Learn by Decoding: An MDL Framework and Pre-Registered Test for the Lexical Crossing Between Sister Languages

Teach the transformation, not the target — a description-length account of the lexical key between related languages, how to choose and ladder targets across a family, and a design to test it

Alejandro Toledo Martínez — Independent researcher (ORCID: 0009-0000-1277-9697)

Scope & status. This is a framework and a pre-registered protocol, not a validated course: there is no human-learning data yet (§7 designs the test). It formalises the lexical crossing — sound/spelling correspondences over cognate stems — which is the tractable part. Grammar (morphosyntax) is a separate, largely non-transformational module and is explicitly out of the transformation's reach (§3.4). Cognate-coverage numbers are computed on core (Swadesh-type) vocabulary and are an upper bound on real-text transparency (§5.6).

Abstract

Closely-related languages — Dutch and Frisian, Spanish and Portuguese, Czech and Slovak, Swedish and Norwegian — share most of their lexicon and grammar, yet are almost always taught as if the learner started from zero. We argue this is a pedagogical error with a precise diagnosis. For a speaker of language A who wants a related B, the *lexical forms* of B are, to a first approximation, a regular transformation of A's ($w_B \approx T(w_A)$ for cognate stems). The learner already possesses A — the plaintext — so recovering that transformation is a favourable known-plaintext situation, not blind decipherment. The efficient object to teach is therefore not the target vocabulary but the transformation T: a few dozen high-frequency correspondence rules, a bounded set of exceptions, and the genuinely different words. We formalise the burden as $C = C_{\text{transform}} + C_{\text{exceptions}} + C_{\text{new}}$ versus the traditional $C = C_{\text{all}}$, and define a pedagogical distance as the description length (MDL) of the key — distinct from historical or raw phonological distance, directional, and split into a *reading* and a *speaking* distance. Four contributions follow: (i) we compute cognate coverage across 20 Indo-European languages and use it to show how to choose a target and ladder across a family (Figure 1), formalising that keys compose ($T_{A\to C} \approx T_{B\to C} \circ T_{A\to B}$); (ii) we redefine "genuinely new" against the learner's lexical reachability set (synonyms, regionalisms, archaisms, derivational relatives), which shrinks — and makes dialect-indexed — the residual to memorise; (iii) we separate the tractable lexical crossing from grammar, which the transformation does *not* reduce (§3.4); (iv) we pre-register a test whose decisive measure is generalisation to unseen words. A computational result from the parent programme (Paper 2) shows the correspondence system is information-theoretically compact — cheap to *infer from data*; whether it is cheap for a *human* to *learn* is exactly the open question the test addresses, and we are careful not to conflate the two.

1. Introduction — the sister-language paradox

Two facts sit uneasily together. First, closely-related languages are largely mutually transparent: a Dutch speaker reading Frisian, or a Spanish speaker reading Portuguese, understands a great deal

on first contact (the mutual intelligibility of closely-related languages is well documented; Gooskens 2007). Second, learners of these same pairs routinely report the target as surprisingly hard, and courses treat it as a new language to be built word by word. The paradox is pedagogical, not linguistic: the transparency is *receptive and passive*, riding on cognate recognition, while courses aim at *productive* competence and teach it from scratch, discarding the very structure that made the language transparent in the first place.

The thesis of this paper is a change of object. Traditional courses teach *"this is how you say it in B."* We propose teaching *"this is how A turns into B."* Formally, where the learner already commands language A and wants a closely-related B,

$$L_B \;\approx\; T(L_A),$$

and the efficient thing to learn is the transformation function T, not the image L_B term by term. The learner's job is to become a *decoder*: given a known word of A, apply T and produce (or recognise) its B counterpart — including for words never explicitly taught.

This is not a metaphor imported for flavour. It is the learner's-side instance of a research programme that models sound change itself as a cipher (§2). What makes the pedagogical case especially favourable is a point cryptanalysts know well: the learner is not decoding an unknown message. They already hold the plaintext — their own language — so this is a known-plaintext ("crib") attack, the easiest kind. The rest of the paper turns that observation into a method, a cost model, honest conditions, computational support, and a test.

2. Theoretical foundation — learning as known-plaintext decoding

2.1 The correspondence system is the key

The parent programme (*The Cipher of Sound Change*) treats two related languages as two encipherments of a common source and the system of regular correspondences between them as the key $K_{A\to B}$. Historically the key is recovered by comparison; pedagogically it is *taught*. Crucially, the key is not a table of word translations but a compact system of rules over sounds and spellings — precisely the object that generalises to words the learner has never met.

The key is honestly a stochastic channel, not a clean bijection: for a given unit a of A it induces a distribution over the B-outcomes,

$$K_{A\to B}(b \mid a).$$

This single fact organises the whole method, because it splits the learning target into three naturally different parts:

1. Regular rules — where $K_{A\to B}(\cdot \mid a)$ is concentrated on one outcome (high probability mass). These are the productive correspondences (Dutch *aa* $\to$ Frisian *ie*; Spanish *-ción* $\to$ Portuguese *-ção*). Learn the rule once, apply it to thousands of words.
2. Conditioned rules and exceptions — where the outcome depends on context or is simply irregular (a residual spread of the distribution). A bounded list to memorise.
3. Genuinely new lexicon — words with no usable correspondence (Dutch *kinderen* vs Frisian *bern* "children"; Dutch *raam* vs Frisian *finster* "window"). These must be learned as in any course — but they are a *minority* between close languages.

2.2 Why the learner has it easy: the crib

In cryptanalysis, a *crib* (known plaintext) collapses the problem: instead of inferring both the message and the key, you already have the message and only recover the key (the known-plaintext attack; Friedman 1922; Shannon 1949). The sister-language learner is in exactly this position. They are not deciphering B from nothing; they hold A (tens of thousands of words, full grammar) and need only the *relation* $K_{A\to B}$. Learning a close language is thus key recovery with the plaintext in hand — a favourable situation, not the blank slate a from-scratch syllabus implicitly assumes. This crib picture is an intuition pump, not the paper's machinery: the formal work (what counts as a rule, how to rank rules, how to measure a pair's difficulty) is done by minimum description length (§4), and the quantitative claims below are MDL claims, not cryptographic ones.

Two further properties of the framing carry straight into pedagogy:

- Reconstruction-free. The learner never needs the proto-language or any historical reconstruction. They need only the *pairwise* transformation $A \to B$, read off attested modern forms. The method is history-*supported* but history-*free* for the student.
- Two channels (consonants vs vowels). Consonants carry most of a word's identity and change conservatively; vowels shift more. Pedagogically this yields a concrete heuristic: anchor recognition on the consonant skeleton (which usually survives the crossing) and treat vowel correspondences as the main body of rules to drill. "Same skeleton, shifted vowels" is the shape of most transparent cognate pairs.

2.3 What the learner is really acquiring

We *hypothesise* — this is the paper's central untested claim, not an established result — that a productive learner internalises $T = K_{A\to B}$ as a *generative* competence: an operator they apply on the fly, not a memorised lexicon. If so, the success signature is correct production of never-seen words; but note (§7) that succeeding at an offline $A \to B$ *conversion* task is weaker than genuine productive competence in B, and the two must be measured apart.

3. The method — Decoding-Based Language Learning (DBLL)

The method decomposes acquisition into the three parts of §2.1 and teaches them in order of leverage.

Step 1 — Establish the crib. Confirm and activate what the learner already transfers for free: shared cognate vocabulary and shared grammar. This is motivational and real: it makes the learner's existing A an asset, not a source of interference.

Step 2 — Teach the transformation as productive rules. Present each high-frequency correspondence as an *operator* with a drill, not as isolated vocabulary. The canonical exercise is *conversion*: give A-forms, ask for B-forms.

> Rule (Dutch→Frisian). `aa` → `ie` in many frequent words. *staat* → *stiet*, *gaat* → *giet*. Drill: convert *slaapt, laat, kwaad, ...*

Because the rule is productive, we *predict* a learner who has drilled it can produce the correct Frisian form of a Dutch word never on any vocabulary list — the decisive difference from list-learning, and precisely what §7 tests.

Step 3 — Teach the bounded exceptions. The residual spread of $K_{A\to B}(\cdot \mid a)$ — context-conditioned variants and irregulars — is a finite list, explicitly flagged as "do not apply the rule here."

Step 4 — Shrink the "new" list against the learner's *reachability set*. "Genuinely new" is not a property of the language pair; it is a property of the *learner's whole accessible lexicon* against B. Before filing a word as new, check whether any form the learner can reach — a standard synonym, a regional/colloquial word, an archaism, or a derivational relative — is the cognate of the B target. Most "new" words dissolve:

- Portuguese *lutar* 'to fight' looks new against Spanish *pelear*, but Spanish luchar is the everyday cognate — not new at all.
- Portuguese *garra* 'claw' looks new against Spanish *uña / zarpa*, but Spanish garra is fully current (especially in Mexican/American Spanish): the learner may already own the target word through their own regional variety.
- Where no everyday form reaches, a cultism/technical/archaic cognate still gives an access point: Spanish *rojo* / Portuguese *vermelho* via Spanish bermellón 'vermilion (pigment)'.

Formally, a word is new only if no accessible A-form is a cognate of the B target, and access points are *mined from that reachability set*, not hand-picked. Two consequences: (i) C_{new} is smaller than a one-headword count implies — the method is stronger than it first looks; and (ii) C_{new} is dialect-indexed (a Mexican and a Peninsular speaker have different residuals), so a manual must know *which variety of* A the learner speaks (§5.5). Regional/synonym access points carry the same false-friend hazard and get the same flag (Spanish *garra* also means 'guts/nerve' figuratively — the form transfers, the full meaning may not).

Step 5 — False-friend inoculation. Where A and B share a form but not a meaning, the transformation would mislead. These are taught adversarially, as the *failures* of the otherwise-reliable key (§5.2).

3.1 Worked example A — Dutch → Frisian

From a parallel paragraph (Dutch / West Frisian):

> NL: *Elke ochtend staat Anna vroeg op. Ze zet koffie en doet het raam open. Buiten lopen de mensen naar hun werk en fietsen de kinderen naar school.* FY: *Elke moarn stiet Anna betiid op. Se set kofje en docht it finster iepen. Bûten rinne de minsken nei har wurk en fytse de bern nei skoalle.*

Aligning the two exposes the three parts at once (illustrative — a production manual would validate each rule against a corpus and a native expert):

type	pattern	examples
regular rule	`aa` → `ie`	staat→stiet, gaat→giet
regular rule	initial `z` → `s`	ze→se, zet→set
regular rule	`aa/aar` → `ea/ei`	paar→pear, naar→nei
regular rule	`ie` → `y`	fietsen→fytse
exception / function word	`het` → `it`, `een` → `in`	(short, high-frequency; memorise)
genuinely new	no correspondence	ochtend/*moarn*, vroeg/*betiid*, raam/*finster*, kinderen/*bern*, lopen/*rinne*

A learner drilling four or five such rules converts most of the paragraph unaided; only a handful of items are true "new vocabulary." The consonant skeleton is the anchor throughout (*st-t*, *g-t*, *p-r*, *f-ts*), with the vowels doing the rule-work — the two-channel heuristic in action.

3.2 Worked example B — Spanish → Portuguese

Almost entirely orthographic-phonological rules over a shared lexicon:

pattern	examples
`-ción → -ção`	nación→nação, educación→educação, información→informação
`ll → lh`	(structural grapheme rule)
`ñ → nh`	España→Espanha
`j → x/g`	viaje→viagem

Here the *orthographic* transformation is dense and regular, so the reading distance (§4.2) is small: a Spanish speaker can be brought to broad Portuguese reading largely by internalising a spelling-rule list plus false-friend warnings. Speaking is a different story — Portuguese's vowel/nasal system is an *expansion* of Spanish's (a large D_{speak}, §4.2) — and productive *writing* also needs the morphosyntax the rule list omits (§3.4). These spelling rules are the *reading* face of the phonological correspondences a manual states in IPA.

3.3 The bridge sentence, and learning the *mode* not just the forms

A rule list teaches the transformation of *forms*. But a language also has a mode — its rhythm, vowel quality, nasality, the melody it is spoken in — and productive competence needs that too. The bridge sentence is the device that carries it. Present each sentence three ways: the target; a bridge — the learner's own language *bent toward* the target, using real but sometimes less-usual words chosen so the two line up almost word-for-word; and the natural phrasing. For a Spanish speaker meeting Portuguese *O transeunte tenta pegar o ônibus*, the bridge *El transeúnte intenta tomar el autobús* (all real Spanish) aligns almost token-for-token, where the natural *El peatón quiere subirse al camión* does not. Reading target-then-bridge, and listening to them in parallel, lets the learner cross into the target's form *and mode* in the smallest step — the bridge is a self-made crib at the sentence level.

Two design points follow. (i) Choose the reference accent closest to the learner. Varieties differ in how far their mode sits from the learner's: for a Spanish speaker, Northeastern Brazilian or cultured northern-Portuguese pronunciation (fuller, less-reduced vowels) is a shorter step than Rio or Lisbon. Fixing one such accent from the start minimises the mode gap while the rules do the form work. (ii) Pronunciation is not decoration. Because the consonant skeleton often maps almost one-to-one, most of what separates two close languages *audibly* is the mode — so parallel listening in the chosen accent is where much of the real crossing happens, not in the rule table alone.

3.4 What decodes and what does not — the lexical crossing vs the grammar

The transformation T is a map over lexical forms: it turns a cognate A-stem into its B-reflex. That is the tractable part, and it is genuinely most of the *vocabulary*. But a language is not only its lexicon, and the known-plaintext picture breaks exactly where B encodes a category A does not — there is no plaintext for a distinction your L1 never makes. Three regimes must be taught differently:

- Substitutable (DBLL's home turf): phonological/orthographic correspondences over cognate stems — a genuine channel $K_{A\to B}(b \mid a)$. Teach as productive rules (§3).
- Re-mappable: categories present in both but *distributed* differently. Spanish→Portuguese verb endings largely correspond, but *ser/estar* boundaries, clitic placement (Spanish proclisis vs European-Portuguese enclisis), differential object marking (Spanish personal *a*), and article+possessive (*o meu livro*) diverge. Teach as paradigm maps plus a short contrast catalogue, with heavy caveats — not as a clean substitution.

- Novel (no transformation exists): a B-category absent from A. Portuguese has a personal (inflected) infinitive (*para eles verem*), a living future subjunctive (*quando fores*), and mesoclisis (*dar-lhe-ei*) — none derivable from any Spanish form. These are learned conventionally; the method offers nothing special here and should say so.

The honest consequence: DBLL is a theory of the lexical crossing; the 10–15 morphosyntactic contrasts that actually diverge for a given close pair are a separate module, taught as an explicit contrast list, not as T. The failure mode of ignoring this is the *portuñol* plateau — fast receptive transparency, then production fossilised on the morphosyntactic residual. Accordingly the trial's production endpoint must be sentence-level, not word-form conversion (§7).

3.5 Fading the crib — from transcoding to production

Drilling $A \to B$ conversion installs a useful habit but also a risk: routing every B utterance *through* A. That is transcoding, not productive competence, and in close pairs the over-mapping habit is the main source of persistent, hard-to-detect interference (comprehensible-but-non-idiomatic collocations, preposition and aspect choices). The method must therefore include a scaffold-removal sequence, from skill-acquisition theory (declarative $\to$ procedural via practice; DeKeyser 2007): (1) A-prompted conversion; (2) B-cued production with A available; (3) *timed* B production with the A prompt withheld; (4) meaning-driven B production (picture/oral elicitation) with no A at all. Parallel listening in the reference accent (§3.3) belongs here as ear-training that weans the learner off orthographic A-mediation. The endpoint that matters is stage (4); a learner who only reaches stage (1) has learned a conversion trick, not B.

4. A cost model and a pedagogical distance

4.1 Learning cost

Traditional acquisition treats the target as new: $C_{\text{trad}} \approx C_{\text{all}}$, the cost of the whole lexicon and grammar. DBLL replaces this with

$$C_{\text{DBLL}} \;=\; C_{\text{transform}} \;+\; C_{\text{exceptions}} \;+\; C_{\text{new}},$$

where $C_{\text{transform}}$ is the cost of a few dozen productive rules, $C_{\text{exceptions}}$ a bounded list, and C_{new} the genuinely non-corresponding lexicon. The method wins exactly when $C_{\text{transform}} + C_{\text{exceptions}} \ll C_{\text{all}} - C_{\text{new}}$ — i.e. when a large, transparent share of the lexicon is *covered by a short rule system rather than memorised word by word*.

4.2 Pedagogical distance — a two-part description length

To be a well-defined, comparable quantity the distance must be all in bits. We use a two-part (MDL) code for "how much must the learner store to turn A into B":

$$D_{\text{ped}}(A \to B) \;=\; \underbrace{L(\text{rules}) + L(\text{exceptions})}_{\text{the key, in bits}} \;+\; \underbrace{L(\text{residual lexicon} \mid \text{reachability})}_{\text{bits to list the non-derivable words}}.$$

The third term is the *reachability-relative* new lexicon (§3, Step 4) coded in bits, so the sum is dimensionally consistent (an earlier draft added bits to a bare proportion — an error). Fixing the code means committing to a segment inventory and a rule grammar (per-rule cost, context conditions, exception list); the manual pipeline (§8) specifies these. A code-free companion needing no such choice is the

frequency-weighted conditional entropy $H(B \mid A)$ of the target segment given the source — directly the "stochastic channel" of §2.1, comparable across pairs out of the box.

Three properties matter. (i) It is not historical distance nor raw phonological distance: two pairs equally distant in years differ in D_{ped} if one accumulated *regular* changes (short key) and the other *irregular* ones (long key). (ii) It is directional, and for *speaking* the expensive direction is the one that demands a new distinction B has and A lacks — a one-to-many split the rules cannot resolve (a merger one way is a guess the other; a Spanish speaker facing Portuguese's nasal and open/close vowels faces a *perceptual* split, not just a code cost). So $L(K)$ under-estimates human difficulty exactly where B adds contrasts. (iii) We therefore report two distances: a reading distance D_{read} (orthographic key, recognition) and a speaking distance D_{speak} (phonological key *plus* the cost of inventory expansions). Spanish→Portuguese is the sharp case: small D_{read} (transparent spelling rules) yet large D_{speak} (Portuguese's vowel/nasal system is an expansion of Spanish's) — the flagship "easy" pair is easy to *read* and hard to *speak*, and a manual must declare which it teaches.

4.3 Choosing a target and laddering across a family (cluster learning)

The distance turns "which language should I learn next?" into an answerable question. As a first, code-free proxy we estimate cognate coverage — the share of shared basic concepts that are cognate, an inverse of ρ_{new} — for 20 Indo-European languages from IE-CoR (Figure 1). Three findings:

- The families are clusters, and the numbers pick your target. Within-family coverage is high (Romance, Germanic, Slavic blocks at 56–91 %), cross-family 16–25 %. From any A, the highest-coverage B is the cheapest first target: Spanish→Portuguese 84 %, German→Dutch 84 %, Czech→Slovak 91 %.
- English is a poor Germanic starting point (51–59 % with its cousins, vs German↔Dutch 84 %): its heavy Romance borrowing and idiosyncratic sound changes lengthen its key — a concrete, actionable warning.
- Keys compose, so learning ladders. After learning B, the step to a third language C is *pre-paid*: $T_{A\to C} \approx T_{B\to C} \circ T_{A\to B}$, so the *residual* key at each rung shrinks, and the optimal order over a family is the least-cost path / minimum-cost arborescence on the distance graph. A Spanish speaker's Romance ladder runs roughly Spanish → Portuguese (84 %) → Catalan (78 % from Portuguese) → Italian (80 % from Catalan) → French, each rung a short marginal key rather than a fresh language. This *productive, cost-ordered ladder* is what distinguishes DBLL from receptive intercomprehension, which is unsequenced.

Caveats: coverage is a coarse proxy for the full two-part D_{ped} (it ignores rule-count, morphosyntax, and reading-vs-speaking), and is computed on core vocabulary (§5.6). It suffices to *rank* candidates and draw the ladder; the full MDL distance is future work.

5. When it works, and when it does not (boundary conditions)

The method is powerful but conditional, and stating the conditions is part of its honesty.

5.1 It scales with shared-vocabulary density

DBLL's leverage is the fraction of B reachable from A by a short key. That fraction is essentially the cognate/transparent-vocabulary coverage between the languages. When coverage is high

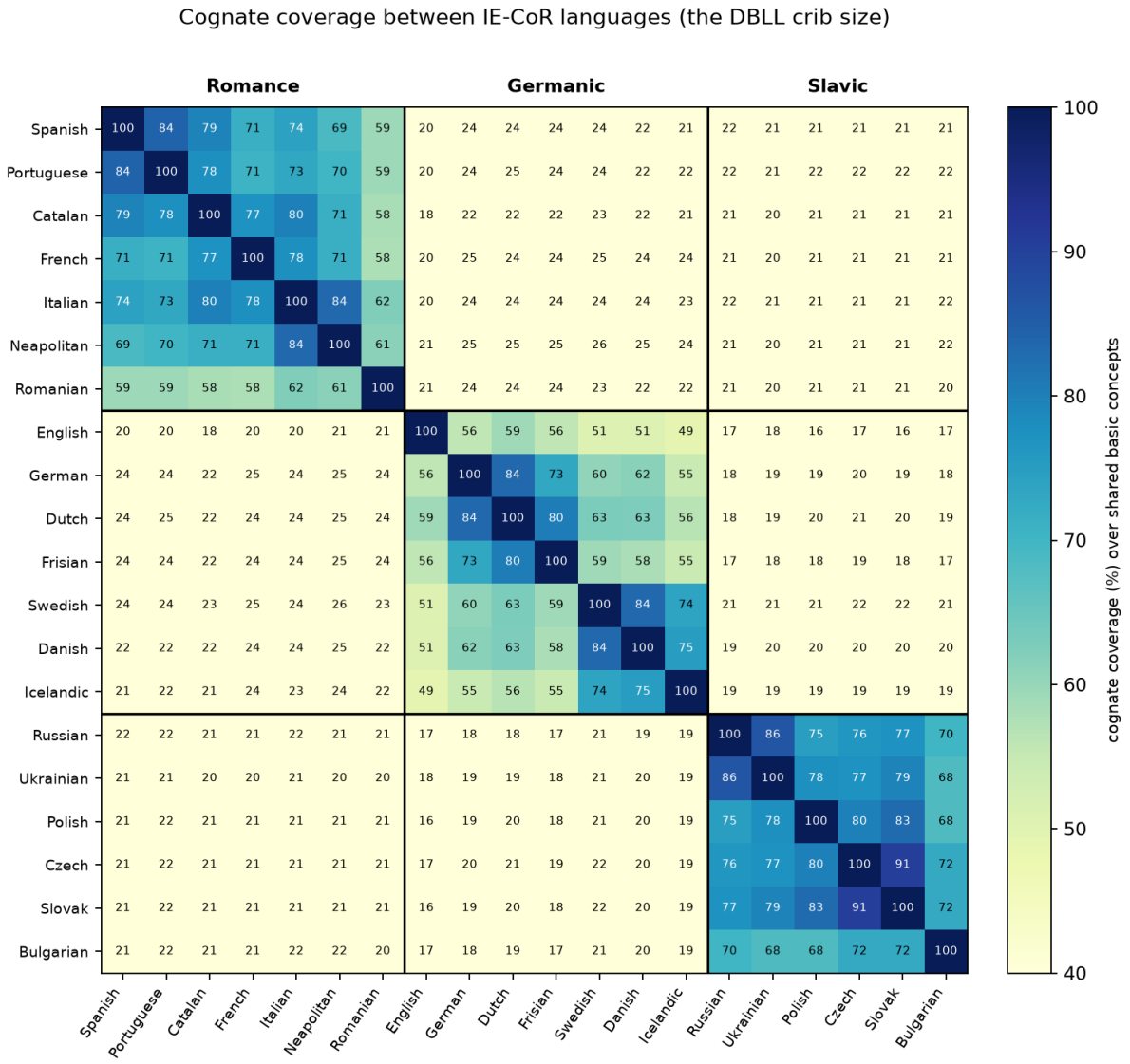

Figure 1: Cognate coverage between 20 IE-CoR languages, grouped by family — the DBLL crib size and the basis for pair-selection and family-laddering. The Romance, Germanic and Slavic blocks are dense (within-family 56–91 %); cross-family coverage is 16–25 %; English is the Germanic outlier.

(Czech↔Slovak, Spanish↔Portuguese, Dutch↔Frisian, Scandinavian mainland trio), the transformation carries most of the load and the method is at its strongest. As coverage falls, C_{new} grows and the method degrades gracefully toward traditional learning — it never does worse than "learn the transparent part by rule, the rest as usual," but its *advantage* shrinks. For unrelated languages (A shares little with B) there is no crib and no meaningful T; the method reduces to ordinary vocabulary learning and offers nothing special. This is the honest ceiling: DBLL is a method for related languages, and its gain is a function of how related.

5.2 False friends are the characteristic hazard

The same regularity that makes the method powerful makes its *failures* systematic: a reliable key invites the learner to over-apply it, producing errors exactly where form-correspondence and meaning-correspondence diverge (false friends; Spanish *embarazada* ≠ Portuguese *embaraçada*). In the parent programme's terms this is the learner's version of apophenia — seeing the rule where it does not hold. DBLL must therefore teach false friends *as first-class content*, the labelled exceptions to an otherwise-trusted transformation, rather than hope they do not arise.

5.3 Why sister languages still feel hard

If close languages are so transparent, why do learners struggle? Three reasons the method directly addresses: (i) traditional courses hide the regularities, teaching transparent words one by one as if opaque, so the learner never forms the productive rule; (ii) the residual — exceptions and false friends — is unlabelled, so learners distrust the transparency they do sense; and (iii) crossing from *passive recognition* to *active production* requires drilling the transformation as an operator, which conventional syllabi rarely do. DBLL is, in effect, the claim that much of the felt difficulty *may be* substantially an artefact of teaching method, recoverable by making T explicit and trainable — a claim §7 is designed to test, not an established fact.

5.4 Orthography is not phonology

Many "rules" a beginner needs are spelling correspondences (Spanish→Portuguese ñ→nh), which are convenient but not identical to the underlying sound correspondences; and a rule reliable in writing may hide exceptions in speech (and vice versa). A production manual must state, per rule, whether it is orthographic, phonological, or both — and the pedagogical distance (§4) differs for reading vs speaking.

5.5 Who the method is for (a learner-population condition)

DBLL is an explicit, symbol-heavy, literacy-dependent method: correspondence tables, orthographic-vs-phonological rules, metalinguistic reflection. That restricts its addressable population to literate, instructed, analytically inclined adult learners, and even among them, gains from *explicit* instruction interact with language-analytic aptitude. Heritage speakers, young children, and oral-culture or low-literacy learners are out of scope or need a different delivery. This is a first-class boundary condition alongside coverage (§5.1): the pedagogical-distance framing describes the *material*, but the *method* of teaching it is not universal. The trial (§7) should treat aptitude and literacy as pre-registered moderators — the interesting question is *for whom* transformation- teaching pays off.

5.6 Core vocabulary is not the full lexicon (and reading is not speaking)

The coverage figures (§4.3, §5.1) come from core / Swadesh-type lists, deliberately selected to be borrowing-resistant and stable — the *most* cognate-dense slice of a language. Real running vocabulary diverges more, driven by differential borrowing (Ibero-Romance over Arabic: Spanish *alfombra* vs Portuguese *tapete*; Germanic over Frankish/French/English strata). So the 80–91 % headline is an upper bound on real-text transparency; a manual should report coverage by frequency band on running text, not only core lists. Differential borrowing is not noise but a *systematic* sub-key ("where the two languages chose different etyma"), teachable as its own rule family. A further discount: part of the "cognate" numerator is false-friend-loaded (Spanish *largo* 'long' vs Portuguese *largo* 'wide'; *esquisito*, *oficina*) — form-cognate, meaning-diverged, so not free. And because most showcase rules are *orthographic*, the coverage story is strongest for reading; the speaking distance (§4.2) is larger wherever B's phonology expands A's.

6. How compact is the key? A premise-plausibility check (not human evidence)

DBLL's premise is that the transformation between close languages carries little information — few, regular rules. The parent programme (Paper 2) bears on this premise, and *only* the premise. Using expert Indo-European cognacy (IE-CoR; Heggarty et al. 2023), it estimates a posterior-concentration curve — how many cognate observations before a per-pair correspondence concentrates on one outcome — and finds that, *conditional on cognate sets*, correspondences that concentrate do so after only a handful of observations, and most concentrate. (We use "posterior concentration / sample complexity"; this is *not* Shannon unicity distance, and we drop that label.)

What this does and does not license, stated plainly:

- It supports $C_{\text{transform}}$ being small — in bits, not in study-hours. It shows the correspondence system is information-theoretically *compact*: cheap to infer from data by an estimator with perfect memory and no interference. It says nothing about how many examples a *human* needs to encode, retain, retrieve, and productively generalise a rule — a different quantity governed by proceduralisation, frequency, and L1 interference (§3.5, §7). We therefore attach no number of examples to human learning, and drop the earlier "≈4" as a claim about learners.
- It does not measure the memory load. The concentration analysis lives *inside* cognate sets; a learner's real load is dominated by the non-cognate share of running text and by grammar (§3.4, §5.6), neither of which this figure measures. We do not equate the residual-of-concentration with $C_{\text{exceptions}} + C_{\text{new}}$ — they have different denominators.
- Honest bounds. The estimate is on curated, core-vocabulary cognacy with small per-correspondence counts and no confidence intervals reported here; take it as a *plausibility* signal for the premise, nothing more. Human learnability is argued only from the trial (§7), never from the same MDL that selected the rules (which would be circular).

7. Empirical validation — a pre-registered design

The core hypothesis is causal and testable:

> H. For a closely-related pair $A \to B$, teaching the transformation T produces faster and more *generalisable* productive competence than teaching B from scratch, at equal study time.

Three arms (native A-speakers wanting B), because a two-arm "new method vs generic course" confounds the *rule* with materials, novelty, and instructor enthusiasm:

- Arm D (DBLL): rules, exceptions, false friends, reachability-mined access points, and the crib-fading sequence (§3.5).
- Arm L (item control): list-learning of *the same cognate pairs* Arm D drills — isolates the rule ingredient from mere exposure to the same items (the clean test of the thesis: rule > instances).
- Arm C (conventional): a strong, credible standard course.

Blinded, self-administered materials; instructors and scorers blind to arm; fixed study budget. A pre-test of untrained transfer (unseen cognates before any teaching) locates every effect *above* the spontaneous cognate transfer §1 says learners already have.

Outcomes (held-out; blinded double coding, reported κ, pre-registered rubric):

1. Diagnostic generalisation (primary). Production of B forms of A-words in *neither* syllabus — crucially, items where the taught rule's output diverges from the naïve cognate guess, so the endpoint isolates *applied the operator* from *read a transparent cognate* (this also defeats the ceiling for very close pairs).
2. Sentence-level production, not just word-forms — to expose the morphosyntactic residual (§3.4) that word conversion hides.
3. Auditory comprehension and timed oral production with the A-prompt withheld (§3.5, stage 4) — the test of whether B competence, not $A \to B$ transcoding, was learned; the §3.3 "mode" is *measured* here, not assumed.
4. False-friend and over-mapping error rates (the predicted failure modes): Arm D must not be *worse* on idiomatic/collocational acceptability, where the over-mapping habit (§3.5) predicts risk.

Power, retention, moderators. A pilot fixes effect-size/variance $\to$ power $\to n$ (not left as "20 hours"). Immediate *and* delayed (2–4 week) post-tests, because rule-teaching predicts a U-shaped dip (early over-regularisation) that an immediate test hides. Aptitude and literacy enter as pre-registered moderators (§5.5), powered for the interaction. Replication on a second pair.

Honest nulls. If Arm D does not beat Arm L on diagnostic generalisation, the *rule* adds nothing over the same items. If D beats L on paper conversion but not on withheld-prompt oral production, the method taught *transcoding*, not B. Either outcome is reported as the finding.

8. The manual series — generating T from data

Because the transformation is the content, and the programme already has tools to *recover* correspondences from parallel/cognate data (alignment, statistical cognacy, MDL-ranked rules), a manual can be semi-automatically drafted for any close pair:

1. Align a parallel/cognate word list $A \leftrightarrow B$.
2. Extract candidate correspondences and rank them by coverage $\times$ reliability (the high-mass, low-residual rules first — the same MDL logic that decides "is this a rule?").
3. Emit the ordered rule list, the exception list, the false-friend list, and the residual-lexicon list — i.e. the four parts of §3, sorted by pedagogical leverage.

This yields a uniform series — *From Dutch to Frisian: Learn by Decoding*, *From Spanish to Portuguese*, *From Czech to Slovak*, *From Swedish to Norwegian* — where only the transformation T changes and the method, the exercise templates, and the cost/distance reporting are shared. Each manual could even publish its measured pedagogical distance D_{ped} and expected coverage, so learners choose targets with eyes open.

9. Relation to existing pedagogy

DBLL is not the first use of the learner's first language: *intercomprehension* pedagogies (e.g. EuroCom for Romance/Germanic/Slavic families; Klein & Stegmann 2000) deliberately exploit cognates for receptive multilingual competence, and contrastive analysis and cross-linguistic-influence research have long studied transfer (Lado 1957; Odlin 1989; Ringbom 2007), including the cognate advantage in vocabulary learning (de Groot & Keijzer 2000). DBLL's specific contributions are (i) making the transformation itself the explicit, drilled object aimed at *productive* competence, not only receptive intercomprehension; (ii) a quantitative cost model and pedagogical distance grounded in the description length of the key; and (iii) a principled account of the boundary conditions and failure mode (coverage-scaling; false friends as systematic over-application; the lexical/grammar split), grounded in a description-length distance. The delta is the packaging — not the idea of exploiting cognates, which is old — as the table makes explicit:

dimension	Intercomprehension (EuroCom)	Contrastive analysis	Cognate-based vocab	DBLL
target competence	receptive	(analysis, not a course)	receptive (vocab)	productive
explicit correspondence rules, drilled?	partial ("sieves")	descriptive only	no	yes
cost model / distance metric?	no	no	no	yes (two-part MDL)
directional reading vs speaking distance?	no	no	no	yes
generalisation-to-unseen endpoint?	no	no	no	yes (primary)
family laddering / key composition?	no (unsequenced)	no	no	yes
reachability-defined "new" lexicon?	no	no	no	yes
lexical vs grammar module split?	implicit	yes	n/a	yes (explicit)

Two honest reads of the table: DBLL is not new in *using* the L1's cognates; it is new in making the correspondence an explicit, drilled, cost-measured, family-sequenced object aimed at production. And it should *build on* intercomprehension rather than oppose it: use the cheap, well-supported receptive phase (EuroCom-style) to assemble the crib, then layer transformation-drilling and crib-fading (§3.5) for production on top of comprehended input.

10. Limitations and honest scope

- Lexical, not grammatical. T reduces the *lexical crossing*; morphosyntax is a separate module the method does not decode (§3.4), so word-level demonstrations do not establish sentence-level

competence.

- Machine compactness $\neq$ human learnability. §6 shows the key is cheap to *infer from data*, not evidence about the human acquisition rate; all efficacy claims are hypotheses pending §7.
- Reading $\neq$ speaking. Coverage and orthographic transparency flatter the *reading* case; the flagship Spanish→Portuguese pair is easy to read and hard to speak (§4.2, §5.6).
- Related languages only. The gain is a function of shared-vocabulary coverage and vanishes for unrelated pairs (§5.1).
- Efficacy is a hypothesis. The computational support (§6) concerns the learnability of the key, not human learning; the causal claim awaits the trial (§7).
- Productive vs receptive. Intercomprehension already delivers receptive gains cheaply; DBLL's distinctive bet is on *production and generalisation*, which is harder and must be shown, not assumed.
- Orthography $\neq$ phonology; register and dialect. Rules differ across the written/spoken channel and across standard varieties (§5.4); a manual must scope each rule.
- False friends and interference. The reliable key can mislead; inoculation mitigates but does not eliminate the hazard (§5.2).
- Individual variation. Learners differ in metalinguistic comfort; rule-based decoding may suit some more than others — itself a measurable moderator in the trial.

11. Conclusion

Between sister languages, the learner is not facing an unknown code but holding the plaintext to a cipher they have most of already. The economical thing to teach is the key — the transformation T that turns what they know into what they want — as a small system of productive rules, a bounded set of exceptions and false friends, and a minority of genuinely new words. This reframing turns language *distance* into a *learning* quantity, predicts exactly where the method wins and where it degrades, and inherits from the parent programme a measured reason to believe the key is information-theoretically compact — cheap to *infer from data*. Whether that compactness converts into faster *human* learning is the one thing this paper does not claim but pre-registers a way to find out — with production of never-seen words, under a withheld A-prompt, as the test that a learner has stopped transcoding one language and started producing another.

References

Cryptography & information theory. al-Kindī (9th c.). *Risāla fī istikhrāj al-kutub al-muʿammāh* [Manuscript on Deciphering Cryptographic Messages]. · Friedman, W. F. (1922). *The Index of Coincidence and Its Applications in Cryptanalysis*. Riverbank Laboratories. · Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4), 656–715. · Chadwick, J. (1958). *The Decipherment of Linear B*. Cambridge University Press. · Knight, K., Megyesi, B., & Schaefer, C. (2011). The Copiale cipher. In *Proc. 4th Workshop on Building and Using Comparable Corpora*, 2–9.

Historical & comparative linguistics. Osthoff, H., & Brugmann, K. (1878). *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*, vol. 1 (preface, on the exceptionlessness of sound laws). S. Hirzel. · Campbell, L. (2013). *Historical Linguistics: An Introduction* (3rd ed.). Edinburgh University Press.

Language data. Heggarty, P., et al. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656), eabg0818. [IE-CoR cognacy database.] · List, J.-M., et al. (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9, 316.

Second-language acquisition & cross-linguistic influence. Lado, R. (1957). *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press. · Odlin, T. (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press. · Ringbom, H. (2007). *Cross-linguistic Similarity in Foreign Language Learning*. Multilingual Matters. · de Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: the roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, 50(1), 1–56. · DeKeyser, R. (2007). *Practice in a Second Language: Perspectives from Applied Linguistics and Cognitive Psychology*. Cambridge University Press. [declarative→procedural; explicit/implicit interface.] · Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, 24(2), 143–188. [usage-based / frequency.] · Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford University Press. [aptitude.]

Speech perception & close-pair interference. Flege, J. E. (1995). Second-language speech learning: theory, findings, and problems. In W. Strange (ed.), *Speech Perception and Linguistic Experience*, 233–277. York Press. [Speech Learning Model.] · Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception. In O.-S. Bohn & M. J. Munro (eds.), *Language Experience in Second Language Speech Learning*, 13–34. Benjamins. [PAM-L2.] · Lipski, J. M. (2006). Too close for comfort? The genesis of "portuñol/portunhol." In *Selected Proceedings of the 8th Hispanic Linguistics Symposium*, 1–22. Cascadilla. [fossilization in a close pair.]

Intercomprehension & mutual intelligibility. Klein, H. G., & Stegmann, T. D. (2000). *EuroComRom — Die sieben Siebe: Romanische Sprachen sofort lesen können*. Shaker Verlag. · Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and Multicultural Development*, 28(6), 445–467.

Model selection. Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471. · Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press.

This programme. Toledo Martínez, A. (2026). *The Cipher of Sound Change* (framing paper); *The Directed, Context-Conditioned Layer of Phonological Correspondences* (Paper 2, unicity-distance result); *The Permutation Dimension: A Mathematics of Metathesis* (Paper 3). The transformations programme.