Toward a Geometric View of Multilingual Intent Detection: Linguistic Packaging vs. Semantic Resolution

Hacia una Visión Geométrica del Reconocimiento Multilingüe de Intenciones: Empaquetado Lingüístico vs. Resolución Semántica

Eduardo Sanchez-Karhunen, Jose F. Quesada-Moreno, Miguel A. Gutiérrez-Naranjo
Dep. Computer Science and Artificial Intelligence. Universidad de Sevilla, 41004, Sevilla, Spain
{fesanchez, jquesada, magutier}@us.es

Abstract: Trustworthy intent detection is hindered by lack of interpretability in deep learning models. While recent research indicates Recurrent Neural Networks (RNNs) solve high-cardinality intent detection tasks via low-dimensional manifolds quantified by a Functional Dimensionality (*FD*), it remains unclear if these findings represent general principles or linguistic artifacts. We investigate *FD* stability using MASSIVE dataset across a diverse set of languages: English, Spanish, Finnish, and Turkish. Our results suggest a dual geometric regularity. First, we observe inter-lingual stability within a fixed embedding paradigm, RNNs converge to statistically stable Semantic Resolution subspaces despite differences in Linguistic Packaging. Second, we observe inter-paradigm robustness of distillation process: while transitioning from static to contextualized embeddings induces massive expansion of geometric volume (*ID*), Distillation Efficiency ($\Omega = FD/ID$) remains statistically stable. These findings provide preliminary evidence for a geometric view of multilingual intent detection, where semantic resolution appears to be governed by task-mandated constraints rather than Linguistic Packaging or representational schema.
Keywords: Intent Detection, Trustworthy AI, RNNs.

Resumen: La falta de interpretabilidad en modelos de deep learning compromete la confiabilidad en la detección de intenciones. Investigaciones recientes sugieren que las RNNs resuelven tareas de alta cardinalidad mediante variedades de baja dimensión, cuya eficiencia se cuantifica usando una Dimensionalidad Funcional. Se desconoce si estos hallazgos son principios generales o artefactos lingüísticos. Analizamos la estabilidad de esta *FD* sobre el corpus MASSIVE en inglés, español, finés y turco. Nuestros resultados sugieren una doble regularidad geométrica. En primer lugar, observamos estabilidad interlengua: dado un esquema de embeddings, pese a las diferencias morfológicas en el empaquetado lingüístico, las RNNs convergen en subespacios semánticos estadísticamente estables. Asimismo, observamos una robustez interparadigma: la transición de embeddings estáticos a contextuales induce una explosión del volumen geométrico total, pero la eficiencia de destilación permanece estadísticamente estable. Todo ello proporciona evidencias preliminares de una visión geométrica de la detección multilingüe de intenciones, donde la resolución semántica depende de la tarea en lugar del formato lingüístico o de embeddings.
Palabras clave: Detección de Intenciones, IA Confiable, RNNs.

1 Introduction

Spoken Language Understanding (SLU) is a cornerstone of modern human-computer interaction, enabling digital assistants to map inputs into actionable semantic intents (Tur and De Mori, 2011). While deep learning models achieve remarkable performance in these tasks (Ravuri and Stolcke, 2015),(Abro et al., 2022), their internal operations remain largely opaque. This lack of transparency creates a critical tension: as models grow more complex, they become less interpretable, yet high-stakes applications require the accountability by design essential for building trustworthy dialogue systems (Arrieta et al., 2020).

Dynamical systems theory offers a framework for analyzing sequential processes. The recurrence of RNNs allows sequence processing to be viewed as a non-linear dynamical system, where sentences are represented as trajectories through a state space (Sussillo and Barak, 2013). This perspective facilitates a mechanistic framework for interpretability by utilizing RNNs as a methodological laboratory. Once this geometric view is solidified, its principles could be adapted to non-sequential self-attention of Transformer-based models (Vaswani et al., 2023).

Recent breakthroughs in this paradigm show that despite vast dimensionality of RNN architectures, task-solving logic often collapses into hyper-efficient, low-dimensional manifolds (Aitken et al., 2021). In intent detection, evidence indicates that RNNs resolve semantic intent by partitioning state space into well-separated clusters (Sanchez-Karhunen, Quesada-Moreno, and Gutiérrez-Naranjo, 2026a). This organization is quantified by *Functional Dimensionality* (FD), a task-aware metric identifying minimum subspace required to preserve semantic integrity of classification. Crucially, functional manifold remains anchored even as task complexity scales, suggesting an increasingly efficient geometric strategy (Sanchez-Karhunen, Quesada-Moreno, and Gutiérrez-Naranjo, 2026b).

Despite these findings, current evidence is limited by two critical factors. First, existing research remains almost exclusively English-centric, leaving stability of manifold across disparate linguistic structures unexplored. Second, prior works utilized embeddings trained from scratch, allowing input representation to co-evolve alongside recurrent layers. This creates a gap: is low-dimensional manifold a language-independent requirement of Semantic Resolution, or is it an artifact of representational co-evolution?

To address this, we leverage MASSIVE dataset (FitzGerald et al., 2023) to conduct a *Typological Stress Test* evaluating recurrent core across a quartet of languages with diverse morphological formats (English, Spanish, Finnish, and Turkish), alongside a *Representational Stress Test*. This dual approach allows us to investigate if semantic bottleneck is a core mechanistic principle by evaluating stability of the manifold across both linguistic structures and embedding paradigms.

To clarify the intersection between this geometric analysis and SLU concepts, we explicitly define Linguistic Packaging as the surface-level structural formatting, token sequence lengths, and morphological characteristics through which an utterance encodes its underlying semantic intent.

By moving beyond English-centric scalability, we provide preliminary evidence that *Semantic Resolution* is a task-mandated geometric constraint rather than a byproduct of the structural format or the representational schema of the input.

2 Our Contributions and Paper Organization

This paper shifts established geometric interpretability framework from discovery and scalability of semantic manifold to a typological investigation of its stability across linguistic and representational landscape. Our main contributions are:

Inter-lingual Stability (O_1): We provide initial empirical evidence suggesting that, within a fixed embedding paradigm, low-dimensional organization of hidden state space exhibits language-independent characteristics. Our analysis indicates that despite radical differences in Linguistic Packaging across English, Spanish, Finnish, and Turkish, RNNs converge to stable Semantic Resolution subspaces of low dimensionality. These findings are supported by a One-Way Anova, which finds no significant differences in FD across typologies, suggesting that semantic resolution may function as a task-mandated regularity across diverse morphological systems.

Inter-paradigm Robustness (O_2): We show that while transition from static embeddings to contextual representations induces an extreme expansion of the total geometric volume of state space (ID), the functional core (FD) remains largely anchored. Our results indicate that *Distillation Efficiency* ($\Omega = FD/ID$) remains statistically stable across paradigms, identifying the recurrent core as a consistent filter that discards representational volume as linguistic noise.

Trajectory-Length Independence: We decouple Semantic Resolution from computational path length. Regression analysis suggests FD dimensionality is primarily influenced by task's semantic cardinality rather than input sentence length.

Rest of the paper is structured as follows: Section 3 reviews related work; Section 4 defines mathematical framework for dynamical systems; Section 5 formalizes our research objectives; Section 6 describes experimental setup; Section 7 presents our results and statistical validation; Section 8 discuss applications and Section 9 presents conclusions.

3 Related Work

Early interpretability research in RNNs identified neurons capable of detecting localized patterns, such as quotation marks or parentheses (Karpathy, Johnson, and Fei-Fei, 2016). However, these localized findings did not necessarily correlate with generalization, suggesting that RNNs encode information in a distributed fashion across the hidden state (Morcos et al., 2018).

By framing RNNs as dynamical systems, researchers have shifted focus toward the global geometry of the state space, uncovering that models utilize low-dimensional manifolds to execute complex computations. In sentiment analysis, for instance, RNNs learn 1D-line attractors to accumulate emotional evidence over time (Maheswaranathan et al., 2019). This framework was later generalized to multiclass text classification, proposing that an N-class problem forms an $(N-1)$-dimensional manifold (Aitken et al., 2021).

Recent studies on intent detection revealed a distinct geometric strategy: unlike continuous evidence-integration, intent detection favors the formation of discrete, well-separated clusters in the hidden space (Sanchez-Karhunen, Quesada-Moreno, and Gutiérrez-Naranjo, 2026a). High-cardinality scaling was addressed using 150-intent benchmarks, discovering networks maintain robust, intent-specific clusters within a low *FD* that grows sub-linearly with task complexity (Sanchez-Karhunen, Quesada-Moreno, and Gutiérrez-Naranjo, 2026b).

Despite robust findings, geometric interpretability remains almost exclusively English-centric. Whether Semantic Resolution is a fundamental mechanistic requirement or merely a byproduct of English syntax remains unknown. Our work extends this lineage by conducting a Typological and Representational Stress Test, investigating if low-dimensional manifolds remain stable across disparate Linguistic Packaging formats and embedding paradigms.

4 Background

4.1 RNNs as Dynamical Systems

RNNs process input sequences $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ by maintaining an internal memory through recurrent connections. As shown in Figure 1, the *hidden state* $\mathbf{h}_t \in \mathbb{R}^n$ is updated at each timestep based on current input $\mathbf{x}_t$ and previous state $\mathbf{h}_{t-1}$ via a nonlinear function $\mathbf{F}$:

$$\mathbf{h}_t = \mathbf{F}(\mathbf{h}_{t-1}, \mathbf{x}_t) \tag{1}$$

This architecture defines a *discrete-time nonlinear dynamical system*. Under this framework, processing of a sentence is represented as a *trajectory* $(\mathbf{h}_1, \ldots, \mathbf{h}_T)$ through an n-dimensional *state space* (Martelli, 1999).

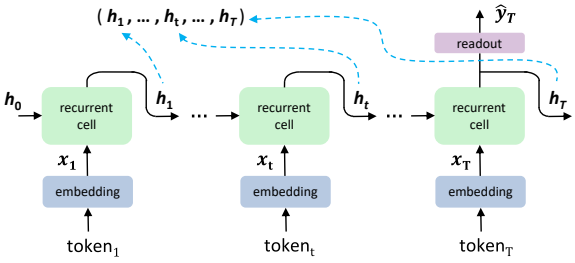

Figure 1: Unrolled RNN for a many-to-one classification task. Illustration of the hidden state trajectory formation $(\mathbf{h}_1, \ldots, \mathbf{h}_T)$.

4.2 Inference: Cluster formation and Readout Alignment

As illustrated in Figure 1, for many-to-one classification tasks, like intent detection, only final hidden state $\mathbf{h}_T$ is utilized for prediction. This state is passed to a linear readout layer to produce an N-logits vector, $\hat{\mathbf{y}}_T \in \mathbb{R}^N$, where N denotes the number of intents:

$$\hat{\mathbf{y}}_T = \mathbf{W}\mathbf{h}_T + \mathbf{b} \tag{2}$$

Here, $\mathbf{W} \in \mathbb{R}^{Nxn}$ and $\mathbf{b} \in \mathbb{R}^N$ are the weight matrix and bias, respectively. Prediction is determined by the dot product of $\mathbf{h}_T$ and rows $\mathbf{r}_i$ of $\mathbf{W}$, known as *readout vectors*.

Mechanistically, model resolves intent-detection by ensuring that intent-specific clusters in hidden space align with their corresponding readout vectors $\mathbf{r}_i$ (Figure 2). *Semantic Resolution* is thus characterized by the network's capacity to steer disparate trajectories toward these stable, well-separated regions. Hence, $\mathbf{h}_T$ provides a stable representation for the readout layer, regardless of the specific trajectory taken (Sanchez-Karhunen, Quesada-Moreno, and Gutiérrez-Naranjo, 2024).

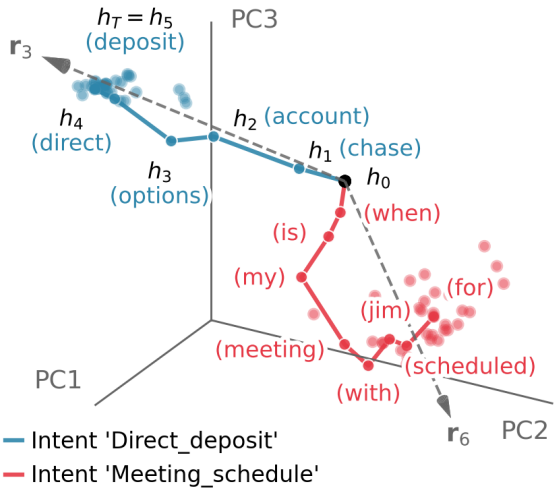

Figure 2: PCA-projected trajectories for sample utterances of two intents. Readout vectors $\mathbf{r}_j$, clusters of final states $\{\mathbf{h}_T\}$, and input tokens are annotated along each path.

4.3 Functional and Intrinsic Dimensionality

To map and quantify geometry of internal representation space, we employ two distinct, complementary dimensionality metrics: Intrinsic Dimensionality (*ID*) and Functional Dimensionality (*FD*). Together, these metrics disentangle total capacity a model allocates for processing a sequence from hyper-efficient subspace used to solve the task:

Intrinsic Dimensionality identifies minimal linear subspace required to retain a specific threshold of total variance within network's hidden layer trajectories. Geometrically, *ID* serves as a proxy for *Linguistic Packaging*: "computational room" that an RNN allocates for processing surface-level syntax, token-level sequences, and variations in morphological density across languages.

To estimate ID, we construct an activation matrix $\mathbf{H}$ by concatenating hidden states $\mathbf{h}_t$ visited across all sequence trajectories and discrete timesteps within train split. We then apply Principal Component Analysis (PCA) to $\mathbf{H}$ to extract the ordered spectrum of eigenvalues. Grounded in variance-explained heuristic established by (Aitken et al., 2021), *ID* is formalized as the minimum number of principal components required to capture 95% of total cumulative variance.

In contrast, **Functional dimensionality** (*FD*) is a task-aware metric that maps Semantic Resolution. It isolates the minimum subspace required to preserve semantic integrity and separation of final intent clusters. While *ID* measures ambient structural volume over paths, *FD* measures target inference geometry over final-state representations $\{\mathbf{h}_T\}$ utilized by linear readout layer.

FD is computed following an iterative process (Sanchez-Karhunen, Quesada-Moreno, and Gutiérrez-Naranjo, 2026b). We establish a baseline partition quality metric (S_{base}) by calculating Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) between ground-truth labels and a set of K-Means clusters (C_{full}) in full-dimensional final states $\{\mathbf{h}_T\}$. Next, we apply PCA to $\{\mathbf{h}_T\}$ and iteratively perform K-means to vectors projected onto top-k principal components. *FD* is formally defined as the smallest k that satisfies $S_k \geq \tau S_{base}$, where $\tau = 0.95$ serves as our functional retention threshold. This approach explicitly quantifies minimum dimensionality to guarantee cluster persistence.

Finally, we introduce **Distillation Efficiency** ($\Omega = FD/ID$) to bridge these two metrics and evaluate the network's capacity to isolate its functional task core from ambient structural packaging. This ratio tracks how effectively the recurrent core acts as a compressor, discarding surface-level linguistic variation to isolate a stable task manifold.

5 Extending Interpretability to Multi-lingual Scenarios

Building on previous findings that RNNs utilize low-dimensional geometric manifolds for English-centric intent detection, we investigate potential generalizability of this mechanism across linguistic and representational boundaries. We hypothesize that if semantic bottleneck is a candidate mechanistic principle of information distillation, it should exhibit a dual regularity, remaining robust against both morphological structure of language and complexity of input representation. To investigate this, we define two primary research objectives:

***O*$_1$: Inter-lingual Stability** investigates stability of low-dimensional organization of hidden space across languages within a fixed representational paradigm. By subjecting recurrent core to a Typological Stress Test (Section 6.4.1) across a quartet: English, Spanish, Turkish, and Finnish, we evaluate whether Semantic Resolution (*FD*) remains stable regardless input Linguistic Packaging.

We hypothesize that if Semantic Resolution is governed by task semantic complexity rather than input structure, *FD* should show no significant differences across linguistic families. We aim to provide preliminary evidence for the *Constant Resolution Principle*, suggesting that manifold complexity is dictated by semantic goal rather than morphological traits (Section 7.1).

O_2: **Inter-paradigm Robustness** investigates how the transition from static embeddings to high-dimensional contextual representations affects the functional manifold. By subjecting the model to a Representational Stress Test (Section 6.4.2) we seek to observe if Ω remains stable, even as the total geometric volume (*ID*) expands to accommodate high-variance contextual signals.

We propose a mechanism where RNNs function as *Semantic Distillers* (i.e. filters), consistently discarding representational volume as extraneous noise to preserve the functional semantic core. This perspective allows us to investigate the *Functional Anchoring* of distillation logic across embedding paradigms, framing the geometric bottleneck as a task-mandated constraint rather than a mere architectural artifact.

6 Experimental Design

6.1 Datasets

To evaluate cross-lingual stability of hidden state manifold, we utilize the MASSIVE (Multilingual Amazon SLU resource package) dataset (FitzGerald et al., 2023). This parallel corpus, containing one million utterances across 51 languages, offers two critical advantages for conducting a typological investigation of geometric regularity:

Semantic Parallelism: Professional translations across languages ensure identical Semantic Resolution while Linguistic Packaging varies, isolating task-mandated constraints from structural input noise.

High Intent Cardinality: 60 intents grouped across 18 domains, providing the representational pressure necessary to validate the geometric manifold as a stable, language-independent solution rather than an artifact of task simplicity.

We utilize the standard parallel splits of the benchmark dataset, comprising 11,514 train, 2,033 validation, 2,974 test utterances uniformly parallel across each language.

6.2 Typological Diversity and Language Selection

Our Typological Stress Test utilizes four languages, challenging RNNs to resolve identical semantic intents across radically disparate structural formats:

English (Indo-European): Serves as an analytic baseline and word-order dependent language with low morphological density.

Spanish (Indo-European, Romance): A fusional language that provides a medium-density morphological comparison within the Indo-European family.

Turkish (Turkic): An agglutinative language used to evaluate the stability of the functional core against morphological expansion, where semantic information is compressed into high-variance tokens.

Finnish (Uralic): A highly agglutinative language representing the extreme end of complexity gradient to test limits of information density within the structural packaging.

6.3 Model and Training Setup

The following section details the specific architectural components and training procedures to facilitate our Stress Tests.

We implement a four-layer architecture consisting of: (1) a text vectorization and frozen pre-trained embedding layer; (2) a unidirectional recurrent layer (75 units); (3) a dropout layer (rate=0.2); and (4) a dense output layer with a softmax.

A central methodological decision is the use of frozen parameters for both static FastText (FT) (Bojanowski et al., 2017) and contextualized BERT embeddings (Devlin et al., 2019). This achieves two primary objectives:

Functional Isolation: Freezing input layer forces recurrent core to perform the entirety of the semantic resolution. This ensures that our measured *FD* reflects model's active computational strategy (distillation logic) rather than the lookup efficiency or co-evolution of a trainable embedding layer.

Linguistic leveling: By utilizing pre-trained sub-word information, we ensure morphologically rich languages like Turkish and Finnish do not suffer from a data-disparity disadvantage. Model is not forced to learn complex morphemic representations from scratch, preventing training-time artifacts from obscuring geometric analysis.

To handle morphological complexity of Turkish and Finnish without normalizing

away their typological traits, we represent each lexical word as a single vector (300-d for FT; 768-d for BERT). We utilize FastText internal character n-gram aggregation and apply a sub-word mean-pooling strategy over BERT's WordPiece tokens (Reimers and Gurevych, 2019). This maintains original word-level token counts, ensuring trajectory length reflects language's natural sequence. This approach preserves Linguistic Packaging by allowing individual tokens to contain dense information while keeping computational path lengths.

Models were trained with ten random seeds by minimizing cross-entropy via Adam optimizer (Kingma and Ba, 2015) (batch size of 16). Learning rate ($\eta = 2.5 \times 10^{-3}$) was halved after second epoch, and early stopping (patience = 2) was applied. We only analyze models that achieve a performance threshold ($\approx$ 80%). This methodological choice ensures that our interpretability analysis is grounded in functional competence. We focus on state space of models that have successfully discovered a stable mechanistic strategy for the task, rather than a degenerate or under-trained representation.

While both LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho et al., 2014) cells achieved equivalent training performance, we restrict our geometric evaluations exclusively to GRUs to preserve conciseness. Preliminary state-space tracking confirmed that both gated variants exhibit highly consistent topological regularities. Consequently, our empirical conclusions are formally qualified as reflections of GRUs rather than generalized recurrent networks.

6.4 Analysis Procedure

To evaluate Section 5 hypotheses, our pipeline isolates structural properties of hidden state activations from their functional Semantic Resolution. We use statistical validation to distinguish architectural artifacts from task-mandated geometric patterns.

6.4.1 Analysis for O_1: Inter-lingual Stability

We investigate whether low-dimensional semantic manifold operates as a language-independent principle within fixed representational paradigms. By subjecting recurrent core to a Typological Stress Test across a diverse quartet (English, Spanish, Finnish, and Turkish), we isolate morphological packaging from functional resolution. This stage evaluates state space by distinguishing between model's learned structural capacity and its functional generalization to ensure findings reflect a stable mechanistic strategy:

(1) Linguistic Packaging (*ID*): Computed using training hidden states to quantify learned structural "footprint" required for morphological processing. This captures total geometric volume allocated during learning of disparate morphological formats.

(2) Semantic Resolution (*FD*): Derived from test hidden states to identify minimal subspace required to preserve the semantic integrity and separation of intent clusters.

Under proposed *Constant Resolution Principle*, this functional core should exhibit stability across languages, reflecting semantic complexity of task rather than input morphology. To statistically validate these observations, we employ a One-Way ANOVA (Dror et al., 2020) to evaluate whether *FD* exhibits language-independent stability across fixed embedding paradigms. Finally, we utilize linear regression to correlate average sequence length (T_{avg}) with *FD* to investigate *Trajectory-Length Independence*, ensuring manifold's complexity is mandated by semantic goal rather than path length.

6.4.2 Analysis for O_2: Inter-paradigm Robustness

This stage seeks to provide evidence on whether geometric bottleneck is an emergent property of task or an artifact inherited from input representation. We subject RNN to a Representational Stress Test by replacing 300-d static FastText embeddings with 768-d frozen BERT representations. We utilize language-specific BERT models for English (Devlin et al., 2019), Spanish (Cañete et al., 2023), Finnish (Virtanen et al., 2019) and Turkish (Schweter, 2020).

This transition forces recurrent layer to manage a 2.6$\times$ expansion in representational complexity, introducing high-variance signals. We monitor network's reaction through two opposing geometric forces:

(1) Structural Expansion: We measure the expansion of the total geometric volume occupied by activations (*ID*), assessing the extent to which these high-variance signals propagate into the recurrent layer.

(2) Functional Anchoring: We track the stability of the semantic core (*FD*) to investigate if the model maintains its low-

dimensional task solution despite the expansion of the surrounding structural space.

To bridge these dimensions, we use Ω to quantify network's capacity to filter task-irrelevant noise while preserving semantic core of an utterance. We evaluate Ω as a potential mechanistic regularity, essentially a functional signature that characterizes how the model distills signal from variability. Robustness of this process across paradigms is statistically assessed using Welch's T-test (Gorman and Bedrick, 2019) to account for potential variance inequality induced by structural expansion of state space. Effect sizes are measured using Cohen's d (Lin, Lucas, and Shmueli, 2013).

7 Results

7.1 Results for O_1: Inter-lingual Stability

Our analysis provides preliminary evidence for a task-mandated Constant Resolution Principle, observing a decoupling between how a language is "packaged" and how it is "resolved" by recurrent architecture. As illustrated in Figure 3 and Table 1, functional subspace remains statistically stable across disparate typologies within fixed representational paradigms when processing held-out test trajectories.

Intra-Paradigm Stability: Regardless of typological distance between selected languages, RNNs converge to a stable Semantic Resolution with *FD* for 60-intent task anchored within a narrow range (11-12 for FastText; 17-20 for BERT) (Figures 3a and 4b). Our One-Way ANOVA finds no significant differences in *FD* across the quartet within FastText ($p = 0.380$) or BERT ($p = 0.169$) paradigms (Table 2). These results suggest that dimensionality of semantic manifold is primarily influenced by semantic cardinality of task rather than specific morphological traits.

Sub-linear Scaling and Mechanistic Convergence: The top-25 intents subplot (Figure 3a) reveals a potentially hyper-efficient representational strategy: doubling task complexity from 25 to 60 intents does not linearly double *FD*, requiring only a marginal increase (e.g. from 9 to 12 in English). Mechanistic evidence for this trend is provided by the ARI saturation curves (Figure 3b), where every language reaches a plateau within a narrow *Convergent Zone*.

Trajectory Length Independence: Our regression analysis (Figure 3c) suggests a decoupling between input sequence length and *FD* ($r = 0.29, p = 0.712$, Table 2), providing support for the hypothesis that manifold complexity depends on task's semantic goals rather than the structural path length.

Linguistic Packaging: While semantic resolution (*FD*) is observed to be stable, Linguistic Packaging (*ID*) required to reach it varies significantly ($F = 280.68, p < 0.001$, Table 2), reflecting distinct structural demands of each language. As shown in Figure 3d, total geometric volume fluctuates, highlighting varying structural requirements across quartet.

Consistent Distillation: Despite these structural fluctuations, Distillation Efficiency (Ω) remains consistently at $\approx 0.32\times$. This identifies recurrent core as a filter, discarding extraneous linguistic noise to preserve functional semantic core in these settings.

7.2 Results for O_2: Inter-paradigm Robustness

Transitioning from static to contextualized embeddings provides further evidence that geometric bottleneck is a task-mandated necessity, not a structural artifact. This characterizes recurrent core as a Semantic Filter, consistently preserving a functional core despite significant representational shifts.

Structural Expansion: Transitioning to BERT induces a substantial expansion of the total geometric volume occupied by activations (*ID*) (Figure 4a), which grows by $\approx 60\%$. This expansion likely reflects the network's need to accommodate the high-variance signals inherent in Transformer-based embeddings. Statistical validation via Welch's T-test ($t = 42.91, p < 0.001$) identifies this as an extreme effect ($d = 9.60$), confirming the intensity of the representational stress applied (see Table 2).

Functional Anchoring: While absolute Semantic Resolution (*FD*) increases (e.g. in Spanish from ≈ 12 to 18), it does so strictly sub-linearly compared to the expansion of surrounding structural space. This Functional Anchoring ($d = 2.72$, Table 2) suggests model maintains a hyper-efficient bottleneck for final inference (Figures 4b and 4d).

Stability of Distillation Efficiency (Ω): A key observation from our data is the relative stability of Distillation Efficiency.

Paradigm	Language	Acc. (%)	T_{avg}	Packaging (*ID*)	Resolution (*FD*)	Ratio (Ω)
Static (FastText)	English (EN)	83.3	6.9	38.2 ± 0.6	12.0 ± 1.5	0.31
	Spanish (ES)	80.8	7.2	36.8 ± 0.4	11.9 ± 1.0	0.32
	Finnish (FI)	79.9	4.8	32.4 ± 0.5	11.4 ± 1.1	0.35
	Turkish (TR)	79.8	5.5	38.4 ± 0.5	12.4 ± 1.4	0.32
Context (BERT)	English (EN)	85.6	6.9	57.7 ± 0.5	17.3 ± 3.4	0.30
	Spanish (ES)	82.8	7.2	57.3 ± 0.5	17.6 ± 3.6	0.31
	Finnish (FI)	83.2	4.8	61.8 ± 0.4	$19.0 \pm 2,4$	0.31
	Turkish (TR)	84.2	5.5	62.3 ± 0.5	20.1 ± 2.8	0.32

Table 1: **Linguistic Packaging vs. Semantic Resolution** across typologies (EN, ES, FI, TR) and embedding paradigms. T_{avg} denotes average sequence length (tokens per sentence).

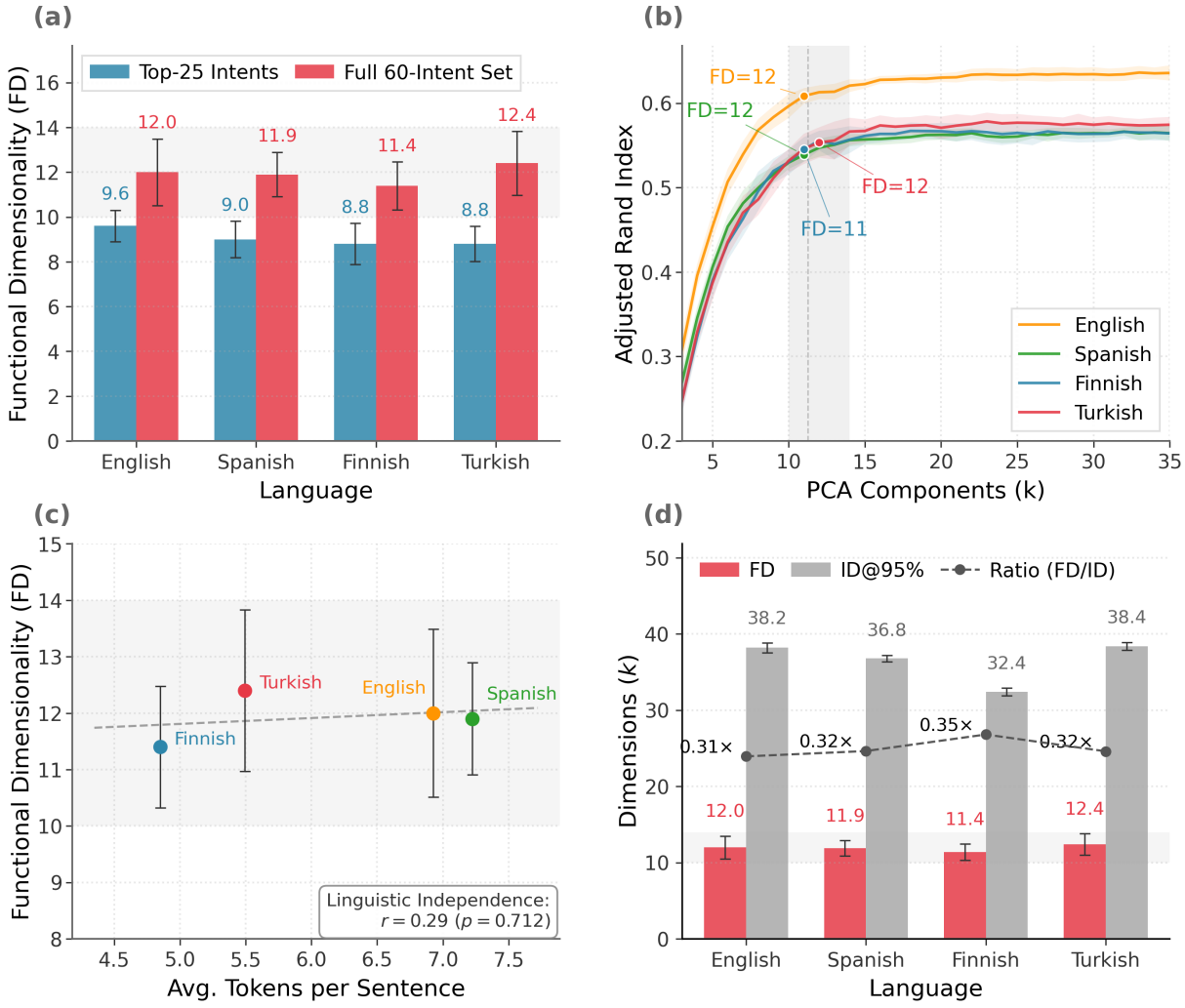

Figure 3: **Geometric Stability across Typologies using FastText. (a)** *FD* comparison across intent cardinalities. **(b)** ARI saturation curves. **(c)** Trajectory-Length Independence. **(d)** *ID* fluctuation with morphological density while FD/ID remains consistent.

The ratio Ω remains statistically stable across paradigms ($p = 0.056, d = 0.43$, Table 2). This suggests that the RNN consistently discards $\approx 70\%$ of the representational volume as linguistic noise to preserve the functional semantic core (Figure 4c).

Finnish Stress Boundary: We observe a localized phenomenon in FastText paradigm, where Finnish exhibits a lower *ID* than rest of quartet. This reduction is absent in Turkish, despite similar morphological density, nor does it persist under BERT,

Hypothesis	Comparison	Stat ($F/r/\|t\|$)	p-value	Effect (d)	Conclusion
Objective O_1: Inter-lingual Stability					
Resolution	FD_{FT} (EN/ES/FI/TR)	$F = 1.06$	0.380	—	Constant Resolution
Packaging	ID_{FT} (EN/ES/FI/TR)	$F = 280.68$	<0.001	—	Linguistic Noise
Length	*FD* vs. Length	$r = 0.29$	0.712	—	Task-Driven
Resolution	FD_{BERT} (EN/ES/FI/TR)	$F = 1.78$	0.169	—	Constant Resolution
Packaging	ID_{BERT} (EN/ES/FI/TR)	$F = 317.58$	<0.001	—	Linguistic Noise
Length	*FD* vs. Length	$r = -0.81$	0.188	—	Task-Driven
Objective O_2: Inter-paradigm Robustness					
Structure	*ID* (FT vs BERT)	$t = 42.91$	<0.001	9.60 [Ext]	Structural Expansion
Anchor	*FD* (FT vs BERT)	$t = 12.18$	<0.001	2.72 [Lrg]	Functional Anchor
Efficiency	Ω_{Pool} (FT vs BERT)	$t = 1.94$	0.056	0.43 [Sml]	Stable Efficiency
Efficiency	Ω_{EN} (FT vs BERT)	$t = 0.66$	0.517	0.30 [Sml]	Stable Efficiency
Efficiency	Ω_{ES} (FT vs BERT)	$t = 0.77$	0.455	0.35 [Sml]	Stable Efficiency
Efficiency	Ω_{FI} (FT vs BERT)	$t = 2.71$	0.015	1.21 [Lrg]	Stress Boundary
Efficiency	Ω_{TR} (FT vs BERT)	$t = 0.03$	0.980	0.01 [Neg]	Stable Efficiency

Table 2: **Statistical validation of geometric regularity.** Analysis of typological stability (O_1) via ANOVA and representational robustness (O_2) via Welch's T-test and Cohen's d.

where Finnish aligns with general structural expansion (Figure 4b). We interpret this as a Representational Stress Boundary: a localized interaction between static sub-word aggregation and typological traits rather than a general property of dense languages.

8 Potential Applications

The observed task-mandated *FD* and stable Ω suggest that geometric analysis may transition from a descriptive framework toward predictive utility. While preliminary, these findings highlight areas where geometric interpretability could inform development of efficient and trustworthy multilingual systems.

Geometric Zero-Shot Cross-Lingual Alignment: O_1 results indicate that disparate languages converge toward Semantic Resolution subspaces of nearly identical dimensionality. This suggests "semantic gap" in multilingual RNNs may be primarily a matter of state space orientation, not dimensionality. Given the stability of these subspaces across quartet, cross-lingual transfer may be investigated as a Procrustes Alignment problem (Wang and Mahadevan, 2008).

Mechanistic Model Compression: The identified $\approx 70\%$ discard rate offers a potential geometrically-grounded ceiling for model pruning. Measured *FD* could inform rank selection for Singular Value Decomposition (SVD) compression. For instance, an observed *FD* of $k = 12$ suggests a projection layer could be compressed to that specific rank with minimal semantic degradation.

9 Conclusion and Future Work

We propose a geometric framework for understanding computational logic of recurrent architectures in multilingual intent detection. By introducing a distinction between Linguistic Packaging (surface-level syntax, token-level sequences, and morphological density) and Semantic Resolution (task-mandated functional requirements), we provide preliminary evidence for a dual geometric regularity that appears to govern information distillation.

Our analysis across a typologically diverse quartet of languages (English, Spanish, Finnish, and Turkish) reveals that the Functional Dimensionality (*FD*) of the intent manifold behaves as a language-invariant principle within our experimental settings. Despite radical differences in morphological structure and sequence length, the RNN converges to a statistically stable semantic resolution. This suggests that geometry of hidden state is not merely a reflection of input syntax, but rather primarily governed by task's semantic cardinality.

Furthermore, we show this geometric bottleneck is remarkably robust to extreme representational paradigm shifts. While transitioning from static FastText to contextual BERT embeddings induces an "Extreme" expansion of total geometric volume (*ID*), Distillation Efficiency (Ω) remains statistically constant. This identifies RNNs as Semantic Filters that discard $\approx 70\%$ of representational volume as packaging noise to preserve

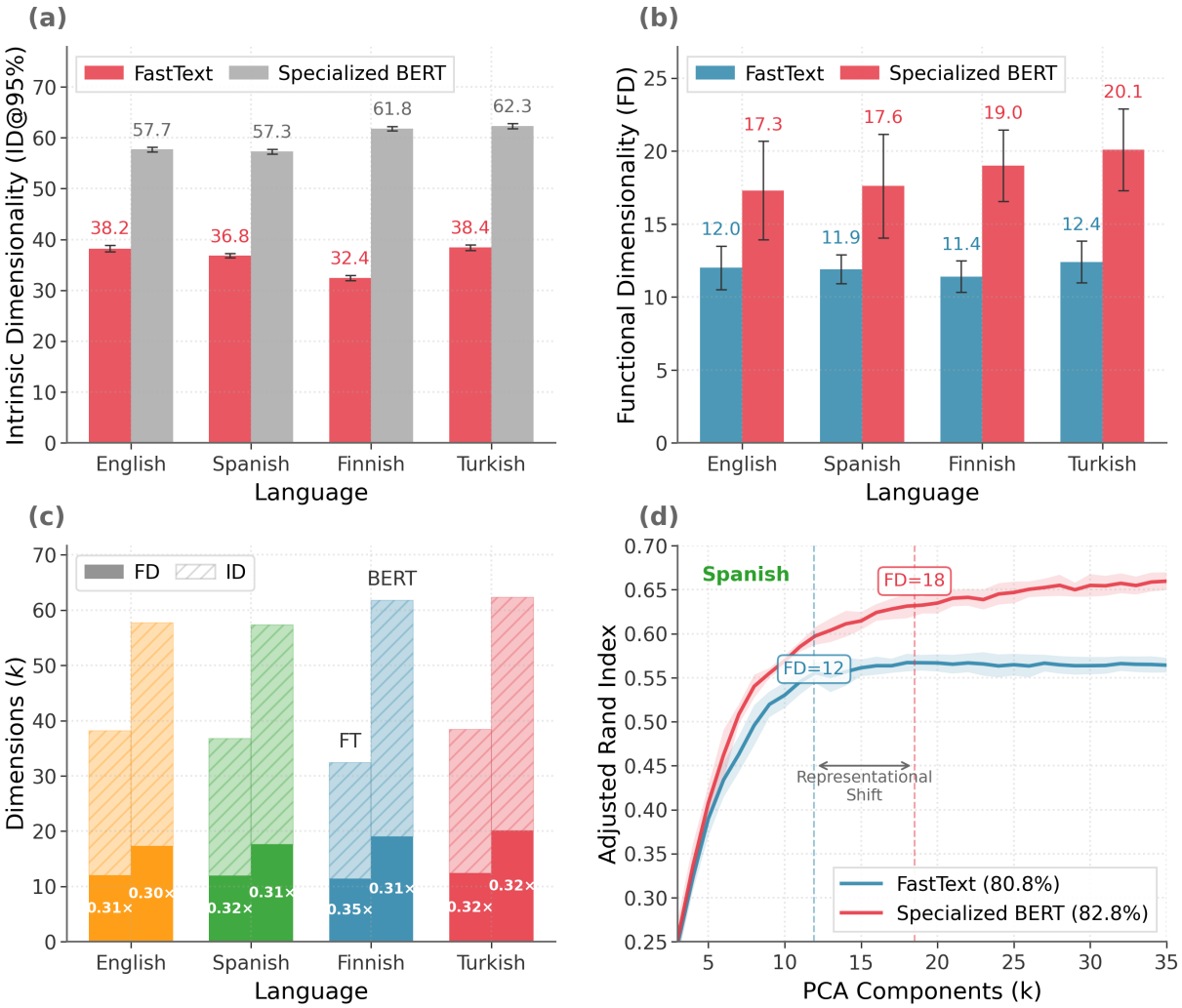

Figure 4: **Representational Robustness across Embedding Paradigms.** **(a)** *ID* expands to accommodate high-variance contextual BERT. **(b)** *FD* remains anchored across paradigms. **(c)** *FD* scales with total manifold volume (*ID*), maintaining a stable Distillation Efficiency ($\Omega \approx 0.32$). **(d)** PCA-ARI saturation for Spanish illustrates the representational shift.

functional core necessary for inference.

While these findings offer significant steps toward a unified geometric view, they represent an initial investigation into stability of neural manifolds. Our results, while robust across the evaluated quartet, should be viewed as a candidate principle requiring further validation. Future research will focus on:

Global Typological Expansion: We intend to extend this analysis to a broader set of languages, including non-Eurasian structures such as Arabic (introflexive) and Vietnamese (isolating), to verify if *FD* remains statistically stable across fundamentally different morphological paradigms.

Granular Per-Intent Metrics: Developing class-specific geometric metrics to complement our global *FD* framework, enabling localized error analysis to determine how low-frequency intents scale geometrically under multi-class pressure.

Architectural Generalization: Extending this geometric framework to Transformer-based architectures will reveal if self-attention mechanisms employ a similar distillation process or favor a different geometric strategy for semantic resolution.

Through these efforts, we seek to transition geometric interpretability from a descriptive science to a predictive tool for building more transparent, efficient, and trustworthy multilingual dialogue systems.

References

Abro, W. A., G. Qi, M. Aamir, and Z. Ali. 2022. Joint Intent Detection and Slot Fill-

ing Using Weighted Finite State Transducer and BERT. *Applied Intelligence*, 52(15):17356–17370.

Aitken, K., V. V. Ramasesh, A. Garg, Y. Cao, D. Sussillo, and N. Maheswaranathan. 2021. The Geometry of Integration in Text Classification RNNs. In *Proceedings of the 9th International Conference on Learning Representation (ICLR 2021)*.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2023. Spanish Pre-trained BERT Model and Evaluation Data. In *Proceedings of the Practical ML for Developing Countries (PML4DC) Workshop (ICLR 2020)*.

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. ACL.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186. ACL.

Dror, R., L. Peled-Cohen, S. Shlomov, and R. Reichart. 2020. *Statistical Significance Testing for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer International Publishing.

FitzGerald, J., C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, and P. Natarajan. 2023. MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 4277–4302. ACL.

Gorman, K. and S. Bedrick. 2019. We Need to Talk about Standard Splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791. ACL.

Hochreiter, S. and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hubert, L. and P. Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Karpathy, A., J. Johnson, and L. Fei-Fei. 2016. Visualizing and Understanding Recurrent Networks. In *Proceedings of the 4th International Conference on Learning Representation (ICLR 2016)*.

Kingma, D. P. and J. Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR 2015)*.

Lin, M., H. C. Lucas, and G. Shmueli. 2013. Research Commentary: Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research*, 24(4):906–917.

Maheswaranathan, N., A. Williams, M. Golub, S. Ganguli, and D. Sussillo. 2019. Reverse engineering Recurrent Networks for Sentiment Classification Reveals Line Attractor Dynamics. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, volume 32, pages 15696–15705. Curran Associates.

Martelli, M. 1999. *Introduction to Discrete Dynamical Systems and Chaos*. Wiley Series in Discrete Mathematics and Optimization. Wiley-Interscience, 1st edition.

Morcos, A. S., D. G. T. Barrett, N. C. Rabinowitz, and M. Botvinick. 2018. On the Importance of Single Directions for Generalization. In *Proceedings of the 6th In-*

ternational Conference on Learning Representation (ICLR 2018).

Ravuri, S. and A. Stolcke. 2015. Recurrent Neural Network and LSTM Models for Lexical Utterance Classification. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, pages 135–139. ISCA.

Reimers, N. and I. Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990. ACL.

Sanchez-Karhunen, E., J. F. Quesada-Moreno, and M. A. Gutiérrez-Naranjo. 2024. Interpretation of the Intent Detection Problem as Dynamics in a Low-Dimensional Space. In *ECAI 2024: 27th European Conference on Artificial Intelligence*, volume 392, pages 3693–3700.

Sanchez-Karhunen, E., J. F. Quesada-Moreno, and M. A. Gutiérrez-Naranjo. 2026a. Interpretability of the Intent Detection Problem: A New Approach. *European Journal on Artificial Intelligence*, 39(2):336–357.

Sanchez-Karhunen, E., J. F. Quesada-Moreno, and M. A. Gutiérrez-Naranjo. 2026b. Interpretable Intent Detection in High-Cardinality Scenarios via Dynamical Systems Analysis. *Procesamiento del Lenguaje Natural*, 76:25–37.

Schweter, S. 2020. BERTurk - BERT models for Turkish. Zenodo.

Sussillo, D. and O. Barak. 2013. Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation*, 25(3):626–649.

Tur, G. and R. De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2023. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, August.

Virtanen, A., J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo. 2019. Multilingual is not enough: BERT for Finnish.

Wang, C. and S. Mahadevan. 2008. Manifold alignment using Procrustes analysis. In *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, pages 1120–1127. ACM Press.