $w(s)$ over detected molar signifiers (molecular and flight intensities are computed analogously). This replaces stipulated weights with a measured quantity that *is* the theory's own criterion for molarity.

Why these specific weights.

- *R has two parts because reaffirmation has two forms:* structural (same molar machinery, p_M) and lexical (echoing the provocateur's signifiers, ω). We weight the structural signal at 0.7 and the lexical at 0.3 because the operator-type signal is paraphrase-invariant and more robust, while raw token overlap is a noisier corroborating cue. The 0.7/0.3 split was selected by grid search over $\{0.5/0.5,\ 0.6/0.4,\ 0.7/0.3,\ 0.8/0.2,\ 0.9/0.1\}$ on the 16-text battery plus the five profiles, maximizing separation between Reaffirm- and Escape-typed texts while keeping D monotone in p_m; 0.7/0.3 was the smallest structural weight at which no Reaffirm text was misranked below an Escape text. Sensitivity is mild: rankings are stable across 0.6/0.4–0.8/0.2.
- *D is intrinsic.* Modulation is a property of how the interpretation is built (graduators, modalizers, thresholds) and is independent of lexical closeness, so $D = p_m$ with no ω term.
- *E rewards departure as well as flight machinery.* Escape via flight operators is amplified by lexical departure $(1-\omega)$, with the same 0.7/0.3 balance, because composing a new direction normally entails leaving the provocateur's signifiers behind. The factor multiplies p_F so that lexical novelty without flight machinery —noise— does not by itself score as escape.
- *G is a centroid on a movement axis.* Reproduction contributes no displacement (0), modulation half (0.5), escape full (1.0); G is the intensity-weighted mean position on $[0, 1]$. The poles are definitional, not empirical knobs.

A signed gradient. G folds anchoring and flight into a single non-negative magnitude. We add a complementary signed variant,

$$G^{\pm} = -1.0\cdot R + 0.5\cdot D + 1.0\cdot E \in [-2, +2], \tag{6}$$

which renders territorial anchoring (negative) versus active flight (positive) on one axis, with modulation near the middle. $G^{\pm}$ is reported alongside G; it does not replace the RDE triplet, which remains the primary, decomposable output.

On the non-orthogonality of R, D, and E. A reviewer correctly noted that the theory treats the three lines as interdependent modes of becoming, not independent axes. The equations honor this: because the intensities enter through shares with $p_M+p_m+p_F = 1$, the triplet is *compositional* and lives on the 2-simplex, which is exactly what Figure 2 displays. R, D, and E are therefore constrained, not orthogonal — raising one share lowers the others— so the design encodes the trade-off the theory predicts rather than assuming independence. Across the profiles and the 16-text battery, R and E are strongly negatively associated and D occupies the interior; we report the full empirical correlation matrix with the released code rather than claiming an independence the construction does not have.

4.5 Velocity for Short Texts

A naive velocity (edges per length) saturates on short interpretations. We refine it as a bounded weighted combination of four normalized terms, each in $[0, 1]$:

$$v = w_1\, T + w_2\, V + w_3\, A + w_4\, L, \qquad \textstyle\sum_i w_i = 1, \tag{7}$$

where T is subtype **turnover** (distinct operator subtypes over operator tokens); V is signifier **diversity** (type–token ratio over content signifiers); A is molar↔molecular↔flight **alternation** (machine switches along linear order, over $\max(1, \text{operators}-1)$); and $L = \log(1+\text{operators})/\log(1+\text{tokens})$ is **logarithmic density**, which replaces the linear ratio that caused saturation. We use $w = (0.3, 0.2, 0.3, 0.2)$, fixed across runs.

Worked example. The two-clause interpretation "*the family is sacred, but perhaps not for everyone*" yields operators {`naturalizer+religious_authority` (molar), `modalizer` *perhaps* + `partial_marker` *not for everyone* (molecular)}: 4 operator tokens across 3 distinct subtypes ($T = 0.75$); content signifiers *family, sacred, everyone* over a 7-token span give $V \approx 0.43$; the order molar→molar→molecular→molecular has one machine switch over 3 gaps ($A \approx 0.33$); $L = \log 5/\log 8 \approx 0.77$. Then $v \approx 0.3(0.75)+0.2(0.43)+0.3(0.33)+0.2(0.77) \approx 0.56$, in the reported short-text band (0.4–0.7) rather than the saturated ≈ 1.0 a naive metric returns.

4.6 Deployment and Reproducibility

The prototype runs as a **static single-file frontend** plus an **edge worker** that proxies the analysis API with protected credentials and CORS handling. This gives a zero-cost free tier (∼100k requests/day) with no cold starts and no server to provision. The edge worker holds all credentials; the frontend never sees them, and the same worker fronts the `/api/ngrams` endpoint and the dynamic-generation route. For end-to-end reproducibility we release: the 526-pattern library with per-language and per-subtype counts; the seed corpora and the centroid script; the embedding checkpoints (LaBSE, XLM-R); the RDE and velocity implementations; the curated provocateur corpus with expert-review notes; the n-gram weighting tables; and the full deployed frontend/worker. All weights are fixed across runs with no per-input tuning.

Table 2: R/D/E scores, composite gradient G, and signed gradient $G^{\pm}$ across five profiles.

Profile	R	D	E	G	$G^{\pm}$
Russian conservative	0.61	0.00	0.00	0.00	−0.61
Russian diaspora	0.00	0.74	0.00	0.50	+0.37
Trans Russian exile	0.29	0.00	0.79	0.73	+0.50
Mexican *malinch.*	0.00	0.29	0.71	0.86	+0.86
Mapuche speaker	0.00	0.72	0.43	0.69	+0.79

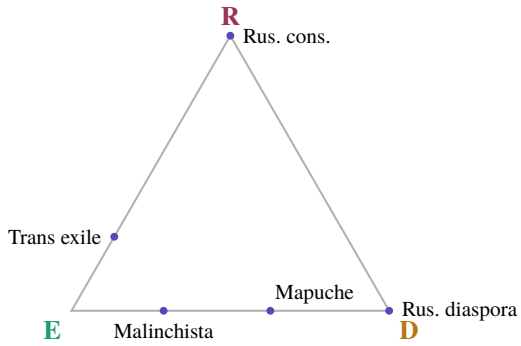

Figure 2: Five profiles in the ternary R/D/E simplex. The compositional constraint $p_M{+}p_m{+}p_F{=}1$ places every profile inside the triangle.

5 Five Extreme User Profiles

We illustrate the protocol on five users facing the same theme (sexism or racism, language-matched). Each scores distinctly (Table 2, Figure 2). The system distinguishes the Russian conservative (who reaffirms their own territory) from the *malinchista* (who scores high on E but toward an external field) by examining **chain provenance**: where each signifier comes from. The *malinchista*'s escape lands in the provocateur's adjacent territory (assimilation); the trans exile's escape lands in flight markers from a distinct semiotic regime (decomposition). Same gradient class, different navigational meaning —a distinction $G^{\pm}$ surfaces partially and provenance resolves fully.

The system distinguishes assimilationist E (*malinchista*) from cosmological-flight E (Mapuche) through chain provenance, not the scalar gradient alone. We stress that this distinction is the *authors'* operational interpretation, not an adjudicated fact: whether introducing *ñuke mapu* constitutes a line of flight or a *reterritorialization* onto an indigenous molar code is a question only Mapudungun speakers and indigenous-cosmology scholars can settle. Asserting it unilaterally would reproduce the colonizing move the protocol is meant to refuse. Section 7 therefore lists participatory validation with the relevant communities as a required step, and the score is presented to

users as a hypothesis to be examined, not a verdict.

6 Deployment Tiers

6.1 Free Tier

The Free dashboard presents the composite gradient, the R/D/E breakdown, the signed gradient $G^{\pm}$, the user's interpretation with operators highlighted in machine-specific colors, and a one-sentence summary. Every number ships with a plain-language explanation, and the interface uses an accessible earth-tone palette. Crucially, the breakdown is now **operator-attributed and actionable**: each of R, D, and E lists the specific operators (and weights) that produced it, so a user learns not merely *that* they reaffirm but *which* signifiers and subtypes drive the score. For a user who wants to move along the gradient, the dashboard surfaces the highest-weight operators to revise and, where applicable, the nearest molecular or flight reformulation of a flagged molar operator (e.g. replacing a `naturalizer` with a `graduator`), turning the score into a concrete editing target rather than an opaque verdict.

6.2 Pro Tier

Pro users inspect the provocateur's own RDE decomposition, manipulate the gradient sliders (including the $G^{\pm}$ axis), and retrieve real-world discourses from the open web that match the manipulated profile. This serves researchers studying how specific discursive territories circulate in real corpora.

6.3 Admin / Corporate Tier

Organizations whose personnel engage in intercultural translation and interpretation —diplomatic services, NGOs, multilingual support teams, content moderation operations, journalism organizations— face a practical question: how much territorial reaffirmation, modulation, or escape is acceptable from their personnel when handling sensitive cross-cultural material?

An aid organization in a conflict zone may need translators to score low on R (high reproduction of either side's molar chains compromises neutrality), high on D (modulation is professionally appropriate), and moderate on E (creative reframing welcome but not unbounded). A journalism organization may want different tolerances. The Admin tier lets the organization define tolerance bands on each axis and visualize how each team member's recent interpretations fall within those bands. This is descriptive of how the team's semantic flow distributes, not prescriptive of which translations are correct.

This tier addresses a problem the StereACuLT call names explicitly: misaligned safety behaviors when culture-agnostic moderation is deployed in culturally varied contexts. A single global threshold for "acceptable stereotype neutralization" is itself a colonizing move. Tolerance bands let each deployment context define its own acceptable distribution, with the protocol providing the measurement instrument.

7 Evaluation

We separate two things the original submission ran together: a **construct demonstration** (which we have) and an **external validation** (which we do not yet have, and no longer claim).

What the constructed examples establish. On the five profiles and a battery of 16 short and long texts across the 8 languages, each text was constructed to exhibit predominantly molar, molecular, flight, or mixed activity; the detector classified each consistently with that design, with no tuning between runs and the weights held fixed. Detector v2 reproduced the v1 classifications and additionally handled the Arabic and German constructions that v1 missed. We are explicit that this shows *internal consistency* —the instrument behaves as the theory predicts on inputs built to the theory— and *not* that it tracks how real users interpret discourse. A theory confirming its own constructed cases is a sanity check, not evidence of validity.

What is therefore still unvalidated, and the protocol that addresses it. The following studies are specified and underway. (a) *Inter-annotator agreement.* Multilingual annotators independently label operators on naturally occurring interpretations; we report Krippendorff's α per subtype and treat low-agreement subtypes as unreliable rather than averaging the disagreement away. (b) *Ablation and error metrics.* Layer A alone, Layer B alone, and A+B are compared against the annotated gold set with per-subtype precision, recall, and F1, plus a confusion analysis. (c) *The irony/quotation confound.* A dedicated subset of ironic and quoted molar uses measures the false-positive rate the detector currently cannot avoid. (d) *Participatory validation.* The flight-vs-modulation classification of indigenous and minoritized signifiers —*ñuke mapu* foremost— is adjudicated *with* Mapudungun speakers and indigenous-cosmology scholars, not by the authors; their disagreement is reported as a finding, not corrected toward the model. (e) *Field study.* Professional translators, interpreters, journalists, and aid workers use the instrument on real material and assess, via structured interviews

and task outcomes, whether the operator-attributed RDE feedback is meaningful and actionable.

The n-gram grounding was checked for face validity on Spanish: high-transversal signifiers (*familia*, *naturaleza*, *orden*) received the highest molar weights, consistent with the theory; this too is face validity, not external validity, pending the corpus study extended to the other seven languages.

We do not claim the protocol replaces stereotype benchmarks; we claim it adds an instrument that captures what users *do* with stereotype material, not just what models output. The two are complementary.

8 Discussion: Why Automated Measurement, and What It Does Not Replace

A reviewer asked whether organizations could get the same self-reflection more cheaply by having translators self-assess on R/D/E, or more richly through expert qualitative code review. The objection is fair and worth answering directly.

Self-assessment fails on three counts the instrument is designed for. It is introspectively biased on exactly the sensitive themes at issue (a user reaffirming a territory is the least likely to report doing so); it is **not comparable across languages and people**, since each respondent applies a private rubric; and it cannot deliver operator-level attribution. Expert qualitative review is the gold standard for nuance but does not scale to ongoing, multilingual workflows and is itself **not reproducible** across reviewers —the very property Section 7's agreement study interrogates. The automated detector contributes what neither alternative does: a reproducible, auditable, language-matched measurement with operator-level granularity and a molar weighting grounded in measured transversal presence, producible at the rate of incoming work.

We do *not* claim the detector is more nuanced than a human expert or that it should replace expert review. The honest positioning is complementary: the instrument is a screening and reflection layer that surfaces distributional patterns and concrete editing targets, which experts and the users themselves then interpret. Its value is realized only if the field study shows the operator-attributed feedback changes what practitioners can see and do; absent that result, the case for the overhead remains a hypothesis, and we mark it as one.

9 Conclusion

We have presented a rhizomatic protocol for measuring user interpretation of stereotype-bearing discourse, grounded in the distinction between molar, molecular, and flight machines. In this revision the weights are derived rather than asserted, molar weighting is anchored in corpus-measured transversal presence, the theory→computation mapping is stated term by term, the velocity metric and Layer B are fully specified, and the deployment is reproducible end to end as a zero-cost serverless system. We have also drawn a sharp line between what the constructed examples demonstrate (internal consistency) and what remains to be validated (external validity, organizational utility, and —above all— the classification of minoritized signifiers, which must be adjudicated with the relevant communities), specifying the protocol that addresses each. The instrument is direction-agnostic, multilingual at the design level, deployed, and designed for both individual reflective use and organizational calibration, with that utility framed as a hypothesis the field study will test rather than a result in hand. It refuses two assumptions widespread in stereotype evaluation: that meaning is hierarchically derived (Chomsky's tree) and that bias has a single corrective direction (English-centric benchmarking). The instrument is deployed publicly as the *Semantic Symbiont (Gradients of Alterity)* and is being integrated into the *Computing Multiplicity* platform alongside a live *Observatory of Monolinguism* that tracks language extinction and epistemic loss, targeting the international translation and interpretation industry. We release the code, the pattern library, the curated corpus, and the deployed prototype.

Limitations

The Layer A pattern library is hand-curated and reflects the authors' linguistic competence; with 8 languages and 526 patterns, coverage remains partial. The n-gram grounding of molar weight is currently computed for **Spanish only** (Google Books Ngram, 2000–2019); molar weights in the other seven languages still rely on the v2 pattern inventory, and the Google Books register over-represents edited print, a known bias in any transversal-presence proxy. Layer B requires seed corpora whose curation is consequential; biased seeds produce biased machine attributions. The dynamic-generation route introduces the generating model's own biases into the provocateur, which is why content warnings and expert review are retained. The detector cannot distinguish ironic or quoted use of molar operators from sincere use; the irony/quotation false-positive rate is a quantified weakness, not a solved one. The operator-attributed feedback is new

and its practical usefulness is itself unvalidated, pending the field study. The Layer B threshold, the velocity weights, and the geopolitical-relevance coefficients are released configurations chosen on development data, not globally optimal values, and should be re-tuned per deployment. Organizations using the Admin tier must guard against using tolerance bands as performance metrics applied to individual workers; treating a descriptive, contextual measurement as evaluative risks reproducing exactly the segmental, normalizing control that Deleuze and Guattari critique. Most importantly, our present evidence is a construct demonstration on curated inputs, **not** external validation: the inter-annotator, ablation/error, and field studies are necessary before any claim of validity, and the classification of indigenous and other minoritized signifiers must be adjudicated with the relevant communities before deployment in contexts that affect them.

Ethical Considerations

The protocol measures personal interpretive movement on sensitive themes (racism, sexism, etc.). User data must not be retained without explicit consent and must never be used to assess individuals against organizational tolerance bands without their knowledge. The Admin tier is designed to surface organizational distribution patterns, not to evaluate workers. We deliberately avoid identity-based demographic questions in calibration; the survey asks about relative position in power systems, not about identity categories. The provocateur corpus contains harmful discourse by construction; in deployment, content warnings precede every provocateur (curated and generated alike), users can skip themes, and the corpus is reviewed by domain experts.

References

Louise Amoore. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Duke University Press.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proc. ACL-IJCNLP*, pages 1004–1015.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalBench: A robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of LLMs. arXiv:2410.02677.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Gilles Deleuze and Félix Guattari. 1987. *A Thousand Plateaus: Capitalism and Schizophrenia*. University of Minnesota Press. Trans. Brian Massumi; orig. 1980 as *Mille Plateaux*.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. arXiv:2306.16388.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CulturePark: Boosting cross-cultural understanding in large language models. In *NeurIPS*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proc. ACL-IJCNLP*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proc. EMNLP*, pages 1953–1967.

Luciana Parisi. 2019. The alien subject of AI. *Subjectivity*, 12(1):27–48.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of ACL 2022*, pages 2086–2105.