IOCMA 2026 Conference
The 2nd International Online Conference on Mathematics and Applications
10–12 June 2026 | Online
MDPI

Invariance of Semantic Projections under Changes of Word Universes in NLP

Ana Coronado Ferrer
IUMPA, Universitat Politècnica de València, Spain | Florida Universitaria, Catarroja, Spain

Funding: Generalitat Valenciana PROMETEO 2024 CIPROM/2023/32

INTRODUCTION & AIM

Semantic projections are indices in [0,1] that quantify how much a term shares contextual meaning with words of a given universe. Defined as co-occurrence ratios over document corpora, they yield interpretable, low-dimensional embeddings grounded in set-theoretic measure spaces.

Two stability questions:
- Projection coherence: are projections from different sources (DOAJ, Scholar, Google, Arxiv) mutually consistent for a fixed universe?
- Universe transfer: can projections be reliably estimated when the semantic universe changes across vocabularies or corpora?

Formal setting (Definition 1):
For A, B ∈ Σ(S): $P_a(B) := \mu(A \cap B) / \mu(A) \in [0,1]$
where (S, Σ, μ) is a finite measure space and each term maps to a measurable subset.

METHOD

Coherence analysis — multiple sources, fixed universe:
- Normalise projection vectors: $N^i(t) = P^i(t) / \|P^i(t)\|$
- Pairwise Pearson correlation matrix
- Chi-squared distance heatmap
- PCA dimensionality reduction
- K-means clustering + Elbow method

Universe transfer — two similar universes U_1, U_2:
- Averaged Weights: convex combination weighted by inverse distance from u^2_j to each u^1_i
- Closer Weights: hierarchical weighting; proved to sum to 1 (Theorem 2); acts as Lipschitz extension
- McShane–Whitney: classical minimal/maximal Lipschitz extensions $F^M(x) = \sup\{f(s) - L \cdot d(x,s)\}$ and $F^w(x) = \inf\{f(s) + L \cdot d(x,s)\}$; estimate = $(F^M + F^w)/2$

Theoretical guarantee (Theorem 1):
If each $P\_u^1\_i$ is Lipschitz in t with constant L, then all three estimators $\hat{P}\_u^2\_j$ are also Lipschitz in t with the same constant L.

Word embedding for distances:
- GloVe 6B-50d (Pennington et al., 2014) for Euclidean distances between universe terms

RESULTS & DISCUSSION

Case study 1 — Projection coherence
Term: t = "plowing" | Universe: U = {crop, farmland, farming, harvest, irrigation, orchard, soil, tractor}

- DOAJ ↔ Scholar: correlation = 0.92 (strong consistency)
- Google ↔ DOAJ: correlation = −0.84 (strong divergence)
- Arxiv ↔ DOAJ: correlation = 0.64 (moderate)

- PCA: PC1 explains 72.4% of total variance
- K-means (k = 2) cleanly separates Google from {DOAJ, Scholar, Arxiv}
- Chi-squared distances confirm: academic repositories form a coherent cluster

Interpretation:
Academic repositories (DOAJ, Scholar, Arxiv) index domain-specific literature and produce semantically coherent projections. Google indexes heterogeneous content, yielding inconsistent semantic patterns for specialised terms. Source selection significantly impacts semantic conclusions.

Case study 2 — Universe transfer
Term: t = "rice"
U_1 = {water source, reservoir, irrigation, crop, orchard}
U_2 = {wetland, farming, water resources, watering}
Distances: GloVe 6B-50d Euclidean

Real values vs. estimates ($P\_u^2\_i$("rice")):
- wetland: real 0.036 | Avg 0.035 | Closer 0.076 | Mc-W 0.055
- farming: real 0.064 | Avg 0.035 | Closer 0.046 | Mc-W 0.039
- water resources: real 0.016 | Avg 0.035 | Closer 0.039 | Mc-W 0.004
- watering: real 0.037 | Avg 0.035 | Closer 0.018 | Mc-W 0.048

RMSE comparison:
- Averaged Weights: 0.0172
- McShane–Whitney: 0.0175
- Closer Weights: 0.0268

McShane–Whitney best reproduces real trends. Averaged Weights provides uniform but reasonable estimates. Closer Weights underperforms when inter-element distances are similar (disproportionate weight on nearest neighbor).

CONCLUSION

- Semantic stability is tractable: both projection coherence and universe transfer admit principled quantitative analysis using statistical and Lipschitz-based tools
- Source selection matters: domain-specific repositories (DOAJ, Scholar) are preferable for technical NLP analysis
- Lipschitz continuity guarantees bounded transfer errors regardless of universe size

FUTURE WORK / REFERENCES

- Extend to multilingual NLP (machine translation, cross-lingual retrieval)
- Integrate with large language models and knowledge graphs
- Apply to healthcare, legal, and environmental domains

Related paper:
Arnau et al., Axioms 2025, 14, 389. doi: 10.3390/axioms14050389
Funding: Generalitat Valenciana, PROMETEO 2024 CIPROM/2023/32