CLIB 2026 · Computational Linguistics in Bulgaria

Genetic Algorithms with Normalized Fitness Functions for Low-Resource Language Identification

A Case Study on Archaic Romanian and Medieval Slavonic Liturgical Manuscripts

Brandon Caillahua Mendoza
brandoncmkot01@gmail.com · ORCID: **0009-0007-2528-9305**

Motivation & Corpus

Hundreds of Cyrillic liturgical manuscripts (13th–17th c.) from Romanian, Moldavian, and Transylvanian archives lack secure identification of their linguistic substrate. Large labeled datasets are unavailable — and Old Church Slavonic / archaic Romanian are absent from multilingual transformer pretraining (mBERT), placing them in a permanent zero-shot regime. Archivists and philologists therefore rely on slow manual paleographic comparison, a process that does not scale to the hundreds of undated fragments awaiting classification across regional collections.

Why identification is hard

- Orthographic variability: the same word appears in multiple spellings within one codex
- Complex morphology distinguishes Romanian sharply from Slavonic and Hungarian neighbours
- Mixed lexicon: liturgical vocabulary draws simultaneously on Latin, Church Slavonic and vernacular Romanian roots
- Scribal inconsistency: individual copyists introduce idiosyncratic abbreviations and diacritics that further fragment the surface form of a single lexeme

Distinctive Romanian markers

- **Enclitic definite article** (omul, cartea) — absent in Hungarian (proclitic article a/az) and Church Slavonic (no article)
- **Regular verbal paradigms** (cântu, cânți, cântă…) produce statistically detectable bigram signatures
- **Analytical-synthetic verbal morphology** — intermediate between Latin synthesis and Romance analysis
- **Three superimposed lexical strata:** Latin core (biserică), Slavic borrowings (glas, molitvă), Hungarian loans (oraș, chip)
- **Taken together,** these markers form a converging bundle of evidence rather than any single diagnostic feature

Related work gap

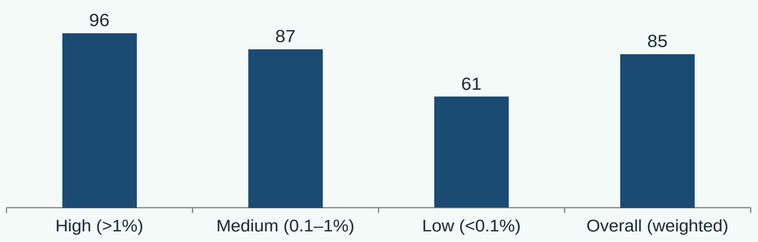
Stylometry needs thousands of function-word occurrences; n-gram SVMs need labeled data per class. Neither condition holds here — the verified reference material totals only **~20,000 tokens** across three manuscript traditions. mBERT does not include Old Church Slavonic or archaic Romanian in pretraining — a permanent zero-shot regime. This gap motivates a search-based, data-frugal alternative rather than a fine-tuned neural classifier.

Control corpus (Gaster, 1891)

Three genres — psalmody, liturgical prose, narrative prose — giving diverse coverage across **20,570 tokens** of verified control text drawn from securely dated, unambiguously Romanian sources.

Only **~20,000 tokens** of verified reference text required — orders of magnitude below the annotation budget of any historical manuscript project.

Token-level concordance by frequency stratum



High-frequency tokens (articles, verbs, prepositions) reach 96.3% concordance — these form the structural backbone the algorithm relies on most. Concordance decays gracefully rather than collapsing as frequency drops, which is consistent with a genuine linguistic signal rather than overfitting to a handful of common forms.

Inter-annotator agreement (first 1,000 tokens)

Comparison	Agreement	κ
Expert A vs. Expert B	87.2%	0.82
Expert A vs. Expert C	85.9%	0.79
Expert B vs. Expert C	88.1%	0.84
Mean (experts)	87.1%	0.82
Algorithm vs. consensus	**84.7%**	**0.80**

The algorithm's agreement with expert consensus (84.7%, κ=0.80) closely approaches mean inter-expert agreement (87.1%, κ=0.82) — comparable to a human expert, in minutes rather than weeks, and well above chance-corrected thresholds conventionally treated as substantial agreement.

Method

A steady-state genetic algorithm searches morpheme-class assignments, guided by a normalized multi-objective fitness function that resolves scale incompatibility between heterogeneous metrics:

$$F(G) = 0.4 \cdot f_{Zipf} + 0.35 \cdot f_{KL}^{norm} + 0.25 \cdot f_{cooc}^{norm}$$

f_{Zipf} — squared Pearson correlation between token-frequency ranks and Zipf's law. A universal, substrate-independent linguistic filter; ensures any natural language scores high.

f_{KL}^{norm} = $\exp(-D_{KL}/\tau)$, τ=3.0 — maps unbounded KL divergence to (0,1). This is the primary substrate-specific discriminant; it carries the signal that distinguishes Romanian from Church Slavonic.

f_{cooc}^{norm} — proportion of adjacent morpheme bigrams matching a verified liturgical bigram lexicon. Provides domain-specific regularisation sensitive to morphological transitions.

Together the three terms trade off universality, substrate specificity, and local morphological plausibility, so no single metric can dominate the search on its own.

Weight rationale: ω_1=0.4 (Zipf, universal filter) · ω_2=0.35 (KL, substrate discriminant) · ω_3=0.25 (bigrams, domain regulariser). Weights encode theoretical priority, not scale.

GA configuration

Population	500 individuals
Selection	Binary tournament, p = 0.75
Crossover	Single-point, p = 0.8
Mutation	Swap, p = 0.05
Replacement	Steady-state, 2 worst / gen.
Convergence	Δfitness < 0.01 over 500 gens (≤15,000 max)

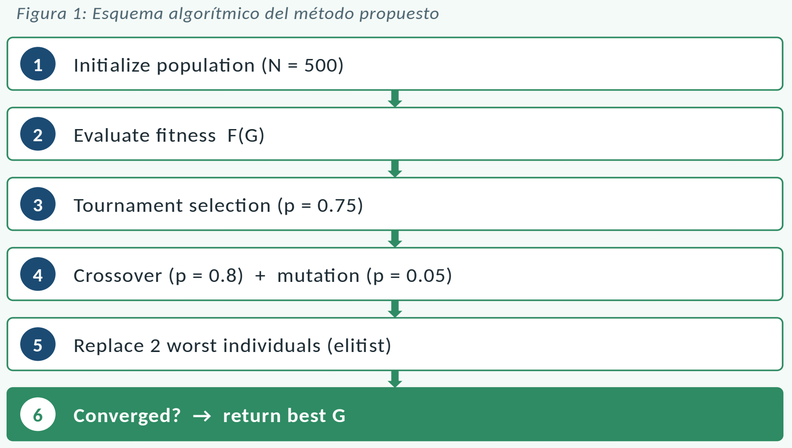
Pre-registered falsifiability criteria

R^2Zipf < 0.70 · wrong-direction ΔDKL · stability < 50% · positive convergence on cipher control

None of these conditions were met.

Steady-state GA — algorithm flow

Figura 1: Esquema algorítmico del método propuesto



Hyperparameter sensitivity (±20% weight perturbation)

Configuration	R^2Zipf	DKL	Stability
Baseline	0.89	1.87	94.2%
ω1 +20%	0.91	1.91	92.8%
ω1 −20%	0.86	1.94	91.3%
ω2 +20%	0.88	1.72	95.1%
ω2 −20%	0.89	2.03	93.4%

All metrics stay within one standard deviation of baseline under ±20% variation — the framework does not depend on precise tuning, so results are not an artefact of a lucky weight choice.

Recurrent error patterns

1. **Slavic roots + Romanian clitics** (e.g. molitvă-a) misread as Slavic because the root frequency matches Church Slavonic while the enclitic article is not visible as a separate token — ~38% of medium-frequency errors

2. **Orthographic variants** of high-frequency tokens (că / că / ce) appear under multiple graphemic realizations; rare variants receive different classes — ~29% of medium-frequency errors

3. **Calque structures** — Hungarian/Slavonic syntax with Romanian lexis create partial matches — ~21% of low-frequency errors. These are systematic and amenable to targeted preprocessing, such as a normalization pass that folds orthographic variants to a canonical grapheme before tokenization.

Results

Validation on Romanian control manuscripts

Manuscript	R^2Zipf	Stability
Psaltirea lui Coresi	0.89 ± 0.03	94.2%
Liturghia română	0.87 ± 0.04	91.7%
Evangheliarul	0.91 ± 0.02	96.1%

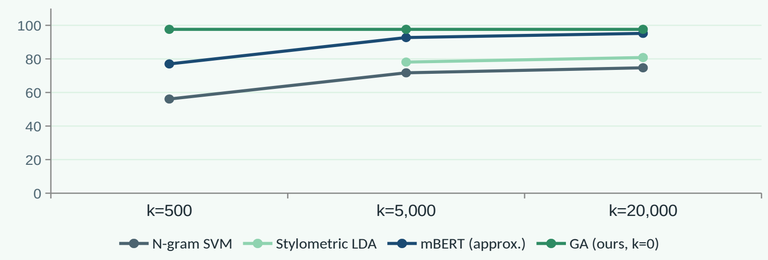
Substrate selectivity: vs. medieval Hungarian texts ΔD_{KL} = +126%. On a Bellaso cipher negative control: R^2Zipf = 0.34, stability 11.7% — correctly rejected, confirming the algorithm does not simply fire on any structured symbol sequence.

Boundary test — Draganov Menaion (Zogr. 54): deliberately run without appropriate base patterns. The GA returns flat fitness landscape and 0% stability, while the correct-direction signal still exists at ground-truth level (ΔD_{KL} = +23.8%) — confirming the null result reflects missing input, not a flawed algorithm.

Substrate differential: Romanian- vs. Hungarian-optimized

Metric	Romanian opt.	Hungarian opt.
DKL	4.23 ± 0.37	1.92 ± 0.24
R^2Zipf	0.81 ± 0.06	0.88 ± 0.04
Stability	68.3 ± 5.7%	93.1 ± 2.8%

Accuracy vs. labeled tokens (PROIEL Codex Marianus)



The GA achieves 97.6% POS-class accuracy with **zero labeled tokens** from the target corpus, exceeding all supervised baselines even at their maximum training budget (k=20,000) — a regime where every competing method still needs thousands of hand-labeled examples to catch up.

Ablation study (Psaltirea, n = 10 runs)

Condition	R^2Zipf	DKL	Stability
Full model	0.89	1.87	94.2%
Without Zipf (ω1=0)	0.74	1.92	87.3%
Without KL (ω2=0)	0.86	2.41	82.6%
Without bigram (ω3=0)	0.88	1.89	89.7%

Removing KL divergence causes the largest degradation (stability −11.6 pts) — confirming it as the **primary substrate-specific discriminant**, exactly as its weight (ω_2 = 0.35) presumes. Removing the Zipf term causes the largest drop in R^2, showing that the two objectives protect against different failure modes.

Application — Ms. A-II-19 (unattributed codex)

Stability under Romanian reference (88.3%) is markedly higher than under Hungarian (71.4%) or Slavonic (63.7%) references — consistent with three independent expert philologists' assessments (mean confidence Romanian: 3.67/5). This convergence between an unsupervised statistical method and independent human philological judgment is the strongest evidence for the codex's Romanian substrate.

Summary & Conclusions

- Normalizing heterogeneous metrics to a common scale removes an unaddressed source of bias in prior fitness-based approaches.
- Zero labeled data from the target manuscript; only ~20,000 tokens of verified reference text required.
- Validated across control manuscripts, negative controls (cipher), a boundary test (Draganov Menaion), and real supervised benchmark (PROIEL) — outperforming all baselines with zero labels.
- 97.6% POS-class accuracy vs. 74.7%–95.2% for supervised baselines at k=20,000.
- Directly extensible to other low-resource ecclesiastical languages: medieval Slavic substrates, Dead Sea Scroll fragments, original-language detection in translations.
- Because pre-registered falsifiability criteria were fixed before analysis and none were met, the positive result cannot be dismissed as a post-hoc fit to convenient data.