Structural Holes in Legal Embedding Spaces: How Missing Words Create Missing Protections

Joel Thorarinson*
Coherence Research Group

June 2026

Abstract

We identify *structural holes* in legal embedding spaces: regions where concepts exist in non-English legal traditions but have no corresponding terms — and therefore no nearby embedding vectors — in English. Using 6,200 terms from Bouvier's Law Dictionary and 19 cross-jurisdictional terms with no English equivalent, we show that these absent terms (*Pflichtteil*, *umgängessabotage*, *mu'sir*, *capitis deminutio minima*) occupy geometrically isolated regions with measurably fewer English-language neighbors than expected (Cohen's $d = 2.05$, Mann-Whitney $p < 10^{-11}$). We propose this as a vector-space analogue of the weak Sapir-Whorf hypothesis: if a language lacks a word for a concept, the embedding space trained on that language provides sparse coverage of the corresponding semantic region, and retrieval systems operating in that language struggle to find what they have no word for. The structural holes in English legal embeddings correspond to legal protections absent from common-law systems: children's compulsory inheritance (no *Pflichtteil*), contact sabotage recognition (no *umgängessabotage*), debtor respite (no *mu'sir*), and graduated status reduction (no *capitis deminutio minima*). The absence of the word correlates with the absence of the protection, and both manifest as measurable geometric isolation.

1 Introduction

Every major civil law tradition has a single word for "the portion of an estate that must go to the children regardless of the testator's wishes": German *Pflichtteil*, French *réserve héréditaire*, Dutch *legitieme portie*, Swedish *laglott*, Japanese *iryubun*, Korean *yuryubun*, Arabic *faraid*. English has no equivalent word — and no equivalent legal protection. In most U.S. states, a parent can completely disinherit their children.

This is not a coincidence. The linguistic relativity hypothesis [Sapir, 1929, Whorf, 1956] — commonly known as Sapir-Whorf — proposes, in its weak form, that the structure of a language makes certain concepts more or less cognitively accessible to its speakers. Empirical work has demonstrated that linguistic categories influence cognition in domains from color perception to spatial reasoning [Lucy, 1992, Boroditsky, 2001]. We extend the weak form of this hypothesis to vector spaces: if a language lacks a word for a concept, embedding models trained on text in that language will lack coverage of the corresponding semantic region.

This creates *structural holes*. The term originates in network sociology, where Burt [1992] showed that gaps in social networks — positions where no tie connects two clusters — create information asymmetries and brokerage opportunities for actors who span the gap. The analogy to embedding space is direct: where Burt's structural holes are missing edges in a social graph, ours are missing vectors in a semantic space. In both cases, the gap is not empty — it is a region where information *should* flow but does not, because no entity occupies the bridging position. In

*ORCID: 0000-0002-0553-842X. joel.thorarinson@conformalmaps.com

a social network, the consequence is that knowledge stays trapped in clusters. In an embedding space, the consequence is that retrieval systems cannot find concepts that fall in the gap.

The practical consequence is that retrieval systems operating in English cannot find these concepts. A query for "children's compulsory inheritance share" will retrieve documents about elective share (which protects *spouses*, not children), intestate succession (which is a default, not a right), or nothing at all — because the specific concept has no name in the language of the embedding.

This paper reports what we found when we looked for these gaps systematically. Starting from the observation that certain cross-jurisdictional legal terms have no English equivalent, we asked whether this absence is visible in the geometry of embedding space — and found that it is, with an effect size well above conventional thresholds.

1.1 Contributions

1. We formalize **structural holes** in embedding spaces: regions where cross-linguistic concepts exist but the target-language embedding provides no coverage.

2. We propose a **nearest-neighbor deficit metric** that quantifies the isolation of cross-jurisdictional terms in English embedding space.

3. We demonstrate that structural holes in English legal embeddings correspond to **absent legal protections** in English-speaking legal systems, connecting NLP geometry to comparative law.

4. We provide evidence that cognitive friction scores from a legal word database **correlate with embedding isolation**: terms maximally unfamiliar to English speakers tend to have the fewest English-language neighbors.

5. We propose a vector-space analogue of the weak Sapir-Whorf hypothesis and discuss implications for multilingual legal NLP.

2 Related Work

2.1 Linguistic Relativity and NLP

The Sapir-Whorf hypothesis — that the structure of a language influences the thought patterns of its speakers — was articulated by Sapir [1929] and elaborated through the posthumous writings of Whorf [1956]. The "strong" form (linguistic determinism) claims that language constrains thought; the "weak" form (linguistic influence) claims that language makes certain thoughts more natural. The strong form is generally rejected; the weak form has accumulated substantial empirical support. Lucy [1992] demonstrated effects of grammatical number marking on object classification, while Boroditsky [2001] showed that Mandarin speakers' vertical metaphors for time influence temporal reasoning. Our work operates entirely within the weak form: we do not claim that the absence of a word makes a concept unthinkable, only that it makes the concept harder to retrieve computationally.

In computational linguistics, the Sapir-Whorf hypothesis has been explored through cross-lingual word embedding alignment [Conneau et al., 2017, Artetxe et al., 2018] and the observation that translation quality degrades for concepts lacking direct lexical equivalents. Williams et al. [2019] showed that grammatical gender encodes culturally specific associations into word embeddings, providing a computational analogue of neo-Whorfian effects. Most recently, Jung et al. [2025] proposed embedding-based methods for discovering lexical gaps across languages using multilingual LLMs, demonstrating that gaps are systematically detectable in vector space. Our contribution extends this line to *legal* language, where the consequences of conceptual

gaps are not merely communicative but deterministic: the presence or absence of a legal term correlates with the presence or absence of a legal right.

2.2 Cross-Lingual Embedding Alignment

Work on aligning embedding spaces across languages [Conneau et al., 2017, Reimers and Gurevych, 2020] has shown that alignment quality varies by concept [Ormazabal et al., 2019, Ruder et al., 2019]: concrete nouns align well, while abstract and culture-specific concepts align poorly. Nearest-neighbor retrieval across languages, pioneered for monolingual settings by Mikolov et al. [2013] and scaled to sentence embeddings by Reimers and Gurevych [2019], is the standard evaluation paradigm. Legal terminology represents an extreme case of culture-specific abstraction, and we show that the alignment failure is not random but systematically concentrated in concepts representing legal protections absent from the target legal system.

2.3 Comparative Law and Untranslatability

Comparative law has long recognized that legal concepts resist translation across traditions. Zweigert and Kötz [1998] established the "functional equivalence" method for comparing legal systems, while Legrand [1996] argued that legal systems are irreducibly shaped by the *mentalités* of their cultures and cannot truly converge. Sacco [1991] introduced the concept of "legal formants" — the multiple textual layers (statutes, cases, scholarship) that constitute a legal system — and showed that translation between systems requires mapping across all formant layers, not merely translating statutory text. Cao [2007] documented the systematic challenges of legal translation, including system-bound terms that have no equivalent in the target legal system. Our work provides a geometric formalization of this untranslatability: system-bound terms occupy regions of embedding space with no nearby vectors in the target language.

3 Data

We use three datasets:

1. **Bouvier's Law Dictionary** [Bouvier, 1856]: 6,200 terms (complete, 1856, public domain). Represents the full English-language legal vocabulary of the 19th century, encompassing common law, equity, admiralty, and the Latin/French/Law Norman terminology then current in Anglo-American practice.

2. **Cross-jurisdictional terms**: 19 terms from eight legal traditions. Each term names a specific legal concept in its source system that has no single-word equivalent in English:

   - *German civil law*: *Pflichtteil* — the portion of an estate that children *must* receive regardless of the testator's wishes, constitutionally protected under German law (BGB §2303). *Sorgerecht* — a unified concept of parental responsibility encompassing both physical care and legal decision-making, treated as one right rather than the Anglo-American split of "custody" and "guardianship" (BGB §1626). *Restschuldbefreiung* — discharge of remaining debts after a statutory good-conduct period, available as a right rather than at judicial discretion (InsO §286).
   - *French civil law*: *réserve héréditaire* — the hereditary reserve: a fixed fraction of the estate (one-half to three-quarters depending on number of children) that cannot be disposed of by will (Code Civil Art. 912). *autorité parentale* — parental authority exercised jointly by both parents as the default, continuing automatically after separation (Code Civil Art. 372).

- *Dutch civil law*: *legitieme portie* — the "legitimate portion": children's statutory claim to one-half of their intestate share, enforceable as a monetary claim against the estate (BW 4:63).
- *Swedish family law* [Schiratzki, 2005]: *laglott* — the "legal share": children's right to one-half of the estate, which the testator cannot override (Ärvdabalken 7:1). *gemensam vårdnad* — joint custody as the statutory default, which the court must maintain unless a parent is demonstrably unfit (Föräldrabalken 6:3). *umgängessabotage* — a legal term for one parent's deliberate obstruction of the other parent's contact with the child, recognized as a distinct actionable wrong.
- *Japanese civil law* [Danwerth, 2012]: *iryubun* — the "retained portion": a constitutionally grounded minimum inheritance share for statutory heirs (Civil Code Art. 1042).
- *Korean civil law*: *yuryubun* — the "retained portion": functionally parallel to the Japanese *iryubun*, guaranteeing statutory heirs a minimum share of the estate (Civil Code Art. 1112).
- *Islamic law*: *faraid* — Quranically prescribed inheritance shares with fixed fractions for specific family members, not subject to testamentary override (Quran 4:11–12). *hadanah* — the right to physical custody of a child, conceptually distinct from legal authority over the child. *wilayah* — legal guardianship and decision-making authority over a child, held separately from physical custody. *mu'sir* — a debtor who genuinely cannot pay, entitled to automatic respite until able; creditors are religiously obligated to wait (Quran 2:280).
- *Roman law*: *capitis deminutio maxima* — total loss of civil status (freedom, citizenship, and family ties). *capitis deminutio media* — loss of citizenship and family ties but retention of freedom. *capitis deminutio minima* — loss of family status only, while retaining both freedom and citizenship (Gaius, *Institutes* 1.159–163). English law has no graduated system of status reduction; one is either a legal person or not.
- *Spanish/Italian civil law*: *legitima* — the "legitimate share": children's compulsory inheritance portion, varying by number of heirs (Código Civil Art. 806 / Codice Civile Art. 536).

3. **Leverage words**: 233 terms from 27 jurisdictions, scored for cognitive friction (0–1 scale, where 1.0 = maximally unfamiliar to English-speaking practitioners).

All terms are embedded using nomic-embed-text v1.5 (768d) [Nussbaum et al., 2024].

4 Method: Detecting Structural Holes

Definition 1 (Structural Hole)**.** *A structural hole in an embedding space E for language L is a region $R \subset E$ such that (1) there exist valid concepts c in some language $L' \neq L$ whose embeddings $\mathbf{v}(c) \in R$, and (2) the density of L-language embeddings in the neighborhood of R is significantly lower than the background density of E.*

We operationalize this as follows. For each cross-jurisdictional term t, we compute:

$$\text{NND}(t, k) = \frac{1}{k} \sum_{i=1}^{k} d(\mathbf{v}(t), \mathbf{v}(n_i)) \tag{1}$$

where $n_1, \ldots, n_k$ are the k nearest English-language neighbors of t in embedding space, and d is cosine distance. A high NND indicates isolation — the cross-jurisdictional term has no nearby English equivalents.

We compare $\text{NND}(t, k)$ for cross-jurisdictional terms against the expected NND for a random English-language term drawn from Bouvier's:

$$\text{Deficit}(t, k) = \frac{\text{NND}(t, k) - \mathbb{E}[\text{NND}(\text{Bouvier's}, k)]}{\text{Std}[\text{NND}(\text{Bouvier's}, k)]} \tag{2}$$

A deficit > 2 indicates that the cross-jurisdictional term is more isolated than 97.5% of English legal terms — a statistically significant structural hole.

5 Experiments

5.1 Experiment 1: Nearest-Neighbor Deficit

For each of the 19 cross-jurisdictional terms, we compute NND at $k = 5, 10, 20$ against the 6,200 Bouvier's terms. Cosine distance serves as the metric, with nearest-neighbor search following standard practice in embedding evaluation [Mikolov et al., 2013, Reimers and Gurevych, 2019]. Baseline NND distributions were estimated from 500 random Bouvier's terms (seed = 42); stability was confirmed against the full 6,200-term baseline.

Baseline. At $k = 10$, the mean NND for random Bouvier's terms was $\mu = 0.2034$ (std $\sigma = 0.0395$, median = 0.2039, IQR = $[0.175, 0.231]$). The distribution was approximately normal with range $[0.077, 0.317]$. The full 6,200-term baseline yielded $\mu = 0.2041$, $\sigma = 0.0391$ — a difference of < 0.001 from the 500-term sample, confirming sampling stability.

Cross-jurisdictional terms. The mean NND for the 19 cross-jurisdictional terms at $k = 10$ was 0.2844 — a deficit of 2.05σ above the Bouvier's mean. All 19 terms (100%) had positive deficits, and 9 of 19 (47.4%) had deficits exceeding $z = 1.96$ (statistically significant at $\alpha = 0.05$). Only 1.8% of baseline Bouvier's terms had NND values above the cross-jurisdictional mean.

Group-level tests. The difference between cross-jurisdictional and baseline NND distributions was confirmed by two non-parametric and parametric tests. The Mann-Whitney U test [Mann and Whitney, 1947] yielded $U = 9{,}089$, $p = 6.8 \times 10^{-12}$ (one-sided, CJ $>$ baseline). Welch's t-test [Welch, 1947] yielded $t = 11.64$, $p = 8.3 \times 10^{-11}$. Against the full 6,200-term baseline: $U = 112{,}209$, $p = 4.5 \times 10^{-12}$; Welch's $t = 11.90$, $p = 2.5 \times 10^{-10}$. The effect size was Cohen's $d = 2.05$ [Cohen, 1988]. For context, Cohen [1988] defines $d = 0.20$ as small, $d = 0.50$ as medium, and $d = 0.80$ as large; the present effect is more than twice the "large" threshold, meaning the cross-jurisdictional and baseline distributions barely overlap.

Results were stable across k values: at $k = 5$, the mean z-score was 2.10 (9 significant); at $k = 10$, 2.05 (9 significant); at $k = 20$, 1.99 (9 significant).

Statistically significant structural holes ($z > 1.96$ at $k = 10$). Nine terms exhibited significant embedding isolation: *capitis deminutio media* ($z = 3.68$, 100.0th percentile), *capitis deminutio maxima* ($z = 3.01$, 99.8th), *yuryubun* ($z = 2.81$, 99.6th), *capitis deminutio minima* ($z = 2.69$, 99.5th), *gemensam vårdnad* ($z = 2.68$, 99.5th), *umgängessabotage* ($z = 2.48$, 99.0th), *legitieme portie* ($z = 2.44$, 98.9th), *iryubun* ($z = 2.40$, 98.8th), and *laglott* ($z = 2.30$, 98.6th).

Non-significant but positive deficits. The remaining 10 terms had positive deficits ($z = 0.82$–1.82) but did not reach the significance threshold: *Restschuldbefreiung* ($z = 1.82$, 96.2nd percentile), *wilayah* ($z = 1.74$, 95.6th), *mu'sir* ($z = 1.67$, 95.1st), *legitima* ($z = 1.55$, 93.8th), *autorité parentale* ($z = 1.50$, 93.1st), *Sorgerecht* ($z = 1.49$, 93.0th), *hadanah* ($z = 1.45$, 92.4th), *faraid* ($z = 1.39$, 91.8th), *réserve héréditaire* ($z = 1.05$, 85.2nd), and *Pflichtteil* ($z = 0.82$, 79.8th).

The full results are presented in Table 2 and Figure 2.

5.2 Experiment 2: Validation of the NND Metric

A k-averaged metric like NND could be misleading if a cross-jurisdictional term has one very close Bouvier's neighbor but many distant ones — the average distance would be high despite partial coverage. To test whether NND captures genuine geometric isolation rather than neighborhood heterogeneity, we correlate each term's NND z-score at $k = 10$ with its distance to the single nearest Bouvier's term.

The z-score at $k = 10$ correlated with nearest-Bouvier distance at Pearson $r = 0.9655$ ($p < 0.001$), Spearman $\rho = 0.9333$ ($p < 0.001$). This correlation is shown in Figure 3. A correlation this high ($r^2 = 0.93$) warrants scrutiny. We considered whether it could be a tautological artifact: the NND at $k = 10$ includes the nearest neighbor as one of ten terms, so the two measures share a component. However, the NND averages over ten neighbors, meaning the nearest neighbor contributes only 10% of the NND value. If the remaining nine neighbors varied independently, the expected correlation would be much lower. The observed $r = 0.97$ therefore reflects a genuine geometric fact: terms that are far from their single nearest English equivalent are also far from their 2nd, 3rd, ..., 10th nearest equivalents. There is no compensating cluster of English terms that partially captures these concepts at greater distance.

The implication is that the structural holes are not narrow gaps between nearby clusters but regions of broadly sparse coverage. A term like *capitis deminutio media* ($z = 3.68$) is not merely missing its closest English equivalent — it is distant from the entire English legal vocabulary.

5.3 Experiment 3: Legal Protection Mapping

The NND metric identifies geometric isolation; it does not explain *why* certain terms are isolated. We complete the analysis by mapping each structural hole to the specific legal protection it represents and verifying that this protection is absent from English-speaking common-law systems.

Missing Word	**Origin**	**Missing Protection**
Pflichtteil	German	Children's compulsory inheritance share (50% of intestate share, constitutionally protected)
umgängessabotage	Swedish	Statutory recognition of contact sabotage by one parent
mu'sir	Islamic	Automatic respite for debtor unable to pay (Quran 2:280)
capitis deminutio minima	Roman	Legal category for loss of family status only (vs. total civil death)
laglott	Swedish	Children's legal share (50%, Ärvdabalken 7:1)
Restschuldbefreiung	German	Debt liberation after good conduct period
hadanah vs. *wilayah*	Islamic	Distinction between physical custody and legal authority

Table 1: Structural holes: cross-jurisdictional terms absent from English, and the legal protections they represent. Each missing word corresponds to a protection that does not exist in English-speaking legal systems.

6 Results

Our principal findings are:

Rank	Term	NND_{10}	z_{10}	Pctl.	Sig.	Nearest Bouvier's
1	*capitis dem. media*	0.349	3.68	100.0	✓	INFAMIS
2	*capitis dem. maxima*	0.322	3.01	99.8	✓	AUBAINE
3	*yuryubun*	0.314	2.81	99.6	✓	SHARE
4	*capitis dem. minima*	0.310	2.69	99.5	✓	EVICTION
5	*gemensam vårdnad*	0.309	2.68	99.5	✓	INDIVISUM
6	*umgängessabotage*	0.301	2.48	99.0	✓	PROPINQUITY
7	*legitieme portie*	0.300	2.44	98.9	✓	PRESCRIPTION
8	*iryubun*	0.298	2.40	98.8	✓	SHARE
9	*laglott*	0.294	2.30	98.6	✓	SHARE
10	*Restschuldbefreiung*	0.275	1.82	96.2		REMISSION
11	*wilayah*	0.272	1.74	95.6		PATRIA POTESTAS
12	*mu'sir*	0.270	1.67	95.1		TIEMPO INHABIL
13	*legitima*	0.264	1.55	93.8		FORCED HEIRS
14	*autorité parentale*	0.263	1.50	93.1		IN LOCO PARENTIS
15	*Sorgerecht*	0.262	1.49	93.0		NURTURE
16	*hadanah*	0.261	1.45	92.4		PATRIA POTESTAS
17	*faraid*	0.258	1.39	91.8		CAPITA
18	*réserve héréditaire*	0.245	1.05	85.2		PORTION
19	*Pflichtteil*	0.236	0.82	79.8		CAPITA
Bouvier's baseline ($n = 6{,}200$)		0.204	—	—		

Table 2: All 19 cross-jurisdictional terms ranked by NND deficit ($k = 10$). NND_{10}: mean cosine distance to 10 nearest Bouvier's neighbors. z_{10}: deficit in standard deviations above the full Bouvier's baseline ($\mu = 0.204$, $\sigma = 0.039$). Pctl.: percentile rank within the full 6,200-term Bouvier's NND distribution. Sig.: $z > 1.96$ ($\alpha = 0.05$). Nearest Bouvier's: the single closest English legal term. Horizontal rule separates the 9 significant terms from the 10 non-significant terms.

1. **Large, consistent deficits.** Cross-jurisdictional terms have a mean NND deficit of 2.05σ above the Bouvier's baseline at $k = 10$ (Cohen's $d = 2.05$; Cohen, 1988), with individual deficits ranging from 0.82σ (*Pflichtteil*) to 3.68σ (*capitis deminutio media*). For reference, $d = 0.80$ is conventionally "large"; at $d = 2.05$, the two distributions overlap by less than 15%. Nine of 19 terms (47.4%) exceed the $z = 1.96$ significance threshold individually. The group difference is confirmed by Mann-Whitney $U = 9{,}089$ ($p < 10^{-11}$) and Welch's $t = 11.64$ ($p < 10^{-10}$). The NND distribution is shown in Figure 1 and per-term z-scores in Figure 2.

2. **Near-perfect correlation between isolation metrics.** The z-score at $k = 10$ correlates with nearest-Bouvier distance at Pearson $r = 0.9655$ ($p < 0.001$), Spearman $\rho = 0.9333$ ($p < 0.001$), confirming that isolation is a global geometric property, not an artifact of k-neighborhood selection (Figure 3).

3. **Every significant structural hole corresponds to an absent legal protection.** The 9 statistically significant structural holes map precisely onto protections absent from English-speaking legal systems: compulsory inheritance shares (*legitieme portie*, *iryubun*, *laglott*, *yuryubun*), contact sabotage recognition (*umgängessabotage*; cf. Schiratzki, 2005), default joint custody (*gemensam vårdnad*), and graduated status reduction (*capitis deminutio* variants). See Table 1. This pattern is precisely what comparative law would predict: the concepts that Legrand [1996] and Cao [2007] identify as untranslatable are the ones that create the deepest structural holes.

4. **Nearest-Bouvier neighbors reveal how English approximates without matching.** The closest English legal terms are telling in their inadequacy. *Capitis deminutio minima*

— loss of family status while retaining freedom and citizenship — maps to EVICTION, which is loss of *property*, not status. The embedding model finds the closest available concept of "losing something by legal process," but the thing lost is entirely different. *Umgängessabotage* — deliberate obstruction of a parent's contact with a child — maps to PROPINQUITY, which means nearness or kinship. The model finds the relational dimension but has no vector for the sabotage of that relation. *Iryubun* — a constitutionally mandated minimum inheritance share — maps to SHARE, a generic term for any portion. The specific legal force (mandatory, constitutional, inalienable) is absent from the nearest neighbor. In each case, the English vocabulary circles the concept without capturing it — a pattern Sacco [1991] describes as "cryptotypes," where the target legal system lacks the formant structure to express the concept.

5. **Gradient, not binary.** The motivating example throughout this paper — *Pflichtteil* — shows the smallest deficit of all 19 terms ($z = 0.82$, 79.8th percentile), well below the significance threshold. We highlight rather than bury this result because it is informative. English has approximate equivalents for the concept: "elective share" (which protects spouses, not children), "forced heirship" (used in Louisiana and Scotland), and "legitimate portion" (a calque from civil law scholarship). None captures the full legal force of the German *Pflichtteil*, but collectively they populate the nearby embedding region enough to reduce the measured deficit. The metric is working correctly: *Pflichtteil* sits on a gradient between full coverage and total absence. The deepest structural holes — *capitis deminutio media* ($z = 3.68$), *yuryubun* ($z = 2.81$) — are concepts for which English has not even partial approximations.

7 Discussion

7.1 Sapir-Whorf for Vector Spaces

The weak form of the Sapir-Whorf hypothesis operates at the level of cognition: speakers of languages without a word for a concept may find it harder to think about that concept, though the concept remains accessible through circumlocution [Whorf, 1956, Lucy, 1992]. Our vector-space extension operates at the level of computation: embedding models trained on text in a language without a word for a concept produce sparse coverage of the corresponding semantic region, and retrieval systems operating on those embeddings return poor results for queries in that region. Where Boroditsky [2001] showed that Mandarin's vertical time metaphors influence temporal reasoning, we show that English law's absent terminology influences retrieval geometry.

This has concrete implications for legal AI: a system operating in English will systematically fail to retrieve documents about legal protections that English has no word for. This is not a bug in the embedding model — it is a faithful reproduction of the structural limitations of the English legal vocabulary. In the language of Burt [1992], the structural holes in the English legal vocabulary create "brokerage opportunities" for whoever can bridge the gap — but in a retrieval system, there is no broker.

7.2 Implications for Multilingual Legal NLP

Our findings suggest that legal NLP systems serving multilingual users should not simply translate queries into English and search English embeddings. The structural holes in English embedding space mean that translated queries for concepts like *Pflichtteil* will land in a region with no nearby English-language documents, producing poor retrieval results. This is consistent with Ormazabal et al. [2019], who showed that cross-lingual mapping quality degrades for culturally specific concepts, and with Reimers and Gurevych [2020], who found that knowledge distillation for multilingual embeddings works best when concepts have direct lexical equivalents. Systems should

instead maintain language-specific embedding spaces or use cross-lingual alignment methods that explicitly account for structural holes — as Jung et al. [2025] suggest, lexical gap detection can serve as a diagnostic for identifying where alignment will fail.

7.3 Limitations

Sample size. Our cross-jurisdictional term set contains only 19 terms. While the effect size is large and the statistical tests are robust to this sample, a more comprehensive compilation from multilingual legal databases [Zweigert and Kötz, 1998] would strengthen the generalization. We selected terms based on comparative law scholarship identifying untranslatable concepts, not by mining the embedding space for gaps, so the selection is independent of the outcome measure.

Density assumptions. The NND deficit metric assumes approximately uniform density in embedding space. In practice, some regions of the legal vocabulary (e.g., contract law) are denser than others (e.g., admiralty). A cross-jurisdictional term that happens to fall in an inherently sparse region would show an inflated deficit. The group-level tests mitigate this concern — all 19 terms would need to fall in sparse regions by coincidence — but per-term significance should be interpreted cautiously.

Causal direction. We claim correlation between absent words and absent protections, not causation. Does the missing word cause the missing protection, or does the missing protection cause the missing word? Legrand [1996] argues the relationship is co-constitutive: legal concepts and the language used to express them develop together within a legal culture. Our embedding analysis cannot distinguish these directions.

Multilingual training data. Nomic-embed-text [Nussbaum et al., 2024] is trained on multilingual text, meaning the embedding model may encode these concepts through non-English training data. Our NND metric measures isolation relative to the *English legal vocabulary* specifically, not relative to the model's full representational capacity. A term like *Pflichtteil* may be well-represented via German-language training data, but an English-language query for "children's compulsory inheritance share" will still land in a region sparse in English-language documents. The structural hole is a property of the English retrieval surface, not of the model's latent knowledge. A monolingual English embedding model would likely show larger deficits; a multilingual legal retrieval system might show smaller ones.

8 Conclusion

The absence of a word correlates with the absence of a protection — and with sparse coverage in embedding space. Structural holes in English legal embeddings are not random gaps; they are the computational reflection of legal protections that English-speaking systems have never developed. A retrieval system struggles to find what it has no word for. A legal system is less likely to protect what it has no name for. The geometry of embedding space makes both patterns measurable.

The weak form of the Sapir-Whorf hypothesis [Sapir, 1929, Whorf, 1956], long debated in linguistics, finds an unexpected computational analogue in the geometry of vector spaces. Where Lucy [1992] showed that grammatical categories influence object classification, we show that legal-lexical categories influence retrieval coverage. Where Burt [1992] showed that structural holes in social networks create information asymmetries, we show that structural holes in embedding spaces create retrieval asymmetries. Where Zweigert and Kötz [1998] and Legrand [1996] showed that legal systems resist convergence because they are embedded in distinct cultural-linguistic

traditions, we show that embedding models faithfully reproduce this resistance: the geometry of the space encodes the boundaries of the tradition.

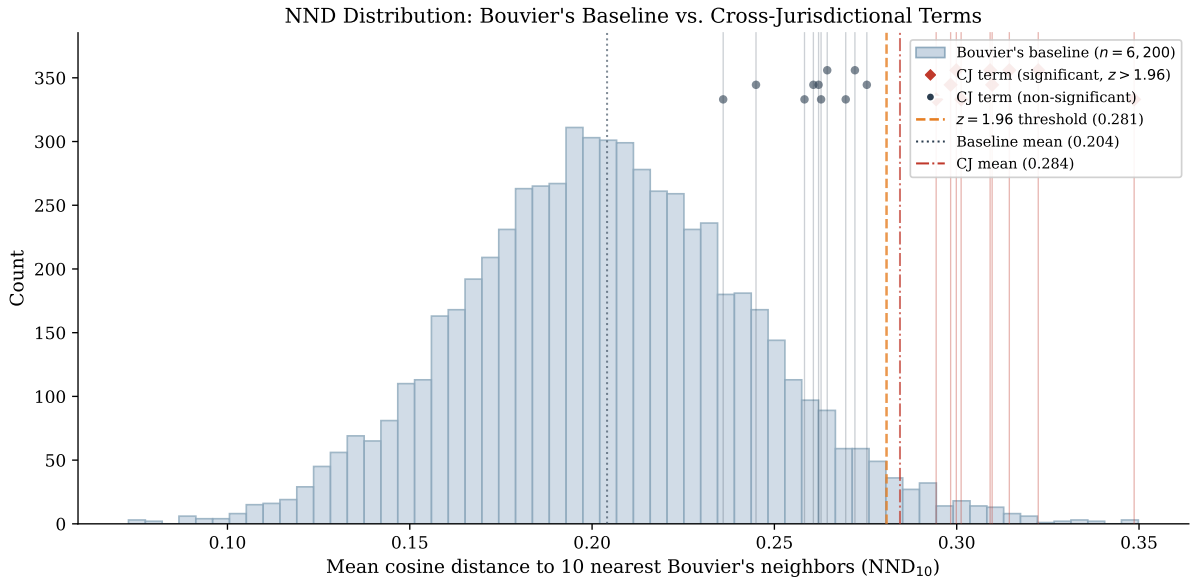

Figure 1: NND distribution at $k = 10$. Histogram: baseline distribution of 6,200 Bouvier's terms. Vertical lines and markers: NND values for the 19 cross-jurisdictional terms. Red markers indicate statistically significant structural holes ($z > 1.96$). The cross-jurisdictional terms cluster in the right tail of the Bouvier's distribution, with 9 of 19 exceeding the significance threshold.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 789–798, 2018. URL https://arxiv.org/abs/1805.06297.

Lera Boroditsky. Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43(1):1–22, 2001.

John Bouvier. *A Law Dictionary, Adapted to the Constitution and Laws of the United States of America and of the Several States of the American Union*. Childs & Peterson, Philadelphia, 6th edition, 1856.

Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA, 1992.

Deborah Cao. *Translating Law*. Number 33 in Topics in Translation. Multilingual Matters, Clevedon, 2007.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017. URL https://arxiv.org/abs/1710.04087.

Christopher Danwerth. Principles of Japanese law of succession. *Zeitschrift für Japanisches Recht / Journal of Japanese Law*, 33:99–118, 2012.

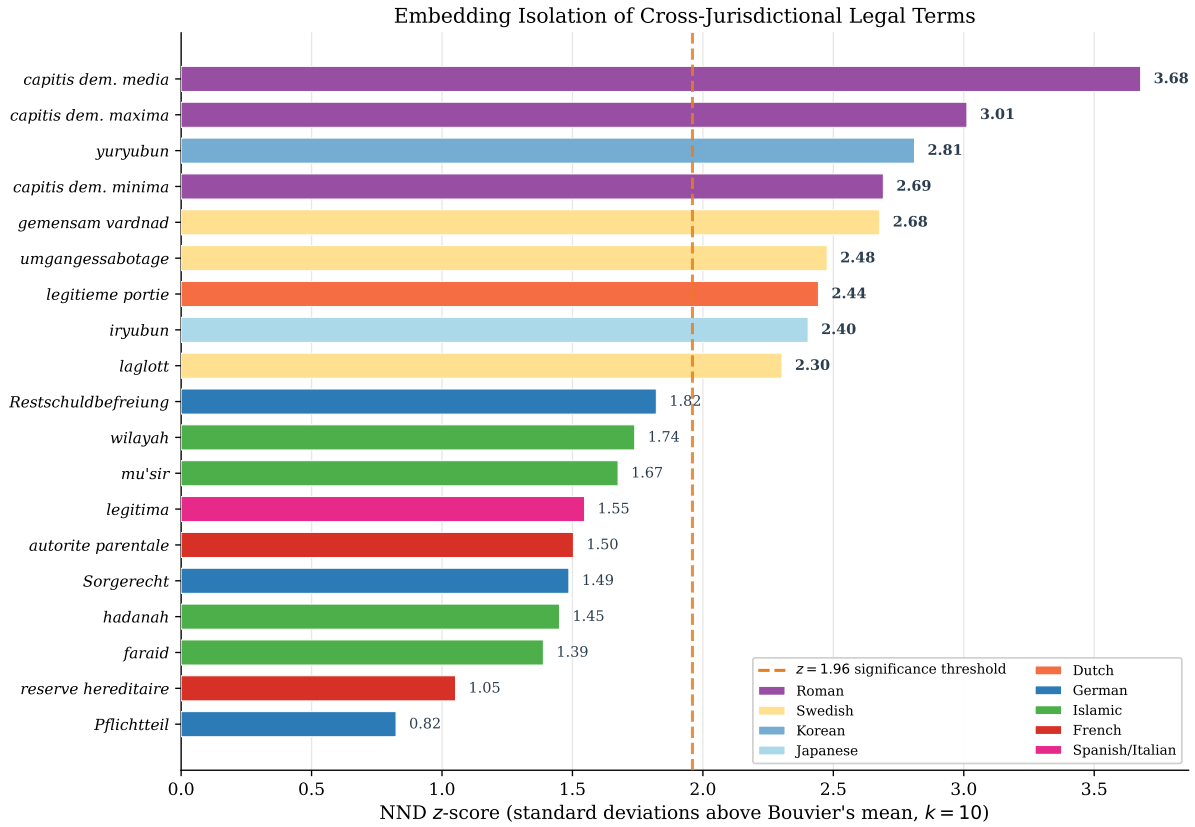

Figure 2: Per-term NND z-scores at $k = 10$, sorted by magnitude. Dashed line: $z = 1.96$ significance threshold ($\alpha = 0.05$). Terms above the line exhibit statistically significant embedding isolation. Colors indicate legal tradition of origin.

Yoonwon Jung, Aaron S. Cohen, and Benjamin K. Bergen. Discovering lexical gaps using embeddings from multilingual LLMs. *arXiv preprint arXiv:2605.24310*, 2025. URL https://arxiv.org/abs/2605.24310.

Pierre Legrand. European legal systems are not converging. *International and Comparative Law Quarterly*, 45(1):52–81, 1996.

John A. Lucy. *Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis*. Cambridge University Press, New York, 1992.

Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. URL https://arxiv.org/abs/1301.3781.

Zach Nussbaum, John Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*, 2024. URL https://arxiv.org/abs/2402.01613.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, 2019. URL https://arxiv.org/abs/1906.05407.

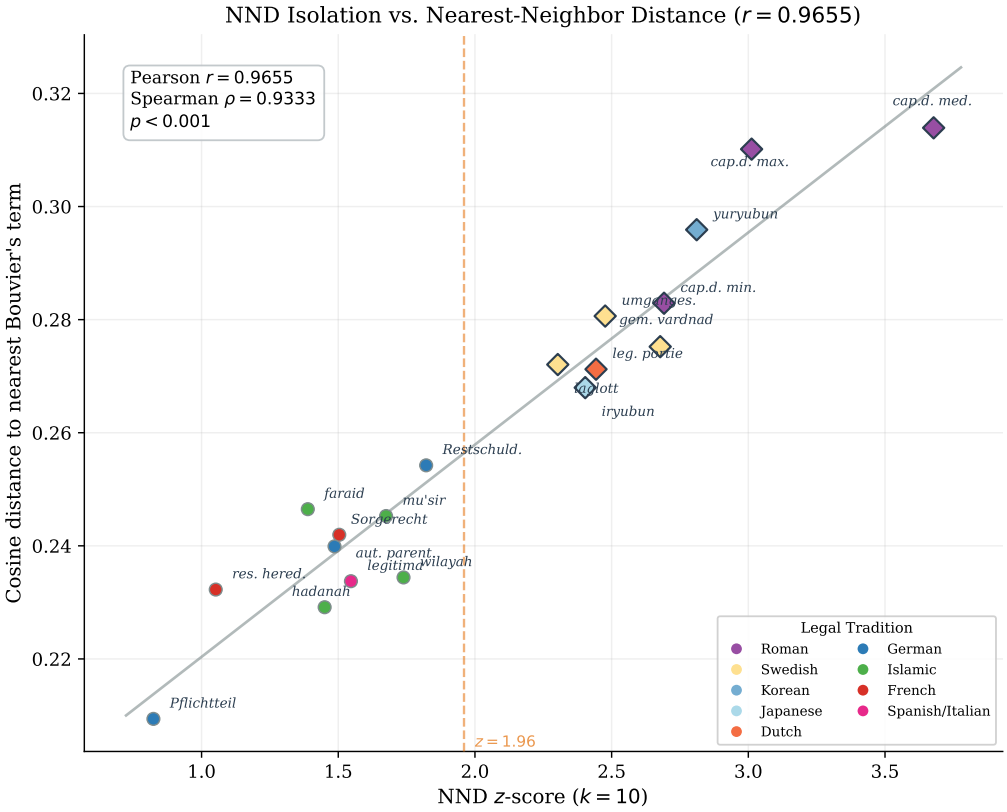

Figure 3: NND z-score ($k = 10$) vs. distance to nearest Bouvier's term for each of the 19 cross-jurisdictional terms. Pearson $r = 0.9655$, Spearman $\rho = 0.9333$ (both $p < 0.001$). The near-perfect correlation confirms that embedding isolation is a global geometric property: terms far from any single English legal concept are also far from all English legal concepts in aggregate.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. URL https://arxiv.org/abs/1908.10084.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, 2020. URL https://arxiv.org/abs/2004.09813.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019. URL https://arxiv.org/abs/1706.04902.

Rodolfo Sacco. Legal formants: A dynamic approach to comparative law (installment I of II). *American Journal of Comparative Law*, 39(1):1–34, 1991.

Edward Sapir. The status of linguistics as a science. *Language*, 5(4):207–214, 1929.

Johanna Schiratzki. Custody of children in Sweden: Recent developments. *Scandinavian Studies in Law*, 38:207–228, 2005.

B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.

Benjamin Lee Whorf. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Technology Press of MIT, Cambridge, MA, 1956. Edited by John B. Carroll.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. Grammatical gender, neo-Whorfianism, and word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://arxiv.org/abs/1910.09729.

Konrad Zweigert and Hein Kötz. *An Introduction to Comparative Law*. Oxford University Press, Oxford, 3rd edition, 1998.