

Typological Features Are Multiple-Choice Questions, Not Yes/No Questions: Lessons from Typological Feature-Value Embeddings

Tung-Le Pan Shu-Kai Hsieh

Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan

r12142006@ntu.edu.tw

shukaihsieh@ntu.edu.tw

Abstract

WALS and Grambank encode typological information in different forms: WALS assigns one value to each multi-valued feature, whereas Grambank mostly records the availability of constructions. Computational pipelines commonly expand both into independent binary columns. We compare that binary encoding with a categorical objective that jointly normalizes the values of each feature. The models otherwise share architecture, data, optimization and capacity. On WALS, the categorical objective recovers more of the structure implicit in feature definitions: distances track ordinal scales, compositional dimensions become consistent directions, and 45% of cross-feature nearest neighbours share a grammatical domain, compared with 20% for the binary model. The advantage persists under reparameterization-invariant diagnostics. External evaluations show the same pattern: categorical embeddings rank independently supported implicational universals more strongly, recover the two clause orders associated with languages lacking a single dominant order even when the corresponding feature is withheld, and correlate more strongly with corpus word-order variability among languages assigned the same WALS label. Held-out-branch prediction differs less between objectives, especially on Grambank. Preserving feature-level competition therefore changes representation geometry even when predictive accuracy changes little.

1 Introduction

WALS feature 81A asks which order of subject, object and verb dominates in a language’s basic clause, and offers seven answers: SOV, SVO, VSO, VOS, OVS, OSV, and *no dominant order*. Japanese is SOV, English SVO, Classical Arabic VSO. Exactly one answer is correct per language, which is what makes 81A a multiple-choice question.

How a question like 81A is encoded matters because typological vectors are used in transfer,

imputation and typology-conditioned models to select informative languages or features (Ponti et al., 2019; Ji et al., 2023). These uses depend on neighbourhood structure as well as cell-level prediction. Under the categorical objective, 45% of a value’s nearest neighbours from other features share its grammatical domain, against 20% under the binary objective; among languages assigned the same database label, predicted uncertainty also correlates more strongly with treebank word-order variability ($r = 0.68$ against 0.19).

Most computational work converts 81A into seven yes/no questions. The encoding introduced with URIEL and lang2vec and continued in URIEL+ (Littell et al., 2017; Khan et al., 2025) gives each value its own binary column and marks the column matching the language. Typological data in this form is used for transfer, probing and database completion (Malaviya et al., 2017; Ponti et al., 2019; Bjerva and Augenstein, 2018; Östling and Tiedemann, 2017; Choenni and Shutova, 2022; Buis and Hulden, 2019; Ji et al., 2023), including recent completion methods based on large language models (Wang et al., 2026; Hus and Anastasopoulos, 2026). This encoding contains no relation among the seven columns: it does not encode that SOV and OSV both place the verb last, that SOV and SVO both begin with the subject, or that VSO and VOS both begin with the verb. In the binary representation, SOV is no closer to OVS than to *tone system absent*.

A categorical objective instead preserves the feature’s answer set and normalizes all of its values jointly with a softmax. The appropriate objective depends on how a database was constructed. WALS assigns one value per feature, so binary expansion is an additional transformation that removes explicit relations among alternatives. Grambank mostly asks independent yes/no questions about construction availability, for which a binary column is appropriate. Its multi-valued parameters

undergo the same expansion as WALS features, so the comparison also applies to that subset of Grambank. Figure 1 summarizes the pipeline and evaluations.

Both objectives yield vector spaces, which we evaluate directly. This extends work that learns typological representations primarily for prediction (Malaviya et al., 2017; Bjerva et al., 2019a,b; Murawaki, 2015). Our diagnostics test whether the geometry represents graded relations among alternatives that independent binary columns do not encode explicitly.

This comparison also bears on gradient typology (Bickel, 2007; Levshina, 2019; Round and Corbett, 2020; Baylor et al., 2023, 2024). Typological databases use discrete labels even when the underlying distributions are gradient. We quantify how much of that structure is recoverable from discrete annotations under each objective and identify the feature types that account for the difference. Ponti et al. (2019) motivate continuous representations of broad typological categories, and Bjerva (2024) discusses the mismatch between database labels and gradient typology.

Contributions. We compare two models that differ only in whether values compete (§3), and measure the effect on held-out-branch prediction in the setting of Bjerva et al. (2019a) (§4). We test the value geometry against an annotation of 61 WALS features in four structural kinds, using a diagnostic matched to each (§5), and examine how values from different features are arranged relative to one another (§6). We then compare the space against evidence outside the annotation: implicational universals, the two orders WALS names for alternating languages, and word-order variability measured from treebanks (§7).

2 Background

Two databases, two kinds of question. WALS (Dryer and Haspelmath, 2013) is a reference work of specialist-written chapters. Each chapter poses one question about a broad property and assigns every language a single value. Because one author designs the answer set as a whole, the values usually stand in a definite relation to each other: points on a scale, cells of a two-way classification, or subtypes of a shared property. Grambank (Skirgård et al., 2023) is organized as a questionnaire, and each parameter asks whether a construction or element is available, answered independently of the

others. Evidence that WALS gathers into one multi-valued question is therefore spread across several Grambank parameters. We use both because they are the two most prominent general-purpose typological databases.

Competing values against independent columns.

Independent columns allow any subset of values to be active, which is appropriate when the columns record availability. Even dominance need not have a single answer, as WALS acknowledges with *no dominant order*. However, independent columns do not explicitly encode relations among alternatives. Activating two columns does not distinguish a language with no dominant order from one with a dominant order plus a second attested order, and the representation does not encode that SOV and SVO share subject-initiality while differing in object position.

Gradient typology. Many typological studies quantify category strength continuously from text, for example by measuring how often an object follows the verb or how mixed a language’s ordering is (Bickel, 2007; Levshina, 2019; Levshina et al., 2023; Round and Corbett, 2020; Baylor et al., 2023, 2024). We use word-order entropy from this work: it is near zero when a language consistently uses one direction and high when both directions occur at similar rates. A useful representation of a discrete database should preserve such intermediate relations when they are supported by cross-linguistic co-occurrence.

Comparative concepts and semantic maps.

Haspelmath (2010) separates language-particular descriptive categories from comparative concepts constructed for cross-linguistic comparison; “SOV order” is one such comparative concept. Typology has long represented comparative concepts spatially. In the semantic-map method, grammatical functions form a graph, and the functions covered by one form in a language must constitute a connected region (Haspelmath, 2003; Georgakopoulos and Polis, 2018; Croft, 2003; van der Auwera and Plungian, 1998). A newly described language in which one form covers two distant functions with no intervening functions would provide evidence against that map (Cysouw, 2007). We estimate analogous relations among database values from their cross-linguistic co-occurrence.

Language vectors and value vectors. Much NLP represents *languages* as vectors, built from

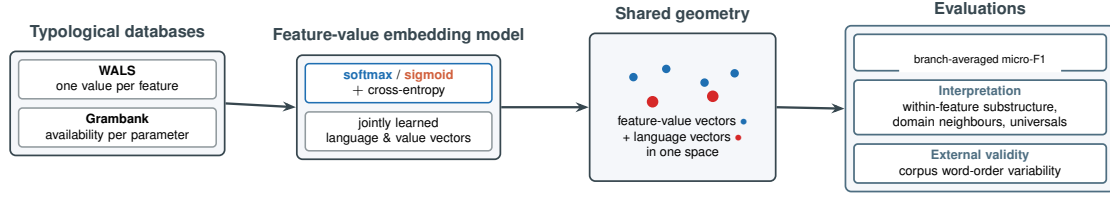


Figure 1: Overview. Both databases are provided to a single model that learns one vector per language and one per feature-value. The two settings we compare differ only in the boxed component: a softmax normalizes a feature’s values against each other, a sigmoid scores each independently. Language and value vectors occupy one space but are analysed separately throughout. The three evaluations are held-out-branch prediction (§4), the structure of the value geometry (§5, §6), and agreement with evidence from outside the databases (§7).

databases (Littell et al., 2017; Khan et al., 2025) or learned from multilingual models and probed afterwards (Östling and Tiedemann, 2017; Bjerva and Augenstein, 2018; Choenni and Shutova, 2022; Murawaki (2015) also learns continuous typological spaces for phylogenetic inference. Our primary objects are feature *values*, so that SOV and *definite affix* each have a vector. Section 6 tests whether values that a typologist would group together occupy nearby regions of this space.

3 Models

Both models learn a vector $\ell_i \in \mathbb{R}^d$ per language and a vector per typological value, and score a value against a language by the dot product $\ell_i^\top v$. Training fits the observed cells of a language-by-value table and skips the rest. The two differ only in how scores become probabilities. Both are trained from scratch on the database tables; no pretrained language model and no text enter either.

Independent binary (BIN). Following the factorization of Bjerva et al. (2019a), each value j has an embedding f_j and is predicted on its own,

$$p(Y_{ij} = 1 \mid \ell_i) = \sigma(\ell_i^\top f_j), \quad (1)$$

with σ the logistic function and a binary cross-entropy loss. The probability of SOV is computed without reference to the probability of SVO.

Categorical (CAT). A feature j with K_j values gets a table $V^{(j)} \in \mathbb{R}^{K_j \times d}$ and one distribution over the set,

$$p(Y_{ij} = k \mid \ell_i) = \frac{\exp(\ell_i^\top V_k^{(j)})}{\sum_{k'=1}^{K_j} \exp(\ell_i^\top V_{k'}^{(j)})}. \quad (2)$$

Normalizing over the feature’s values couples them: raising the score of one lowers the probability of

the rest, so an example asserting that a language is SOV also asserts that it is not SVO and not VSO, and every value vector in the feature moves. Under Eq. 1 the same example moves one.

Training and filtering. Both models use $d=64$, batch size 64, Adam at 10^{-3} with an L2 penalty of 0.1 on the vectors used in each batch, early stopping with patience 10, and a 200-epoch cap. We keep languages with at least ten observed values and values attested in at least ten languages, which leaves 1,683 WALS languages (1,641 with a Glottocode) over 640 values and 2,360 Grambank languages over 392 (Appendix A). The prediction experiments in §4 evaluate on genealogical branches with at least five languages: WALS genera, groups of evidently related languages about as deep as Germanic or Romance, and for Grambank a comparable Glottolog clade. We train five seeds, 42 to 46, and report means with standard deviations; where we compare seeds pairwise we use all ten pairs the five seeds allow. BIN matches CAT in architecture, data, optimizer, regularization and capacity, with the details and limits of that match in Appendix B.

Comparing value geometry. We compare value vectors by cosine distance after subtracting each feature’s mean value vector in both models. Softmax probabilities are invariant to adding the same vector to every value of a feature. Scores also identify the vectors only up to an invertible map, $\ell_i \mapsto A\ell_i$ and $v \mapsto A^{-\top}v$, which can change cosines (Roeder et al., 2021; Nielsen et al., 2026). The L2 penalty favors balanced solutions, which are unique up to rotations that preserve cosine (Srebro et al., 2005). Because early stopping need not reach that optimum, Appendix H repeats the measurements with whitened vectors and score profiles,

both invariant to the relevant reparameterization.

4 Predicting held-out branches

We first evaluate held-out-branch feature prediction using the design of Bjerva et al. (2019a). A genealogical branch is held out, the model trains on every observed cell outside it, a fraction ρ of the held-out branch’s cells is added as in-branch supervision, and the remaining cells are predicted. Each run is scored by micro-F1 over held-out cells, equivalent to accuracy for these single-label cells; we report the mean over branches and five repeats. Branch-level holdout prevents the model from relying on a close relative of the target language and is standard in typological completion (Bjerva et al., 2019a; Buis and Hulden, 2019; Bjerva, 2024).

Branches are defined separately for each database. We use the WALS genus column directly. Grambank has no equivalent, so we use a Glottolog clade two levels below the family root, chosen to approximate WALS-genus granularity but not intended as an exact equivalent.

We do not replicate the accuracy reported by Bjerva et al. (2019a). Their frequency baseline is far lower than the one public WALS sources yield (0.294 in their Table 1 against 0.524 here), and on the same measure their reported figure at $\rho = 0.20$ is close to 0.88 against 0.580 for our matched binary model, a gap of roughly 0.30 that we could not close and cannot account for from the published description. Everything below therefore compares CAT with BIN under identical data, splits and scoring, and not with a published number.

Figure 2 plots both models against the frequency baseline across supervision levels. On WALS, both fall well below that baseline when in-branch evidence is scarce, and BIN is ahead of CAT at $\rho \leq 0.05$; the ordering reverses at $\rho = 0.10$ and widens at $\rho = 0.20$, where CAT reaches 0.622 against 0.580. Over the 81 held-out branches, the $\rho = 0.20$ difference lies between +0.034 and +0.051 (CAT ahead on 70), while the difference at $\rho = 0.05$ lies between -0.023 and -0.008 (Appendix H). The early BIN advantage follows partly from the number of training targets: one observed cell of a K -valued feature supplies K binary targets to BIN but one categorical target to CAT. On Grambank, both models remain near the high frequency baseline (0.723); at $\rho = 0.20$, CAT exceeds BIN by 0.017 (+0.014 to +0.021 over 52 branches). Most Grambank parameters are binary, so only the multi-

valued parameters differ substantively between the two objectives. CAT’s mean is higher at every supervision level, and from $\rho = 0.05$ onward the branch-resampling intervals exclude zero.

The predictive margins are modest. Bjerva (2024) argues that filling database gaps does not by itself constitute a contribution to typological knowledge because the predictions largely recover existing correlations. The remaining analyses therefore focus on representation structure.

5 Within-feature structure

To evaluate within-feature geometry, we annotated 61 WALS features, assigning each to one of four structural types and using a diagnostic matched to that type.

An *ordinal scale* orders its values, as 22A does for inflectional synthesis, from two categories per word up to nine. Here we correlate the distance between two values with the gap in their annotated rank, and score the result against a null that permutes the value labels within the feature.

A *presence-to-subtype* feature offers an absent value and several ways of being present. Feature 9A distinguishes no velar nasal, a velar nasal barred from initial position, and one that occurs initially. The diagnostic is triplet accuracy: for each triple consisting of the absent value and two subtypes, the two subtypes should be closer to each other than either is to absence. Numeral classifiers (55A) is a clear case, where *optional* and *obligatory* classifiers both presuppose a classifier system and should group against *no classifiers*.

The final two types are grid-like but encode different factor structures. A *Boolean lattice* crosses presence against presence: 4A records whether a voicing contrast exists among plosives and whether one exists among fricatives, and a value is a conjunction of two independent yes/no facts. A *cross-product* crosses one set of competing strategies against another: 34A crosses which nouns take plural marking, human or all, against whether the marking is optional or obligatory, and a language selects one level from each set. For a lattice we ask whether each component is carried by a direction: every pair of values differing in that component alone should be displaced the same way, so we measure how well those difference vectors align. For a cross-product we ask whether each axis is recoverable as a grouping, classifying each value by nearest centroid among the groups on that axis,

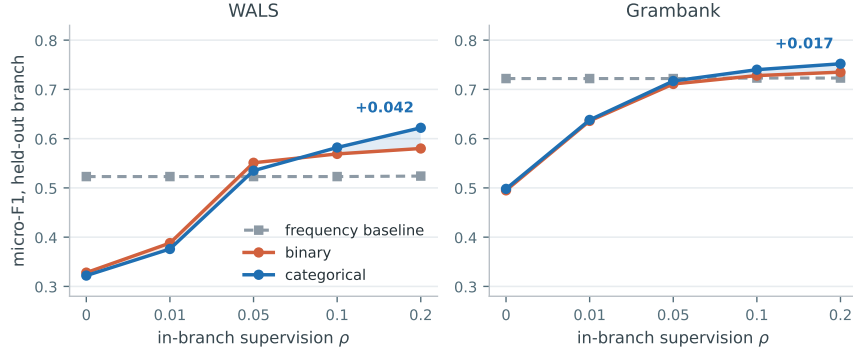


Figure 2: Held-out-branch performance against in-branch supervision ρ (branch-averaged micro-F1), with the frequency baseline for each database. Shading marks where the two models differ; BIN leads on WALs while supervision is thin, and CAT leads from $\rho=0.10$. On Grambank the frequency baseline is high and both models stay close to it.

with the value held out of its own centroid.

Table 1 reports all four diagnostics. CAT has the higher mean for every type, with the largest difference on presence-to-subtype (+0.189 in triplet accuracy). The effect occurs across multiple features: CAT wins 12 of 20 ordinal features and ties 4, wins 8 of 15 presence features and ties 6, and wins 6 of 7 lattices. At the feature level, CAT wins 32, ties 12 and loses 8 of the 52 (sign test $p \approx 2 \times 10^{-4}$); by type, only presence-to-subtype reaches $p < .05$ on its own (Appendix H). Figure 3 shows 22A, where BIN places the ends of the ordinal scale closer together than the middle, and 37A, where it weakens the separation between absence and presence.

Lattices are the weakest case for both models; the difference there reaches only $p = 0.06$, so evidence for a model difference is weaker.

6 Cross-feature organization

We next test whether cross-feature neighbours share linguistic domains. Taking each of the 640 WALs values and asking what share of its nearest neighbours from other features fall in the same WALs area, CAT reaches 0.450 at the nearest neighbour against a chance rate of 0.162, while BIN reaches 0.195, which is close to chance; Figure 5 breaks the same measurement down by area, and Table 3 gives the rates at larger k . The rates are computed within each seed and then averaged, and no single BIN seed reaches 0.23, so the mean is not an unstable structure averaged into noise. The comparison also survives the reparameterization concern of §3: on whitened vectors CAT reaches 0.550 against 0.351, and on score profiles across languages 0.401 against 0.212 (Appendix H). Whiten-

Structure	n	CAT	BIN	W/T/L
Ordinal scale	20	.252	.157	12/4/4
Presence to subtype	15	.693	.504	8/6/1
Cross-product	10	.695	.598	6/2/2
Boolean lattice	7	.084	.004	6/0/1

Table 1: Within-feature recovery by structure, five-seed means, with wins, ties and losses counted per feature. A tie is a feature on which the two five-seed means are equal; most are three-valued features that both models score identically on every seed. Each row uses the diagnostic described in the text, so rows are not comparable with each other. Ordinal and lattice rows are margins over a within-feature label-permutation null; presence and cross-product rows are accuracies. All four differences favour CAT at $p < 0.01$ on a paired t -test across the five seeds, except the lattice row at $p = 0.06$; that test measures optimization variance, and the text gives tests over features. Cross-product features number 18 before filtering and 10 after.

ing helps BIN most, because its raw vectors share one dominant direction, the narrow band of Figure 4, which hides some of the domain structure it does carry.

Figure 7 lists the nearest cross-feature neighbours for four values from different areas. Under CAT, *strongly suffixing* morphology (26A) is nearest to tense-aspect suffixes (69A), case suffixes (51A) and plural suffixes (33A), grouping affix-position properties across verbal and nominal morphology; under BIN, its neighbours include long vowels and locational marking. Ergative-absolutive alignment in full noun phrases (98A) is adjacent to the same alignment in pronouns (99A) and to oblique-patient constructions (108A). *Object before verb* (83A) has the combined OV-and-

Within-feature value distances

model cells show the five-seed mean and standard deviation; diagonals are blank because self-distance is zero

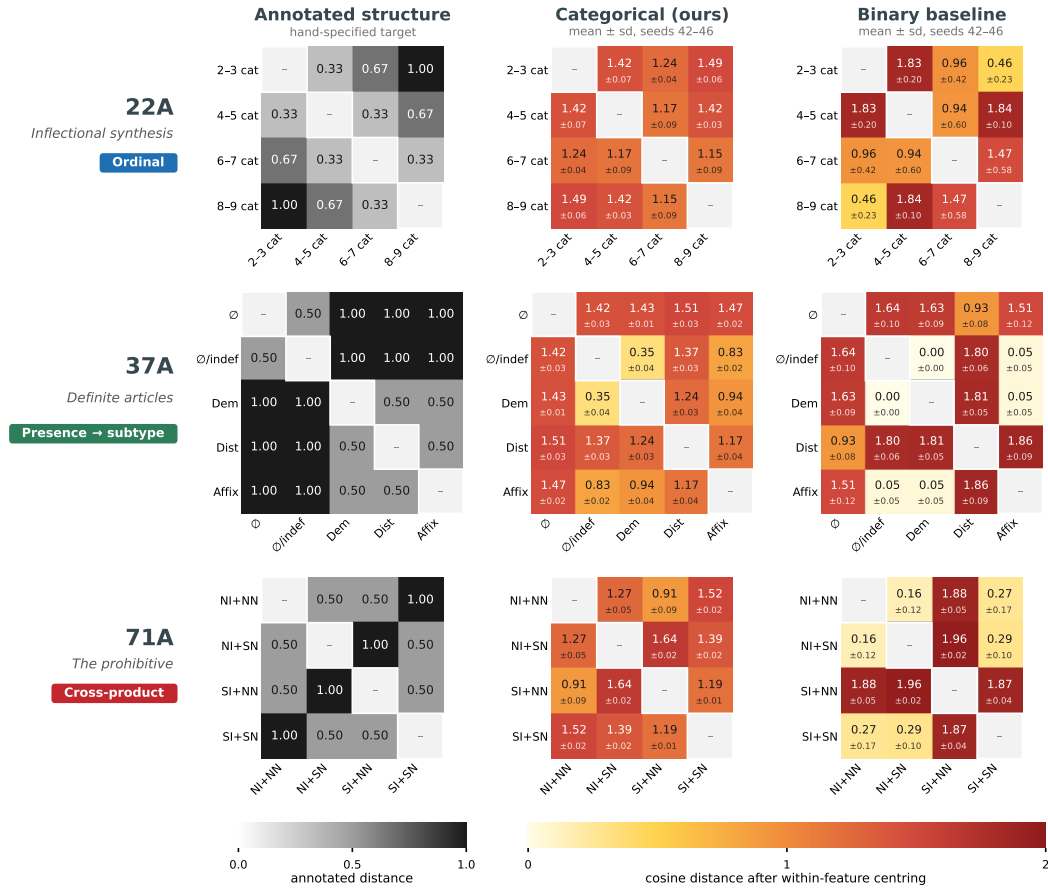


Figure 3: Distances among the values of three features: 22A (ordinal), 37A (presence to subtype) and 71A (cross-product). Left, the annotated structure; middle, CAT averaged over five seeds; right, BIN. Dark cells are pairs the model places close together. BIN inverts the 22A scale and blurs the 37A boundary between absence and presence.

postposition value of 95A, SOV and postpositions among its nearest neighbours. A complex tone system (13A) is nearest to exclusively isolating morphology and zero case-marking, plausibly reflecting the areal profile of mainland Southeast Asia without implying a direct dependency between those features.

Feature centroids show similar domain organization. After collapsing each feature to the mean direction of its values, object–verb order is adjacent to genitive–noun, adposition and adjective–noun order, and the two clause-order features, 81A and 82A, are adjacent. Such neighbourhoods matter for transfer-language selection, imputation and typology-conditioned models: a representation that placed adposition order among phonological fea-

tures would provide poorly matched sources for these uses.

Some established typological patterns are recovered under both objectives. Within the word-order values, both models separate head-final from head-initial almost perfectly, with 0.96 of the five nearest neighbours on the same pole in each case, consistent with strong Greenbergian correlations in the co-occurrence data (Greenberg, 1963; Dryer, 1992). Seed-to-seed stability differs more: across the ten seed pairs, Jaccard overlap among a value’s ten nearest neighbours is 0.706 under CAT and 0.204 under BIN on WALS; Figure 4 gives a qualitative projection. Language vectors also recover Glotlog family membership above chance under both objectives and about twice as strongly under CAT.

Because typological and genealogical similarity covary, we use this only as a representation-quality diagnostic (Appendix D).

7 External validation

7.1 Implicational universals

An implicational universal is a cross-linguistic conditional: object before verb tends to go with postpositions. We convert 191 entries from the Konstanz Universals Archive into implication formulas over Grambank values, following the operationalizations of Verkerk et al. (2026). Of these, 172 can be evaluated in our space; the remainder either contain aggregates our scoring does not handle, carry no implication arrow, or refer to values that our attestation threshold removed from the Grambank matrix, and we call this last restriction the coverage filter.

Proposed universals can reflect genealogical or areal confounds: a statement may hold in a sample because a few large families share the trait or because neighbouring languages borrowed it. Verkerk et al. (2026) test each candidate with a spatiophylogenetic regression and a co-evolutionary analysis. Sixty formulas pass both tests, 29 pass the spatiophylogenetic test alone, and 102 pass neither; after our coverage filter the groups contain 55, 27 and 90 formulas. If the representation encodes the dependencies supported by these external analyses, fully supported formulas should score above unsupported formulas, with the intermediate group between them.

CAT exhibits this ordering. Formulas passing both tests reach a structured implication cosine of +0.684, with 52 of 55 above the null; those passing one reach +0.262; those passing neither reach +0.185 (Mann–Whitney test of well-supported against unsupported, $p < 0.0001$). BIN separates the classes far less, scoring unsupported formulas at +0.552 against supported at +0.766. Its raw cosines sit on a higher baseline, since random cross-feature pairs average +0.328 under BIN and 0.000 under CAT, so the two models are best compared on a scale-free measure. BIN can evaluate only 146 of the 172 formulas, because a two-valued feature gives it a single column and no vector for the absent value; on those 146, the probability that a fully supported formula outscores an unsupported one is 0.81 for CAT and 0.64 for BIN, a difference of +0.17 (+0.08 to +0.27 under a bootstrap over formulas). Agreement is strongest

for the narrow word-order and hierarchy universals that Verkerk et al. (2026) find best supported, and near zero for the group they label *other*, which collects formulas from several unrelated domains and has no shared subject matter for the geometry to reflect (Table 4); Figure 9 shows the separation for both models.

7.2 Languages with two dominant orders

WALS asks a second clause-order question that standard binary expansion does not connect to 81A. Feature 81B names the two dominant orders of languages that have two, and its 66 languages are almost all recorded in 81A as having no dominant order; the exception is Suri, recorded as SVO in 81A with an SOV alternative in 81B. Binary expansion retains 81B as separate columns but encodes no constraint linking an 81B pair to *no dominant order* in 81A.

We ask whether the predicted distribution over 81A values concentrates on the pair that 81B names. For each language with an 81B value, we measure the share of predicted order-mass falling on the two orders it names. CAT places 0.613 ± 0.002 there against a chance share of 0.400, and BIN places 0.496 ± 0.004 (paired $p = 1.3 \times 10^{-6}$). Three 81B values pass the attestation threshold and were in training, but retraining both models without 81B leaves the result unchanged (CAT 0.612 ± 0.002 , BIN 0.500 ± 0.006). The pair can be inferred from elsewhere in the table: for 40 of the 65 languages with 82A (subject and verb) and 83A (object and verb) recorded, those two values single out the same pair, yet withholding them as well lowers CAT only to 0.600 (BIN 0.493). The categorical distribution over 81A therefore agrees more closely with the pair implied by related features and distinguishes the specific alternatives associated with each language.

We also expected two-order languages to fall between fixed and fully free languages in predicted entropy, but their entropy is 1.52, above the 1.38 for languages with no 81B value. This comparison does not test the intended scale. A language may lack an 81B value because its freedom extends beyond two orders or because no second order was established; the latter is common. Languages without an 81B value have 46 observed WALS features on average, compared with 64 for languages with an 81B value, so they cannot be interpreted as a uniformly free comparison group.

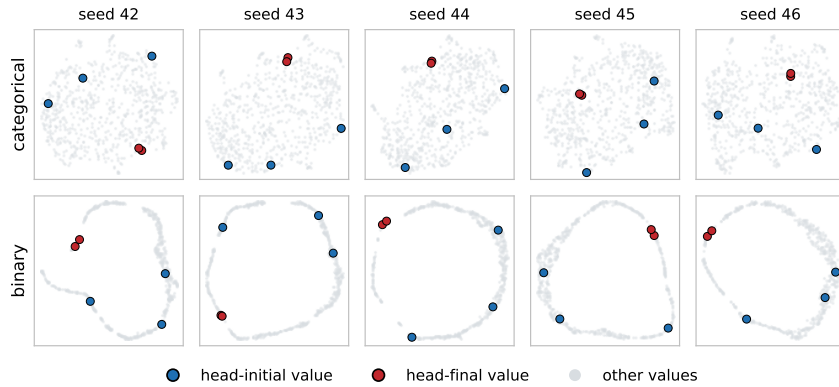


Figure 4: UMAP projections of the two WALS value spaces (cosine metric, $n_neighbors=20$, $min_dist=0.30$), one column per training seed, with five head-directionality values marked. Head-final values fall together under both objectives, and the separation between the head-directionality poles is if anything wider for the binary model (2.59 ± 0.34 against 2.02 ± 0.14). The principal difference concerns the remaining values: categorical values occupy the interior of the space, whereas binary values concentrate on a narrow band. UMAP is a visualization and no measurement in the paper depends on it.

7.3 Word-order variability in corpora

We next compare model predictions with token-level word-order frequencies from Universal Dependencies treebanks (Nivre et al., 2020). We compute, for each relation, how often the head precedes its dependent and the entropy of that choice. Our measurements cover 95 languages: those with a UD treebank that survive filtering in the WALS matrix and carry the relevant value. We computed head-direction profiles for 54 of them directly from the treebank repositories, extending an earlier set of 41.

For each model, we obtain a predicted distribution over the relevant WALS values and test whether probability assigned to the head-initial value tracks the corpus proportion and whether predicted entropy tracks corpus entropy. We apply the same tests to URIEL, a maintained binary resource in wide use (Khan et al., 2025).

Both models track direction, with mean $|r| = 0.885$ for CAT and 0.744 for BIN; variability correlations are $r = 0.365$ and 0.100 , respectively (Table 2; paired $p = 8.6 \times 10^{-6}$ and 5.1×10^{-5}). On adjective–noun order, BIN is anti-correlated with corpus variability. URIEL’s normalized order columns reach 0.894 on direction and 0.341 on variability. One possible source of the latter signal is database disagreement: URIEL+ merges sources using their maximum or average (Khan et al., 2025), so conflicting source labels distribute mass across

multiple order columns. For CAT, softmax outputs are normalized probability distributions and their entropy directly measures predicted uncertainty; renormalized independent sigmoid columns do not have the same probabilistic interpretation.

Restricting the test to the 84 languages WALS records with a single dominant order removes variation in the database label as a possible source of the signal. On this subset, CAT reaches $r = 0.682 \pm 0.015$ and BIN 0.191 ± 0.094 (Figure 10). The 95 languages come from 28 Glottolog families; after resampling families, the interval for the CAT–BIN difference excludes zero on the single-order subset ($+0.31$ to $+0.71$) and on the full-sample variability and direction measures (Appendix H).

Although geography is absent from training, predicted variability is spatially structured. Neighbouring languages differ about a quarter less than randomly paired ones (999-shuffle permutation test, $p < 0.001$), same-family pairs less again, and among languages WALS records as fixed the Australian mean is roughly twice the global one ($p < 10^{-14}$). This pattern is consistent with descriptive and theoretical work on flexible constituent order in Australian languages (Dixon, 1980; Hale, 1983) (Appendix G).

8 Discussion

Across the representation analyses, CAT recovers more of the structure implicit in WALS feature

Feat	UD rel.	Variability r	
		CAT	BIN
81A	obj	.428	.165
83A	obj	.601	.349
85A	adp	.360	.059
86A	gen	.169	.062
87A	adj	.269	-.137
<i>Mean</i>		.365	.100
<i>Direction</i> $ r $, mean		.885	.744

Table 2: Correlation with corpus word order over the 95 treebank languages, five-seed means. *Variability* compares the entropy of a model’s predicted distribution with the corpus entropy of the corresponding relation; *Direction* compares the probability placed on the head-initial value with the corpus proportion. URIEL is reported separately in the text because its columns are not in one-to-one correspondence with the WALS features (Appendix F).

definitions. Within features, ordinal scales appear as ordered distances, presence separates from subtypes, and lattice components are partly recoverable. Across features, same-domain nearest neighbours occur at more than twice the BIN rate and are more stable across seeds. External tests show the same direction of difference: CAT orders implicational universals by independent support, recovers the two alternatives associated with WALS 81B, and correlates more strongly with corpus word-order variability among languages that share the same WALS label. These differences are much smaller in held-out-cell prediction, where the objectives remain close and both approach the high Grambank frequency baseline.

For NLP applications, the relevant comparison depends on how typological embeddings are used. Transfer, imputation and conditioning use neighbourhood relationships, whereas database completion evaluates individual cells (Ponti et al., 2019; Ji et al., 2023). Preserving WALS feature structure requires no additional annotations; it changes the training objective. In held-out-branch prediction, CAT surpasses BIN only after the branch contributes sufficient supervision, while BIN performs better when in-branch evidence is extremely sparse. Prediction accuracy and representation geometry should therefore be evaluated separately. The present model is a 64-dimensional embedding over two databases. Future work could incorporate additional databases and the descriptive material underlying feature values. Another extension is to assign provisional coordinates to newly proposed

values from partial evidence without introducing a separate predeclared binary column for each value.

For linguistic typology, the embedding provides quantitative relations among comparative concepts. Semantic maps already encode such relations using graph connectivity (Haspelmath, 2003; Georgakopoulos and Polis, 2018); the learned space extends this idea to the retained database values using cross-linguistic co-occurrence. The corpus analysis further shows that co-occurrence among discrete labels retains information about gradient word-order variability (Levshina, 2019; Levshina et al., 2023; Baylor et al., 2024). The universals analysis should not, however, be interpreted as discovering new universals from database correlations. Establishing a universal requires language-level descriptive evidence and controls for ancestry and contact (Newmeyer, 2005; Ploeger et al., 2024).

Limitations

The main result concerns representation. A categorically trained value space recovers more structure from the database annotations than the matched binary encoding, but this does not establish that grammatical categories themselves are continuous.

BIN is matched to CAT in architecture, data, optimization and capacity; it was not tuned independently for maximum binary-system performance. Public WALS data also yields a much easier frequency baseline than the setting reported by Bjerva et al. (2019a), so BIN is not a numerical replication of that published score. A separately tuned binary system, for example one with per-value bias terms, could reduce part of the observed gap (Appendix B). URIEL+ cannot serve as an independent predictive baseline because it is built from WALS and Grambank, so we use it only for the corpus comparison in §7.3.

The substructure annotation is single-author and checked against WALS chapter prose, with less certain cells marked provisional. It has not been independently re-annotated; a blind re-annotation protocol and Cohen’s κ script were prepared for that purpose. At the feature level, only the presence-to-subtype advantage is significant on its own, the ordinal result is borderline, and cross-products and lattices contain too few retained features for a clear type-specific result. Features with only two retained values provide no within-feature direction to recover.

The corpus analyses cover word order alone,

where usage-based measures are well established; extending the test to other grammatical domains would require domain-specific continuous measures. The 95 languages are those whose treebanks resolve to a Glottocode in the WALS matrix and yield a usable profile; 18 treebank languages remain unresolved, and a hand-checked alias table could add some of them. Some treebanks are translations. Kann (2025) finds that translated parliamentary text follows the source in subject–verb order proportions but shifts toward the target language’s dominant order, which could understate variability for languages measured mainly from translated material. The universals probe also inherits the coverage and operationalization decisions of Verkerk et al. (2026); similarity between value vectors does not directly test whether two strategies can co-occur.

Ethical Considerations

This work uses published typological databases (WALS, Grambank, URIEL+), the Konstanz Universals Archive and Universal Dependencies treebanks under their licenses; the models are trained on database tables alone, and no text enters them. Typological databases encode expert judgments that can be incomplete or contested, especially for under-documented languages. Our methods represent and rank existing annotations; they do not establish the descriptive validity of those annotations. Model-derived structure should therefore be checked against descriptive sources before use.

Acknowledgments

We thank the anonymous MRL reviewers for helpful comments. Claude (Anthropic; chat and Claude Code) was used to assist with analysis-code development and debugging and to draft or edit portions throughout the manuscript, including the main text and appendices. The authors verified the code, numerical results, citations and final text and take responsibility for the published work.

References

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. [The past, present, and future of typological databases in NLP](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1163–1169. Association for Computational Linguistics.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. [Multilingual gradient word-order typology from Uni-](#)

[versal Dependencies](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–49, St. Julian’s, Malta. Association for Computational Linguistics.

- Balthasar Bickel. 2007. [Typology in the 21st century: Major current developments](#). *Linguistic Typology*, 11(1):239–251.
- Johannes Bjerva. 2024. [The role of typological feature prediction in NLP and linguistics](#). *Computational Linguistics*, 50(2):781–794.
- Johannes Bjerva and Isabelle Augenstein. 2018. [From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916. Association for Computational Linguistics.
- Johannes Bjerva, Yova Kementchedjheva, Ryan Cotterell, and Isabelle Augenstein. 2019a. [A probabilistic generative model of linguistic typology](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics.
- Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019b. [What do language representations really represent?](#) *Computational Linguistics*, 45(2):381–389.
- Annebeth Buis and Mans Hulden. 2019. [Typological feature prediction with matrix completion](#). In *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP*, pages 13–15.
- Rochelle Choenni and Ekaterina Shutova. 2022. [Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology](#). *Computational Linguistics*, 48(3):635–672.
- William Croft. 2003. *Typology and Universals*, 2 edition. Cambridge University Press.
- Michael Cysouw. 2007. [Building semantic maps: The case of person marking](#). In Matti Miestamo and Bernhard Wälchli, editors, *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*, pages 225–248. Mouton de Gruyter, Berlin.
- R. M. W. Dixon. 1980. *The Languages of Australia*. Cambridge University Press.
- Matthew S. Dryer. 1992. [The Greenbergian word order correlations](#). *Language*, 68(1):81–138.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

- Thanasis Georgakopoulos and Stéphane Polis. 2018. [The semantic map model: State of the art and future avenues for linguistic research](#). *Language and Linguistics Compass*, 12(2):e12270.
- Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press.
- Kenneth Hale. 1983. [Warlpiri and the grammar of non-configurational languages](#). *Natural Language and Linguistic Theory*, 1(1):5–47.
- Martin Haspelmath. 2003. [The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison](#). In Michael Tomasello, editor, *The New Psychology of Language*, volume 2, pages 211–243. Erlbaum, New York.
- Martin Haspelmath. 2010. [Comparative concepts and descriptive categories in crosslinguistic studies](#). *Language*, 86(3):663–687.
- Jonathan Hus and Antonios Anastasopoulos. 2026. [A RAG approach for typological database completion](#). In *Proceedings of the 8th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 39–49, Rabat, Morocco. Association for Computational Linguistics.
- Tao Ji, Yuanbin Wu, and Xiaoling Wang. 2023. [Typology guided multilingual position representations: Case on dependency parsing](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13524–13541. Association for Computational Linguistics.
- Amanda Kann. 2025. [Are translated texts useful for gradient word order extraction?](#) In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 177–182, Vienna, Austria. Association for Computational Linguistics.
- Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. [URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952. Association for Computational Linguistics.
- Natalia Levshina. 2019. [Token-based typology and word order entropy: A study based on Universal Dependencies](#). *Linguistic Typology*, 23(3):533–572.
- Natalia Levshina, Savithry Namboodiripad, Marc Allasonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoyanova. 2023. [Why we need a gradient approach to word order](#). *Linguistics*, 61(4):825–883.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. [URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.
- Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. [Learning language representations for typology prediction](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535.
- Yugo Murawaki. 2015. [Continuous space representations of linguistic typology and their application to phylogenetic inference](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334, Denver, Colorado. Association for Computational Linguistics.
- Frederick J. Newmeyer. 2005. *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. Oxford University Press.
- Beatrix M. G. Nielsen, Emanuele Marconato, Luigi Gresele, Andrea Dittadi, and Simon Buchholz. 2026. [Logit distance bounds representational similarity](#). In *Proceedings of the 43rd International Conference on Machine Learning*. ArXiv:2602.15438.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.
- Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. [What is “typological diversity” in NLP?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700.
- Edoardo Maria Ponti, Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.
- Geoffrey Roeder, Luke Metz, and Durk Kingma. 2021. [On linear identifiability of learned representations](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9030–9039. PMLR.
- Erich R. Round and Greville G. Corbett. 2020. [Comparability and measurement in typological science: The bright future for linguistics](#). *Linguistic Typology*, 24(3):489–525.

- Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latache, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. [Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss](#). *Science Advances*, 9(16):eadg6175.
- Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. 2005. [Maximum-margin matrix factorization](#). In *Advances in Neural Information Processing Systems 17*. MIT Press.
- Johan van der Auwera and Vladimir A. Plungian. 1998. [Modality’s semantic map](#). *Linguistic Typology*, 2(1):79–124.
- Annemarie Verkerk, Olena Shcherbakova, Hannah J. Haynie, Hedvig Skirgård, Christoph Rzymiski, Quentin D. Atkinson, Simon J. Greenhill, and Russell D. Gray. 2026. [Enduring constraints on grammar revealed by Bayesian spatiophylogenetic analyses](#). *Nature Human Behaviour*, 10:126–136.
- Qianwen Wang, York Hay Ng, Aditya Khan, and En-Shiun Annie Lee. 2026. [Typological feature prediction with large language models: An in-context learning approach](#). arXiv:2609.03775. *Preprint*, arXiv:2609.03775.
- Robert Östling and Jörg Tiedemann. 2017. [Continuous multilinguality with language vectors](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649. Association for Computational Linguistics.

A Training details

Embedding dimension $d=64$, batch size 64, Adam ($\beta_1=0.9$, $\beta_2=0.999$) at learning rate 10^{-3} , early stopping on a 5% validation split with patience 10, a 200-epoch cap, and seeds 42–46. Adam’s built-in weight decay would shrink every row at every step, including the many rows absent from a batch, and its adaptive step turns that into a constant-rate pull toward zero. We therefore set it to zero and add an explicit L2 penalty, with coefficient 0.1 as in Bjerva et al. (2019a), on the language and value vectors that appear in each batch. Both tables start from $\mathcal{N}(0, 0.01^2)$, and neither model has bias terms.

Filtering retains languages with at least ten observed values and values attested in at least ten languages. The retained matrices hold 1,683 WALS languages (1,641 of them with a Glottocode, the set used wherever a Glottocode is needed) over 640 values and 2,360 Grambank languages over 392 values. Parameter counts are 147k for BIN against 149k for CAT on WALS, and 164k against 176k on Grambank, so CAT carries 1% and 7% more.

Branches for the held-out-branch experiments, those with at least five languages, come from the WALS genus column for WALS, and for Grambank from a Glottolog clade taken two levels below the family root, since Grambank has no genus column. The frequency baseline in Figure 2 predicts each feature’s most frequent value under the same splits, giving 0.524 on WALS and 0.723 on Grambank at $\rho=0.20$. For reference, the global majority-class rate over the retained matrices is 0.543 across 169 WALS features and 66,472 observed cells, and 0.741 across 192 Grambank features and 342,621 observed cells.

B The binary baseline

Public WALS sources yield a much higher frequency baseline than the setting reported by Bjerva et al. (2019a), and we could not reconstruct from the published description the evaluation protocol that produced the reported score. We therefore do not present BIN as a numerical replication. It is an architecturally matched control: the same embeddings, the same scoring function, the same data, optimizer and L2 penalty, differing only in Eq. 1 against Eq. 2. Because a two-valued feature needs only one binary column, BIN has 609 value vectors on WALS where CAT has 640. Its seed-42 run on WALS stops at epoch 17 and keeps its epoch-7 weights, where training and validation loss are both 0.440, so it is neither underfitting nor overfitting at the retained point. Across seeds BIN’s retained epochs are 6, 7, 14, 14 and 29 against 56 to 83 for CAT; the binary objective reaches its validation optimum sooner, and its geometry is fixed earlier in training. It carries slightly fewer parameters than CAT. We did not test a binary system tuned specifically for peak predictive performance; such tuning could reduce part of the observed gap.

C Substructure annotation

Coverage. Sixty-one WALS features carry a schema, distributed as ordinal 20, presence-to-subtype 15, cross-product 18 and Boolean lattice 8. Table 1 reports the 52 that retain at least three values after filtering, three values being the minimum for a distance profile. The annotation files list each feature, its structural type, its values, and the coding of its ranks, components or dimensions.

What the annotation records. An ordinal feature receives an integer rank per value, taken from the ordering the chapter itself uses; 22A counts

inflectional categories per word, so its four values are ranked 1 to 4. A presence-to-subtype feature receives a presence bit per value and, where present, a subtype label; 37A distinguishes no article at all and an indefinite article only on the absent side, against a definite affix, a definite word distinct from the demonstrative, and a demonstrative used as an article on the present side. A cross-product receives a coordinate per dimension; 34A assigns a domain (human nouns, all nouns) and an obligatoriness level (optional, obligatory, mixed), so *only human nouns*, *optional* is the pair (human, optional). A Boolean lattice receives a vector of component bits; 4A assigns voicing in plosives and voicing in fricatives, so *in both* is [1, 1], *in plosives alone* is [1, 0] and *no voicing contrast* is [0, 0].

Reference distances and diagnostics. Scales use the absolute difference in rank. Presence-to-subtype features use a three-level distance, zero within a subtype, one half between subtypes on the same side of the presence boundary, and one across it, and the reported diagnostic is the mean across-boundary distance minus the mean same-side distance. Cross-products and lattices use the alignment of difference vectors described in §5: for each dimension or component we take every pair of values differing along it while agreeing on all others, normalize the difference between their vectors, and average the pairwise cosine among those differences, reversing the sign for pairs that traverse the same transition in opposite directions. Nulls permute value labels onto vectors within a feature, 200 times per feature, and the reported score is the observed value minus the mean of the null.

Provenance. Annotation is single-author and checked against the WALS chapter text, with medium-confidence cells marked provisional. A blind re-annotation packet lists every feature with its values but not its type, together with a stratified sample for a second annotator and a Cohen’s κ script.

D Neighbourhoods, stability and language vectors

Same-domain neighbour rates over all 640 WALS values (609 for BIN) are 0.450, 0.342 and 0.294 at $k = 1, 5, 10$ for CAT, against 0.195, 0.187 and 0.183 for BIN, with a chance rate of 0.162 set by the distribution of area sizes. Restricted to the word-order values, the share of neighbours that are

Property	Data	CAT	BIN
Same-domain neighbour $k=1$, chance .162	WALS	.450 $\pm.011$	.195 $\pm.017$
Same-domain neighbour $k=5$	WALS	.342 $\pm.006$	.187 $\pm.007$
Seed stability Jaccard@10	WALS	.706 $\pm.133$	.204 $\pm.144$
Seed stability	Grambank	.672 $\pm.138$	.430 $\pm.190$
Family kNN lang. vectors, chance .044	WALS	.277 $\pm.004$	.125 $\pm.005$
Family kNN chance .085	Grambank	.448 $\pm.003$	.230 $\pm.031$

Table 3: Properties of the learned space, five-seed means. Same-domain neighbour: the share of a value’s k nearest neighbours from other features that share its WALS area, computed over all 640 values (609 for BIN, which has one column per two-valued feature). Family kNN clears a 500-permutation baseline at $p < 0.002$.

themselves word-order values runs 0.78, 0.66 and 0.47 at $k = 1, 5, 10$ for CAT against 0.42, 0.26 and 0.21 for BIN (Figure 6).

Seed stability measured by Procrustes-aligning whole value clouds between seed pairs, where lower disparity means more reproducible (Figure 8a), is 0.043 ± 0.001 on WALS and 0.072 ± 0.003 on Grambank for CAT, against 0.126 ± 0.005 and 0.222 ± 0.029 for BIN.

Language and value vectors are coupled only through the score $\ell_i^T v$. Principal angles between the top-ten subspaces of the two clouds average 16.9° for CAT, where two random ten-dimensional subspaces average about 70° , and a linear boundary separates the clouds at 88%. Ranking languages by cosine to a value vector retrieves genuine carriers of that value at precision@10 of 0.364 for CAT, against a base rate of 0.063 ($5.8\times$), and at 0.066 for BIN, 1.06 times its base rate.

E Universals probe

Source and coverage. Verkerk et al. (2026) operationalize 191 proposed Universals Archive statements as formulas over Grambank parameters and test each one. We take their formulas and support labels as given. After excluding formulas containing sum aggregates, formulas with no implication arrow, and formulas referring to values absent from our retained Grambank space, 172 remain testable.

The three grades of support. Two independent tests are reported per formula. A spatiophylogenetic regression asks whether the implication sur-

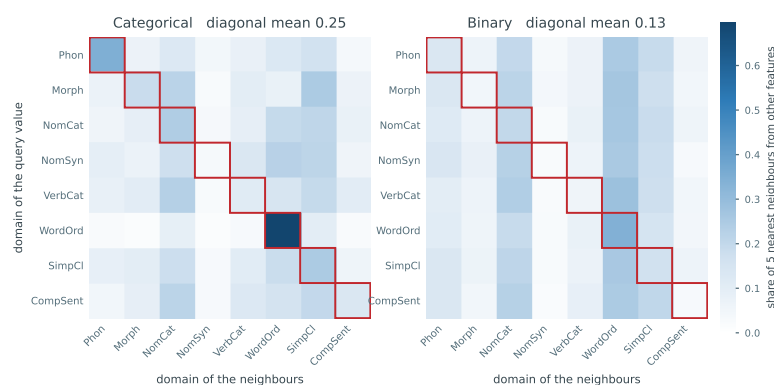


Figure 5: Where a value’s neighbours come from, by domain. Each row is the WALS area of a query value, and the row shows how its five nearest neighbours from other features are spread across areas, averaged over five seeds. Outlined cells on the diagonal are neighbours from the query’s own area.

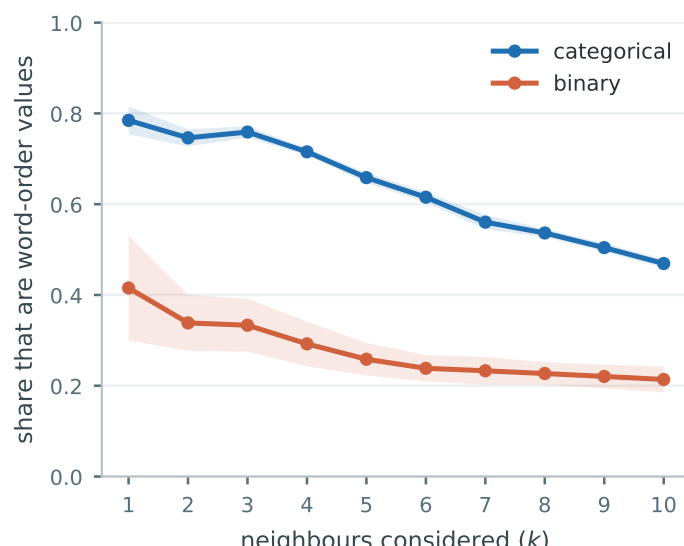


Figure 6: Share of a word-order value’s k nearest cross-feature neighbours that are themselves word-order values, five-seed means with one-standard-deviation bands.

vives control for shared ancestry and geographic proximity, which addresses the two standard confounds: a trait pair may co-occur because the languages showing it descend from one ancestor, or because they are neighbours. A co-evolutionary analysis asks the complementary question of whether the two traits change together along the phylogeny, which speaks to whether one state tends to bring the other about. Of the 191 formulas, 60 pass both, 29 pass the spatiophylogenetic test alone, and 102 pass neither; after our coverage filter the groups hold 55, 27 and 90. We call these fully supported, partially supported and unsupported. These support labels are defined independently of our representations from phylogenetic and spatial analyses of the Grambank data.

Scoring a formula. For a formula $A \rightarrow C$ we resolve antecedent and consequent into Grambank feature-value tokens and compare their directions in the value space. A simple formula, with one value on each side, reduces to a single cosine between the antecedent direction and the consequent direction. Compound formulas keep their logical structure. A conjunctive antecedent, requiring several values at once, is represented by the centroid of its tokens; a purely disjunctive antecedent, satisfied by any of several values, takes the minimum over its triggers, so the least-supported disjunct determines the antecedent score. On the consequent side a disjunction takes a maximum, since satisfying any disjunct satisfies the implication, a conjunction takes a minimum, since all conjuncts

Categorical		Binary	
strongly suffixing	morph	strongly suffixing	morph
69A Tense-aspect suffixes	verb.cat	69A Tense-aspect suffixes	verb.cat
51A Case suffixes	nom.cat	16A Long vowel	phon
33A Plural suffix	nom.cat	43A Related for all demonstratives	nom.cat
99A Nominative - accusative (stand...)	simp.cl	33A Plural suffix	nom.cat
43A Related for all demonstratives	nom.cat	121A Locational	simp.cl
ergative-absolutive	simp.cl	ergative-absolutive	simp.cl
99A erg-abs	simp.cl	15A Left-edge: First or second	phon
108A Oblique patient	simp.cl	74A Affixes on verbs	verb.cat
30A Four	nom.cat	8A No laterals	phon
127A Deranked	comp.sent	144D S[V-Neg]O	word.ord
144H NoDoubleNeg	word.ord	122A Non-reduction	comp.sent
object before verb	word.ord	object before verb	word.ord
95A OV and Postpositions	word.ord	118A Nonverbal encoding	simp.cl
81A SOV	word.ord	97A OV and NAdj	word.ord
85A Postpositions	word.ord	143A [V-Neg]	word.ord
86A Genitive-Noun	word.ord	144S Suffix&NoDoubleNeg	word.ord
144Q No SNegOV	word.ord	143F [V-Neg]	word.ord
complex tone	phon	complex tone	phon
68A From 'finish', 'already'	verb.cat	133A 7-7.5	lex
20A Exclusively isolating	morph	81A VSO	word.ord
25B Zero-marking	morph	26A Weakly prefixing	morph
72A Both types of system	verb.cat	69A Mixed type	verb.cat
7A Implosives only	phon	43A Related by gender markers	nom.cat

■ neighbour from the query's own domain ■ neighbour from a different domain

Figure 7: The five nearest values from other features, for four query values drawn from different WALS areas. Shaded rows are neighbours from the query's own area, with the area abbreviated at the right of each row. Representative seed.

must hold, and a mixed consequent falls back to a centroid. The result is M1, the structured implication cosine reported in §7.1. We compare M1 with a null distribution built from 30,000 random cross-feature token pairs.

Limitations of the probe. Two classes of formula fall outside its scope. Co-existence statements ask whether two strategies can appear in one language, which is not the question of whether their value profiles are distributionally similar. Antecedents defined by absence pool heterogeneous language types, so a token such as *no prefixal object indexing* does not correspond to a coherent direction. We exclude both classes because cosine similarity does not encode co-existence directly and absence-valued antecedents pool heterogeneous language types.

Domain	#	M1	margin
Narrow word order	65	+0.50	+1.03
Hierarchy	25	+0.61	+1.29
Broad word order	64	+0.21	+0.50
Other	18	+0.03	+0.10

Table 4: CAT by universal domain, five-seed means. Above-null rates are 56/65, 23/25, 41/64 and 10/18.

F Corpus variability details

Corpus profiles give, for each dependency relation, the share of tokens whose dependent precedes its head and the binary entropy of that share in bits. We use six core relations (obj, nsubj, case, obl, nmod, amod) and require at least twenty observations of a relation before scoring it. Profiles for 54 languages were computed for this paper from the treebank repositories in the same format as the pre-existing 41; where a language has several treebanks, all of them are pooled. Averaged over the five word-order features, CAT exceeds BIN on direction at $p = 8.6 \times 10^{-6}$ and on variability at $p = 5.1 \times 10^{-5}$ under a paired test across seeds; Appendix H gives intervals over families. URIEL+ coverage varies by feature and its scores are computed on whatever subset it covers. Figure 10 plots predicted uncertainty against measured variability for the 95 languages and for the 84 with a single dominant order.

G Geography of predicted variability

Predicted $P(\text{no dominant order})$ varies systematically across geographic regions among the 1,641 WALS languages with a Glottocode. Nearest-neighbour pairs differ by 0.089 against a permutation null of 0.120 (999 shuffles, $p < 0.001$,

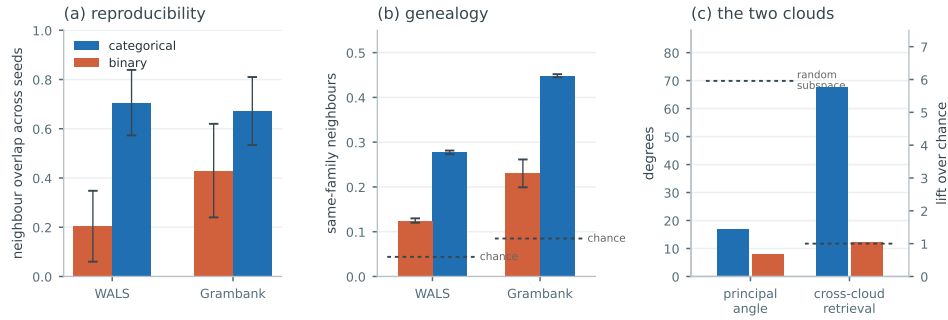


Figure 8: (a) Neighbourhood overlap between seed pairs, with whole-cloud Procrustes disparity inset. (b) Share of a language’s ten nearest neighbours sharing its Glottolog family, against a permutation baseline. (c) Relation between the language and value clouds.

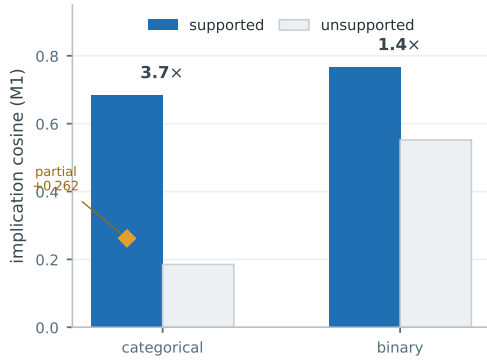


Figure 9: Formulas surviving the phylogenetic and spatial controls against those surviving neither, in each model’s raw cosine. The ratios are not comparable across models, because random cross-feature pairs average +0.33 under BIN and 0.00 under CAT; on a common scale the separation is an area under the curve of 0.81 for CAT against 0.64 for BIN (§7.1).

$n = 900$), and within-family pairs differ by 0.069 against 0.120 across families. Macroarea means are Australia 0.325, North America 0.223, South America 0.157, Papunesia 0.095, Africa 0.089 and Eurasia 0.082. Restricted to languages WALs records with a single dominant order, Australia averages 0.159 against 0.079 elsewhere (Mann–Whitney $p = 9.2 \times 10^{-15}$).

The pattern is not a held-out prediction, since 81A is observed during training for most of these languages and both models reconstruct the discrete distinction essentially perfectly, with areas under the curve of 1.000 and 0.996. These geographic

differences concern graded predictions; the discrete 81A labels themselves do not encode their magnitude.

H Additional robustness checks

Reparameterization. Table 5 gives the cross-feature same-domain rate per seed in three forms. *Centred* is the measurement of Table 3. *Whitened* multiplies the centred vectors by the inverse square root of the value covariance, which gives the same cosines, up to a rotation, under any invertible reparameterization. *Score profiles* represent each value by its scores over all languages, centred over languages, which depend on the scores alone. CAT leads in every seed under all three. BIN’s raw vectors share a dominant direction (mean cosine between values of different features 0.448, against 0.020 for CAT); whitening removes it and nearly doubles BIN’s rate, so the dominant direction accounts for part of BIN’s deficit in Table 3; the remaining vectors still contain some domain structure.

Prediction over branches. Resampling held-out branches (10,000 draws), the CAT–BIN difference on WALs is -0.006 at $\rho=0$ (95% interval -0.017 to $+0.005$), -0.015 at 0.05 (-0.023 to -0.008), $+0.013$ at 0.10 ($+0.006$ to $+0.021$) and $+0.042$ at 0.20 ($+0.034$ to $+0.051$), over 81 branches. On Grambank it is $+0.003$ at $\rho=0$ (-0.009 to $+0.015$), then $+0.006$, $+0.011$ and $+0.017$ at 0.05, 0.10 and 0.20, each interval above zero, over 52 branches.

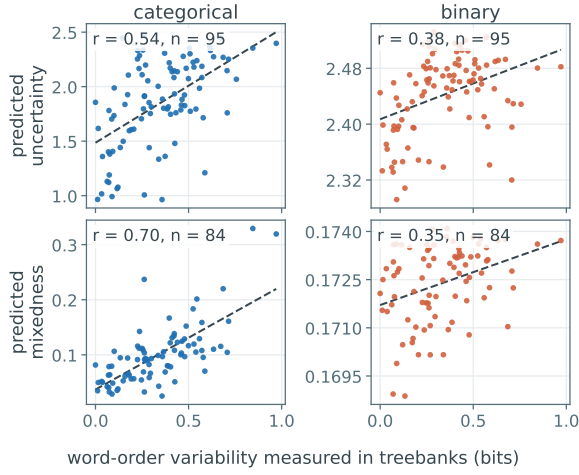


Figure 10: Predicted uncertainty against word-order variability measured in treebanks (mean binary entropy over the six relations with at least 20 observations), seed 42. Top row, the 95 languages, using the entropy of the predicted clause-order distribution. Bottom row, the 84 languages WALS records with a single dominant order, using the predicted probability of *no dominant order*; here the database gives the same verdict for every language, so the label supplies no gradience. Seed 42 is BIN’s strongest on the bottom row: across the five seeds its correlation there runs from 0.05 to 0.35, and CAT’s from 0.69 to 0.74.

Two dominant orders. Retraining without 81B gives 0.612 ± 0.002 for CAT and 0.500 ± 0.006 for BIN, compared with 0.613 and 0.496 with 81B included. For 40 of the 65 languages with both 82A and 83A recorded, those two values alone determine the 81B pair among the retained orders. Withholding 82A and 83A as well gives 0.600 ± 0.003 for CAT and 0.493 ± 0.007 for BIN. The CAT–BIN difference is therefore 0.117 with 81B, 0.112 without 81B, and 0.107 when 82A and 83A are also withheld.

Universals on a common scale. On the 146 formulas both models can evaluate, fully supported formulas outscore unsupported ones with probability 0.809 for CAT (0.726 to 0.883) and 0.636 for BIN (0.536 to 0.733); the difference is +0.17 (+0.08 to +0.27; 5,000 draws over formulas). Measured as the supported-minus-unsupported gap divided by the standard deviation of each model’s null, the separation is 0.78 for CAT and 0.29 for BIN.

Corpus comparison over families. The 95 treebank languages belong to 28 Glottolog families, so we resample families (1,000 draws). A family-level bootstrap matches Table 2 to within 0.02 on the

	42	43	44	45	46	mean
<i>Centred</i>						
CAT	.444	.439	.444	.469	.453	.450
BIN	.213	.161	.189	.218	.207	.198
<i>Whitened</i>						
CAT	.553	.528	.555	.547	.566	.550
BIN	.396	.443	.299	.319	.297	.351
<i>Score profiles</i>						
CAT	.405	.408	.395	.403	.394	.401
BIN	.174	.199	.253	.243	.190	.212

Table 5: Share of a value’s nearest neighbour from another feature that falls in its WALS area, per seed. Chance is .157 over CAT’s 640 values and .162 over BIN’s 609. The centred rows reproduce Table 3 exactly for CAT and within .003 for BIN.

means and 0.05 on single features. The CAT–BIN difference is +0.14 on direction (+0.08 to +0.19), +0.26 on variability (+0.14 to +0.32) and +0.49 on the single-order subset (+0.31 to +0.71).

Feature-level inference. Over the 52 features of Table 1, CAT wins 32, ties 12 and loses 8 (two-sided sign test $p = 1.8 \times 10^{-4}$); by type the sign-test p is .077 for ordinal scales, .039 for presence-to-subtype, .29 for cross-products and .125 for lattices.