The Forms of the Endolinguistic Code

Five axes, and why no claim about ”the code” is interpretable until they are declared

Alejandro Toledo Martínez

What kind of document this is. It is not speculative and it proposes no new theses. It orders. The concept of the endolinguistic code has grown — of the word, of the root, of the network, with subclasses, with provenance, with states of loss — and today the sentence ”the code of this word is Λ·Ξ” can mean five different things. This text sets the axes side by side, states which rule governs each, and leaves a checklist. Every figure cited comes from a measurement on the corpus, with its source.

1. Why it is needed

In 2025 the code was a simple object to state: *a bounded, ordered combination of two or three OAS classes, extracted from the normalised root of a word by way of its pronunciation.* That remains true, and it remains insufficient — because practice revealed that the same definition admits readings that do not yield the same object.

Three real examples, all three from this corpus:

- Latin *rīvus* is Λ·Φ·Σ read whole, and Λ in its descendant Spanish *río*. Which is ”the code”?
- English *ignite* is Χ·Ξ·Θ, but the Latin *ignis* it comes from is Ξ·Ξ·Σ: the velar had nasalised ([ˈɪŋ.nɪs]) and English restored it by reading the letter. Which is ”the code” of that network?
- English *tooth* is Θ·Θ, but its co-derivative network is predominantly Θ·Ξ·Θ. Which is taught?

None of the three has an answer until one declares which axis is being spoken on. Hence this document.

2. What a code is, and what it is not

The long definition, unchanged:

> An endolinguistic code is a bounded, ordered combination of two or three OAS classes — themselves discrete, articulatorily anchored and meaningless partitions, indexed to the psychic system under study — extracted from the normalised root of a word by way of its pronunciation, not its spelling. Its unordered set of classes is its nucleus. Order carries quality. A code means nothing: it orients a field of semantic possibility, observable only through the co-derivative networks that instantiate it.

And the negations, which in practice are more useful:

It is not a consonant string · it is not read off the spelling · it is not a morpheme or an etymology · it is not a meaning · it is not the sum of its classes’ profiles · it is not a set (order counts) · it does not belong to a word: it belongs to a network.

3. Axis I · Level of abstraction — the ladder

The same material can be read at three heights, and they are the three the discipline already had:

level	what it is	example	specificity	coverage
skeleton	the actual canonical consonants	`d·n·t`	maximum	minimum
code	the ordered classes	`θ·Ξ·θ`	medium	medium
nucleus	the classes without order	`{θ,Ξ}`	minimum	maximum

Measured over 19,623 networks with ≥3 languages: the mean coverage of the modal value is 56.4 % at the skeleton, 65.1 % at the code and 73.7 % at the nucleus. *(`network_code_profile`.)*

> Coverage rises as one climbs and specificity falls. There is no correct height: there is one per question. And every claim must say at which it was made.

Below the skeleton lies the qualo — the originating consonantal sensation, continuous and pre-symbolic — and above the code lies the molecule, the bonded system of codes that forms a compound word. Neither is a rung of this ladder: they are its floors.

4. Axis II · Extent read — word or root

The same object yields different codes depending on how much of the word is read.

	ROOT code	HEARD-WORD code
contains	the root without affixes	what reaches the ear, affixes included
serves	comparative and descent claims	claims about psychic legibility
the affix is	noise, specific to each language	part of the datum: the hearer does not segment before resonating
they mix	never within one test	

And the affix is not simply noise, for two reasons, the second the stronger: the hearer receives the whole word; and this is how the themes of roots arise — an extension attaches and ceases to be an affix, so the root/affix boundary is not ontological but diachronic.

State of the data: 539,456 forms have **`core_skeleton`**, and in 100 % of the networks where it is computed the root code differs from the word code. The choice is never innocuous.

And there is a third state that is neither: the bare root — the word that *is* its root, where the word code is already the right one. Measured: 592,527 forms are certified as such, and recognising them took the usable population of the emergence test from 3,232 to 70,260 networks.

5. Axis III · Bearer — the form or the network

A word does not have a code. It belongs to a co-derivative network, and the code is what is preserved across it. German *Teil* is Θ·Λ because it stands in the network *Teil·deel·del·dails*. A consonant pattern without a network is not a code: it is a consonant pattern.

Hence the measurable object is not a value but a network profile: the modal value at each level, its coverage, and how many distinct values coexist.

Worked example — TOOTH, collapsed network, 138 languages:

level	modal	coverage	distinct values
skeleton	`d·n·t`	21.0 %	39
code	`θ·Ξ·θ`	34.8 %	24
nucleus	`{θ,Ξ}`	58.7 %	12

English *tooth* (`θ·θ`) comes second, at 17.4 %.

Two counting units, and a third that does not exist. One counts by language — how many languages it serves — and by branch — whether the modal crosses branches or is concentrated in one. By family one cannot: measured, no co-derivative network crosses two families (63,321 in a single family, zero in more than one). This is definitional, not a threshold: a cognate set is intra-family by construction. A-07's "family" unit governs typological claims, which are a different thing.

And networks must be collapsed before anything is measured. The same etymon appears split across several sets: 157,327 cognate sets reduce to 69,416 networks. Without that collapse a test can come out significant merely by counting one etymon several times — which has already happened ($p = 0.0155 \rightarrow 0.133$).

6. Axis IV · Provenance — where the network's code came from

A first-class attribute, with four states:

provenance	what it means	witness
inherited	the code descends from the root	Slavic **bělъ* ? *bʰelH-
supervened	the code arose during the language's history	Lat. *bellus* ? **duenelos*, where the B-L was not there
borrowed	the code arrived with the word	Sp. *bello* ? Occitan *bel*
restored	a class the spoken language had lost returns by reading the spelling	Eng. *ignite* X·Ξ·Θ against Lat. *ignis* Ξ·Ξ·Σ

The fourth is new, and its signature had to be corrected once computed — worth saying why. The first version read "the loan has more classes than its etymon". It fails on its own witness: *ignite* X·Ξ·Θ against *ignis* Ξ·Ξ·Σ has the same three classes. Restitution is not a lengthening: it is a substitution — the X returns where spoken Latin had put Ξ ([ˈɪŋ.nɪs]).

> Correct signature, and it requires THREE generations: the child recovers a class the GRANDPARENT had and the PARENT had lost. It is the classic pattern of *agreeing with the ancestor against the intermediate*.

Computed (`code_provenance`, 183,250 three-generation chains with codes): 3,369 restored forms, of which 21 follow the exact Latin *gn* pattern. And the signature finds the phenomenon on its own: Eng. magnes — grandparent **M**·X·Ξ·Σ → Lat. *magnes* [ˈmaŋ.neːs] **M**·Ξ·Ξ·Σ → Eng. **M**·X·Ξ·Σ. An independent instance of *ignis*, found without looking for it.

And its recall is limited by the data: *ignite* itself does not fall under the signature, because its particular chain has no coded grandparent. The signature is specific but incomplete, and is declared so.

"Convergent" enters no category's name. Convergence is the hypothesis, not a property of the object; putting it in the name bakes the conclusion into the vocabulary.

7. Axis V · State of each position — present, hidden, destructured, absent

A code can shorten for three different reasons that the model used to write identically. Two are events in the language; one is an event of ours.

notation	state	in the word?	in the code?	where it happened
`Φ`	present	yes	yes	—
`[Φ]`	hidden — the instrument does not show it	yes	yes, invisibly	in us
`(Φ)`	destructured — passed into the qualic layer	yes	no	in the language
∅	absent — no longer in the form	no	no	in the language

Examples: Sp. *río* = Λ·(Φ) — the *w* of *rīvus*, set face to face with the vowel, dissolved into it; Ger. *rein* ? *hrainiz = ∅·Λ·Ξ — the *h* lost without transit; and every /j/ in the corpus is **`[hidden]`**, because the J cell has no assigned OAS class: 320,845 segments invisible to the code, which is blindness of ours and not silence of the object.

8. A transversal refinement · subclasses

Within a class there may be a distinction the language does transmit and which the class flattens. Decided and measured for Λ:

$$\Lambda \supset \{\Lambda_r, \Lambda_l\}$$

Luna and *runa* are treated as distinct, connected in the parent. There are still seven classes: a subclass creates no class — it makes nameable, under its parent, a distinction the skeleton already held.

Measured: the subclass is preserved in 94 % of co-derivative networks (0.983 against a null of 0.733). And its 6 % of exceptions is not noise: it is a catalogue of 1,245 cases of L↔R change, sorting itself into Indo-Iranian merger, inherited heteroclisis and suffix alternation.

And the methodological rule that follows: splitting a class can only lower or equal the parent code's preservation — never raise it. The test that applies is not "does preservation improve?" but "is the subclass preserved?".

9. The full grid

A claim about a code is determined by five coordinates:

```
   level        ×   extent   ×   bearer   ×  provenance  ×    state
skeleton           word          form         inherited       present
code               root          network      supervened      hidden
```

```
nucleus          (bare root)              borrowed        destructured
                                          restored        absent
```

And most disagreements about codes are disagreements of coordinate, not of fact. *Tooth* is Θ·Θ and its network is Θ·Ξ·Θ: both true, on different bearers.

10. The rules governing each axis

axis	rule
level	C-10′ the code is a ladder; the rung is declared
level	C-25 the code is binary or ternary, because roots are; 4+ classes is a word with affixes
level	C-21 order is load-bearing: with 69,244 pairs the shuffled condition is flat — a code is not a set
extent	C-10 the two codes are never mixed within one test
bearer	C-19 the code is the property of a network, not of a word
bearer	C-22 collapse duplicate networks before pairing; without that control a false positive already occurred
bearer	C-20 the null is drawn from the same class of object: if networks are compared, networks are drawn
provenance	C-18 four states; ”convergent” enters no name
state	PE-01…04 every lost position declares its state and where the loss occurred
all	C-15 a two-class code is carried by 513 words: sharing it is not evidence
all	C-26 but C-15 weighs evidence, it does not define the object: filtering to $\geq$4 classes excludes the real codes

C-15 and C-25 coexist thus, and it is the architecture the theory always had: the code is binary or ternary and therefore does not identify on its own. What identifies are the network and the field.

11. Checklist

Any code reading should be able to answer:

1. At which level — skeleton, code or nucleus?
2. On word or root? And if word, is it a certified bare root?
3. Is the bearer a form or a network? If a network, were duplicates collapsed?
4. What is the coverage of the modal value, and in which unit — languages or branches?
5. What provenance does the code have in that network?
6. Is any position hidden, destructured or absent? Was the loss the language’s or the instrument’s?
7. How many words carry that value — that is, how much does sharing it weigh?

All seven are answerable today with the corpus. The seventh is not precomputed but is calculated in 115 ms per value; an index on `skeleton.code` is missing.

12. What remains open

- The didactic optimum. Coverage is in hand (axis I $\times$ III) and specificity can be computed (C-15); crossing them is what is missing, in order to choose a rung by criterion rather than by eye.
- Language de-duplication. 120 glottocodes carry several `lect_id`s and 1,387 lects have no glottocode: coverage by language is inflated by an amount not yet measured.
- The J cell. With no OAS class, every /j/ is invisible. It is the largest hole in the inventory.
- And the underlying question, which no axis answers: the five axes say how a code is read. None says what it detects. That remains the discipline's question: what persists through genealogy is not the code, but a structural continuity the code allows us to observe.