

Lexical Resource Consistency and Learned Transformation Vocabulary in Multilingual Stemming

Leo Galambos

Abstract—Lexicon-trained stemmers convert lexical assignments into an executable vocabulary of transformations, so inconsistencies in a training resource can survive as exceptional learned commands. We study this propagation in Radixor using a frozen audit of 12 resource configurations whose baseline encoder produced more than 500 distinct patch commands. The ordered audit combines source-structural sanitation with resource-internal assignment dominance; it does not use an external lexicon to decide multilingual removals. Ten non-destructive candidates are materialized. Nine use the standard grouped rule and German uses a deliberately less conservative individual rule. Among the grouped-rule candidates, 8 of 9 reduce distinct command vocabulary; the sum of per-resource vocabularies changes from 20,281 to 19,842 commands (2.165%), with a median per-resource reduction of 1.334%. German is reported separately and reduces its command vocabulary by 23.948%, accounting for 79.2% of the unstratified command drop. The informative result is the non-one-to-one response: Finnish removes 3 dominance-qualified assignments but loses no command because the affected commands remain supported elsewhere, whereas Hebrew removes no dominance-qualified assignment yet loses 41 commands through source-structural sanitation. Polish UniMorph loses 3 commands while its packaged GZip artifact grows by +121 bytes; this is a packaging effect because candidate materialization adds a fixed 237-byte UTF-8 provenance block and reserializes parser-accepted rows. German is evaluated separately against two related CELEX2-derived CISTEM policies; micro CISTEM-style F1 increases by 0.001973 and 0.002769. The bounded conclusion is that lexical evidence affects learned transformation vocabulary through command-support topology. Command count can prioritize audit work, but threshold crossing alone does not establish resource error, and command vocabulary is not a substitute for trie size, memory, latency, or downstream retrieval effectiveness.

Index Terms—stemming, lexical resources, data quality, morphology, patch commands, intrinsic evaluation, UniMorph, CISTEM, CELEX2, Radixor

I. INTRODUCTION

LEXICAL resources are often described as static inputs to an algorithm. In a lexicon-trained system that separation is incomplete: the resource determines which transformations the learner must represent. A questionable form-to-target assignment can therefore become an executable exception, but the relationship is not one-to-one. Several assignments may induce the same transformation, and a transformation disappears only when its last supporting assignment disappears or is recoded. Understanding that support structure is necessary before using learned complexity as evidence about data quality.

This paper studies that problem in lexicon-trained stemming, an unusually inspectable setting. Classical stemmers such as Porter execute hand-specified suffix rules [1]. The Egothor lineage instead learns edit commands from lexical mappings

and stores their evidence in a trie [2], [3]. Radixor modernizes that design around independently versioned lexical artifacts and compiled patch-command tries [4]. The training dictionary is not merely queried at runtime; it induces the transformation vocabulary from which runtime behavior is constructed.

The audit problem has two opposing risks. A naive cleaner can mistake legitimate morphology, homography, suppletion, lexicalized derivation, or annotation policy for error. A naive complexity analysis can make the opposite mistake and treat every large command vocabulary as evidence of a defective resource. The frozen Radixor audit therefore uses an evidence hierarchy. Source-structural sanitation acts only on directly inspectable source artifacts. Assignment-dominance removal requires a competing assignment for the same surface form inside the same resource, a substantially smaller orthographic edit distance, and command-frequency support. Except for German, at least three dominated forms from one source row must agree on the same alternative target. Cases without sufficient evidence remain in a review queue; this is an abstention mechanism rather than a declaration that retained assignments are correct.

The study covers 12 resource configurations selected by one operational criterion: the fixed backward *D1I1R1M0* encoder produced more than 500 distinct commands. This threshold defines an audit queue, not an error threshold. Ten candidates are materialized, while Dutch and Persian admit no safe automatic change and are no-safe-change cases showing that the $K > 500$ threshold is insufficient, by itself, to justify automatic modification.

The research questions are:

- RQ1: How often and by how much does conservative resource-local auditing change the distinct transformation vocabulary?
- RQ2: What support mechanisms explain non-intuitive outcomes such as mapping removal without command reduction, sanitation-only command reduction, and command reduction accompanied by packaged-file growth?
- RQ3: For German, where the individual dominance rule is intentionally less conservative, does the frozen candidate regress under external-source CELEX2/CISTEM relations?
- RQ4: Which negative and ambiguous cases constrain the use of command vocabulary as an audit signal?

The principal result is a support-topology result, and intervention strata must be kept separate. Under the standard grouped rule, 8 of 9 materialized candidates reduce K ; the sum of the per-resource vocabularies falls from 20,281 to 19,842 (2.165%), and the median per-resource reduction is 1.334%. German uses

a different individual rule and is therefore not pooled into that headline; its own reduction is 23.948%. Finnish removes three dominance-qualified assignments with $\Delta K = 0$ because no affected command loses its last support. Hebrew removes no dominance-qualified assignment yet loses 41 commands through source-structural sanitation. German’s ordered pipeline removes 2,055 dominance-qualified assignments and yields 1,673 fewer commands, but the frozen evidence does not stage-decompose that change. Polish UniMorph produces the most important packaging paradox: three commands disappear while the checked-in candidate GZip grows by 121 bytes. Inspection of the frozen generator explains why: it adds a fixed 237-byte provenance block before compression and reconstructs the file from parser-accepted rows. The byte delta therefore cannot be interpreted as lexical-payload growth, trie growth, or failed reduction.

German supplies the only external-source relation-level check in this paper. The two CISTEM gold standards are related policies derived from the same CELEX2 source, not independent replications. The filtered candidate changes scorer TP/FP/FN incidences by +762/-18/-762 on GS1 and +1,478/-21/-1,478 on GS2, increasing micro CISTEM-style F1 by 0.001973 and 0.002769. This is directional corroboration across two related conflation policies; it does not establish that every removed lexical assignment is linguistically incorrect or that the two standards constitute independent validation datasets.

The paper contributes: (i) a formal command-support view of resource-induced transformation vocabulary; (ii) a reproducible 12-resource audit census with positive, zero-change, and no-safe-change outcomes; (iii) stratified analysis separating the standard grouped rule from the German individual rule; (iv) mechanistic explanations for the major non-intuitive results; (v) a correction to the tempting but invalid use of packaged GZip size as a controlled lexical-footprint metric; and (vi) an external-source German case study with explicit claim boundaries.

II. BACKGROUND AND NOVELTY BOUNDARY

A. Lexicon-Trained Stemming

Porter’s algorithm established a compact rule-based approach to English stemming [1]. Rule systems are attractive because the executable logic is directly authored and its size is largely independent of the corpus or dictionary used to evaluate it. In lexicon-trained stemming, the direction is reversed: lexical examples define desired mappings, an edit encoder converts those mappings into operations, and a data structure generalizes the induced operations to unseen forms [2]. Galambos also emphasized evaluation against lexical relationships rather than treating output strings in isolation [3].

Radixor represents each lexical row as a canonical target followed by associated forms. Under fixed edit costs and backward traversal, a form-to-target pair is encoded as a patch command. Repeated identical commands are shared, while the trie organizes the evidence used to select a preferred command. This makes the number of distinct commands a meaningful, deterministic property of the learned transformation vocabulary under a fixed encoder. It is not, however, equivalent to trie size, serialized dictionary size, or runtime latency.

B. Morphological Resource Provenance

The default Radixor resource family is predominantly derived from UniMorph datasets [5]; the audited Polish alternative uses PoliMorf [6]. These resources are scientifically valuable precisely because they expose large-scale morphological paradigms, but their provenance is heterogeneous. The frozen Radixor model catalog records model IDs, licenses, upstream repositories, source identities where known, and checksums [4]. Several legacy UniMorph imports lack the exact historical upstream commit used at initial conversion, a limitation retained in the model notices rather than retroactively reconstructed.

Direct precedent exists for semi-automatic morphological-resource vetting. UniMorph 3.0 introduced validation tooling that flags suspicious observations such as non-lexical fields, forms repeated across paradigms, and large lemma-form edit distance for human inspection rather than automatic truth assignment [7]. Language-specific reanalysis has also documented systematic errors that require linguistic knowledge unavailable to generic heuristics, for example in Central Kurdish UniMorph data [8]. The present audit therefore treats orthographic and frequency signals as evidence constraints, not semantic adjudication. Its additional contribution is downstream traceability: suspicious resource-local assignments are connected to the patch commands they support in a concrete learned stemmer.

Dataset documentation frameworks argue that provenance, construction decisions, intended uses, and limitations are necessary parts of reproducible NLP [9], [10]. The present study adds an executable perspective: if a resource is used to induce transformation logic, one can ask which source observations are responsible for structurally exceptional learned behavior. This is not equivalent to proving that the source annotation is wrong. An internal inconsistency can reflect homography or policy rather than error, and an unusual command can reflect legitimate morphology.

C. Stemming Evaluation

Stemming quality has long been evaluated through over-stemming and under-stemming errors or through more task-specific criteria [11], [12]. Intrinsic stemming quality and downstream retrieval effectiveness are related but not interchangeable [13]. CISTEM provides two German gold standards constructed from CELEX2 and illustrates a further issue: there is no unique universally correct conflation policy for all retrieval applications [14], [15]. That makes external validation valuable, but only when the scorer and policy are described precisely.

D. Novelty Boundary

This paper does not propose a general lexical-error detector, a new morphological annotation standard, or a new stemming metric. It also does not claim that a model with fewer commands is linguistically better. Its contribution is narrower: a traceable case study of how conservative, input-internal consistency evidence propagates into the transformation vocabulary of a modern lexicon-trained stemmer, with explicit no-safe-change outcomes and an external-source German validation

layer. The unit of structural analysis is the distinct patch-command vocabulary under one fixed encoder. Candidate trie-node/edge counts are outside the frozen article evidence and are therefore not reported; the measured structural quantity is transformation-vocabulary size.

III. FROM LEXICAL MAPPING TO LEARNED COMMAND

Let a parsed lexical resource D contain rows whose first field is a canonical target s and whose remaining fields are associated forms x . After the production case policy used by the audit, define the multiset of observed assignments

$$M(D) = \{(x, s)\}. \quad (1)$$

For the fixed backward encoder E with delete, insert, replace, and match costs $(1, 1, 1, 0)$, an assignment induces a command

$$q(x, s) = E(x, s). \quad (2)$$

For a command c , define its support count

$$n_D(c) = \sum_{(x,s) \in M(D)} \mathbf{1}[q(x, s) = c]. \quad (3)$$

The learned transformation vocabulary is then

$$C(D) = \{c : n_D(c) > 0\}, \quad K(D) = |C(D)|. \quad (4)$$

$K(D)$ is the primary structural observable in this paper. It counts distinct transformations required by the fixed training representation. It is not a count of rows, trie nodes, bytes, or operations executed per token.

If an intervention were a pure deletion under an otherwise unchanged encoding, a command c would disappear exactly when $n_D(c) > 0$ and $n_{D'}(c) = 0$. Thus one removed mapping need not change K , while many removed mappings can collapse to fewer disappearing commands because they share support. The frozen audit is slightly more general than pure deletion: source sanitation can remove or normalize assignments before command frequencies are recomputed. Consequently, a whole-audit command change can also arise from recoding or command collapse. The article does not possess an immutable staged census $K_{raw} \rightarrow K_{sanitized} \rightarrow K_{filtered}$, so mixed cases are not decomposed causally between sanitation and assignment-dominance filtering.

For the final candidate D' , transformation reduction is

$$R_K(D, D') = 100 \frac{K(D) - K(D')}{K(D)}. \quad (5)$$

For a resource stratum $\mathcal{S}$, we also report the resource-command incidence total and its relative change,

$$S_K(\mathcal{S}) = \sum_{D \in \mathcal{S}} K(D), \quad (6)$$

$$R_\Sigma(\mathcal{S}) = 100 \frac{S_{K,0}(\mathcal{S}) - S_{K,1}(\mathcal{S})}{S_{K,0}(\mathcal{S})}. \quad (7)$$

The same command string appearing in two resources contributes twice to S_K ; this is a sum of per-resource vocabularies, not a global union. We pair R_Σ with the median of per-resource R_K values. Because German uses a different automatic-removal rule, the grouped-rule candidates and German are reported as

separate intervention strata. Neither summary is a population estimator.

The repository also records the byte sizes of baseline and candidate `stemmer.gz` files. We denote these observed package sizes by $B_{pack}(D)$. They are *not* a controlled lexical-payload metric. Candidate materialization preserves the leading source header, adds a fixed provenance block, and reserializes production-parser-accepted rows before GZip compression. Raw B_{pack} deltas are therefore reported only to explain packaging behavior, never as a measure of transformation complexity, compiled-trie size, or memory. A controlled footprint experiment would require both baseline and candidate to pass through the same canonical serializer with identical metadata policy.

Figure 1 summarizes the ordered evidence path. The registered resource remains immutable; auditing creates a separate candidate; K is recomputed from the final candidate; German alone receives an external-source evaluation layer.

IV. AUDIT METHODOLOGY

A. Frozen Audit Universe

The article uses repository snapshot `0c3b13f` [4]. The Gradle task `dictionaryMeaningAudit` explicitly enumerates the 12 resource configurations in Table I; each had more than 500 distinct baseline commands under the same backward `D1I1R1M0` encoder. The universe is a purposive high-command audit set, not a random sample. Polish appears twice because `pl-pl-unimorph` and `pl-pl-polimorf` are distinct model inputs.

The audit does not modify registered model inputs. Ten configurations have a destination candidate module; Dutch and Persian are intentionally destination-free because no automatic change met the conservative rule. Candidate modules are excluded from published defaults and packages [4].

B. Ordered Audit Pipeline

The intervention order is part of the method:

$$\begin{aligned} \text{parse} &\rightarrow \text{source-structural sanitation} \rightarrow \text{lower-case} \\ &\rightarrow \text{command frequencies} \rightarrow \text{dominance} \rightarrow \text{group rule}. \end{aligned} \quad (8)$$

This ordering matters because sanitation can alter which commands exist and how often they occur; command frequency is then one of the dominance criteria. The stages are not assumed to commute.

C. Source-Structural Sanitation

Before assignment-dominance analysis, the audit handles source properties that are directly evidenced by the checked-in resource rather than by an external lexicon. The frozen rules remove Unicode format controls without lexical content, isolated export markers, an embedded `Citations:` source label, non-lexical dash placeholders, known metadata fields embedded in one Czech row, and a Czech punctuation artifact. These decisions are deterministic and source-specific; “source-structural” does not mean language-independent or universally correct.

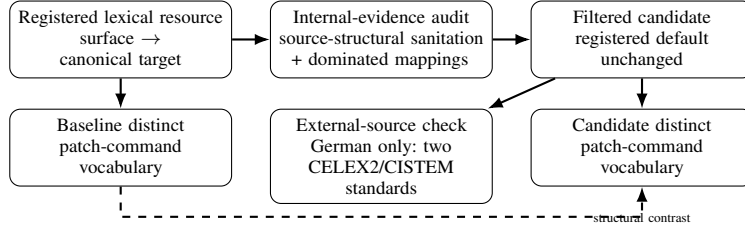


Fig. 1. Ordered evidence path used in the study. Source-structural sanitation precedes command-frequency analysis and assignment dominance; the stages are therefore not commutative. Candidate modules are non-destructive. External-source validation is available only for German.

TABLE I

FROZEN INTERNAL-EVIDENCE AUDIT CENSUS. K_0 AND K_1 ARE DISTINCT PATCH-COMMAND COUNTS BEFORE AND AFTER THE COMPLETE ORDERED AUDIT. “REMOVED” COUNTS DOMINANCE-QUALIFIED LEXICAL ASSIGNMENTS REMOVED BY THE AUDIT RULE. SOURCE ARTIFACTS USE HETEROGENEOUS FIELD/ROW UNITS AND ARE NOT SUMMED. “REVIEW” COUNTS RETAINED CASES QUEUED FOR INSPECTION, NOT CONFIRMED ERRORS.

Resource	K_0	K_1	ΔK	Removed	Review
Czech	537	468	-69	26	98
Dutch	538	538	+0	0	176
Persian	704	704	+0	0	385
Italian	750	705	-45	56	193
Polish UniMorph	846	843	-3	5	224
Spanish	1,496	1,458	-38	8	183
Russian	1,840	1,821	-19	28	286
Finnish	2,683	2,683	+0	3	604
French	2,730	2,576	-154	75	1,115
Hebrew	4,150	4,109	-41	0	2,247
Polish PoliMorf	5,249	5,179	-70	104	1,686
German	6,986	5,313	-1,673	2,055	960

The production parser also reports whitespace-bearing rows or variants it would reject before command encoding. In PoliMorf, 50 rows have whitespace in the canonical field; they contribute to neither baseline nor filtered command vocabulary [4]. Because candidate materialization rewrites only parser-accepted rows, such rejected raw data can nevertheless change packaged-file bytes. This is one reason package size is not used as a lexical-payload metric.

The source documentation’s “Artifacts” column mixes field and row units. The paper therefore does not sum artifact counts or model them as a continuous variable.

D. Internal Assignment Dominance

After sanitation and production-equivalent lower-casing, the audit computes command frequencies over the sanitized resource. For each surface form x , it selects an internally observed best assignment using, in order, smaller edit distance, synthetic identity preference, higher command frequency, and deterministic target ordering. Let the current mapping be (x, s) and the best competing internal assignment be (x, s^*) .

The current mapping is strongly dominated only if

$$s \neq s^*, \quad (9)$$

$$f(q(x, s)) \leq 2, \quad (10)$$

$$d(x, s) \geq 4, \quad (11)$$

$$d(x, s) - d(x, s^*) \geq 2, \quad (12)$$

$$2d(x, s^*) \leq d(x, s), \quad (13)$$

$$f(q(x, s^*)) \geq f(q(x, s)). \quad (14)$$

The frozen implementation computes d as Levenshtein distance over Java UTF-16 `char` units. It is an orthographic implementation criterion, not grapheme, morpheme, phonological, or semantic distance. The thresholds are frozen constants rather than values fitted in this article [4]. This design is intentionally conservative relative to generic suspiciousness heuristics. UniMorph’s own semi-automatic vetting treats frequent cross-paradigm forms and large lemma–form edit distance as reasons for manual inspection [7]; the Radixor rule additionally requires a competing assignment of the *same surface form*, command-frequency support, and—outside German—multi-form agreement within one source row. Even these stronger conditions identify resource-internal dominance, not linguistic truth.

For every resource except German, automatic removal additionally requires at least three dominance-qualified forms from one source row to agree on the same alternative target. This is a group-consensus heuristic, not proof that the row is a linguistically coherent paradigm or that the alternative target is canonical. German uses minimum group size one and is analyzed separately. Rare divergent cases without sufficient evidence are retained for review. The 523 German quote-plus-vowel spellings are also retained because earlier normalization changed command frequencies and ranking, reinforcing that pipeline ordering is substantive [4].

E. Candidate Materialization and Packaging Boundary

A candidate preserves the leading source comment header, applies the ordered audit, and writes a new GZip lexical artifact. The writer then inserts four candidate-specific provenance lines

whose UTF-8 representation is 237 bytes before compression [4]. It also serializes parsed rows rather than reproducing the original raw byte stream. Hence baseline and candidate GZip sizes answer a packaging question, not a controlled payload-size question.

The “Review” counts in Table I are algorithmically retained candidates. They are not estimated error counts; human adjudication is outside this paper.

V. MEASURES AND EXTERNAL VALIDATION

A. Transformation-Vocabulary Measures

Equation (5) is evaluated for all ten materialized candidates. Candidate command counts are recomputed under the same fixed encoder as the baseline. The primary aggregate summaries are stratified by intervention rule. For the nine grouped-rule candidates we report R_Σ from Eq. (7) and the median per-resource R_K ; German’s individual-rule result is reported separately. The unstratified ten-candidate total is retained only as a reproducibility diagnostic because German contributes 79.2% of its command drop. None of these finite-census summaries is a success probability or a language-population estimate.

The frozen implementation can generate trie nodes, edges, value references, and logical path statistics, but the repository snapshot does not version the corresponding generated `audit-summary.csv`. Those counters are therefore unavailable as immutable article inputs. We do not reconstruct them from unrelated experiments. The measured quantity is transformation-vocabulary size K , not complete compiled-model complexity.

Observed baseline/candidate GZip sizes are retained in the derived package only as B_{pack} diagnostics. No aggregate byte-reduction estimand is used in the scientific conclusions because the serializers are not controlled to be identical.

B. German External-Source Case Study

German receives a stronger evaluation because its individual dominance-removal rule omits the three-form group condition. The two CISTEM gold standards are derived from CELEX2 rather than from Radixor’s UniMorph-derived German training input [14], [15]. This separates the evaluation source from the lexical input used to select removals, although the frozen evidence does not establish a preregistered rule-selection chronology. Both standards arise from the same CELEX2 resource and encode related, differently aggressive conflation policies; they are therefore two policy probes, not independent replications. The result is treated as case-study corroboration, not as an unbiased estimate of a general filtering rule.

The benchmark uses a documented micro CISTEM-style adaptation. For each gold cluster it identifies the most frequent predicted stem as the main stem; cluster members mapped to it contribute TP, members mapped elsewhere contribute FN,

and words from other clusters mapped to the same selected stem contribute FP. We recompute

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad (15)$$

$$F_1 = \frac{2TP}{2TP + FP + FN}. \quad (16)$$

The TP/FP/FN quantities are scorer-defined cluster incidences, not a conventional unique-token confusion matrix. A predicted stem can contribute false-positive incidence to more than one gold cluster when cluster main stems collide. The formulas in Eq. (16) are therefore interpreted only within this frozen scorer. This scorer is not numerically interchangeable with the macro evaluation in the original CISTEM paper. CISTEM itself is contextual only.

C. Inferential Scope

The resource set is fixed and purposively selected. There is no random-language sampling model and no population-level p -value. Command counts and German quality deltas are deterministic functions of the frozen resources and code. Claims are finite-census and case-study statements. Agreement in delta direction across GS1 and GS2 is robustness across two related CELEX2-derived policies, not replication across independent samples.

VI. RESULTS

A. Multilingual Audit Outcomes

The audit produces ten candidate resources and two no-safe-change outcomes. Nine candidates use the standard grouped rule; German uses the individual rule. Within the grouped-rule stratum, 8 of 9 candidates reduce distinct command vocabulary. The resource-command incidence total S_K changes from 20,281 to 19,842, giving $R_\Sigma = 2.165\%$, while the median per-resource reduction is 1.334%. German is reported separately at 23.948%. If the two intervention strata are naively pooled, German accounts for 79.2% of the total command drop; this dominance is precisely why the pooled 7.746% figure is retained only as a reproducibility diagnostic rather than a headline estimand.

Dutch and Persian begin with 538 and 704 commands but admit no automatic change under the frozen rule. This shows that threshold crossing is insufficient evidence for automatic intervention. It does not establish that either resource is error-free, because the audit can abstain when its evidence conditions are not met.

B. Why the Non-Intuitive Cases Occur

Table II summarizes the cases that are most informative mechanistically.

Finnish is the cleanest illustration of Eq. (3): 3 dominance-qualified assignments are removed but $\Delta K = 0$ because every affected command retains at least one supporting assignment. German demonstrates the same unit mismatch at scale: its ordered pipeline removes 2,055 dominance-qualified assignments and finishes with 1,673 fewer distinct commands. Because the frozen projection lacks staged K values, that final

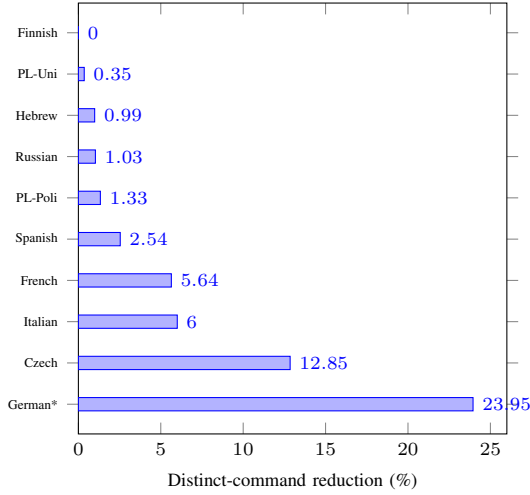


Fig. 2. Per-resource reduction in distinct patch-command vocabulary. German* uses the individual rule and is not pooled with the grouped-rule headline. Each value is a whole-audit effect, not a decomposition between sanitation and assignment-dominance filtering.

TABLE II
MECHANISMS BEHIND THE PRINCIPAL NON-INTUITIVE OUTCOMES.

Case	Mechanistic interpretation
Finnish	3 dominance-qualified assignments are removed but $\Delta K = 0$: each affected command remains supported by retained assignments.
Hebrew	No dominance-qualified assignment is removed, yet K falls by 41; the change is attributable to source-structural sanitation rather than the dominance predicate.
PL-UniMorph	K falls by 3, while the packaged GZip grows by +121 B. Candidate materialization adds a 237-byte UTF-8 provenance block before compression and reserializes parser-accepted rows, so package bytes are not a payload-size metric.
German	The ordered pipeline removes 2,055 dominance-qualified assignments and yields a final vocabulary with 1,673 fewer distinct commands; the frozen data do not stage-decompose the drop.
Dutch/Persian	Both exceed the audit-selection threshold but admit no safe automatic change, showing that threshold crossing alone is insufficient evidence for automatic intervention; it does not prove either resource error-free.

German command change is not attributed solely to assignment removal. Mapping count and command count remain different units because support is many-to-one and sanitation precedes command-frequency analysis.

Hebrew shows the converse source of confusion. No dominance-qualified assignment is removed, yet K falls by 41. Its command change is produced by the source-structural sanitation stage. More generally, Czech, Spanish, and French combine sanitation with dominance-qualified assignment removal, and the frozen article projection lacks a staged K value between those stages. Their total command reductions therefore cannot be quantitatively attributed to assignment-dominance filtering alone.

Polish UniMorph explains the apparent file-size paradox. Its command vocabulary falls by 3, while the checked-in candidate GZip grows by +121 bytes. The candidate writer adds a fixed 237-byte UTF-8 provenance block before compression and reconstructs parser-accepted rows rather than preserving the original byte stream. In addition, DEFLATE output depends on the resulting serialized sequence, not only on its semantic

TABLE III
PACKAGING DIAGNOSTICS. $\Delta B_{pack} = B_{candidate} - B_{baseline}$; POSITIVE MEANS THE CHECKED-IN CANDIDATE GZIP IS LARGER. CANDIDATE AND BASELINE SERIALIZATIONS ARE NOT CONTROLLED TO BE IDENTICAL, SO THESE BYTE DELTAS MUST NOT BE INTERPRETED AS LEXICAL-PAYLOAD OR COMPILED-MODEL REDUCTIONS.

Candidate	ΔK	ΔB_{pack} (B)	Dominance removals
German	-1,673	-4,504	2,055
Czech	-69	-10	26
Italian	-45	-855	56
French	-154	-1,451	75
Spanish	-38	-4,098	8
Polish PoliMorf	-70	-15,299	104
Russian	-19	-1,265	28
Hebrew	-41	-786	0
Polish UniMorph	-3	+121	5
Finnish	+0	-2,006	3

content. The +121-byte observation is therefore a packaging consequence under changed serialization. It is *not* evidence that the lexical payload, trie, memory image, or “data structure after reduction” grew.

The same boundary applies in the other direction. PoliMorf’s candidate packaging omits 50 raw whitespace-canonical rows that the production parser already ignored. Such bytes can disappear from the package without changing either baseline or candidate command vocabulary. Packaged file size and learned transformation vocabulary are therefore deliberately separated in the revised analysis.

For completeness, Table III preserves the exact package-byte deltas as provenance diagnostics, not as a structural metric.

C. German External-Source Evaluation

German is the largest dominance-qualified assignment intervention. Table IV recomputes all values from frozen TP/FP/FN counts.

On GS1, the filtered candidate changes scorer TP/FP/FN incidences by +762/-18/-762, with $\Delta P = 0.000178$, $\Delta R = 0.002400$, and $\Delta F_1 = 0.001973$. On GS2, the corresponding incidence changes are +1,478/-21/-1,478, with $\Delta P = 0.000438$, $\Delta R = 0.004767$, and $\Delta F_1 = 0.002769$. Thus the scorer improvement is primarily recall recovery while precision also increases slightly. The same direction under GS1 and GS2 is corroboration across two related CELEX2-derived policies, not two independent replications.

VII. DISCUSSION

A. Support Topology Governs Command Vocabulary

The central systems result is now more precise than “cleaner data gives a smaller model.” Distinct command vocabulary is governed by support topology. A removed assignment changes K only when it eliminates the last support of a command, while sanitation can recode assignments and collapse commands. Finnish, German, and Hebrew expose three different mechanisms. This is why the number of resource edits cannot be read directly as the number of learned transformations removed. The stratified analysis adds a second caution: German’s individual

TABLE IV
GERMAN EXTERNAL-SOURCE CELEX2/CISTEM EVALUATION. VALUES USE THE FROZEN RADIXOR MICRO CISTEM-STYLE SCORER DESCRIBED IN SEC. V-B; THEY ARE NOT THE ORIGINAL PUBLICATION’S MACRO CISTEM SCORES.

Gold	System	TP	FP	FN	Precision	Recall	F1
GS1	Registered Radixor	179,879	3,638	137,562	0.980176	0.566653	0.718140
GS1	de-de-filtered	180,641	3,620	136,800	0.980354	0.569054	0.720113
GS1	CISTEM	192,893	26,500	124,548	0.879212	0.607650	0.718632
GS2	Registered Radixor	266,098	20,724	43,969	0.927746	0.858195	0.891616
GS2	de-de-filtered	267,576	20,703	42,491	0.928184	0.862962	0.894386
GS2	CISTEM	286,002	83,103	24,065	0.774853	0.922388	0.842208

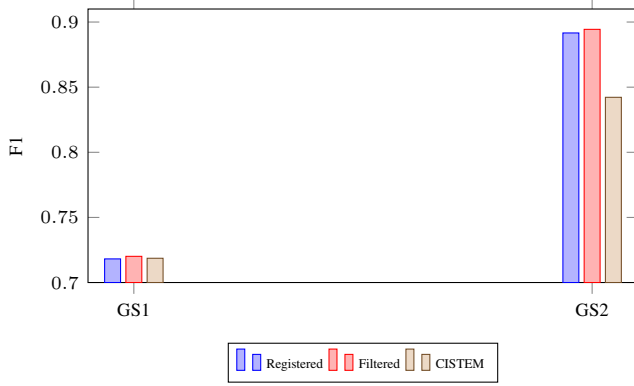


Fig. 3. German micro CISTEM-style F1 under the two external-source CELEX2-derived standards. The figure is a case-study corroboration across two related CELEX2-derived policies, not cross-language validation or independent replication of the audit rule.

rule is substantially more permissive than the grouped rule and accounts for 79.2% of the unstratified command drop, so pooled totals obscure intervention heterogeneity.

This support view also clarifies what the audit can and cannot claim. A rare command may be perfectly legitimate. A common command may be supported partly by questionable mappings. Command vocabulary is therefore an inspectable intermediate representation, not a correctness oracle.

B. Command Vocabulary Is a Triage Signal

Dutch and Persian are scientifically important because they fail to produce a candidate despite crossing the same $K > 500$ audit threshold. The threshold prioritizes inspection; the dominance and group rules decide whether any change is sufficiently supported. Their no-safe-change outcome demonstrates only that $K > 500$ is insufficient evidence for this automatic intervention, not that the resources contain no defects. The large retained review queues in several resources make the same point: review volume measures unresolved evidence, not error prevalence.

C. Why Packaged Bytes Were Demoted

The original analysis compared command reduction with compressed `stemmer.gz` reduction. Hard review showed that this comparison was not controlled: candidate generation adds metadata, reserializes parser-accepted rows, and can omit raw rows that runtime had already ignored. The Polish UniMorph +121-byte result is therefore valuable precisely because it

exposes the measurement flaw. Rather than forcing a post-hoc explanation that “compression behaves strangely,” the revised paper treats package bytes only as packaging provenance.

A future source-footprint experiment should canonicalize baseline and candidate through the same serializer, remove or equalize comments, and then compare uncompressed canonical payload bytes and deterministic recompression under the same compressor settings. A future executable-structure experiment should separately version trie nodes, edges, paths, value references, and runtime memory. Those are different questions.

D. Evidence Hierarchy and Data Governance

Dataset documentation frameworks emphasize provenance and transformation history [9], [10]. Morphological-resource work also shows why automatic certainty is inappropriate: UniMorph’s validation tooling flags suspicious observations for human inspection [7], and later language-specific auditing has required linguistic repair beyond generic heuristics [8]. Lexicon-trained systems add another useful layer: maintainers can connect source observations to the commands they support. The Radixor audit does this non-destructively and keeps uncertain cases in a review queue. This traceability is useful even when a candidate is never promoted.

The model catalog also preserves a limitation that the article cannot repair: several legacy UniMorph imports lack the exact historical upstream revision used at original conversion. The frozen Radixor bytes and present audit code are reproducible even when upstream historical reconstruction is incomplete.

E. What This Paper Does Not Establish

The article does not transfer runtime results from registered defaults to filtered candidates. It has no immutable before/after candidate trie census, matched candidate JMH run, or memory measurement. A 23.948% reduction in German transformation vocabulary is not a 23.948% reduction in runtime work, memory, or compiled bytes.

Likewise, German intrinsic improvement under two external-source stemming policies does not imply higher MAP, nDCG, recall, or index compression. The two policies share a CELEX2 source and are not independent validation replications. Nor does relation-level improvement prove that each deleted training assignment was linguistically erroneous: the scorer evaluates conflation behavior, not lexical-sense correctness. Intrinsic stemming agreement and downstream IR effectiveness are not interchangeable [13]. Promotion of a candidate requires language review and task-relevant evaluation beyond this audit.

VIII. THREATS TO VALIDITY

A. Internal Evidence Is Not Independent Annotation

Dominance-qualified removals are selected using the same lexical resource that trains the model. “Internally dominated” is not synonymous with “linguistically wrong.” Homography, sense distinctions, derivation, annotation policy, and suppletion can create distant assignments legitimately.

B. Purposive Audit Selection

The 12 configurations are selected by $K > 500$ and are not representative samples of languages or lexicons. Reported reduction rates are finite-census descriptions only.

C. Intervention Heterogeneity and Aggregation

German uses an individual-removal rule whereas the other materialized candidates use the grouped rule. The paper therefore stratifies aggregate summaries by rule. The sum S_K counts resource-command incidences rather than a global union of command strings, and neither the number of candidates with $\Delta K < 0$ nor a median reduction is interpreted as a success probability.

D. Ordered Pipeline and Missing Staged Counts

Sanitation precedes frequency analysis and dominance, so changing sanitation can change later decisions. The frozen article projection does not contain staged command counts after sanitation but before assignment-dominance filtering. Whole-audit ΔK values for mixed cases therefore cannot be decomposed causally between stages.

E. Orthographic Distance Criterion

The dominance predicate uses Levenshtein distance over Java UTF-16 code units. It is neither grapheme-cluster distance nor a linguistic similarity metric. Its value is reproducible but language-dependent in interpretation.

F. German Uses a Different Rule

German uses minimum dominated-group size one, whereas other dominance-removal candidates require at least three agreeing forms. The three-form condition is itself only a group-consensus heuristic; it does not prove morphological coherence or resolve lexical ambiguity. German is therefore a distinct case study. The frozen package does not document preregistration of this exception relative to external evaluation, so its CELEX2/CISTEM results are corroborative rather than a population generalization estimate.

G. External Evaluation Is German-Only

Only German has external-source CELEX2-derived evaluation in the frozen evidence. GS1 and GS2 come from the same CELEX2 source and represent related conflation policies, so agreement in their delta direction is not independent replication. Other candidates establish resource-internal audit changes and transformation-vocabulary consequences, not external linguistic improvement.

H. Scorer Adaptation

The German benchmark uses a micro CISTEM-style scorer whereas the original CISTEM publication macro-averaged cluster scores. Its TP/FP/FN values are cluster-incidence aggregates; predicted-stem collisions can contribute FP incidence to more than one gold cluster, so the counts are not a conventional unique-token confusion matrix. Cross-paper numerical substitution is invalid.

I. Representation Metric Boundary

Distinct command count captures transformation-vocabulary diversity only. It does not measure trie topology, path sharing, encoded command length, memory, latency, or throughput. The absent versioned before/after trie statistics prevent stronger executable-structure claims.

J. Packaging Diagnostic Boundary

Baseline and candidate GZip files are not produced by an identical serialization policy. Candidate files add a provenance block and reconstruct parser-accepted rows. Package-byte deltas are therefore not controlled estimates of lexical payload size. PoliMorf parser-rejected rows illustrate the same boundary from the opposite direction.

K. Frozen Converted Artifacts versus Current Upstream Resources

The audit operates on Radixor’s frozen converted runtime dictionaries, not on current upstream UniMorph repositories. A finding therefore characterizes the checked-in Radixor lexical artifact and its conversion history. It must not be projected retroactively onto a current upstream language repository without a separate source-level reconciliation.

L. Source Heterogeneity and Review Queue

The audit spans heterogeneous UniMorph-derived resources and one PoliMorf resource; historical source provenance is incomplete for some legacy imports. Review counts are algorithmic candidates, not human labels, and source-artifact counts use heterogeneous units.

IX. REPRODUCIBILITY

The article package separates frozen source projections from derived analysis. The evidence anchor is Radixor repository snapshot 0c3b13f [4]; the bibliographic citation intentionally points to the repository root because article-release integration occurs after manuscript freeze. Provenance records the exact snapshot and relevant source identities.

The audit census is matched to the 12 configurations enumerated by the frozen Gradle task. The implementation defines the ordered sanitation and dominance pipeline. German TP/FP/FN are projected from the frozen comparison table and independently recomputed by the article analysis. Candidate package sizes are retained only as packaging diagnostics. The analysis additionally verifies the fixed 237-byte UTF-8 candidate provenance block documented by the generator.

The package contains source projections, deterministic analysis generation, an independently coded verifier, generated tables/figures, the review record and responses, and nested/global SHA-256 manifests. The verifier requires the 12-resource census, ten candidates, two no-safe-change outcomes, the nine-candidate grouped stratum (8 with command reduction), the grouped S_K change from 20,281 to 19,842, Finnish assignment removal with zero command drop, Hebrew sanitation-only command reduction, the Polish UniMorph +121-byte packaging case with the 237-byte added provenance block, the German 2,055-to-1,673 assignment/command response, and both positive German F1 deltas. No absent staged or trie statistics are reconstructed.

X. CONCLUSION

Lexicon-trained stemming exposes a maintenance layer between lexical data and runtime code: the support structure of the learned transformation vocabulary. In the frozen Radixor audit, resource-local evidence creates ten non-destructive candidates from twelve high-command configurations. Under the standard grouped rule, 8 of 9 candidates reduce K , with $R_\Sigma = 2.165\%$ and median per-resource reduction 1.334%. German uses a separate individual rule and reduces K by 23.948%. Two audited resources admit no safe automatic change under the frozen rule.

The most informative results are the exceptions. Finnish removes dominance-qualified assignments without losing a command because support remains elsewhere. Hebrew loses commands without assignment-dominance removal because sanitation changes the accepted evidence. German’s ordered pipeline removes 2,055 dominance-qualified assignments but yields 1,673 fewer commands; the staged contribution of each audit phase is not identifiable from the frozen projection. Polish UniMorph loses commands while its packaged GZip grows; inspection of candidate materialization shows that the package adds metadata and changes serialization, so the byte increase is not model growth. These cases make the mechanism more informative than an aggregate percentage.

German further shows how a less conservative rule should be evaluated: the frozen candidate increases scorer TP incidences and slightly reduces FP incidences under both related CELEX2/CISTEM policies. This corroborates the candidate’s conflation behavior under those policies; it neither provides independent replication nor establishes the universal lexical correctness of every removed assignment.

The broader conclusion is intentionally bounded. Lexical consistency can change learned transformation vocabulary, sometimes substantially, because lexical assignments determine command support. Command vocabulary can prioritize auditing, but it cannot diagnose lexical error by itself, and a no-safe-change outcome does not prove that a resource is clean. Stronger claims about source footprint require controlled canonical serialization; stronger claims about executable model complexity require versioned trie/memory measurements; stronger claims about application value require downstream retrieval experiments.

REFERENCES

- [1] M. F. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [2] 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.
- [3] —, “Semi-automatic stemmer evaluation,” in *Intelligent Information Processing and Web Mining*. Springer, 2004, pp. 209–218.
- [4] —, “Radixor,” GitHub repository, 2026, article evidence uses frozen repository snapshot 0c3b13f485a9ad0b460c0931e4497ea95bed66a1; accessed Sep. 2, 2026. [Online]. Available: <https://github.com/leogalambos/Radixor>
- [5] 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 Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 840–855. [Online]. Available: <https://aclanthology.org/2022.lrec-1.89/>
- [6] M. Woliński, M. Miłkowski, M. Ogródniczuk, and A. Przepiórkowski, “PoliMorf: a (not so) new open morphological dictionary for Polish,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul, Turkey: European Language Resources Association, 2012, pp. 860–864. [Online]. Available: <https://aclanthology.org/L12-1107/>
- [7] A. D. McCarthy, C. Kirov, M. Grella, A. Nidhi, P. Xia, K. Gorman, E. Vylomova, S. J. Mielke, G. Nicolai, M. Silfverberg, T. Arkhangelskiy, N. Krizhanovsky, A. Krizhanovsky, E. Klyachko, A. Sorokin, J. Mansfield, V. Ernstreits, Y. Pinter, C. L. Jacobs, R. Cotterell, M. Hulden, and D. Yarowsky, “UniMorph 3.0: Universal morphology,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, pp. 3922–3931. [Online]. Available: <https://aclanthology.org/2020.lrec-1.483/>
- [8] S. Ahmadi and A. Mahmudi, “Revisiting and amending central kurkish data on UniMorph 4.0,” in *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 38–48. [Online]. Available: <https://aclanthology.org/2023.sigmorphon-1.5/>
- [9] E. M. Bender and B. Friedman, “Data statements for natural language processing: Toward mitigating system bias and enabling better science,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018.
- [10] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford, “Datasheets for datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [11] C. D. Paice, “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.
- [12] 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.
- [13] F. N. Flores and V. P. Moreira, “Assessing the impact of stemming accuracy on information retrieval—a multilingual perspective,” *Information Processing & Management*, vol. 52, no. 5, pp. 840–854, 2016.
- [14] 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.
- [15] R. H. Baayen, R. Piepenbrock, and L. Gulikers, “CELEX2 LDC96L14,” Linguistic Data Consortium, Philadelphia, PA, 1995.