

The Stemming Policy Is Data: Adapting a Lexicon-Trained Stemmer to Alternative Conflation Objectives

Leo Galambos

Abstract—A lexicon-trained stemmer can encode not only a transformation mechanism but also a data-defined decision about which lexical forms should be conflated. We test whether that decision changes when the Radixor learning and runtime mechanism is held fixed and only the training partition is changed. Two German CISTEM gold standards provide a controlled nested contrast: GS1 is broader and derivation-friendly, whereas GS2 is lemma-oriented and more conservative. Because the public CISTEM files preserve equivalence classes but omit the CELEX-derived reference strings that created them, the class partition is treated as the policy and three deterministic class-to-target encodings are evaluated as a nuisance-factor sensitivity. On a common unambiguous universe of 298,259 forms, GS2 strictly refines GS1, creating 6,202,030 one-sided policy-disagreement pairs. Leakage is prevented by five-fold cross-fitting over complete join components. The primary Levenshtein-medoid effect is $\Delta_{\text{macro}} = 0.0493$; pair-micro switch coverage is only 0.61%, but 98.46% of observed switches are policy aligned. Shortest-member and lexical-first encodings retain the positive direction. A post-review relation-stratified diagnostic also shows that the GS2-conditioned model joins substantially more pairs that both policies join while joining far fewer pairs that GS2 specifically separates, arguing against a uniform conservatism explanation. The evidence supports *sparse but directionally robust policy-conditioned relational generalization* along this nested German policy axis. The nested intervention does not isolate conservatism from partition-induced representation geometry, and the study does not establish arbitrary semantic-policy transfer, universal linguistic correctness, or recovery of the omitted CELEX target strings.

Index Terms—stemming, information retrieval, lexical conflation, policy transfer, cross-fitting, data-driven morphology, Radixor, CISTEM

I. INTRODUCTION

STEMMING is not only a string transformation. In information retrieval it defines a *conflation policy*: which lexical forms are treated as equivalent for indexing and matching. Classical algorithms embed much of this policy in rules or affix procedures [1], [2]. Data-driven stemming can infer behavior from corpus evidence or supervised feedback [3], [4]. The Egothor/Radixor line is more specifically lexicon-trained: explicit word-to-target mappings are compiled into reusable edit transformations [5], [6]. That flexibility creates both an opportunity and a methodological problem.

The opportunity is configurability. If the training resource changes from a conservative, lemma-oriented grouping to a broader, derivation-friendly grouping, a genuinely policy-adaptive stemmer should change its behavior accordingly

without changing its runtime algorithm. The methodological problem is circularity. When the same lexical resource defines both training targets and evaluation relations, a high score alone cannot show whether the system learned a reusable transformation policy or merely retained resource-specific associations.

Radixor is especially suitable for isolating this question. Its build path transforms lexical word-to-target mappings into patch commands stored in a trie and subsequently reduced into an immutable runtime model [5], [7]. The runtime mechanism is fixed; the lexical mapping is the variable. Previous Radixor evaluations therefore motivate a stronger question than another leaderboard comparison: *does changing only the declared training policy cause a predictable change in conflation behavior on lexical components that were never used to build either model?*

German provides an unusually useful controlled setting. Weissweiler and Fraser deliberately constructed two CISTEM gold standards because stemming has no single universally correct grouping [8]. GS1 derives from CELEX flat morpheme segmentation: morphemes without an assigned word class are stripped, producing a broader, derivation-friendly grouping. GS2 uses CELEX wordform-to-lemma assignments and therefore preserves more lexical distinctions. The distinction is operationally morphological/lemma-based rather than a direct semantic annotation: the CISTEM paper motivates both resources from information-retrieval meaning, but GS2 is not a synonym or word-sense clustering. The resources encode two defensible conflation policies over substantially overlapping lexical material rather than a “correct” policy and a corrupted alternative.

We test the policy-transfer hypothesis by fitting two models under an otherwise invariant Radixor configuration:

$$G_1 \longrightarrow M_{G_1}, \quad G_2 \longrightarrow M_{G_2}, \quad (1)$$

and cross-evaluating both learned relations against both policies. The key evidence is not the in-policy diagonal alone. It is a held-out *interaction*: on lexical pairs where G_1 and G_2 disagree, does M_{G_1} preferentially reproduce the G_1 decision while M_{G_2} preferentially reproduces the G_2 decision?

The contributions are:

- 1) a leakage-safe formulation of policy transfer for lexicon-trained stemming, using the join partition of two reference policies as the cross-fitting unit;

- 2) a primary disagreement-pair estimand that decomposes adaptation into *switch coverage* and *switch direction*, while avoiding pseudo-replication over lexical pairs;
- 3) a proof that two superficially separate directional contrasts are the same statistic on policy-disagreement pairs, preventing double-counting of evidence;
- 4) a three-encoding sensitivity analysis for the canonical-target confound introduced when equivalence classes must be converted to word-to-target training mappings; and
- 5) an executable JDK 21 reproducibility pipeline with commit- and blob-pinned experimental inputs that retains CELEX-derived lexical material only in private build products and publishes aggregate sufficient statistics.

The scope is intentionally narrow. This study does not claim that either CISTEM policy is linguistically superior, that intrinsic conflation agreement implies better retrieval effectiveness, or that Radixor is universally optimal. It asks whether one fixed lexicon-trained stemming mechanism can generalize *different declared conflation objectives*.

II. RELATED WORK AND NOVELTY BOUNDARY

A. Stemming as Conflation Policy

The conflation interpretation of stemming predates modern learned morphology [9], [10]. Paice formalized under-stemming and over-stemming as distinct error dimensions [11], [12], while Hull showed that aggregate retrieval outcomes can conceal qualitatively different stemming behavior [13]. This literature makes a policy distinction unavoidable: aggressive and conservative stemmers can each be appropriate under different application costs.

CISTEM made that distinction explicit for German. Weissweiler and Fraser constructed two gold standards representing different points on the stemming spectrum and evaluated German stemmers against both [8]. GS1 is based on flat CELEX morpheme segmentation after removing morphemes annotated without a word class, whereas GS2 follows CELEX lemma assignments [14]. The authors describe GS1 as favoring more aggressive stemming and GS2 as favoring more conservative stemming. Their examples make the direction concrete: derivationally related forms may remain together in GS1 while being separated into distinct lemma classes in GS2. Their study therefore supplies the paired reference policies required by our controlled experiment.

B. Automatic and Data-Driven Stemmer Construction

Learning or generating a stemmer from data is established prior art. Xu and Croft induced corpus-specific stemming behavior from co-occurrence evidence [3]. Goldsmith, Higgins, and Soglasnova applied an unsupervised morphology learner to language-specific stemming for retrieval [15]. Bacchin, Ferro, and Melucci formulated language-independent probabilistic stemmer generation [16], and Hammarström studied unsupervised recognition of same-stem words [17]. Karanikolas later treated stemmer construction as supervised learning from information-retrieval expert feedback on the output of a

simple primary stemmer [4]. These lines already establish that stemming behavior can be inferred or generated from data; that proposition is not the novelty claimed here.

The closest distinction is experimental rather than architectural. Xu and Croft adapt stemming from corpus co-occurrence evidence, while Karanikolas uses supervised expert feedback to refine a stemmer builder. The present study instead holds one lexicon-trained compiler/runtime mechanism fixed, supplies two alternative reference-defined conflation partitions over a common vocabulary, and asks whether the *difference between those policies* survives on lexical components excluded in their entirety from model fitting. Leakage is blocked by the join of both reference relations, and the outcome is a relational interaction on policy-disagreement pairs rather than merely exact prediction of held-out target strings.

In lemmatization, edit transformations and context-sensitive mappings have likewise long been learned from annotated examples [18]–[20]. The Egothor line represents lexical transformations as edit commands and generalizes them through trie structure [5], [6]. Again, the contribution here is not learning an edit transformation. It is a paired policy intervention under an invariant mechanism, evaluated with component-safe cross-fitting and an estimand that asks whether policy-conditioned relational differences generalize outside the fitted components.

C. Why Cross-Evaluation Is Necessary

Training on G_1 and scoring only on G_1 cannot distinguish policy adaptation from policy memorization. Even a standard train/test split by individual word is unsafe. If two forms from one morphological family are divided between train and test, the learned trie may receive direct evidence from the same family that later contributes the test pair. A second policy makes the problem stricter: two forms separate in G_1 may be connected in G_2 . Leakage control must therefore respect *both* policies simultaneously.

The experiment below uses the join partition $G_1 \vee G_2$, the coarsest equivalence relation generated by either policy. Holding out a block of this partition removes every reference path connecting its test forms to training forms under either G_1 or G_2 . This formulation is valid even when two policies cross-cut. The exact CISTEM preflight later reveals a simpler special case: on the controlled universe, GS2 is a refinement of GS1 and therefore $G_1 \vee G_2 = G_1$. Retaining the join formulation makes the leakage argument explicit rather than assuming this hierarchy in advance.

III. FORMALIZING POLICY TRANSFER

A. Policies as Partitions

Let W be a finite set of normalized lexical forms. A conflation policy G is represented as a partition of W . For an unordered pair $p = \{u, v\}$, $u \neq v$, define

$$g(p) = \mathbf{1}[u \sim_G v]. \quad (2)$$

A deterministic stemmer model M induces a partition by output equality:

$$m(p) = \mathbf{1}[M(u) = M(v)]. \quad (3)$$

We consider two reference policies G_1, G_2 on the same W , and two models M_{G_1}, M_{G_2} fitted by the same learning mechanism.

The most informative set is the policy-disagreement relation

$$\mathcal{D} = \{p : g_1(p) \neq g_2(p)\}. \quad (4)$$

Pairs outside $\mathcal{D}$ do not identify which policy a learned model follows because both references request the same binary decision.

B. Aligned and Reverse Switches

For $p \in \mathcal{D}$, define an *aligned switch*

$$a(p) = \mathbf{1}[m_1(p) = g_1(p) \wedge m_2(p) = g_2(p)], \quad (5)$$

and a *reverse switch*

$$r(p) = \mathbf{1}[m_1(p) = g_2(p) \wedge m_2(p) = g_1(p)]. \quad (6)$$

Because $g_1(p) \neq g_2(p)$, $a(p) + r(p) = 1$ exactly when the two models make different pair decisions. If both models join the pair or both keep it separate, neither indicator is active.

Let C be one leakage-safe lexical component and $\mathcal{D}_C$ its disagreement pairs. We define

$$d_C = \frac{\sum_{p \in \mathcal{D}_C} (a(p) - r(p))}{|\mathcal{D}_C|}. \quad (7)$$

Then $d_C = 1$ means every disagreement pair is an aligned switch, $d_C = -1$ means every disagreement pair is a reverse switch, and $d_C = 0$ means no net directional adaptation.

Two diagnostics separate magnitude from direction:

$$s_C = \frac{\sum (a + r)}{|\mathcal{D}_C|}, \quad \text{switch coverage}, \quad (8)$$

$$q_C = \frac{\sum a}{\sum (a + r)}, \quad \text{aligned share among switches}. \quad (9)$$

Whenever at least one switch occurs,

$$d_C = s_C(2q_C - 1). \quad (10)$$

Thus a positive but tiny d_C cannot be mistaken for substantial policy control: it may reflect either very few model differences or many differences with weak directional purity.

C. One Interaction, Not Two Independent Effects

A tempting analysis reports the two contrasts

$$\delta_1 = A(M_{G_1}, G_1) - A(M_{G_2}, G_1), \quad (11)$$

$$\delta_2 = A(M_{G_2}, G_2) - A(M_{G_1}, G_2), \quad (12)$$

where A is agreement restricted to $\mathcal{D}$. These expressions look like two directional confirmations. They are not independent.

Proposition 1. *On any fixed set of policy-disagreement pairs, $\delta_1 = \delta_2$ exactly, and both equal the mean aligned-minus-reverse switch score.*

Proof. For every $p \in \mathcal{D}$, $g_2(p) = 1 - g_1(p)$. If $m_1(p) = m_2(p)$, both directional agreement differences are zero. If $m_1(p) \neq m_2(p)$, exactly one model decision equals $g_1(p)$ and the other equals $g_2(p)$. An aligned switch contributes +1 to both contrasts; a reverse switch contributes -1 to both.

Summing and dividing by the same number of disagreement pairs proves the identity. $\square$

The experiment therefore has one primary policy-alignment interaction, not two corroborating tests. The equality is retained in the verification code as an implementation invariant.

Corollary 1 (Nested one-sided policy contrast). *If G_2 strictly refines G_1 on W , then for every $p \in \mathcal{D}$, $g_1(p) = 1$ and $g_2(p) = 0$, so*

$$a(p) - r(p) = m_1(p) - m_2(p). \quad (13)$$

Consequently, for every disagreement-bearing component,

$$d_C = \frac{1}{|\mathcal{D}_C|} \sum_{p \in \mathcal{D}_C} (m_1(p) - m_2(p)), \quad (14)$$

that is, d_C is exactly the difference between the two models' join rates on the one-sided disagreement relation.

This corollary applies to the exact CISTEM resources used below. It makes the primary statistic transparent in the nested case: the primary effect identifies a policy-conditioned join-rate contrast on the relation that GS1 joins and GS2 separates. Evidence against a merely global change in join frequency requires a separate relation-stratified diagnostic.

D. Primary Estimand

Let $\mathcal{C}_D$ be the set of join components containing at least one disagreement pair. The primary statistic is the unweighted component-macro effect

$$\Delta_{\text{macro}} = \frac{1}{|\mathcal{C}_D|} \sum_{C \in \mathcal{C}_D} d_C. \quad (15)$$

Equal component weighting prevents a large family from dominating merely because a component with n forms can contribute $O(n^2)$ pairs. A pair-micro analogue

$$\Delta_{\text{micro}} = \frac{\sum_C \sum_{p \in \mathcal{D}_C} (a(p) - r(p))}{\sum_C |\mathcal{D}_C|} \quad (16)$$

is reported only as a secondary estimand.

IV. DATA AND CONTROLLED POLICY CONSTRUCTION

A. CISTEM Policy Construction and Public Representation

The source resources are the public CISTEM gold standards at repository commit `7c19867c2e062c8a7d44b394c19573845ac4bd89` [21]. The exact Git blob identities are:

GS1: `8627bb28b67429f6488f8d017f510327b2c84d1c`

GS2: `2cb401638a67760f5fec47c8379646bf6d6d1b8e`.

The CISTEM paper reports that its construction produced 30,951 GS1 stems and 47,852 GS2 stems and that 1,000 were sampled from each for the published evaluation [8]. The exact public files pinned above, however, contain 30,746 and 50,362 non-empty cluster lines, respectively. Repository history shows that these exact two blobs first appear publicly at commit `6f9369c3ab18a384e19dc6bb5c75b4caabc61fe3` on 2017-11-14 and remain unchanged through the pinned CISTEM revision; the preceding public commit does not contain the gold-standard files [21]. The discrepancy therefore predates Radixor and is

not later repository drift. Its cause is not inferable from the public evidence, so the paper-reported construction counts are kept distinct from the public-file cluster-line counts. The byte sizes and Git blob hashes define the resources used here.

The distinction between *how CISTEM created a class* and *what the public file retains* is essential for Radixor. In the original procedure, GS1 was formed by stripping CELEX morphemes annotated with no word class, while GS2 used the CELEX lemma assigned to each wordform [8], [14]. Those intermediate stem/lemma labels were then used to group wordforms. The public CISTEM repository intentionally releases the groups without a reference stem for each cluster [21]. Thus the observable gold object used in this experiment is a *partition*: it states which forms belong together, but it does not prescribe the output string a trained stemmer must emit.

Each non-empty source line is parsed as one reference class. Tokens are Unicode NFC-normalized and lowercased without stripping German diacritics or mapping ß to ss. Duplicate normalized forms inside a class are removed. A form appearing in multiple classes within the same policy is ambiguous for a context-free partition and is excluded from the primary experiment.

B. Common Universe and Observed Policy Hierarchy

After within-policy ambiguity removal, the experiment retains only forms occurring unambiguously in *both* resources. This creates one controlled universe W and removes vocabulary-size as an explanation for a policy difference. The exact run yields $|W| = 298,259$, with 30,215 restricted G_1 classes and 49,352 restricted G_2 classes. Relative to the 304,215 normalized forms shared by the two raw public files, the ambiguity restriction removes 5,956 forms (1.96%); the retained universe therefore covers 98.04% of the raw shared vocabulary.

The resulting policies do not cross-cut. Every restricted GS2 class is contained in exactly one restricted GS1 class; hence GS2 is a strict refinement of GS1 on W . This direction is consistent with the construction described by Weissweiler and Fraser: lemma identity in GS2 preserves distinctions that the more aggressive GS1 can collapse. Importantly, “refinement” does not mean a negligible difference. GS2 splits 6,165 of the 30,215 GS1 classes (20.40%). Among those split classes, the mean number of GS2 subclasses is 4.10, the median is 2, and the maximum is 74. GS1 contains 7,687,650 within-class pairs on W ; GS2 separates 6,202,030 of them (80.68%). Consequently all 6,202,030 policy-disagreement pairs have the same orientation: GS1 joins the pair and GS2 separates it.

A separate descriptive audit on the 304,215 normalized forms shared by the two raw public files *before* within-policy ambiguity exclusion likewise finds no GS2-co-clustered pair that lacks a GS1 co-clustering. The observed hierarchy is therefore a property of the pinned public resources, not an artifact introduced by the primary ambiguity filter.

This hierarchy narrows the interpretation. The experiment identifies adaptation along a substantial but one-directional aggressive-to-conservative policy axis. It does not by itself establish transfer between arbitrary cross-cutting conflation policies.

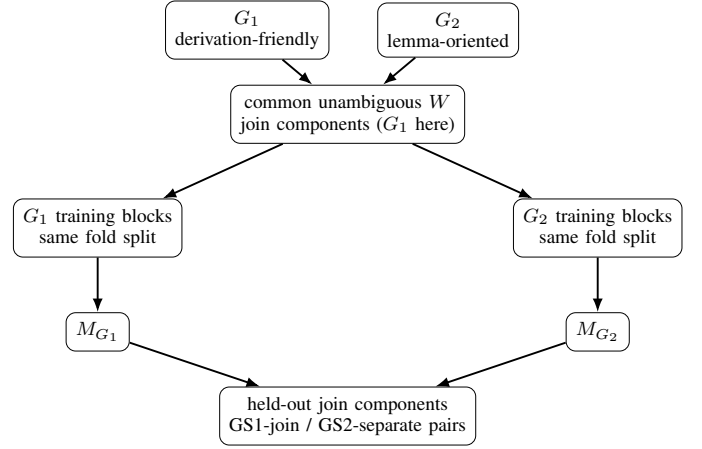


Fig. 1. Controlled policy-transfer design. The mechanism, normalized universe, folds, compiler, reduction, and runtime are fixed. The public CISTEM partitions differ, and deterministic canonicalization converts each partition to Radixor word-to-target mappings.

C. The Join Partition as Leakage Boundary

Define an undirected graph on W with an edge between two forms whenever either G_1 or G_2 places them in the same class. Its connected components are the blocks of the partition join $G_1 \vee G_2$. No join component may be split between training and test data.

Because GS2 refines GS1 on the exact controlled universe, the join is GS1 itself and contains 30,215 components. 6,165 components contain policy disagreement. The largest component has 584 forms, represents only 0.0020 of W , and contributes 0.0268 of disagreement evidence. Thus the controlled contrast has distributed disagreement rather than one giant component that would invalidate five-fold blocking.

V. EXPERIMENTAL DESIGN

A. Deterministic Five-Fold Component Cross-Fitting

The join components are assigned to five folds deterministically. Components containing disagreement evidence are sorted by decreasing disagreement-pair count and then by size; a stable component identifier provides the final tie break. Greedy placement first minimizes accumulated disagreement evidence, then forms, then component count. Components with no policy disagreement are subsequently balanced by forms. Model predictions never influence fold assignment.

For fold f , every component assigned to f is removed from *both* policy training dictionaries. The remaining four folds produce $M_{G_1}^{(f)}$ and $M_{G_2}^{(f)}$. Both are evaluated only on fold f . Thus no single fitted model sees the complete GS1 or GS2 partition: each lexical component participates in training for four fold models but, for its one reported out-of-fold prediction, *every form in that component is absent from fitting*. Across all five folds the complete controlled universe is evaluated exactly once out of fold. Concatenating the held-out predictions gives one out-of-fold prediction per form for each policy-conditioned model, but relational metrics that require arbitrary pair comparisons are computed within folds unless the pair is guaranteed to reside in one join component.

This distinction matters. Every $p \in \mathcal{D}$ belongs to one join component, so the primary disagreement-pair statistic is well defined across all components. In contrast, a predicted conflation between forms from two different folds would compare outputs from two different fitted models; such cross-fold pairs are excluded from secondary whole-relation F1.

B. Converting Partition-Only Supervision to Training Targets

The released CISTEM files do not contain the CELEX-derived reference stem or lemma used during gold-standard construction. They provide only cluster membership [21]. Radixor, in contrast, learns from word-to-target mappings: each dictionary class must have a concrete canonical first column. The experiment must therefore introduce a target string that is *not* part of the public gold relation.

This does not make the training policy undefined. For each class, the canonical target is selected from that class itself. Because the restricted classes are disjoint, assigning one member as the target encodes exactly the same in-sample equivalence relation regardless of which member is chosen: all members of one class receive one target and no two classes share that target. What changes is the *edit geometry* presented to Radixor—the deletion/substitution/insertion commands from surface forms to that target—which can influence how the learned trie generalizes to unseen components.

We therefore treat canonicalization as an explicit nuisance variable and predeclare three deterministic encodings:

- 1) *Levenshtein medoid* (primary): the class member minimizing total exact Levenshtein distance to other members, with shorter length and then lexical order as tie breakers;
- 2) *shortest member*: shortest normalized class member, lexical order on ties;
- 3) *lexical first*: lexicographically first normalized class member.

The sensitivity is substantive rather than nominal. All three rules select the same canonical member for only 12,362 of 30,215 GS1 classes (40.91%) and 22,734 of 49,352 GS2 classes (46.07%). A policy-transfer conclusion is therefore required to retain its direction across materially different target assignments.

Representatives are derived only from training classes. Because whole join components are held out, no test form can influence a training representative. The experiment tests *relational* generalization by output equality; it neither reconstructs nor claims to recover the omitted original CELEX stem/lemma string.

C. Frozen Radixor Mechanism

All models use Radixor commit `0c3b13f` and the exact distribution artifact whose archive SHA-256 is `5d448f9de48aba77f4df1957dd6f4ee79779c10f6685224107fb88f0dd98b6cd`. The binary-model compiler is invoked with:

- backward trie traversal;
- `LOWERCASE_WITH_LOCALE_ROOT` case processing;
- diacritic processing `AS_IS`;
- `-store-original`; and

- the ranked `getAll()`-equivalence reduction. The exact enum value is frozen in the experiment specification.

The two policy paths differ only in their training classes and resulting canonical mappings. There is no runtime policy flag. Persisted models are loaded through the binary compiled-trie loader, queried through the preferred normalized-string path, and executed with the stored compiled patch command.

D. Descriptive Anchors, Not Causal Controls

The frozen Radixor snapshot contains a separate full-model German CISTEM comparison. The registered `de-de-default` Radixor model is trained from German UniMorph data rather than from GS1 or GS2; its model-data preparation groups inflected forms by lemma [7]. Its objective is therefore structurally closer to the lemma-oriented GS2 relation than to the broader derivation-friendly GS1 relation. Under the snapshot’s aggregate cluster TP/FP/FN scorer on the complete public CISTEM resources, registered Radixor obtained F1 values 0.718140 on GS1 and 0.891616 on GS2, while CISTEM obtained 0.718632 and 0.842208. The systems are essentially tied on GS1 and Radixor is higher on GS2 in that independent full-model context, but the lemma-oriented training source prevents interpreting the GS2 advantage as policy-neutral or universal Radixor superiority.

The primary policy-interaction experiment therefore compares only the two cross-fitted Radixor policy models. To separate policy controllability from reference-recovery context, the pinned official CISTEM implementation is additionally evaluated as a fixed-rule secondary reference. Two scopes are retained deliberately. The *fold-restricted* scope scores CISTEM inside the same five held-out subsets used for Radixor and excludes cross-fold relations, matching the relation scope available to cross-fit Radixor. Because CISTEM is one fixed model, we also report its *full restricted-universe* score over all 298,259 forms at once, which includes cross-fold false-positive collisions. Neither CISTEM scope contributes to the primary interaction statistic.

A second post-execution secondary analysis quantifies the information-removal gap within Radixor itself. For each policy and each target encoding, one *full-information reconstruction model* is trained on all restricted classes and evaluated on the same common universe. Every evaluation form is therefore present in that model’s training dictionary. These six models are deliberately in-sample and are not evidence of generalization; they provide a reconstruction ceiling against which the component-blocked cross-fit scores can be compared using exactly the same partition metrics.

VI. STATISTICAL ANALYSIS

A. Experimental Unit and Dependence

Lexical pairs are not independent observations. Pairs share words, and one component of n forms can generate quadratically many pairs. Treating each pair as an IID sample would therefore produce artificially narrow uncertainty estimates. The primary unit is the join component, and Eq. (15) weights each disagreement-bearing component equally.

The dataset is also a fixed lexical resource rather than a random sample from a well-defined population. We therefore avoid presenting a conventional population-significance p -value. The main result is a deterministic finite-resource effect. A fixed-seed, 10,000-replicate component bootstrap [22] is reported as a *resampling stability interval*, not as a population confidence interval. Its purpose is to reveal whether the macro conclusion depends strongly on which components receive repeated weight under resampling.

Additional concentration diagnostics are:

- the five held-out fold-level macro effects;
- the leave-one-component-out range of Δ_{macro} ;
- the pair-micro effect Δ_{micro} ; and
- the largest component’s share of disagreement evidence.

No pair-level significance test is used for the primary claim.

B. Secondary Partition Metrics

For each held-out fold we also measure pairwise TP, FP, and FN for every model-policy cell. Pooled within-fold counts give descriptive precision, recall, and F1. These metrics characterize overall held-out relation recovery, including pairs on which the policies agree, but they do not replace the policy-interaction statistic.

A second secondary metric adapts the class-matching logic of the original CISTEM evaluation to the controlled fold-restricted universe: for each held-out gold class, the most frequent predicted stem is taken as its main output; per-class precision, recall, and F1 are calculated against other held-out classes in that fold and then macro-averaged by class [8]. We refer to this explicitly as CISTEM-style macro F1 rather than an exact reproduction of the historical full-corpus procedure. It is reported separately from pairwise F1 to prevent metric conflation.

The fixed-rule CISTEM reference uses the official Python implementation pinned by Git blob identity. The primary experiment normalizes the controlled universe to NFC lowercase, and CISTEM is intentionally given those same forms. This comparability choice suppresses CISTEM’s capitalization-sensitive branch; it is not intended to represent every ordinary CISTEM operating condition. For fold-restricted CISTEM scores, the existing five folds define relation scopes identical to those used for cross-fit Radixor. A separate full-restricted-universe score evaluates the same fixed CISTEM outputs jointly over all controlled forms. The baseline was added after completion of the predeclared Radixor intervention and does not enter the primary estimand, fold construction, or falsification criteria.

VII. PREDECLARED INTERPRETATION AND FALSIFICATION

The frozen experiment specification separated policy divergence, behavioral differentiation, directional adaptation, and robustness checks. One aspect requires explicit post-execution discipline: “non-trivial switch coverage” was stated qualitatively but no numerical decision threshold was frozen. It would be inappropriate to invent such a threshold after observing the data. Coverage is therefore reported as a magnitude diagnostic rather

than retrospectively classified as a passed or failed hypothesis gate.

Policy divergence requires non-zero $\mathcal{D}$ distributed over enough join components for blocked evaluation. Without this, the experiment cannot identify adaptation.

Behavioral differentiation is quantified by switch coverage and affected-component prevalence. If M_{G_1} and M_{G_2} make almost identical held-out decisions, different model files or command counts alone are not evidence of broad policy control. The observed coverage determines the magnitude language used in the conclusion.

Directional adaptation requires $\Delta_{\text{macro}} > 0$ in the primary medoid encoding. A zero effect means that policy-aligned and reverse evidence cancel; a negative effect falsifies the intended transfer direction.

Directional robustness requires the sign not to reverse under shortest-member or lexical-first encoding and not to be attributable to one giant component or one pathological fold. A resampling stability interval crossing zero would be reported as fragility rather than hidden by a micro-weighted alternative. Size-stratified component diagnostics are additionally reported post review to expose the weighting behavior behind the macro/micro difference; they are sensitivity evidence, not a redefinition of the primary estimand.

No fold definition, representative rule, normalization setting, compiler parameter, or primary metric may be changed after observing real GS1/GS2 model predictions to rescue the hypothesis. If the preflight join structure is infeasible, the correct outcome is a design limitation, not a post-hoc split of connected components. The final claim therefore distinguishes *directional robustness* from effect breadth: stable direction under sensitivity analyses does not imply that a large fraction of all disagreement pairs switches.

VIII. RESULTS

A. Exact-Input Execution and Audit

The frozen experiment completed on the exact pinned CISTEM byte streams; completed is therefore the empirical status represented by this revision. No numerical result is imported from the historical CISTEM evaluation or from the synthetic smoke test. A clean replay from the exact GS1/GS2 blobs and frozen Radixor JAR regenerated all 30 persisted models (three canonicalizations, five folds, two policies) with matching model-file SHA-256 digests and reproduced the public aggregate result tables. CELEX-derived lexical strings remain outside the publication bundle.

B. Policy and Component Preflight

Table I confirms both policy divergence and a feasible leakage boundary. The one-sided disagreement is not weak: 20.40% of GS1 classes are refined by GS2, and the affected refinements separate 80.68% of all within-GS1 pairs. At the same time, the largest component contributes only 0.0268 of disagreement evidence, so the primary component-macro effect is not structurally forced by one family.

TABLE I
EXACT-INPUT POLICY PREFLIGHT ON THE COMMON UNAMBIGUOUS
UNIVERSE.

Quantity	Value
Common unambiguous forms	298,259
Restricted GS1 classes	30,215
Restricted GS2 classes	49,352
GS1 classes split by GS2	6,165
Join components	30,215
Disagreement-bearing components	6,165
GS1-same / GS2-different pairs	6,202,030
GS2-same / GS1-different pairs	0
Total disagreement pairs	6,202,030
Largest component (forms)	584
Largest disagreement share	0.0268

C. Primary Policy-Aligned Switching

The primary medoid encoding yields 37,002 aligned switches and only 579 reverse switches. Thus, conditional on the two learned models making different pair decisions, 0.9846 of changes follow the requested policy direction. The pair-micro switch coverage is nevertheless only 0.0061 (0.61%). The correct interpretation is therefore not broad policy replacement but a sparse, highly directional response.

The component view gives complementary scale. Under the medoid encoding, at least one model switch occurs in 1,287 of 6,165 disagreement-bearing components (20.88%); 95.03% of those affected components have a positive net directional effect. All five held-out medoid fold effects are positive (0.0450–0.0551), and the leave-one-component-out range is [0.0491, 0.0494]. The fixed-seed component-resampling stability interval [0.0447, 0.0541] remains away from zero.

Canonicalization changes the magnitude but not the conclusion. Shortest-member and lexical-first encodings give $\Delta_{\text{macro}} = 0.0720$ and 0.0599 , respectively, with aligned-switch shares above 0.98. Their affected-component shares are 24.93% and 20.05%. Because the three target rules choose identical targets for fewer than half of the classes, this sign stability is evidence that the relational result is not an artifact of one arbitrary replacement string.

D. Post-Review Relation-Stratified Specificity Under the Nested Design

The refinement structure creates an immediate alternative explanation: perhaps M_{G_2} merely becomes more conservative everywhere. This post-review secondary diagnostic tests that explanation using the already frozen medoid predictions; it does not redefine the primary estimand. The observed predictions reject that simple account. Table III partitions held-out relation evidence into pairs that both references join, pairs on which GS1 joins but GS2 separates, and false-positive joins against the GS1 negative relation.

The GS2-conditioned model is therefore not uniformly more conservative. On the 1,485,620 pairs that *both* policies join, M_{G_2} joins 821,599 (55.30%), substantially more than the 537,501 (36.18%) joined by M_{G_1} . Yet on the 6,202,030 pairs where GS2 specifically requests separation, its join rate falls

from 0.61% under M_{G_1} to 0.03% under M_{G_2} . The GS1-negative false-positive join count also falls from 961 to 46. This interaction is inconsistent with explaining the result as a single global “stem less” shift: the GS2-conditioned model recovers more shared positive structure while selectively suppressing joins on the relation that distinguishes GS2 from GS1.

E. Post-Review Component-Size Sensitivity

Equal component weighting prevents one large lexical family from dominating the primary statistic, but the large difference between Δ_{macro} and Δ_{micro} also requires checking the opposite risk: excessive influence from small components. This post-review sensitivity analysis stratifies already frozen component evidence and does not alter the primary weighting rule. Table IV stratifies the primary medoid result by disagreement-pair mass.

Every stratum remains positive, including the >1,000-pair components that contain 88.48% of all disagreement pairs. The macro result is therefore not a singleton or tiny-family artifact. The effect does decay as component pair mass grows, which explains much of the macro–micro contrast and further supports describing the adaptation as sparse rather than broad.

F. Explicit Model-by-Gold Quality Context

Table V makes the model identities explicit. The two 1.0000 diagonal values come from *different* full-information policy-conditioned models: $M_{G_1}^{\text{full}}$ exactly reconstructs GS1, while $M_{G_2}^{\text{full}}$ exactly reconstructs GS2. Their cross-policy cells are not 1.0. The exact diagonal is therefore a real training-set partition-reconstruction ceiling, not a claim that one model is universally perfect or that the selected output strings are linguistically canonical.

The held-out rows reproduce the earlier cross-evaluation without hiding the off-diagonal cells. There is no diagonal dominance: M_{G_2} scores better than M_{G_1} even against GS1 under both whole-partition summaries. That does not contradict the primary interaction. Whole-partition recovery mixes the disagreement relation with the much larger shared structure; Table III shows directly that M_{G_2} recovers more of the shared-positive relation while separating far more of the pairs that distinguish GS2 from GS1.

The two CISTEM rows expose relation scope rather than two versions of the algorithm. Fold-restricted macro F1 is 0.9213 on GS1 and 0.9329 on GS2. When the same fixed predictions are scored jointly over the complete restricted universe, cross-fold false-positive collisions are included and macro F1 becomes 0.9016 and 0.9120. The fold-restricted row is the appropriate structural reference beside cross-fit Radixor; the full-universe row is CISTEM’s unrestricted score on the controlled 298,259-form universe. Neither is a leakage-equivalent fitting control because CISTEM is not refitted per fold.

Table VI supplies scale for the component-blocked scores. GS1 loses more partition recoverability than GS2 under every deterministic target construction. Because the full-information rows are in-sample and the held-out rows exclude every form in the evaluated component from fitting, the difference is a *component-held-out generalization gap*. It must not be interpreted as the degradation expected when a normal production stemmer encounters an arbitrary unseen token.

TABLE II

POLICY-TRANSFER EFFECT AND CANONICAL-TARGET SENSITIVITY. THE MEDOID ROW IS PRIMARY; THE OTHER ROWS ARE PREDECLARED SENSITIVITY ANALYSES. “STABILITY” IS A DESCRIPTIVE COMPONENT-RESAMPLING INTERVAL, NOT A POPULATION CONFIDENCE INTERVAL.

Representative	Δ_{macro}	Δ_{micro}	Switch coverage	Aligned share	Stability 95%	LOO range
Levenshtein medoid	0.0493	0.0059	0.0061	0.9846	[0.0447, 0.0541]	[0.0491, 0.0494]
Shortest member	0.0720	0.0085	0.0088	0.9853	[0.0665, 0.0777]	[0.0718, 0.0722]
Lexical first	0.0599	0.0053	0.0054	0.9826	[0.0547, 0.0653]	[0.0598, 0.0601]

TABLE III

MEDOID RELATION-SPECIFICITY DIAGNOSTIC. SHARED-POSITIVE AND DISAGREEMENT ROWS ARE REFERENCE-DEFINED RELATION STRATA; THE FINAL ROW REPORTS WITHIN-FOLD GS1 FALSE-POSITIVE JOINS.

Relation	Ref. pairs	M_{G_1} joins	M_{G_2} joins
Shared positive	1,485,620	537,501 (36.18%)	821,599 (55.30%)
GS1 join / GS2 split	6,202,030	38,047 (0.61%)	1,624 (0.03%)
GS1-negative FP	—	961	46

TABLE IV

MEDOID DIRECTIONAL EFFECT BY DISAGREEMENT-PAIR MASS PER COMPONENT. PAIR SHARE IS THE FRACTION OF ALL POLICY-DISAGREEMENT PAIRS IN THE STRATUM.

Pairs/component	Components	Pair share	Macro Δ	Micro Δ
1–10	822	0.07%	0.0648	0.0660
11–100	2,892	1.66%	0.0716	0.0610
101–1,000	1,691	9.79%	0.0223	0.0218
>1,000	760	88.48%	0.0075	0.0030

G. Independent Full-Model Quality Context

The frozen external benchmark answers a different practical question. Registered de-de-default Radixor and CISTEM are effectively tied on GS1 (0.718140 versus 0.718632), while registered Radixor is higher on GS2 (0.891616 versus 0.842208). The Radixor model is derived from German UniMorph with forms grouped by lemma, so its objective is naturally closer to GS2. The GS2 advantage is therefore useful full-model context but not evidence of policy-neutral or universal Radixor superiority.

Taken together, Table V, Table VI, and Table VII distinguish three questions: exact reconstruction with complete declared policy information, extrapolation to components absent from fitting, and ordinary quality of an independently trained maintained model. The policy-transfer claim concerns the second question and the interaction between its two policy-conditioned models.

H. Structural Consequences and Model Identity

The policy-bearing training representation also changes model structure. In the primary medoid encoding the mean compiled sizes are 1164250.6 bytes for M_{G_1} and 362320.0 bytes for M_{G_2} , with 126914.6 and 40495.4 mean trie nodes, respectively. Thus the GS2-conditioned models are about 68.88% smaller in persisted compiled size and 68.09% smaller in trie-node count on these fold means. These are descriptive structural consequences of the changed policy-bearing training representation; without a dedicated runtime experiment they imply neither lower latency nor lower end-to-end memory cost.

Persisted model-file SHA-256 is the authoritative reproducibility identifier. A clean replay matched all 30 reported model SHA-256 digests. The frozen Radixor commit also exposes a runtime trie “fingerprint”, but that diagnostic is process-dependent for compiled patch-command values in this snapshot and is therefore not used as a content identity in this study.

IX. DISCUSSION

A. What the Experiment Establishes

The strongest supported statement is directional, not comprehensive. Changing the policy-bearing training partition while holding the mechanism and canonicalization rule fixed changes held-out conflation decisions in the requested direction: the primary component-macro effect is positive in every fold, its component-resampling stability interval remains positive, the leave-one-component-out range is narrow, and the sign survives two materially different canonical-target encodings. These are the senses in which the result is *directionally robust*.

Effect breadth is much smaller. Pair-micro switch coverage is below one percent for all three encodings, so the learned models do not wholesale rewrite the disagreement relation. Roughly one fifth to one quarter of disagreement-bearing components exhibit at least one switch, and about 95% of affected components move net in the intended direction. The post-review size-stratified analysis further shows a positive effect in every component-mass band, including the > 1,000-pair band that contains 88.48% of all disagreement pairs, but with sharply decreasing magnitude as pair mass grows. The pair-micro effect falls from 0.0660 in the 1–10-pair band to 0.0030 in the highest-mass band, indicating that policy-conditioned adaptation is strongest in small-to-medium disagreement components and much weaker in the largest relation masses. The evidence is therefore best summarized as *sparse but directionally robust policy-conditioned relational generalization*.

B. Specificity Beyond a Global Aggressiveness Shift

Because GS2 strictly refines GS1, every disagreement pair requests the same local change: split a relation that GS1 joins. That nested design cannot demonstrate arbitrary cross-cutting policy control. It also creates an obvious scalar alternative—that GS2 training simply makes the stemmer more conservative everywhere.

The post-review relation-stratified result argues against that alternative. The GS2-conditioned model joins substantially *more* shared-positive pairs than the GS1-conditioned model, while joining dramatically fewer policy-disagreement pairs. It also produces fewer false-positive joins against GS1 negatives.

TABLE V

MEDOID MODEL-BY-GOLD QUALITY MATRIX. FULL-INFORMATION ROWS ARE IN-SAMPLE RECONSTRUCTION MODELS TRAINED ON ALL RESTRICTED CLASSES. HELD-OUT ROWS ARE THE FIVE-FOLD COMPONENT-BLOCKED MODELS. CISTEM FOLD-RESTRICTED SCORES OMIT CROSS-FOLD RELATIONS TO MATCH THE RELATION SCOPE AVAILABLE TO CROSS-FIT RADIXOR; THE FULL-UNIVERSE CISTEM ROW EVALUATES THE ONE FIXED MODEL JOINTLY OVER ALL CONTROLLED FORMS. BOLD VALUES MARK POLICY-MATCHED RADIXOR CELLS, NOT A STATISTICAL SIGNIFICANCE CLAIM.

Regime	Model	GS1 macro F1	GS2 macro F1	GS1 pair F1	GS2 pair F1
Full information	$M_{G_1}^{\text{full}}$	1.0000	0.6598	1.0000	0.3239
Full information	$M_{G_2}^{\text{full}}$	0.9411	1.0000	0.3239	1.0000
Component held out	M_{G_1}	0.6962	0.6933	0.1393	0.5213
Component held out	M_{G_2}	0.7985	0.8422	0.1935	0.7117
CISTEM fold scope	CISTEM	0.9213	0.9329	0.2943	0.8836
CISTEM full universe	CISTEM	0.9016	0.9120	0.2915	0.8575

TABLE VI

TARGET-ENCODING SENSITIVITY OF POLICY-MATCHED COMPONENT-HELD-OUT MACRO F1. THE CORRESPONDING FULL-INFORMATION RECONSTRUCTION IS 1.0000 FOR EVERY ENCODING; THE FINAL COLUMNS ARE FULL-INFORMATION-MINUS-HELD-OUT GAPS.

Encoding	GS1 held out	GS2 held out	GS1 gap	GS2 gap
Medoid	0.6962	0.8422	0.3038	0.1578
Shortest	0.7229	0.8777	0.2771	0.1223
Lexical	0.7059	0.8557	0.2941	0.1443

TABLE VII

SEPARATE FROZEN FULL-MODEL BENCHMARK ON THE COMPLETE PUBLIC CISTEM RESOURCES. VALUES ARE THE SNAPSHOT’S AGGREGATE CLUSTER TP/FP/FN F1, NOT THE CISTEM-STYLE MACRO F1 OR PAIRWISE F1 IN TABLE V. THE REGISTERED GERMAN RADIXOR MODEL IS UNIMORPH-TRAINED AND GROUPED BY LEMMA.

Gold	Registered Radixor	CISTEM	Δ Radixor-CISTEM
GS1	0.718140	0.718632	-0.000492
GS2	0.891616	0.842208	+0.049408

The observed behavioral change is therefore selective: M_{G_2} improves recovery of relations that both policies retain while suppressing joins concentrated on the relation that GS2 specifically removes. This specificity is stronger evidence for a policy-conditioned response than a global change in join frequency would be.

C. Information Removal Versus Full-Model Quality

The cross-fit and full-information analyses make the knowledge boundary explicit. A policy-matched model with all restricted classes available reconstructs that supplied partition exactly, while a cross-fit model must infer the same relations without observing any form from the evaluated components. The medoid gaps of 0.3038 on GS1 and 0.1578 on GS2 therefore measure dependence on direct component-specific information under this controlled construction. They are not ordinary deployment error rates.

The larger GS1 gap is substantively informative but should be interpreted cautiously. GS1 asks for broader derivational conflations, whereas GS2 preserves more lexical distinctions. Within this nested CISTEM construction, the GS2 relation is recovered more accurately outside fitted components. The design does not identify whether that difference is attributable to conservatism as such, finer partition granularity, the resulting

class-size distribution, or the edit-command geometry induced by the partition-to-target mappings. It therefore explains the empirical asymmetry between the two fitted relations without establishing that conservative policies are intrinsically easier to learn.

The independent full-model benchmark supplies complementary practical context rather than a direct bridge between metrics. The maintained German Radixor model is trained from UniMorph with forms grouped by lemma, so its objective is naturally closer to GS2. Its tie with CISTEM on GS1 and higher snapshot F1 on GS2 demonstrate strong independent full-model results under that benchmark’s own training source, universe, and aggregate cluster TP/FP/FN scorer. Because all three differ from the component-blocked experiment, those values cannot calibrate the numerical magnitude of the held-out scores and do not establish policy-neutral Radixor superiority.

D. Partition Policy Versus Output String

The public CISTEM resources make an unusually clean distinction between these concepts. Their observable supervision is the equivalence relation; the original CELEX-derived stem/lemma label has been omitted from the release. Radixor nevertheless requires a concrete target string to induce edit commands. Consequently the experiment has two layers:

- 1) the *policy layer*: which forms are equivalent under GS1 or GS2; and
- 2) the *encoding layer*: which deterministic class member represents that equivalence class during Radixor training.

The reported claim concerns the policy-bearing partition intervention, while the second layer is a required representation choice whose influence is measured rather than ignored. Medoid, shortest, and lexical-first canonicalizations frequently choose different targets but preserve the direction of the policy-conditioned effect. This strengthens the relational interpretation without proving invariance to every possible target construction or recovering the CELEX labels omitted by CISTEM.

E. A Nested but Substantial Policy Contrast

The exact public resources instantiate a hierarchy rather than two crossing partitions: GS2 refines GS1. This is compatible with their linguistic construction. A shared lemma implies a relatively narrow lexical identity, while GS1 intentionally removes material that can collapse derivational distinctions. CISTEM’s own examples illustrate the same direction [8].

The hierarchy limits external validity but is not a weak intervention. 20.40% of GS1 classes and 80.68% of its within-class pairs are affected by the refinement. The experiment therefore demonstrates controlled adaptation along a substantial aggressive-to-conservative axis, not movement among arbitrary semantic taxonomies.

F. Policy-Bearing Data as Part of the Model Definition

For a configurable lexicon-trained stemmer, source code alone does not define stemming behavior. Runtime algorithm, compilation settings, lexical partition, and partition-to-target encoding jointly determine the learned model. The clean replay further shows that these choices can be frozen at the binary-model level. Reproducible evaluation should therefore version policy-bearing data and target-construction rules with the same care as implementation code.

X. THREATS TO VALIDITY

A. One Language and a Nested Policy Axis

Both policies are German and derive from CELEX. This is a strength for controlled contrast because orthography and lexical coverage are shared, but it limits external validity. Moreover, the pinned GS2 relation is a refinement of GS1 rather than a cross-cutting alternative. The result establishes adaptation along this German aggressive-to-conservative axis; it does not establish multilingual transfer or adaptation to arbitrary cross-cutting policies.

B. Policy Is an Operational Definition

GS1 and GS2 are legitimate resource-defined conflation objectives, not direct ground truth for all retrieval tasks. GS1's broader morphological grouping and GS2's lemma-oriented grouping operationalize different choices. Although CISTEM motivates evaluation by information-retrieval meaning, GS2 is not a manually sense-annotated semantic partition and neither resource encodes synonymy. Agreement with either is evidence of controllability relative to that resource-defined choice, not universal linguistic correctness.

C. Policy-Representation Entanglement

Changing a partition is the intended intervention, but for a lexicon-trained edit-command learner it necessarily changes more than an abstract scalar notion of aggressiveness. GS2 contains more, finer classes than GS1; after deterministic canonicalization, this changes class-size distributions, form-to-target edit geometry, command frequencies, and the complexity of the compiled trie. The three canonical-target encodings test one important representation degree of freedom, and the relation-stratified diagnostic rules out a simple global "join less" explanation, but the design cannot causally separate conservatism itself from the structural learnability induced by the refined partition. The supported causal object is therefore the complete policy-bearing training representation, not conservatism as an isolated latent property.

D. Canonical Target Strings

The public CISTEM files intentionally omit reference stems, while Radixor requires a target string. The three deterministic encodings expose this nuisance degree of freedom and differ on most classes, but they cannot prove invariance to every possible canonicalization. The stable positive direction across all three supports the reported relational claim; it does not justify a claim about the linguistic quality of any selected output string. Access to licensed CELEX labels could support an additional target-reconstruction sensitivity study, but such a study is neither available from the public CISTEM release nor required for the present partition-level question.

E. Ambiguity and Context

Context-free stemming cannot represent word-sense or morphosyntactic disambiguation. Forms assigned to multiple classes within one normalized policy are therefore excluded from the primary universe. The filter removes 5,956 of the 304,215 normalized forms shared by the raw resources (1.96%), retaining 298,259 forms. This increases internal validity at the cost of removing precisely the ambiguous cases that contextual lemmatizers are designed to address.

F. Join-Component and Weighting Concentration

The general design allows cross-cutting policies to create large join components. In this dataset GS2 refines GS1, so the join reduces to GS1. The largest component contributes only 0.0268 of disagreement evidence, and five-fold plus leave-one-component-out diagnostics show no single-family dependence. Conversely, equal component weighting could overemphasize small families. The post-review size-stratified sensitivity addresses that concern directly: every disagreement-mass stratum remains positive, including the $> 1,000$ -pair stratum containing 88.48% of disagreement pairs. The effect decreases with component mass, so macro and micro estimands answer meaningfully different weighting questions rather than serving as interchangeable confirmations.

G. Finite-Resource Uncertainty

The component bootstrap is a stability analysis conditional on the observed lexical resource and deterministic cross-fitting procedure. It is not justified as a sampling distribution for all German words or all possible policy resources. Fold effects likewise are dependence diagnostics rather than five independent experimental replications.

H. Fixed-Baseline and Historical-Tuning Asymmetry

CISTEM cannot be subjected to the same training-data intervention as Radixor because it is a pre-existing fixed rule system rather than a model fitted by this experiment. The asymmetry is also historical. Weissweiler and Fraser report changing the random 1,000-class gold-standard samples several times while developing CISTEM, including after development for the final evaluation, and describe evaluating individual rule changes against the gold standards to improve F1 [8]. CISTEM

is therefore fixed only in the present run; its rule design benefited from repeated feedback from the same gold-standard construction, whereas each reported Radixor test component is explicitly absent from fitting.

The CISTEM values remain useful contextual references, but not leakage-equivalent evidence about out-of-component learning. Their scope must also be explicit: fold-restricted scores omit cross-fold collisions to match the cross-fit relation scope, while full-universe scores include them. Finally, the experiment supplies NFC-lowercase forms to both systems for a controlled universe, which suppresses CISTEM’s case-sensitive branch. The separate lemma-grouped UniMorph Radixor benchmark is reported only as a distinct full-model context.

I. Intrinsic Versus Retrieval Evaluation

Pairwise conflation is an intrinsic structural criterion. It does not include query frequencies, collection statistics, relevance judgments, or downstream ranking. A stemmer can match a lexical policy more closely without improving a particular search application. Retrieval experiments would answer a different question.

J. Model-Family Scope

The intervention is intentionally within Radixor. Other supervised or data-driven stemmers may also adapt to target changes. This study tests whether *this* fixed patch-command/trie mechanism exhibits controlled policy transfer; it is not a comparative claim that such adaptation is unique to Radixor.

XI. REPRODUCIBILITY

The implementation under test is frozen independently from the publication harness. The experiment fixes:

- Radixor source commit `0c3b13f485a9ad0b460c0931e4497ea95bed66a1`;
- the exact Radixor JAR SHA-256 `ca74d617fc939339a76754324bb7be8301f811bf7b40cc76430bcec70729b95d`;
- CISTEM commit `7c19867c2e062c8a7d44b394c19573845ac4bd89`;
- the official `Cistem.py` Git blob `dbc90836bb6361712b52b2e504b85c702294a29f`; and
- the exact GS1/GS2 Git blob identities listed in Section IV-A.

The source bundle accompanying this revision contains the repository-ready harness at `experiments/policy-transfer/`. It includes the frozen experiment specification, exact-input fetcher, Python orchestration, JDK 21 model runner, independent aggregate verifier, policy-structure audit, fixed CISTEM evaluator, full-information ceiling evaluator, post-review relation-specificity and component-size sensitivity analysis, synthetic smoke test, frozen Radixor JAR, and publication-safe aggregate result tables. The harness revision is distinct from the older Radixor core commit under test [7].

The real runner validates CISTEM byte size and Git blob SHA-1 before analysis. CELEX-derived words, generated fold dictionaries, compiled models, and token predictions are

local build products and are not committed or redistributed. Public result tables contain aggregate component counts, confusion counts, model hashes, and summary statistics only. A clean replay from the pinned inputs regenerated all 30 cross-fit model-file SHA-256 digests and the reported primary aggregate outputs. The full-information evaluator reconstructs six additional all-data models and verifies exact policy-matched reconstruction. The CISTEM evaluator generates both fold-restricted and full-restricted-universe scopes from the same pinned implementation. The post-review sensitivity script derives the relation-specificity and component-size tables from already frozen aggregate predictions/counts and does not alter any primary model or estimand.

The separate maintained-model benchmark is an imported frozen snapshot result because it predates this policy-transfer harness. The experiment documentation includes the exact JMH command that regenerates the registered Radixor-versus-CISTEM comparison from the frozen Radixor source tree. Its metric remains explicitly separate from the CISTEM-style macro and pairwise metrics in this study.

The Radixor repository root is intentionally retained as the persistent project locator [7]. The publication harness is integrated there as part of the article-release workflow, so the resulting integration commit does not exist until the final article bundle is frozen and its hash is not available at manuscript-generation time. Reproducibility of the reported experiment therefore relies on the exact tested core commit, JAR digest, external Git commits/blobs, model-file digests, and the source bundle accompanying the article. After release, repository history supplies the immutable integration commit without changing any scientific content or numerical result reported here.

XII. CONCLUSION

This study asks whether one fixed lexicon-trained stemming mechanism can respond to a change in the conflation policy encoded by its training data and generalize that change to lexical components excluded in their entirety from fitting. On the pinned CISTEM resources, the answer is positive but deliberately qualified.

GS1 and GS2 encode a substantial nested aggressive-to-conservative contrast. The public files expose partitions rather than the CELEX-derived canonical labels used to create them, so the experiment treats relational membership as the policy object and explicitly varies the deterministic target string needed for Radixor training. The primary medoid model pair yields $\Delta_{\text{macro}} = 0.0493$ with a positive stability interval; 98.46% of observed switches are policy aligned, and shortest-member plus lexical-first encodings preserve the sign. Pair-micro switch coverage is only 0.61%, so the claim is *sparse but directionally robust policy-conditioned relational generalization*, not broad control of all held-out conflations.

The nested design does not reduce to a uniform conservatism shift: a post-review specificity diagnostic shows that the GS2-conditioned model recovers more shared-positive relations while joining far fewer pairs on the relation that GS2 specifically separates. The directional effect is also

positive across component-size strata, including the stratum containing most disagreement pairs, although its magnitude declines strongly with disagreement mass. The changed policy-bearing representation also produces substantially different compiled model complexity, with the medoid GS2-conditioned models about 69% smaller by both persisted bytes and trie-node count; this structural observation is not a runtime claim. Full-information policy models exactly reconstruct their own supplied partitions, whereas component blocking creates a measurable generalization gap; fixed CISTEM provides useful but historically and procedurally asymmetric reference scores. A separate lemma-grouped UniMorph Radixor benchmark supplies independent maintained-model context under a different scorer.

The result therefore establishes a specific capability of the frozen Radixor mechanism: changing the policy-bearing lexical training representation can change the direction of relational generalization outside the fitted components without changing runtime code. Because the nested intervention jointly changes partition granularity and the edit geometry presented to the learner, it does not isolate conservatism as the causal property. Nor does it establish multilingual transfer, arbitrary cross-cutting policy control, retrieval effectiveness, universal linguistic correctness, or recovery of the target strings omitted from the public CISTEM release.

REFERENCES

- [1] J. B. Lovins, "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics*, vol. 11, pp. 22–31, 1968.
- [2] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [3] J. Xu and W. B. Croft, "Corpus-based stemming using cooccurrence of word variants," *ACM Transactions on Information Systems*, vol. 16, no. 1, pp. 61–81, 1998.
- [4] N. N. Karanikolas, "Supervised learning for building stemmers," *Journal of Information Science*, vol. 41, no. 3, pp. 315–328, 2015.
- [5] L. Galambos, "Lemmatizer for document information retrieval systems in Java," in *SOFSEM 2001: Theory and Practice of Informatics*, ser. Lecture Notes in Computer Science, vol. 2234. Springer, 2001, pp. 243–252.
- [6] —, "Semi-automatic stemmer evaluation," in *Intelligent Information Processing and Web Mining*. Springer, 2004, pp. 209–218.
- [7] —, "Radixor multilingual stemming-quality snapshot," GitHub repository, commit 0c3b13f485a9ad0b460c0931e4497ea95bed66a1, 2026, checked-in 314-scenario quality matrix and methodology; accessed Aug. 26, 2026. [Online]. Available: <https://github.com/leogalambos/Radixor>
- [8] L. Weissweiler and A. Fraser, "Developing a stemmer for german based on a comparative analysis of publicly available stemmers," in *Language Technologies for the Challenges of the Digital Age*, ser. Lecture Notes in Computer Science. Springer, 2018, vol. 10713, pp. 81–94.
- [9] J. B. Lovins, "Error evaluation for stemming algorithms as clustering algorithms," *Journal of the American Society for Information Science*, vol. 22, no. 1, pp. 28–40, 1971.
- [10] M. Lennon, D. S. Peirce, B. D. Tarry, and P. Willett, "An evaluation of some conflation algorithms for information retrieval," *Journal of Information Science*, vol. 3, no. 4, pp. 177–183, 1981.
- [11] C. D. Paice, "An evaluation method for stemming algorithms," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 42–50.
- [12] —, "Method for evaluation of stemming algorithms based on error counting," *Journal of the American Society for Information Science*, vol. 47, no. 8, pp. 632–649, 1996.
- [13] D. A. Hull, "Stemming algorithms: A case study for detailed evaluation," *Journal of the American Society for Information Science*, vol. 47, no. 1, pp. 70–84, 1996.
- [14] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX lexical database, release 2," Linguistic Data Consortium, LDC96L14, Philadelphia, PA, 1995.
- [15] J. A. Goldsmith, D. Higgins, and S. Soglasnova, "Automatic language-specific stemming in information retrieval," in *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop*, ser. Lecture Notes in Computer Science, vol. 2069. Springer, 2001, pp. 273–283.
- [16] M. Bacchin, N. Ferro, and M. Melucci, "A probabilistic model for stemmer generation," *Information Processing & Management*, vol. 41, no. 1, pp. 121–137, 2005.
- [17] H. Hammarström, "Poor man's stemming: Unsupervised recognition of same-stem words," in *Information Retrieval Technology: Third Asia Information Retrieval Symposium, AIRS 2006*, ser. Lecture Notes in Computer Science, vol. 4182. Springer, 2006, pp. 323–337.
- [18] G. Chrupała, "Simple data-driven context-sensitive lemmatization," *Procesamiento del Lenguaje Natural*, no. 37, pp. 121–127, 2006.
- [19] G. Chrupała, G. Dinu, and J. van Genabith, "Learning morphology with Morfette," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008, pp. 2362–2367.
- [20] M. Juršič, I. Mozetič, T. Erjavec, and N. Lavrač, "LemmaGen: Multilingual lemmatisation with induced ripple-down rules," *Journal of Universal Computer Science*, vol. 16, no. 9, pp. 1190–1214, 2010.
- [21] L. Weissweiler and contributors, "CISTEM: Stemmer for german," GitHub repository, 2022, pinned commit and gold-standard Git blob identities are reported in Sec. IV-A and the reproducibility artifact; accessed Aug. 2026. [Online]. Available: <https://github.com/LeonieWeissweiler/CISTEM>
- [22] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.