

Edit Costs as Structural Hyperparameters in Dictionary-Induced Stemming: A Multilingual Sensitivity Study

Leo Galambos

Abstract—Weighted edit costs are usually interpreted as local parameters of string alignment. In a dictionary-induced stemmer, however, they also determine which reusable transformations are generated and therefore which symbolic values are compiled into the runtime model. This study measures that structural role for Radixor across 20 UniMorph-derived lexical resources. Delete, insert, and replace costs are drawn from $\{1, 2, 3, 5, 10\}$ and match costs from $\{0, 1\}$; removing common positive scaling leaves 234 normalized configurations. We define exact resource-specific equivalence by command-for-command equality over the complete ordered dictionary mapping sequence and prove that this equivalence is invariant under every nested protocol subset. Consequently, 4,680 resource–cost combinations collapse exactly to 334 induced behaviors, reducing a nominal 234,000-observation campaign to 16,700 physically evaluated scenarios without approximation. We then measure the complete sampled-grid quality landscape inside matched resource $\times$ split $\times$ coverage strata. Across all 900 held-out strata, the median range in unseen changed-form exactness is 0.0067 percentage points (interquartile range 0–0.0608; 95th percentile 0.360), while the maximum is 2.577 pp for Czech. The corresponding median $F_{0.5}$ range is 0.000044 and its 95th percentile 0.00695. Thus typical sensitivity is small but not uniformly negligible. A full-grid structural minimum yields a smaller full-dictionary command vocabulary for 17 of 20 resources, with reductions of 24.23% for Finnish and 24.07% for French. Although up to 90 numeric settings tie at a structural minimum, all tied minimum behaviors produce identical changed-form exactness in every matched stratum; the largest median paired exactness difference from the baseline is +0.041 pp. A secondary operating-point rule yields paired-median baseline dominance for 15 resources, a negative paired median for 2, and the baseline for 3. Uniform-over-behavior sensitivity analysis confirms the strong relation between retained-command ratio and stored value references, while showing that some other correlation conclusions depend on equivalence-class weighting. The results support treating edit costs as structural hyperparameters of transformation induction while bounding the claim to the sampled grid, evaluated resources, context-free normalization metrics, and deterministic representation outcomes rather than runtime memory or latency.

Index Terms—stemming, weighted edit distance, morphology, information retrieval, lexical resources, hyperparameter sensitivity, tries, edit scripts

I. INTRODUCTION

STRING edit operations provide a natural representation for mappings between inflected forms and canonical forms. Classical string-to-string correction formalizes a transformation as a minimum-cost sequence of insertions, deletions, and substitutions [1]. In computational morphology, edit-script

abstractions are widely used to define reusable lemmatization or transformation classes rather than predict complete output strings directly [2]–[4]. Weighted edit models further show that changing the cost function can change the preferred correspondence between strings [5].

In a dictionary-induced stemmer, edit costs have an additional role. A minimum-cost path does not merely explain one form–target pair: it determines a serialized transformation that becomes a value in a shared model. If two cost settings choose different paths for many mappings, they can induce different transformation vocabularies, value frequencies, and reduced trie structures even though the lexical supervision is identical. Conversely, different numeric settings may generate exactly the same command for every mapping and are then behaviorally indistinguishable for that resource.

This mechanism is explicit in the Egothor/Radixor patch-command method [6]. A lexical resource provides a canonical target and observed forms. A dynamic-programming encoder converts each form-to-target mapping into a patch command; commands are inserted into a trie, ranked by frequency, reduced, and compiled for runtime lookup. Edit costs therefore sit upstream of the learned symbolic representation. We use the term *structural hyperparameter* for this role: the costs are external to the induced trie, fixed before induction, and capable of changing the model structure while supervision remains constant.

The present study evaluates a normalized 234-point cost grid over 20 checked-in default resources, five protocol-fixed family rankings, and ten dictionary-coverage levels. Cost configurations are compared inside identical language-resource $\times$ split $\times$ coverage strata, so training evidence and evaluation populations are held fixed while only the edit-cost configuration changes. This pairing is central: it separates cost sensitivity from the changing test-population issue that arises when coverage levels themselves are compared.

We ask four research questions:

- **RQ1:** How many distinct induced transformation behaviors are represented by the normalized edit-cost grid for each lexical resource?
- **RQ2:** How wide is the complete sampled-grid quality landscape, and how heterogeneous is that sensitivity across resources and matched strata?
- **RQ3:** What structural slack remains at full-grid command minima and practical lower-complexity operating points?
- **RQ4:** Which secondary representation–quality associations persist when exact behaviors, rather than numeric

grid labels, receive equal weight?

The principal contributions are: (i) an exact equivalence relation over complete dictionary command sequences together with a subset-invariance proposition that makes experiment reduction exact rather than statistical; (ii) a 20-resource census of the partition induced by 234 sampled relative-cost settings; (iii) a direct full-grid characterization of changed-exact and pairwise-quality sensitivity inside matched strata; (iv) full-grid structural minima with complete quality envelopes across all tied minimum behaviors, plus paired operating-point comparisons; and (v) parallel nominal-grid-weighted and uniform-over-behavior association summaries for model-invariant variables. The study does *not* claim a universal best cost setting, population-level inference over languages, contextual lemmatization, downstream retrieval improvement, or runtime gains from structural counts alone.

II. BACKGROUND AND RELATED WORK

A. Weighted String Editing

The string-to-string correction problem assigns costs to primitive edit operations and chooses a minimum-cost path from source to target [1]. Multiplication of all operation costs by a common positive scalar preserves path ordering, whereas changing relative costs can change the selected path and therefore the edit script. Ristad and Yianilos [5] showed that edit-distance behavior itself can be learned probabilistically, establishing that edit costs can encode substantive modeling assumptions rather than merely implementation constants.

Our objective is different. We hold the lexical mappings fixed and measure how a finite normalized set of deterministic relative costs partitions into distinct *induced transformation models*. The central object is therefore the mapping from sampled cost settings to symbolic model behavior, not a learned string-similarity metric.

B. Edit Scripts in Computational Morphology

Edit-script representations are established in data-driven morphology. Morfette learns morphology through automatically induced mapping classes [2]; Lemming uses edit trees in joint lemmatization and tagging [3]; contextual systems likewise predict or exploit edit trees and related edit-script representations [7]. The closest recent methodological precedent is Toporkov and Agerri [4], who compare shortest-edit-script representation methods and measure their effect on contextual lemmatization across morphologically diverse languages.

The distinction from the present study is important. Toporkov and Agerri vary the *edit-script representation method* and evaluate downstream contextual lemmatization. Here the patch-command representation, lexical supervision, trie builder, and reduction semantics are held fixed; we vary only the *relative operation costs* and study exact induced-model equivalence plus compiled structural consequences. The task is deterministic search-oriented stemming rather than contextual morphological analysis.

C. Patch Commands and Dictionary-Induced Stemming

The Egothor patch-command formulation derives compact edit instructions from minimum-cost paths between forms and canonical targets and stores those instructions as trie values [6]. The current Radixor implementation retains this lineage while adding compiled immutable commands, deterministic ranking, uniform-subtree contraction, semantic reduction, and model integrity metadata. A companion family-holdout experiment in the same public snapshot establishes that training-derived transformations can be applied to surfaces absent from the training subset; the present study keeps that evaluation design fixed and varies the edit-cost function instead [8].

This separation matters. Generalization asks whether learned transformations transfer beyond exact training membership. Cost sensitivity asks *which* transformations are induced from the same lexical evidence and how the resulting representation changes. The latter supports matched comparisons because every cost configuration within one stratum sees identical training rows and identical evaluation forms.

D. Task Interpretation: Search-Oriented Stemming

We use *stemming* in the information-retrieval sense of deterministic term conflation. The UniMorph-derived supervision is lemma-like because each lexical family supplies a resource-defined canonical form, but Radixor does not use sentence context, predict morphosyntactic tags, or disambiguate homographic lemmas. Exact canonical-target agreement in this study is therefore an evaluation target for a context-free normalization model, not a claim of general contextual lemmatization. This distinction also separates the present task from contextual edit-script systems such as [4].

The pairwise measures follow the retrieval-oriented stemming tradition in which over-stemming joins gold-unrelated forms and under-stemming separates gold-related forms [9], [10]. We use $F_{0.5}$ only as a secondary precision-favoring objective: in search-oriented conflation, false joins can collapse semantically unrelated terms, so precision receives greater weight than recall. No single metric is treated as universally authoritative.

III. MODEL AND COST-SPACE FORMULATION

A. Lexical Families and Patch-Command Encoding

For one resource let the parsed dictionary contain lexical families

$$F_j = (c_j, V_j), \quad (1)$$

where c_j is the resource-defined canonical form and V_j the ordered variants retained by the production parser. For each $w \in V_j \cup \{c_j\}$, the encoder generates a deterministic patch command transforming w to c_j . The production model uses backward traversal, reflecting the suffix-oriented bias of the historical design, although individual commands can include deletion, insertion, replacement, and preservation operations.

Let a cost configuration be

$$\kappa = (d, i, r, m), \quad (2)$$

with delete, insert, replace, and equal-character match costs. The production baseline is $D1I1R1M0 = (1, 1, 1, 0)$. The experiment uses

$$d, i, r \in \{1, 2, 3, 5, 10\}, \quad (3)$$

$$m \in \{0, 1\}. \quad (4)$$

The Cartesian product contains 250 tuples. Configurations differing only by common positive scaling have identical path-cost ordering and are deduplicated, leaving 234 normalized settings.

B. Exact Resource-Specific Equivalence

For resource D , let

$$E_D(\kappa) = \langle e_{\kappa,1}, e_{\kappa,2}, \dots, e_{\kappa,n} \rangle \quad (5)$$

be the complete ordered sequence of serialized patch commands generated for all accepted form-to-canonical mappings. Define

$$\kappa_a \equiv_D \kappa_b \iff E_D(\kappa_a) = E_D(\kappa_b) \quad (6)$$

by direct command equality at every position. This is stronger than equality of command counts, trie statistics, or output accuracy.

Proposition 1 (Subset invariance of exact command equivalence). *If $\kappa_a \equiv_D \kappa_b$, then for every protocol training subset $S \subseteq D$ obtained by restricting the same ordered dictionary rows, the ordered key–command streams generated under κ_a and κ_b are identical. Consequently, any deterministic trie construction and reduction pipeline whose state is a function only of that ordered stream produces identical compiled tries under the two configurations.*

Proof. Equality of the complete command sequence implies equality of every command associated with every accepted mapping. Restricting the same ordered dictionary to S therefore restricts two equal sequences at identical positions and preserves equality. The resulting ordered key–command streams are identical. A deterministic builder and deterministic reduction applied to identical streams must return identical compiled representations. $\square$

A SHA-256 fingerprint is used only for candidate bucketing; equivalence-class membership is confirmed by direct command equality. Proposition 1 therefore makes the campaign reduction exact, not probabilistic.

IV. EXPERIMENTAL DESIGN

A. Resources and Campaign Scale

The study uses the same 20 default UniMorph-derived lexical resources as the Radixor generalization campaign [8], [11]. Resource sizes range from 69 rows for Persian to 396,939 rows for English; baseline full-dictionary command vocabularies range from 138 distinct commands for Danish to 6,986 for German. The sample spans Latin, Cyrillic, Hebrew, and Perso-Arabic scripts and includes Germanic, Romance, Slavic, Uralic, Semitic, and Iranian languages; these descriptors establish breadth but are not used as typological predictors. Language identity and resource identity are confounded in this

TABLE I
CAMPAIGN IDENTITY AND SCALE.

Quantity	Value
Normalized cost configurations	234
Language–resource configurations	20
Splits per coverage level	5
Coverage levels	10
Exact command classes	334
Physical scenarios	16,700
Expanded logical observations	234,000
Partial-coverage matched strata	900

design because one default resource is evaluated per language. The resources are therefore treated as 20 language–resource configurations, not as a random sample of languages or resource traditions.

For every resource, five deterministic protocol rankings define nested 10%–100% training prefixes. Each physically distinct exact class is evaluated at all ten coverage levels for all five splits. Across resources, 334 exact classes therefore produce

$$334 \times 10 \times 5 = 16,700 \quad (7)$$

physical scenarios. Expansion back to all 234 normalized labels yields

$$20 \times 234 \times 10 \times 5 = 234,000 \quad (8)$$

logical observations. Partial-coverage quality analysis uses 900 language–resource $\times$ split $\times$ 10–90% strata, where a held-out population exists.

B. Matched-Stratum Quality Outcomes

The complete dictionary is evaluated at every coverage level. The primary generalization scope retains only occurrences from withheld families whose parser-normalized surface was absent from training; the headline exactness metric further restricts this scope to forms whose surface differs from the expected canonical target. We denote this rate by A_{chg} .

Pairwise outcomes guard against exact-target agreement masking undesirable conflation. Under the repository’s overlapping-family protocol, two forms are gold-related when their family memberships intersect. From pairwise true positives, false positives, false negatives, and true negatives we use precision P , recall R , and

$$F_{0.5} = \frac{1.25PR}{0.25P + R}, \quad (9)$$

together with over-stemming

$$OI = \frac{FP}{TN + FP}, \quad (10)$$

and under-stemming

$$UI = \frac{FN}{TP + FN}. \quad (11)$$

$F_{0.5}$ favors precision and is used as a search-oriented secondary objective; OI and UI expose the two error directions directly [9], [10]. The unseen-family pairwise scope is a partial-model stress test, not a full-model quality score; at 100% coverage it is empty.

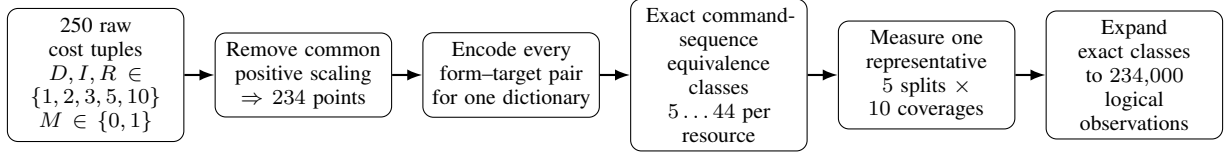


Fig. 1. Experimental reduction of the normalized edit-cost grid. Class membership is exact: two cost settings share a class only when every generated patch command agrees in order for the complete dictionary. The 334 classes across 20 resources therefore require 16,700 physical scenarios rather than 234,000 separately materialized scenarios.

For cost comparisons, all outcomes are paired within the same language–resource $\times$ split $\times$ coverage stratum. Thus $A_{\text{chg}}(\kappa_a) - A_{\text{chg}}(\kappa_b)$ compares deterministic models built from exactly the same selected rows and evaluated on exactly the same unseen occurrences. Coverage-level differences are not interpreted as cost effects because their held-out complements differ.

C. Representation Outcomes

The experiment records distinct retained patch commands, trie nodes, edges, longest and average logical path lengths, dense child-table slots, stored value references, and logical leaf paths. The main normalized structural quantity is

$$R_{\text{cmd}} = \frac{\text{retained commands under } \kappa}{\text{retained commands under D1I1R1M0}} \quad (12)$$

within the matching stratum. These are deterministic representation counts and proxies. They do not measure heap occupancy, serialized bytes, CPU instructions, cache behavior, or lookup latency.

D. Direct Full-Grid Quality Sensitivity

For each partial-coverage stratum s , let $\mathcal{B}_s$ be the set of physically distinct induced behaviors after exact collapse. The complete sampled-grid changed-exact range is

$$G_A(s) = \max_{b \in \mathcal{B}_s} A_{\text{chg}}(b) - \min_{b \in \mathcal{B}_s} A_{\text{chg}}(b), \quad (13)$$

and $G_F(s)$ is defined analogously for $F_{0.5}$. We also retain behavior-unweighted and nominal-grid-weighted interquartile ranges. Minimum and maximum are invariant to duplication of equivalent labels, whereas the two IQR estimands need not coincide. Each resource is summarized by the median and maximum of its 45 matched-stratum ranges. All 16,700 physically distinct scenarios in the published campaign are classified *VARIABLE*; consequently, full-grid and viable-grid ranges coincide for this dataset.

E. Objective-Separated Full-Grid Views

To avoid making one compromise selector carry the scientific argument, the analysis separates three objectives.

Structural minimum: at 100% dictionary knowledge, minimize the number of distinct retained patch commands over all 234 normalized settings. All numeric labels and all distinct behaviors attaining the minimum are retained. Their paired partial-coverage quality deltas are reported as an envelope, so the conclusion does not depend on a favorable representative chosen among ties.

Guarded quality optimum: among configurations with no increase in median unseen-family *OI* or *UI* relative to the baseline, maximize median unseen $F_{0.5}$. This view asks whether alternative costs can change pairwise quality under explicit guards. It is secondary to the full-grid range because selection and reporting use the same matrix.

Practical operating point: find the best median A_{chg} , admit settings within 0.25 percentage points, then minimize median R_{cmd} ; remaining ties prefer $F_{0.5}$ and exactness. The implementation defines ratios above $5\times$ baseline as *MARGINAL* and above $10\times$ as *NOT_VIABLE*; none of the physically distinct classes in this campaign crosses the first threshold.

For the operating-point view, the primary comparison uses medians of paired within-stratum differences rather than a difference of two separately calculated medians. We call a setting *paired-median dominant* over the two reported objectives when

$$\text{median}_s[A_{\text{chg}}(\kappa, s) - A_{\text{chg}}(\text{D1I1R1M0}, s)] \geq 0 \quad (14)$$

and

$$\text{median}_s R_{\text{cmd}}(\kappa, s) < 1. \quad (15)$$

This is dominance on two paired median summaries, not pointwise dominance in every stratum or a complete Pareto frontier over all metrics.

F. Secondary Association Analysis and Weighting Semantics

Spearman correlation is primary and Pearson correlation secondary. Coefficients are computed separately inside each language–resource $\times$ split $\times$ coverage stratum to avoid pooled associations manufactured by resource size or coverage. The distribution of within-stratum coefficients is summarized by its median, central empirical 95% interval, and positive/zero/negative sign counts; the interval is descriptive, not a confidence interval under independence.

Two weighting estimands are reported for variables invariant within an exact class. *Nominal-grid weighting* expands a class back to all of its member labels, corresponding to a uniform draw from the defined 234-point grid. *Uniform-behavior weighting* gives each physically distinct induced behavior one observation, corresponding to a uniform draw from the resource-specific behavior partition. Cost-coordinate predictors do not have a unique class-level value when several numeric labels share one behavior, so the behavior-weighted sensitivity analysis is restricted to representation and quality variables that are constant inside a class.

Across-resource coefficients use 20 observations and remain exploratory. Where two coefficients share the same outcome,

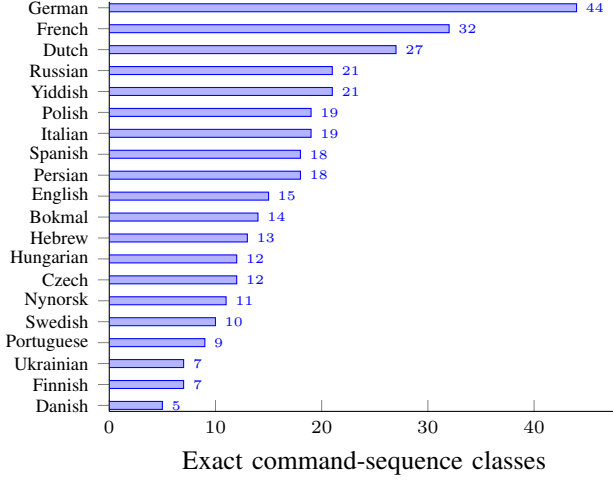


Fig. 2. Dictionary-specific sensitivity of the 234 normalized cost points. Fewer classes mean that more cost settings induce exactly the same command sequence for that resource; the count is not a quality score.

their observed difference is accompanied by a paired resource bootstrap rather than interpreted from point estimates alone.

V. RESULTS

A. RQ1: The Sampled Cost Grid Collapses into Few Exact Behaviors

The 4,680 resource–cost combinations collapse to 334 exact resource-specific command classes, a $14.01\times$ reduction in distinct behavior before any statistical analysis. The mean resource exposes 16.7 classes and the median 14.5, but the range is wide: Danish exposes 5 classes and German 44. Figure 2 shows all 20 resources.

This is not evidence that edit costs are unimportant. It establishes that the mapping from the *sampled* numeric grid to induced symbolic behavior is highly many-to-one and resource-specific. For Danish, 234 normalized settings yield only five exact behaviors; for German, the same grid yields 44. The largest observed equivalence class contains 110 labels for Danish and Finnish. No claim is made about the geometric volume of corresponding regions in the continuous positive cost space.

B. Exploratory Resource-Level Associations

Across the 20 configurations, exact-class count has Spearman $\rho = 0.260$ with dictionary rows and $\rho = 0.088$ with form occurrences. Its observed association with the baseline full-dictionary patch-command vocabulary is $\rho = 0.510$ (Pearson $r = 0.639$). The observed difference between the two Spearman coefficients is 0.250, but a paired 20,000-resample resource bootstrap gives a 95% interval of approximately $[-0.049, 0.564]$. The interval includes zero, so the larger command-vocabulary coefficient is treated as a descriptive hypothesis rather than evidence that it is inferentially stronger.

The association does not establish that command vocabulary causes sensitivity, nor can language morphology be separated from resource construction. It motivates a narrower hypothesis

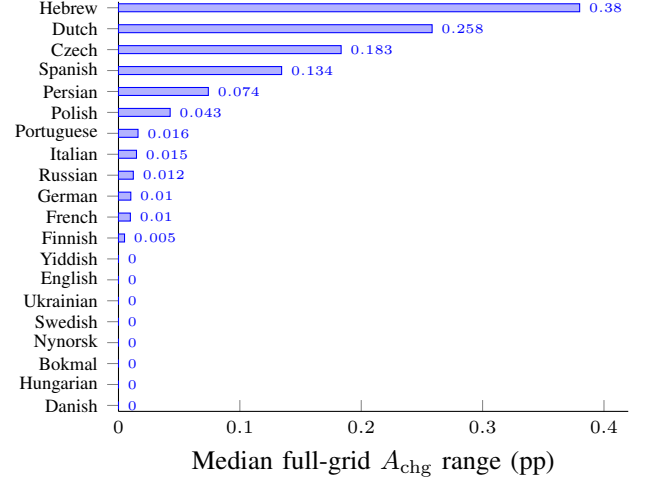


Fig. 3. Median within-stratum range of unseen changed-form exactness across the distinct induced behaviors represented by the 234 sampled cost settings. Eight resources have a zero median range, whereas Hebrew, Dutch, Czech, and Spanish show the largest typical sensitivity. The figure reports medians across 45 matched strata; the global maximum range is 2.577 pp for one Czech 10% stratum.

for future multi-resource studies: transformation-vocabulary complexity may be more informative than raw row count when predicting how many alternative alignments a resource exposes.

C. RQ2: Full-Grid Quality Sensitivity Is Usually Small but Not Uniformly Zero

Direct enumeration of every distinct induced behavior shows a strongly heterogeneous quality landscape. Across all 900 matched strata, the median $G_A(s)$ is 0.0067 pp, the interquartile range is 0–0.0608 pp, and the 95th percentile is 0.3603 pp; 365 strata have zero range. The median across resources of the resource-level median $G_A(s)$ is 0.010 pp, with an interquartile range of 0–0.050 pp. Eight resources have a zero median range across their 45 matched strata. The largest typical ranges are 0.380 pp for Hebrew, 0.258 pp for Dutch, 0.183 pp for Czech, and 0.134 pp for Spanish. Figure 3 shows all resources.

The largest single-stratum range is 2.577 pp for the Czech resource at 10% coverage under one fixed split. Across all 900 strata, the $F_{0.5}$ range has median 0.000044, interquartile range 0–0.000493, 95th percentile 0.00695, and maximum 0.0310. The baseline is tied for maximum changed exactness in 796 strata; an alternative behavior exceeds it in 104, while at least one behavior falls below it in 522. These results reject both extreme interpretations. The sampled cost grid is often linguistically redundant, but it is neither universally quality-invariant nor uniformly safe.

D. RQ3: Full-Grid Structural Extrema Expose Substantial Slack

Seventeen of 20 resources have a strictly smaller full-dictionary structural minimum than the production baseline. The largest reductions are 24.23% for Finnish, 24.07% for French, 20.80% for Italian, 17.34% for Portuguese, and 14.24% for Spanish. Figure 4 shows all resources.

TABLE II

RESOURCE-SPECIFIC EXACT COST SENSITIVITY AND FULL-GRID STRUCTURAL MINIMUM. “GRID RED.” IS 234/EXACT CLASSES. “MIN RED.” IS THE REDUCTION IN THE MINIMUM FULL-DICTIONARY DISTINCT COMMAND VOCABULARY RELATIVE TO D111R1M0. “MED. RANGE” IS THE MEDIAN WITHIN-STRATUM FULL-GRID RANGE OF UNSEEN CHANGED-FORM EXACTNESS IN PERCENTAGE POINTS.

Resource	Base cmd.	Classes	Grid red.	Min red.	Med. range	Resource	Base cmd.	Classes	Grid red.	Min red.	Med. range
Danish	138	5	46.80×	10.87%	0.000	English	337	15	15.60×	0.00%	0.000
Finnish	2,683	7	33.43×	24.23%	0.005	Persian	704	18	13.00×	6.25%	0.074
Ukrainian	286	7	33.43×	4.90%	0.000	Spanish	1,496	18	13.00×	14.24%	0.134
Portuguese	369	9	26.00×	17.34%	0.016	Italian	750	19	12.32×	20.80%	0.015
Swedish	332	10	23.40×	4.52%	0.000	Polish	846	19	12.32×	6.50%	0.043
Nynorsk	299	11	21.27×	0.00%	0.000	Yiddish	202	21	11.14×	0.50%	0.000
Czech	537	12	19.50×	5.21%	0.183	Russian	1,840	21	11.14×	0.00%	0.012
Hungarian	462	12	19.50×	11.04%	0.000	Dutch	538	27	8.67×	6.69%	0.258
Hebrew	4,150	13	18.00×	3.06%	0.380	French	2,730	32	7.31×	24.07%	0.010
Bokmål	325	14	16.71×	3.69%	0.000	German	6,986	44	5.32×	2.89%	0.010

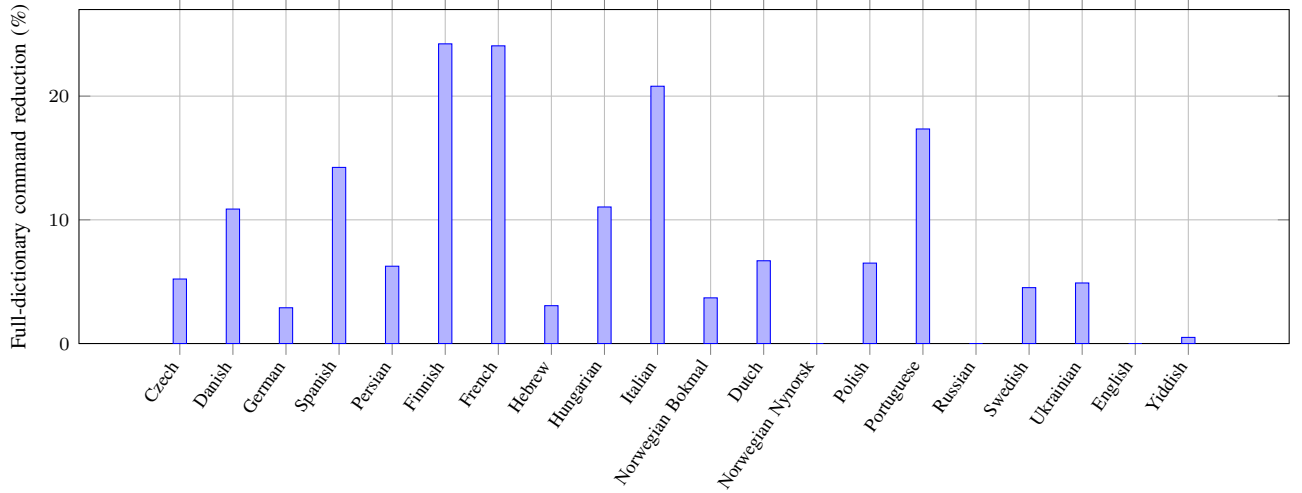


Fig. 4. Structural extreme over the complete normalized grid: reduction in the minimum full-dictionary distinct patch-command vocabulary relative to D111R1M0. Seventeen of 20 resources admit a strictly smaller vocabulary; these are representation counts, not byte, heap, or latency measurements.

Numeric ties at the structural minimum range from 2 to 90 labels, and three resources have more than one exact minimum behavior. Nevertheless, every tied minimum behavior produces identical unseen changed-form exactness in every matched partial-coverage stratum; the maximum within-stratum exactness spread across tied minimum behaviors is exactly zero. Their median paired exactness differences from the baseline are zero for 18 resources, +0.0033 pp for Finnish, and +0.0411 pp for Hebrew. Thus the structural conclusion is not an artifact of choosing a favorable representative among tied minima. Pairwise $F_{0.5}$ differs slightly among tied minimum behaviors only for Nynorsk and Yiddish, with a maximum within-stratum spread of 0.000695; no tied minimum behavior has a negative median paired $F_{0.5}$ difference from the baseline.

These findings establish the existence of materially smaller transformation vocabularies with no negative median paired changed-exact signal in the measured campaign. They do not establish runtime compression because command count is only one representation dimension.

E. Guarded Quality Optima Are Secondary and Small

Under the explicit no-worse-median-OI/UI guards, a non-baseline $F_{0.5}$ optimum is selected for 7 of 20 resources. The largest median $F_{0.5}$ improvement is +0.001308 for Hebrew;

Dutch gains +0.000084 and German +0.000019, while several alternatives are ties at reported precision. These deterministic effect sizes are real within the campaign but small, so they are not used as headline evidence of practical improvement. Their role is to confirm that edit costs can alter pairwise linguistic behavior even when lexical supervision is fixed.

F. Paired Operating Points: 15 Resources Dominate on the Reported Medians

The practical selector is evaluated with medians of paired within-stratum exactness differences. Fifteen resources select a setting with both a lower median command count and a nonnegative median paired A_{chg} difference. Persian and Russian accept negative paired medians of -0.0740 pp and -0.0060 pp, respectively; English, Nynorsk, and Yiddish retain the baseline. Among the 15 paired-median-dominating resources, the candidate is jointly favorable on both objectives in 656 of 675 matched strata overall, in at least 37 of 45 strata for every resource, and in all 45 strata for 10 resources.

Across resources, the selected median command reduction is 5.75%; the largest are 26.64% for French and 26.37% for Finnish. These are descriptive same-matrix operating points, not externally validated defaults.

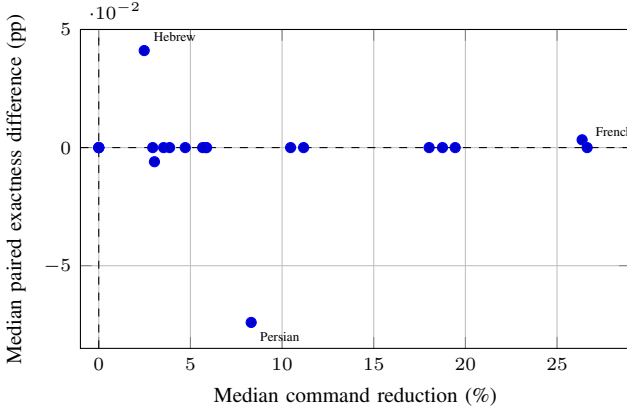


Fig. 5. Secondary operating-point analysis using medians of paired within-stratum differences. Fifteen resources combine a lower median command count with a nonnegative median paired exactness difference, two accept a negative paired median, and three retain the baseline. This is a same-matrix operating-point summary, not an external-test Pareto estimate.

G. RQ4: Weighting Changes Some Association Conclusions

Table III compares nominal-grid and uniform-over-behavior estimands for variables that are constant within an exact class. The relation between retained-command ratio and stored value references is robust: the behavior-weighted median Spearman coefficient is 0.987, the central empirical interval remains positive, and 899 of 900 strata have a positive coefficient. The command-ratio-node relation remains large in median magnitude but is less universal under behavior weighting: its central interval crosses zero and 64 of 900 strata are negative.

Behavior weighting also makes the negative leaf-path-changed-exactness tendency directionally clearer in the 502 strata where both variables vary, while the corresponding positive relation with under-stemming is present in 504 of 517 defined strata. These are descriptive properties of the measured representation class, not universal causal laws: coefficients are undefined in many strata because exact collapse leaves too little variation, and language/resource identity remains confounded.

VI. DISCUSSION

A. Edit Costs Partition the Sampled Grid into Resource-Specific Models

The exact-equivalence census gives the cleanest answer to the paper’s structural question. Relative edit costs choose among alternative minimum-cost alignments, and those choices determine the symbolic patch-command stream supplied to the trie builder. The 234 sampled settings are partitioned into exact induced-model equivalence classes whose number varies from 5 to 44 across the evaluated resources. This is the precise sense in which costs act as structural hyperparameters: they are external induction parameters that can change the learned symbolic representation while lexical supervision and compilation semantics remain fixed.

We intentionally avoid describing class sizes as geometric regions of continuous cost space. The sampled grid is discrete and nonuniform in ratio space. A class containing many labels represents many sampled settings, not a measured volume of a continuous region.

B. The Complete Grid Is Mostly Redundant but Contains Real Quality Sensitivity

The direct range analysis closes an ambiguity left by selector-based summaries. For many resources and strata, every induced behavior produces the same changed-exact score; for eight resources this is true in at least half of the matched strata. Yet Hebrew, Dutch, Czech, and Spanish show nontrivial typical ranges, and one Czech stratum spans 2.577 pp. The correct conclusion is therefore neither “edit costs do not matter” nor “every numeric setting defines a linguistically different model.” Most of the sampled grid is behaviorally redundant, but the remaining distinctions can matter in resource-specific conditions.

C. Structural Slack Is Not a Tie-Breaking Artifact

The structural-minimum envelope is stronger than a single favorable representative. Although some minima contain dozens of numeric labels and multiple exact behaviors, all tied minimum behaviors agree exactly on changed-form exactness within every matched stratum. This demonstrates structural slack with a quality statement that is invariant to minimum tie selection. The result remains representation-specific: fewer commands do not automatically imply fewer bytes, less heap, or lower latency.

D. Weighting Defines the Association Question

Exact collapse creates two legitimate but different populations: sampled numeric cost points and distinct induced behaviors. The strong command-ratio-value-reference relation survives both estimands. The command-ratio-node relation is less uniform when each behavior receives one vote, illustrating why class multiplicity cannot be treated as an innocuous implementation detail. Cost-coordinate associations remain grid-weighted because a behavior class has no unique delete/insert/replace/match tuple.

VII. THREATS TO VALIDITY

A. Language and Resource Identity Are Confounded

All 20 default inputs are UniMorph-derived and predominantly describe inflectional morphology [11]. Because one default resource is evaluated per language, observed differences cannot be uniquely attributed to linguistic typology rather than resource construction, annotation conventions, coverage, or preprocessing. The campaign spans Latin, Cyrillic, Hebrew, and Arabic-derived scripts and several language families, but it is not a controlled typological sample.

B. Stemming, Canonical Targets, and Ambiguity

The supervision supplies lemma-like canonical targets, but the model is context-free and search-oriented. A surface form that legitimately maps to different lemmas in different contexts cannot be fully resolved by a deterministic token-only stemmer. Exact-target differences therefore combine transformation quality with ambiguity and resource policy. The study does not claim contextual lemmatization or morphological analysis.

TABLE III

SECONDARY WITHIN-STRATUM SPEARMAN SUMMARIES UNDER TWO WEIGHTING ESTIMANDS. “POSITIVE” AND “NEGATIVE” COUNT SIGNS AMONG STRATA WHERE THE COEFFICIENT IS DEFINED. INTERVALS ARE EMPIRICAL DISTRIBUTIONS, NOT CONFIDENCE INTERVALS.

Predictor	Outcome	Weighting	Median ρ	Central 95%	Positive	Negative
Command ratio	Value references	Numeric grid	+0.988	[+0.263, +1.000]	894/900	6/900
Command ratio	Value references	Exact behavior	+0.987	[+0.414, +1.000]	899/900	1/900
Command ratio	Trie nodes	Numeric grid	+0.953	[+0.187, +1.000]	890/900	10/900
Command ratio	Trie nodes	Exact behavior	+0.937	[−0.152, +1.000]	836/900	64/900
Leaf paths	Changed exactness	Numeric grid	−1.000	[−1.000, +0.126]	15/502	487/502
Leaf paths	Changed exactness	Exact behavior	−1.000	[−1.000, −0.140]	5/502	495/502
Leaf paths	Under-stemming	Numeric grid	+1.000	[−0.158, +1.000]	487/517	30/517
Leaf paths	Under-stemming	Exact behavior	+1.000	[+0.056, +1.000]	504/517	13/517

C. Model-Class and Traversal Bias

The experiment is conditional on Radixor’s backward-traversed patch-command representation, deterministic ranking, and reduction semantics. The same lexical mappings encoded by a prefix-oriented, bidirectional, context-sensitive, or different edit-script model could partition the cost grid differently. This limitation is especially relevant for resources with substantial prefixal or non-concatenative morphology.

D. Finite and Nonuniform Cost Grid

The 234 normalized settings form a broad but finite and nonuniform discrete grid. Exact classes, quality ranges, structural minima, and operating points are complete for that grid but not for all positive relative costs. Unmeasured ratios may create new behaviors or wider trade-offs. Removal of common positive scaling is exact, but the remaining labels do not define a probability distribution over all meaningful costs.

E. Selection and Reporting Use the Same Matrix

The practical operating-point and guarded-quality selectors use the same campaign for selection and reporting. Their results are descriptive observed optima, not unbiased estimates for a new resource or domain. The 0.25 pp tolerance is an engineering design choice, not a universal indifference threshold. The full-grid range and structural-minimum envelope do not depend on this selector.

F. Weighting, Dependence, and Descriptive Intervals

Nominal-grid and uniform-behavior coefficients answer different questions. Neither turns the 900 partial-coverage strata into independent observations: strata share resources, nested training prefixes, and related cost settings. Central empirical intervals summarize heterogeneity rather than sampling uncertainty. Across-resource correlations use only 20 configurations; the paired bootstrap for their coefficient difference is descriptive and does not create a population sample of languages.

G. Representation Proxies Are Not Runtime Measurements

Distinct commands, trie nodes, value references, and path statistics describe a compiled representation. They do not directly measure heap occupancy, serialized artifact size, CPU time, cache locality, or lookup latency. Any claim that a structurally smaller model is faster or smaller in bytes requires a separate systems experiment.

H. Coverage-Conditioned Generalization and Downstream Utility

Coverage levels use nested training prefixes but different held-out complements. Differences between 10% and 90% coverage combine changes in training evidence with changes in the evaluation population; the paper does not interpret them as causal learning effects. Cost settings are compared only within matched strata. Likewise, exact canonical-target agreement and pairwise conflation metrics are not direct estimates of retrieval MAP, NDCG, or application-specific utility.

VIII. REPRODUCIBILITY

The publication snapshot is Radixor commit `0c3b13f`. Measurements were produced by Radixor 4.2.0-6-g84e57fb at core revision `84e57fb`; the source state is explicitly recorded as *dirty*. The protocol identifier is `radixor-cost-sensitivity-v4`. Full SHA-256 identities for the uncompressed raw campaign, generator source, and repository analysis source are recorded in the accompanying provenance manifest.

The article bundle includes the compressed 16,700-row physical campaign, a deterministic analysis script, and machine-readable outputs for the full-grid quality range, weighting sensitivity, structural-minimum envelopes, paired operating points, and resource-level bootstrap. The script validates that every physical class is `VIABLE`, reconstructs all member labels, and regenerates the reported summaries. Exact equivalence remains auditable from complete command sequences: hashing is not accepted as proof of equality [8].

IX. CONCLUSION

Across 20 UniMorph-derived language–resource configurations, 234 sampled relative-cost settings map to far fewer exact induced behaviors. The 4,680 resource–cost combinations collapse to 334 command-sequence classes, enabling a $14.01 \times$ exact reduction of the experimental model space. This result follows from subset invariance of exact command equivalence rather than from statistical similarity.

The complete sampled-grid quality analysis shows that typical changed-exact sensitivity is small but not universally zero: across 900 matched strata the median range is 0.0067 pp and the 95th percentile 0.360 pp, while the largest single-stratum range is 2.577 pp. At the resource level, eight configurations have a zero median range and the largest resource median is 0.380 pp. The full-grid structural analysis identifies smaller command vocabularies for 17 resources, with reductions above

24%. All tied structural-minimum behaviors agree exactly on changed-form exactness in every matched stratum, and the largest median paired exactness difference from the baseline is +0.041 pp.

A secondary operating-point analysis yields paired-median dominance for 15 resources, negative paired medians for 2, and the baseline for 3. Association sensitivity further shows that conclusions can depend on whether numeric grid labels or distinct induced behaviors receive equal weight, although the relation between retained commands and stored value references is strong under both estimands.

The evidence supports a precise conclusion: edit costs are structural hyperparameters because they determine which reusable transformations are induced and compiled. The sampled grid contains substantial resource-specific structural redundancy together with localized linguistic sensitivity. The study does not establish a universal tuning rule, contextual lemmatization quality, downstream retrieval gains, or runtime benefits from reduced structural counts.

REFERENCES

- [1] R. A. Wagner and M. J. Fischer, “The string-to-string correction problem,” *Journal of the ACM*, vol. 21, no. 1, pp. 168–173, 1974.
- [2] 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, 2008.
- [3] T. Müller, R. Cotterell, A. Fraser, and H. Schütze, “Joint lemmatization and morphological tagging with lemming,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 2268–2274.
- [4] O. Toporkov and R. Agerri, “Evaluating shortest edit script methods for contextual lemmatization,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, 2024, pp. 6451–6463.
- [5] E. S. Ristad and P. N. Yianilos, “Learning string-edit distance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 5, pp. 522–532, 1998.
- [6] 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.
- [7] A. Chakrabarty, O. A. Pandit, and U. Garain, “Context sensitive lemmatization using two successive bidirectional gated recurrent networks,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1481–1491.
- [8] L. Galambos, “Radixor source, generalization, and edit-cost experiment snapshot,” GitHub repository, commit 0c3b13f485a9ad0b460c0931e4497ea95bed66a1, 2026, cost-sensitivity protocol: radixor-cost-sensitivity-v4. Accessed Aug. 26, 2026. [Online]. Available: <https://github.com/leogalambos/Radixor>
- [9] 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.
- [10] 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.
- [11] K. Batsuren, O. Goldman, S. Khalifa, N. Habash, W. Kieraś, G. Bella, B. Leonard, G. Nicolai, K. Gorman *et al.*, “UniMorph 4.0: Universal morphology,” in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Marseille, France: European Language Resources Association, 2022, pp. 840–855.