Integrative Corpus

A multi-layer lexical database of 225 macrosystems, with row-level provenance and its verification apparatus

Alejandro Toledo Martínez · ORCID 0009-0000-1277-9697 · 2026

Abstract

We present the Integrative Corpus, a lexical database integrating eighteen sources into 3,788,294 forms from 4,735 doculects across 225 macrosystems, with phonological segmentation, row-level provenance and a computed layer of 3,091,596 normalised consonant skeletons that no source supplies.

The base layer is aggregation: every form comes from elsewhere. What is contributed is the union built so that one can audit where each figure came from, a phonological representation computed with the same rules for all 225 macrosystems, and the decisions and gaps declared with their cost measured.

We describe in equal detail the verification apparatus published with the corpus: 60 invariants that fail rather than warn — and the five that only warn, enumerated one by one — a schema contract declaring what each column means and which it is confused with, and meta-checks whose object is not the data but the reach of the apparatus itself. Its justification is six real errors from this project — five of which passed every check that existed at the time — which turned out to be the same error six times over: a check that covered less than its name promised and reported conformity exactly when there was a failure.

Coverage is given by layer rather than as a single figure, because none exists: two consumers of the same corpus — a word-history query layer and a catalogue of form–meaning correspondences — need different layers and arrive at different figures, both correct.

The corpus is distributed under CC-BY-SA-4.0, a licence imposed by the 38.6% of forms coming from a ShareAlike source. The repository is the distribution: it is cloned and rebuilt. Access is granted on request — to whoever asks, with the right to fork — because the work is in progress, not out of any reservation about the data: what is managed is access to the repository, not the rights over what it contains.

1 · The problem

Comparing words across languages requires four things at once: the form, the meaning, a comparable phonological representation and the provenance of every datum. Existing resources each give one part, and give it well, because each was designed for a different question (§2).

The result is that a question needing two of those dimensions at once has nowhere to be asked. A catalogue of wordlists covers 225 macrosystems and carries no etymology; a dictionary dump carries all the etymology of one macrosystem and maps none of its entries to concepts; expert cognacy judgements exist for 170 concepts and one family.

Joining those sources is the easy part. Joining them so that afterwards one can audit where each figure came from is the part that decides whether the result is usable, and that is what this work is about.

What this corpus is, and is not

It is a multi-layer lexical database integrating eighteen sources: 3,788,294 forms from 4,735 doculects across 225 macrosystems, with segmentation, row-level provenance, and a computed layer of 3,091,596 normalised consonant skeletons that no source supplies.

The base layer is aggregation. Every form comes from somewhere else. There is no new collection of primary data here, and presenting it otherwise would be false. What is contributed is three things, none of them ”more data”:

1. An auditable union, not merely a union: every row knows where it came from, and that is what allows any figure to be traced back to its source afterwards (§3).
2. A phonological representation common to all 225 macrosystems, computed with the same rules for all, which is what makes languages documented by different projects comparable (§4, §5).
3. The decisions and the gaps, declared with their cost measured (§6, §9).

And one more thing, perhaps the most transferable

A corpus integrating eighteen sources accumulates errors that do not manifest as errors: no exception, no missing row, no failing query. There is a figure slightly different from what it should be, and nothing flags it.

This work publishes, with the corpus, the apparatus built to make that class of failure visible: 53 invariants that fail rather than warn, a schema contract declaring what each column means and which it is confused with, and — what we believe most useful — meta-checks, whose object is not the data but the reach of the apparatus itself (§7).

The justification for that apparatus is §8: six real errors from this project, five of which passed every check that existed at the time, and which turned out to be the same error six times over.

How to read what follows

§2 situates the corpus among existing resources. §3 and §4 describe the model and how each layer is computed. §5 details what is computed. §6 gives coverage and where it stops. §7 and §8 are the verification apparatus and the errors that justify it. §9 covers licensing and distribution, §10 the limits and §11 how to cite it and who uses it.

Anyone about to use the corpus should start with §3.3 — the trap of the two provenance columns — and with §6, which says how far each layer reaches.

2 · Related resources

It is worth saying before anything else: the base layer of this corpus is aggregation. Every form comes from elsewhere, and those places are good. There is no new collection of primary data here, and presenting it otherwise would be false.

What follows describes what each resource does — well, and for what it was designed — and what remains unresolved when they are needed together.

2.1 · The identity infrastructure

Three catalogues that supply no forms and without which none of this could be joined:

Glottolog gives each doculect's stable identifier (the glottocode) and its classification. It is what makes it possible to know that two rows from different sources are the same language — and it is the basis of the mechanical duplicate rule of §3.6.

Concepticon gives each concept's stable identifier, linking the elicitation lists each source uses. Without it, "tres" from one list and "three" from another are two things.

CLTS gives the transcription standard and the articulatory features, which in our corpus cover 95.6% of segments (§6.5).

And CLDF is not a dataset but the format that makes all of the above fit together. It is the reason aggregation is possible at all: without a common packaging, integrating eighteen sources would be eighteen parsers and eighteen sets of tacit assumptions.

2.2 · The form sources, and what each is deep in

Measured on our own ingestion, which is what we can state with confidence:

source	forms	languages	macrosystems	concepts	with skeleton
Lexibank	1,740,092	3,120	225	3,205	100%
Kaikki	1,457,643	37	1	2,452	82%
IDS	437,902	269	58	1,308	0
NorthEuraLex	121,611	107	21	954	100%
IE-CoR	25,731	152	1	170	96%

Lexibank is the breadth: 3,120 languages and all 225 macrosystems, with wordlists elicited against concept lists and consistently segmented. It is what makes any comparative work of scope possible, and that is why the macrosystems at 100% skeleton coverage are exactly those coming wholesale from it. What it does not give — nor claim to — is history: it is a collection of wordlists.

Kaikki / Wiktextract is the depth, in a single macrosystem. 37 languages and one macrosystem, but 1.46 million forms and practically the entire etymological layer: 1.24 of the corpus's 1.26 million etymology edges come from it, plus 72,023 proto-forms and two of the three cognacy networks. It is a dictionary, and brings what a dictionary brings: prose definitions, etymologies, descendants. What it does not bring is concept mapping — 14.8%, §6.2 — because its entries were not elicited against any list.

IDS is breadth of concepts: 269 languages and 58 macrosystems, all mapped. Its CLDF publishes no segmentation, hence its zero skeletons (§6.2).

NorthEuraLex contributes 107 Northern Eurasian languages with its own consistent IPA.

IE-CoR contributes the scarcest thing there is: expert cognacy judgements, 152 Indo-European languages over 170 basic-vocabulary concepts. It is the corpus's only gold cognacy, and therefore the only place where a method designed to work without it can be checked against a known answer.

2.3 · What none of them resolves, and that is the gap

Read the table by columns and the pattern jumps out: each resource is deep in one dimension and shallow in the others, because each was designed for a different question.

- Breadth of languages and historical depth do not coincide: Lexibank covers 225 macrosystems without etymology; Kaikki covers one with all the etymology.
- Concept coverage and phonological coverage do not coincide: IDS maps 100% of its forms to concepts and publishes not one segment; Kaikki segments 80% and maps 15%.
- Gold cognacy exists over 170 concepts and one macrosystem.

None of this is a defect. These are resources that do their own job well. The gap appears when a question needs two dimensions at once, which is the case for any work wanting to compare forms by their meaning in macrosystems where there is no reconstruction: it needs form and meaning and a comparable phonological representation, over 225 families rather than one.

2.4 · What this corpus adds, stated narrowly

Three things, and none of them is ”more data”:

1 · The union, with row-level provenance. Integrating eighteen sources is easy; integrating them so that one can afterwards audit where each figure came from is not, and it is the difference between a usable corpus and a mixture. Hence the model of §3 and the apparatus of §7.

2 · A derived layer that does not exist upstream: 3,091,596 normalised consonant skeletons across 225 macrosystems, computed with the same rules for all (§4). The sources bring segmentation; the canonical consonant by qualic value is our computation, and it is what makes two languages realising the same consonant differently comparable.

3 · The decisions and their gaps, declared and measured. Not ”we exclude IDS” but why, with the figure for what it costs; not ”we withdrew a source” but which claims stop holding when we do. §6 and §9.

What this corpus is NOT: a primary source, a substitute for any of the resources in §2.1 and §2.2, or a dataset broader than Lexibank. It is the union of several with a computed layer on top, and its value lies in that union being auditable.

3 · Data model

87 tables, organised in four layers that stack: each is computed from the previous one and none overwrites what it received.

3.1 · The core

```
lect ──< form >── concept
          │
          ├──< segment      (19,848,118 · one sound per row)
          ├──  skeleton     ( 3,091,596 · the consonant skeleton)
          └──< sense        (the senses the source gives that form)
```

table	unit	what it is
`lect`	a doculect	a language *as documented by one source*. It is not a language in the glottological sense: two rows may be the same language from two sources, and that is why a mechanical duplicate rule exists (§3.4)

table	unit	what it is
`concept`	a concept	the gloss by which forms from different languages are grouped. It is the key to all comparative work
`form`	a form	the word: orthography, source IPA, segmentation, provenance
`segment`	a sound	one row per IPA segment, with its position and annotation
`skeleton`	a skeleton	the form's canonical consonants, computed from `segment` (§4)

3.2 · The historical layer

`form_etymology` (1,262,353 form→ancestor edges) · `ancestry_edge` (lineage between languages) · `cognate_set` / `cognate_member` (157,327 sets, 631,507 members) · `protoform_hypothesis` (72,166).

As §6.4 states, this layer exists only for Indo-European. It is in the model because the model contemplates it for any macrosystem; it is empty in 224 of 225 because none of the currently integrated sources supplies cognacy outside Indo-European — a gap in this integration, not in the field (§6.4).

3.3 · Row-level provenance, and the trap that published two errors

Every form carries its source, and its upstream record identifier. This is not a design preference: it is the compliance mechanism for CC licences, which require attribution, and there is an invariant that fails if any form lacks it.

It is worth stating how far this reaches, because "row-level provenance" can be read as more than is sometimes offered. Here it offers two things: the source (`source_id`) and the upstream record identifier, carried inside `form.id` and preserving the structure of the dataset it came from:

```
lexibank:aaleykusunda-KusundaGM-above-1
   source : dataset - doculect - concept - no.

ids:107-1-100-1
   source : language - chapter - entry - no.

nel:abk-1000_know-1
   source : iso - concept - no.

kaikki:af:'n appeltjie te skil hê:verb
   source : language : entry : part of speech
```

So from a row one can return to the specific record in the source dataset, not merely to the dataset. What is not yet possible is returning to a specific *version* of that dataset, because versions are not pinned (§9.6.1): traceability reaches the record and stops short of the version.

But there are two `source_id` columns, and they mean different things:

	unit	meaning
`form.source_id`	the form	which source *this word* came from. This is the one to use for any per-source statistic
`lect.source_id`	the lect	which source *the language record* came from. It does not say where its forms come from

One language gathers forms from several sources, so the two breakdowns differ, and by a lot:

```
lect.source_id='wiktionary'  with forms from kaikki    133,120
lect.source_id='kaikki'      with forms from ids        45,404
lect.source_id='ids'         with forms from lexibank   19,206
```

The schema allows either to be written and nothing warns which was intended. Confusing them published two errors: a per-source loss table giving 65.6% where it was 100%, and a hygiene check that attributed 85% of Turkic infinitives to the wrong source.

This document declares it because it is a warning to anyone using the corpus, not merely an internal note: the trap is not in our code, it is in the schema, and anyone writing a query can fall into it.

A third thing needs bounding: which tables carry provenance at all. This work's title says "row-level provenance", and that phrase is true of the `form` table — which is what everything above concerns — but not of the whole database. There are today thirteen tables with neither a provenance column nor a link column through which to inherit one. Eleven are our own production: aggregates, statistics, study-output tables; their provenance is us, and there is nothing to attribute. The other two do hold third-party content: `concept` (3,205 rows of glosses and definitions from Concepticon) and `feature` (33,912 rows of articulatory features from CLTS). Both are CC-BY-4.0, both are cited in §12, and in both the provenance is single for the whole table — every row of `concept` is from Concepticon — so a `source_id` column would be a constant. There is no licensing problem; what there is, is an edge to the invariant of §7, and calling it anything else would be describing a guarantee that does not exist.

Since 6 September 2026 that edge is written down: the thirteen tables are declared one by one with their reason in `ci/decisiones.py`, and a meta-check discovers them from the catalogue and fails if one appears that is not declared (§7.4). The guarantee still has a limit; the difference is that the limit is auditable, and that a new table without provenance cannot cross it in silence.

3.4 · The schema contract

Hence an artefact published with the corpus: `ci/esquema.py` declares, for every column the studies use, four things — how it is written in SQL, what it is a property of (lect, form, concept, skeleton), what it means in one sentence, and what it is confused with.

The fourth is the one that matters. The two `source_id` columns are declared mutually confusable, and a test fails if that marking disappears.

The third carries a rule of use: if the meaning cannot be written in one sentence, the column is not understood and should not be used in an analysis.

`analysis/tools/verifica_esquema.py` checks the contract against the live database, so it does not age silently when a migration changes a column.

3.5 · Fields that look like data and are heuristics

Two schema columns look like observations and are conjectures. They are declared together, by name, because naming them is half the work:

`form.is_proper` does not say whether the form is a proper noun: it is an initial-capital detector. Of the 99,698 forms flagged in Indo-European, 94% simply begin with a capital, and German and English are 82,000 of them. `Abzweig`, `Axt`, `Mahl` are common German nouns. Filtering by this column would erase the German common lexicon.

`form.is_loan` is a borrowing marker inherited from the source, with different criteria per source. It is neither comparable across sources nor complete.

3.6 · Declared facts about the lects

The corpus declares what it knows about its own lects, without deciding what to do with them:

class	what it asserts
`duplicate`	the same doculect recorded twice from different sources (same glottocode, high form overlap). Which to keep is indicated
`misassigned`	classification error at source. There is one: a Kam-Sui language assigned to Uralic, detected by three independent checks — it is 6,446 km from the family's centre, it marks tone in digits when no Uralic language does, and its forms are θaːm^{453} "three" and ðɐm^{214} "water"
`reconstructed`	the lect is unattested

The class is a fact; what to do about it is each consumer's policy. A catalogue of correspondences excludes reconstructions because its design is not to look at them; a query layer showing a word's lineage back to Proto-Indo-European displays them, because that is what it does. Both are correct, and that is why the decision cannot live in the corpus.

4 · Ingestion and normalisation

4.1 · The eighteen sources

They are ingested in their native format: CLDF for the ecosystem databases (Lexibank, IDS, NorthEuraLex, IE-CoR, Concepticon, Glottolog, CLTS), JSON dump for Kaikki, and our own scraping for those publishing no structured format.

Each enters with its row in `source`, carrying its citation, URL, type and its licence by name. "Open" is not a licence, and two sources were declared that way — noted in §9 as pending identification or withdrawal.

Where what this corpus can guarantee ends. The `source` table carries no source version, and for `kaikki` — 38% of the forms — it no longer can: it is a rolling dump regenerated from Wiktionary by overwriting, whose date lives on the page and not in the file. What there is instead is the `sha256` fingerprint of the 148 raw files in `docs/MANIFIESTO-FUENTES.md`, so the pipeline's input is identified even though its version is gone. The division of responsibility is in §9.6.1.

4.2 · The computation chain

```
form.ipa_raw          the transcription as the source wrote it. NEVER altered
      ↓
form.segments_raw     the segmentation, from the source when it supplies one
      ↓
segment               one row per sound, with position and annotation
      ↓
skeleton.cons_skeleton   the canonical consonants, separated by ·
```

No layer overwrites the previous one. `ipa_raw` is preserved whole, and that is what made it possible to recover African tone when the segmentation was found to destroy it (§6.6): the data was still there.

And one rule governs the whole chain: the skeleton is computed from **`segments_raw`**, never from the orthography. When a source brings no segmentation there is no skeleton, and that is what leaves IDS at zero (§6.2).

4.3 · Three layers, and a function between each two

The corpus distinguishes three representations and each step between them is a reproducible function, not an interpretation. It is put this way because using the corpus must not require accepting any theoretical hypothesis: anyone who only wants to compare forms can stop at the second layer and ignore the third.

```
segment          the sound as the source segments it
   │  f1 : table IPA -> canonical consonant
   v
cons_skeleton    the canonical consonants, separated by ·
   │  f2 : table consonant -> class symbol
   v
code             the sequence of classes
```

Both tables — `f1` in `ingest/recompute_skeleton.py` — are explicit dictionaries, not procedures: given a segment, the output is determined and can be reproduced without running them. An invariant checks that the two sequences have the same length, so position-by-position correspondence is guaranteed.

The corpus claims nothing about the third layer (§5.1): it is published as a column and its interpretation belongs to other work. The name that literature gives to the criterion behind `f1` — "canonical consonant by qualic value" — is not needed to use the corpus: what must be known is what the table groups, and that is below.

4.3bis · What a consonant skeleton is

The complete sequence of a form's consonants, each replaced by its canonical consonant under `f1`. `kərniajc` → `k·r·n·c`.

We deliberately collapse contrasts that are phonemic in many languages. It is worth putting it that way and no other: the rhotics — `ɾ ɹ ɻ ʀ ʁ ɽ` — are written `r`; the nasals `ŋ ɲ ɴ` are written `n`, and `ɱ` is written `m`; the dorsal fricatives `ɣ χ ç ʝ` are written `x`, and the glottals and pharyngeals `ɦ ʔ ħ ʕ` are written `h`. Aspiration, labialisation and palatalisation are stripped.

It is not claimed that phonology treats those segments as variants: `/n/` and `/ŋ/` contrast in a great many languages, and aspiration and palatalisation are phonemic across whole families. What is done is to choose a scale of comparison in which those contrasts are not represented, which is a decision of the instrument and not a description of the languages.

Why that scale. Without collapsing, two languages saying the same word with different realisations of the same consonant do not group, and surface variation hides the correspondence. There is a price: any question depending on a collapsed contrast cannot be asked at this layer. The method paper quantifies that price for one case — merging the sibilants makes lambdacism invisible, and lambdacism defines the primary split of Turkic.

That is why `ipa_raw`, `segments_raw` and `skeleton_raw` are preserved intact: the normalisation is an added and reversible layer, not a substitution. Whoever needs the contrast has it one layer down.

4.4 · Four decisions that are not obvious

They are declared because each changes what comes out, and three were born from errors.

Prenasalisation ≠ nasalisation. Vowel nasalisation (`ã`) is a modulation of the segment and is stripped. The prenasal of `ⁿt`, `ᵐb`, `ᵑg` is a nasal segment and occupies its position. They were in the same basket,

and the consequence was visible: `mainty` gave `m·t` while its cognate `metan` — unprenasalised — gave `m·t·n`. The same nasal, present in one language and erased in its sibling, by a decision of the instrument. Aspiration, labialisation and palatalisation are still stripped: the asymmetry is deliberate, and is written where it applies.

Notation that is not a segment, and a defect that is. The first version of this paragraph put three different things in one sentence, and one of the three was wrong.

Tone superscripts and parentheses are notation, not segments, and do not enter the skeleton. That is right — and it is exactly what destroys African tone, where tone rides as a diacritic on the vowel: for those languages it must be read from `ipa_raw`, which is preserved (§6.6).

Clicks are consonants, and until 2026-09-06 they were being discarded. `ǀ ǁ ǂ ! ʘ` are core consonants in the families that use them, and they were treated in two wrong ways at once: `ǁ` lived in the lateral group of the class map, so **`ǁao`** gave **`l`** — merging the lateral click with /l/ — and the others sat in the ignored-notation list, so they disappeared. In 39 forms the click was the only consonant and the word was left with no skeleton at all.

The first was worse than the second, and it is worth saying why: an absence can be declared and measured; a wrong value cannot, because nothing distinguishes it from a right one. `l` is a well-formed skeleton, leaves no residual and trips no invariant. It is the class of error §8 is about.

How it was resolved, and why not otherwise. The click enters the skeleton as itself: `ǀo` gives `ǀ`, `ǁk'x'um` gives `ǁ·k·x·m`. No canonical consonant is assigned, because what a click corresponds to is an open question this work has not studied, and assigning one would be writing a hypothesis as if it were data — which is exactly what the previous version did.

In the class layer it carries `[?]`, the symbol the corpus already uses for "the element is in the word and the instrument cannot show it", which already applied to glides for the same reason. It is not a gap: it is a position occupied by something not yet classified.

Result: of 491 forms with a click, 488 keep it in the skeleton and none is left without a skeleton. The remaining three are in languages without clicks — Anamgura, Maleng — where the sign is a source separator and not a consonant.

Concealment marked, not erased. When the element is in the word and the instrument cannot show it, the position is kept marked with `[?]` rather than disappearing. `year` gives `[?]·ʌ`, not `ʌ`. It is our event, not the language's, and so it is written differently from a real absence. An invariant checks that no residual `[?]` reaches `cons_skeleton`, the layer the studies use.

The biclass rule exists in the code and is NOT applied to the data. It is described here because an earlier version of this text presented it as a live decision, and it is not. The rule provides for some segments — the displaced sibilants `ʃ ʒ ʂ ʐ ɕ ʑ` — to contribute two classes rather than one, but the published corpus was built with it switched off: `ʃ` gives a single entry in the skeleton and a single one in the code, and the invariant requiring `#classes = #consonants` passes at zero, which is the check that it is not active.

4.5 · What is NOT normalised, and why

The orthography. `form.orthography` is displayed as is. Only the dot some sources use to separate segments is removed — `b.i.r`, `w.s.w.r`: 296 Turkic forms — and only when it separates single characters, because showing "b.i.r" in a table would make the reader think Turkish says that.

Concepts across sources. Each source brings its gloss; they are not unified by our judgement beyond the Concepticon mapping the source itself declares. Unifying glosses is a semantic decision and has no place in ingestion.

Nothing per family. There is no automatic morphological stripping. Cutting endings that "look like

suffixes" is the quick route to manufacturing exactly the codes one wants to see, so stripping rules are declared by hand, per family, with a double condition — orthography and skeleton tail — and live in the project that uses them, not in the corpus.

5 · The computed layers

On top of the core of §3, layers are computed that no source supplies. This is the corpus's material contribution, and two things it holds mixed are worth separating: the layers the corpus offers to anyone and the tables a particular study left behind.

5.1 · The consonant skeleton and its class layer

table	rows	what it is
`skeleton_raw`	3,034,556	the consonants before normalisation
`skeleton.cons_skeleton`	3,078,840	the canonical consonants by qualic value (§4.3), out of 3,091,596 rows in `skeleton`: 12,756 forms have no consonant at all
`skeleton.code`	3,078,840	the sequence of classes of those consonants
`skeleton_lineage`	461,686	grouping of one code across stages and branches

That `skeleton_raw` is kept separately is not redundancy: it is what allows the normalisation to be undone and what was merged to be checked. A computed layer without its input is a claim with no way to audit it.

And it must be explained why it has FEWER rows than the layer it precedes, because intuition says the opposite. There are 3,034,556 of 3,091,596, or 98.2%, with 57,040 missing. There are none the other way — zero `skeleton_raw` rows without their skeleton.

An earlier version of this section attributed the gap to the pass being additive, covering only what existed when it ran, and concluded that re-running it would close the gap. It was re-run on 6 September 2026 and wrote 92 rows: the gap went from 57,132 to 57,040. The explanation was false, and the true one — three causes, summing exactly — belongs in its place:

	forms	what happens
A	34,190 · 59.9%	the form has no **`segments_raw`**. The pass reads that column, so it excludes them by construction; their skeleton was computed by the orthographic route
B	12,590 · 22.1%	the form has no consonant at all: it is all vowel. There is no raw skeleton to write
C	10,260 · 18.0%	the skeleton consists only of glides (`j`, `w`, `ɥ`)

A and B are not defects: they are the definition of the layer. A form with no `segments_raw` has no "before normalisation" to record, and a form with no consonants has no skeleton in either version. No pass will close them, and saying one would was promising an impossible repair.

C is an inconsistency, and is declared as one. On 31 August 2026 it was settled that a glide is not an exclusion category but a position on the saturation gradient, and since then `skeleton` puts it into the skeleton with the hidden-element symbol (§4.4). The `skeleton_raw` pass never received that change and still discards it. The two layers disagree about what counts as skeleton, and since `skeleton_raw` exists precisely to audit normalisation, the disagreement attributes to normalisation a difference that is really about glide inclusion.

Aligning them would change 499,546 forms — 16.2% — and force the recomputation of `morph_rule` (2,823,128 rows), `morph_rule2` (2,180,499) and `affix_code_lect`, which feed the morphological decomposition and the irreducible nuclei. Decision (2026-09-06): the data is not touched in this release and the divergence is declared, with a check that counts it and fails if it grows. Changing published figures to close 18% of a 1.8% gap does not justify it; concealing it does not either.

About the class layer, a warning that is part of the data. The `code` column reduces each consonant to a class. It is in the corpus, it is published, and the corpus claims nothing about it: it is an instrument whose status requires a study this work does not undertake. It is offered as a column, not as a result.

And it has a declared residual: 303,434 skeletons carry **`?`** in **`code`** — unclassified classes — against zero in **`cons_skeleton`**. The asymmetry is deliberate and is the thing to look at: the layer the studies use is clean; the one that is not interpreted carries its gaps in plain sight. There is an invariant for each, at different severities.

5.2 · Morphology

`morph` (1,307,991) and `morph_rule` (2,823,128) decompose the form into pieces with their role — root, affix — and each piece's code. `affix` (18,853) and `construction` (24,267) collect the resulting inventory.

The origin matters and bounds the reach: the decomposition is not guessed from orthographic resemblance, it is taken from what Kaikki already publishes in the etymology — *"From angosto + -ura"*, *"sub- + marino"*. Where no decomposition is declared, the root is left null, which is the honest outcome. Cutting endings that *look like* suffixes is the quick route to manufacturing the result one is after.

Consequence: this layer is as broad as Kaikki's etymological coverage, that is, Indo-European (§6.4).

5.3 · Meaning and correspondence

table	rows	what it is
`polyseme_link`	1,026,479	forms the source gives more than one sense
`colex`	231,313	pairs of colexified concepts: expressed by the same form in one language
`correspondence`	111,778	sound correspondences between lect pairs, with environment and type
`sound_class`	3,516	the class inventory
`racimo` / `racimo_member`	5,531 / 29,152	groupings by root and stem

`colex` is the material for colexification networks; `correspondence` records, for each lect pair, which class corresponds to which in which environment and how many times, distinguishing retention from change.

5.4 · What is in the database and is NOT a corpus layer

This surfaced while writing this section and is better written down than glossed over. The database also contains tables that are the output of particular studies, not general-purpose layers:

`crypto` (3,061,314) · `code_axes` (800,723) · `code_specificity` (838,128) · `network_code_profile` (209,343) · `code_family` (26,856) · `code_clean` (3,068,519) · `code_provenance` (244,576)

They are recognisable because their schema carries the question inside it: `code_axes` classifies network pairs into boxes such as ”pure resonance”; `network_code_profile` stores a network's modal value and its coverage. That is not a fact about languages: it is the result of one particular way of looking at them.

Since the separation of 2026-09-05 those studies live in a separate repository, and this section said these tables should follow them. On going to move them, on 6 September, the premise turned out to be false, and it is worth correcting here rather than leaving it as debt pending a move that does not apply.

Who queries them was measured, distinguishing who reads them from who builds them. Of the twenty-five that have no foreign keys and could therefore move, **`codigos`** — the studies' repository — reads none, and Meulemans reads seventeen: `morph_rule2`, `net_quality`, `network_code_profile`, the six `ncp_*` tables, `code_specificity`, `code_family_b`, `code_family_dmas`, `form_nucleus2`, `affix_code_lect`, `affix_allomorph` and `f6_bateria`. The remaining six are not loose either: `morph_rule`, `form_nucleus` and `affix_inventory` are the input from which `nucleus_inventory` is built.

So they are not study output awaiting a move: they are what the query layer serves, and their place is where they are, for the same reason the `qf_*` tables stayed when Tatoeba left (§9.6.3). What did move was what really was study output that nobody else read: `construction` and `construction_align`, which went to the `qualicfields` database.

The criterion below stands, and so does the objection that motivated it: **`code_axes`** is still a table whose definition is a study's hypothesis, and its presence in the corpus database is arguable even though there is nowhere to take it today. What changes is that this is no longer called debt from a move, but a tension between the criterion and the fact that the query layer was built on top.

The criterion for telling them apart, in case it helps anyone building something similar: a layer belongs to the corpus if more than one consumer would use it without disputing its definition. `skeleton` does — both the catalogues and the query layer use it. `code_axes` does not: its definition is one study's hypothesis.

6 · Coverage, and where it stops

This section describes the corpus's layers and how far each reaches. It gives no figure for ”usable corpus”, because none exists: what is usable depends on the question, and two consumers of the same corpus — a word-history query layer and a catalogue of form–meaning correspondences — need different layers and would arrive at different figures, both correct. What is given instead is each layer with its coverage, and the intersections that vary most across macrosystems.

6.1 · The corpus, by layer

lects	4,735 (3,350 with glottocode)
macrosystems	225
concepts	3,205
forms	3,788,294
forms with a concept assigned	2,533,666 (66.9%)
consonant skeletons	3,091,596
segments	19,848,118
etymology edges	1,262,353
cognate sets	157,327 (631,507 members)

proto-forms	72,166
sources	18, none carrying a licence that prevents redistribution — with the caveat in §9.6.2

There is no single ”usable corpus” figure, and it is worth saying so first, because the temptation to give one is strong and it produces false descriptions. What is usable depends on the question:

- A browser of a word's history — such as the Meulemans query layer (§13.1) — needs form, sense, etymology and cognates. It is served by 3.79 M forms, and a form without an assigned concept is perfectly good material: you look up the word, you see its lineage.
- A catalogue of form–meaning correspondences — such as the consonantal code programme (§13.1) — needs skeleton and concept together, because its unit is concept × language. There, what is usable is 2,058,119 skeletons, 66.6% of those that exist.
- A phonological study of inventories needs no concept at all, and is served by the whole 3.09 M skeletons.

All three figures are correct. This section describes the layers and their coverage, and whoever uses the corpus composes their own.

6.2 · Two large sources failing in opposite directions

source	forms	with concept		with skeleton
`lexibank`	1,740,092	1,740,092	100%	1,740,092
`kaikki`	1,457,643	215,568	14.8%	1,199,823
`ids`	437,902	437,902	100%	0
`nel`	121,611	114,373	94.0%	121,611
`iecor`	25,731	25,731	100%	24,761

The asymmetry explains the shape of the whole corpus and is worth seeing side by side:

Kaikki brings the form without the meaning. It is a dictionary dump: entries with prose definitions, not elicited against a concept list. Mapping a Wiktionary definition to a Concepticon concept is a semantic matching problem, and today it resolves 14.8%. That is 1.24 M forms with a computed skeleton and no concept: invisible to a catalogue, fully usable for looking up a word.

IDS brings the meaning without the form. Its 437,902 entries come from a concept list, so 100% are mapped — but its CLDF ships the `Segments` column empty at source and its forms are orthography, not IPA. Without segments there is no skeleton: zero of 437,902. Converting that orthography would require an orthography profile per language for 274 languages, and a bad profile does not lose the form: it manufactures a false skeleton that enters looking like data. Declared out of scope, with its reason, in `docs/DECISION-IDS.md`.

6.3 · By macrosystem

The penultimate column is the intersection of two layers — skeleton and concept — not a judgement about what is useful: it is what a concept × language catalogue needs, and it is given because it varies most across macrosystems.

macrosystem	languages	forms	skeleton		skel. + concept		concepts
Indo-European	185	1,554,812	1,270,368	81.7%	323,476	20.8%	2,785
Austronesian	586	233,351	218,840	93.8%	218,840	93.8%	1,825
Nakh-Daghestanian	56	231,624	104,562	45.1%	104,170	45.0%	1,818
Sino-Tibetan	270	200,399	200,399	100%	200,339	100%	2,737
Atlantic-Congo	387	157,076	157,076	100%	157,076	100%	1,787
Uralic	28	101,065	82,952	82.1%	81,142	80.3%	1,929
Nuclear Trans New Guinea	196	94,306	94,306	100%	94,306	100%	897
Tai-Kadai	28	87,799	54,434	62.0%	54,434	62.0%	1,736
Austroasiatic	102	86,305	49,866	57.8%	49,866	57.8%	1,570
Turkic	33	70,853	55,877	78.9%	55,300	78.0%	1,828
Afro-Asiatic	112	69,237	64,460	93.1%	64,323	92.9%	2,034
Tupian	64	48,532	41,269	85.0%	41,269	85.0%	1,514
Pama-Nyungan	155	45,855	45,855	100%	45,855	100%	1,160

Long tail: 204 further macrosystems, 657 languages, 392,008 forms, 328,051 skeletons. Plus 142 lects with 222,204 forms with no macrosystem assigned.

The largest macrosystem is the one that overlaps least, and it is a direct consequence of §6.2. Indo-European has 1.55 M forms and only 20.8% carry skeleton and concept together, because nearly all of Kaikki is Indo-European. The macrosystems at 100% — Sino-Tibetan, Atlantic-Congo, Trans New Guinea, Pama-Nyungan — are there because they come wholesale from Lexibank, which is material elicited against concept lists.

Again: this does not say Indo-European is less useful, it says what each part is useful for. Its 1.23 M forms with skeleton and no concept are exactly where the etymological layer lives (§6.4), which exists in no other macrosystem. The size of a macrosystem does not predict what can be done with it, and neither does any one of its layers.

6.4 · Etymology and cognacy: Indo-European, and nothing else

This is not a shortcoming the corpus should apologise for: it is the field's starting condition, and it is why everything built on top is organised as it is.

macrosystem	forms	with etymology	in cognate sets
Indo-European	1,554,812	424,787 (27.3%)	279,010 (17.9%)
the other 224	—	0 (0.0%)	0 (0.0%)

Exactly zero, not "few". The corpus's three cognacy networks confirm it: all three live in one macrosystem.

network	sets	members	languages	macrosystems
`kaikki-cog`	82,664	356,354	36	1
`kaikki-etymology`	72,023	251,757	37	1
`iecor-gold`	2,640	23,400	152	1

The corpus is broad in form and extremely narrow in history: 225 macrosystems with consonant skeletons, one with integrated kinship judgements.

And it matters to state precisely what this is a claim about, because the short formulation invites reading more into it. What is claimed is a gap in this integration, not in the field: outside Indo-European there are mature comparative traditions and published cognate sets — the *Austronesian Comparative Dictionary*, Uralic cognate databases, Bantu and Sino-Tibetan comparative work, among others — and none of them is integrated here. The correct statement is:

> None of the historical resources currently integrated into this corpus provides comparable cognacy coverage outside Indo-European.

That is a statement about our eighteen sources and about integration work still to be done, not about what historical linguistics has achieved.

What can be done is to build for that condition rather than against it, and that explains the shape of the work using this corpus:

- Catalogues are built per macrosystem and without cognacy, because in 224 of 225 there is none. A method requiring it works on 0.4% of the corpus's languages.
- Indo-European, precisely because it is the only one with gold judgements, is not the object but the test bench: it is where one can measure what a method working without them loses, checking against the known answer.
- And each macrosystem is worked separately, because what must be learned is how much the instrument costs in each — not an average over 225 that would describe none of them.

6.5 · Phonological annotation: what the schema declares and what it currently carries

column	populated	
`segment.prosodic`	9,253,078	46.6%
`segment.syllable`	9,780,772	49.3%
`segment.is_stressed`	9,780,772	49.3%
`segment.dolgo · sca`	9,253,078	46.6%
`segment.tone · role · length`	0	pending
articulatory features (`feature`)	18,977,285	95.6%

Three columns are declared and pending work, not broken: they exist in the schema because the model contemplates them, and they have not been populated yet. The distinction matters and is the same one that governs §7: an empty column known to be empty is a task; one believed to be full is an error waiting to be published.

Articulatory features are the best-covered layer, at 95.6% of segments.

6.6 · Tone: two routes that are not interchangeable

Tone reaches the corpus by two different paths, and confusing them produces false zeros:

As its own segment (`prosodic='T'`), in Southeast Asian transcriptions: 339,543 segments, directly recoverable.

As a diacritic on the vowel, in African orthographies: 41,953 forms in Atlantic-Congo alone. There the segmentation destroys it — `máma dáda` becomes `m a m a d a: d a` — and it must be extracted from the `ipa_raw` field, deciding per language what each mark means. Documented in `../codigos/estudios/DECLARACION-TONO-AFRICANO.md`.

And there is a trap worth declaring because it already produced a wrong reading of ours: the absence of tone is not observable. Gurung and Tamang are textbook tone languages and in this corpus they come out at 0.0% of forms with tone, because their source does not transcribe it. ”This language has no tone” cannot be distinguished from ”this source did not write it”, and any measurement whose contrast depends on that absence will be measuring the transcriber’s practice.

6.7 · What the corpus cannot answer

Listed here so it need not be inferred from the tables:

1. Anything requiring cognacy outside Indo-European. §6.4.
2. Anything requiring skeletons in the 274 IDS languages. §6.2, and `docs/DECISION-IDS.md`.
3. Anything requiring knowledge that a language has NO tone. §6.6.
4. Anything requiring a concept for 1.24 M Kaikki forms, which have form, skeleton and provenance but no mapped meaning. They serve for looking up a word and for questions about form; not for anything whose unit is concept × language. §6.2.
5. Anything prosodic: reliable duration, stress and syllable structure. `is_stressed` and `syllable` are at 49.3% and do not cover the tonal families.
6. Anything morphological per language. There is no morphological segmentation; the stripping rules that exist are per family, declared by hand and limited to citation forms.

7 · Verification

A corpus integrating eighteen sources accumulates errors that do not manifest as errors. There is no exception, no missing row, no failing query: there is a figure slightly different from what it should be, and nothing flags it. This section describes the apparatus published with the corpus to make that class of failure visible, and the next gives the cases that justify it.

7.1 · The principle: declared exception versus silent gap

Everything else follows from here.

> An undeclared absence is indistinguishable from an oversight.

When a source supplies no skeletons, when a lect is left out, when a column is empty, there are two possible worlds: someone decided it, or someone missed it. From outside — and from inside, six months later — they are identical. The only difference is produced by writing it down.

That is why the decisions live in `ci/decisiones.py` and not scattered through the code that applies them, each with its reason and its figure; and why the invariants consult them: a source declared out of scope is reported as a decision, and a new source with zero skeletons still fails.

7.2 · Three layers, and none replaces another

layer	what it checks	how many
unit tests (`tests/unit/`)	the architecture: that the DSN is defined in one place, that nobody opens connections by hand, that no script computes the root by counting levels, that declared debt carries a ceiling	34
schema contract (`ci/esquema.py`)	that every published column states what it means, in what unit, and what it is confused with	11 columns
database invariants (`tests/qa.py`)	the live data: referential integrity, layer coverage, network coherence, licensing	53

The unit tests do not touch the database — an invariant checks this — and so run in five seconds and serve as a pre-commit hook. The invariants do, and take thirty-six.

7.3 · The invariants: 61 in 20 classes

class		what it watches
`SKEL`	10	the skeleton layer: residuals, mismatches, languages and sources with none
`LIC` `COG`	6 each	provenance and licensing · the cognacy networks, each with its own rule
`PROTO`	5	the proto-forms
`INT` `ETY` `CLS` `RAW`	4 each	referential integrity · lineage · sound classes · the `skeleton_raw` gap, separated by cause (§5.1)
`DUP` `VOW` `MARK` `FEAT` `POLY` `SUBS`	2 each	duplicates, vowels, external annotation, features, polysemy, substrates
`CPT` `CORR` `CRY` `COLX` `SPOT` `SEG`	1 each	concepts, correspondences, cryptology, colexification, sampling, segmentation

Each declares a severity: `fail` brings the gate down, `warn` warns, `info` describes. The proportion is deliberate, because a check that only warns is not read, and that is why the five that warn are enumerated one by one in §7.5 rather than left in the program's output.

In the run accompanying this version: 39 green, 5 warnings, 3 declared debts with a ceiling, 14 informational and none red.

When one moves from `info` to `fail`, the reason is recorded. The licensing one did so on 2026-09-05, when it was decided the corpus would be published without restrictions: previously non-redistributable rows could exist because an export script removed them on the way out; now there is one database and it is the distributable one, so any row from a withdrawn source is a violation.

7.4 · The meta-checks, which is what makes this apparatus different

An invariant checks the data. A meta-check checks that the apparatus covers what it says it covers, and there are three:

`COG · network with no declared invariant`. Each cognacy network has its own rule, because they do not share a criterion: an etymon-based network requires the member to belong to the family the set declares, and the `kaikki-cog` network crosses families by design, so there it suffices that the lect exists. Generalising one rule to all produced false positives. But separating them opens a gap: a new network enters with no rule and nobody notices. This check fails if any is left out, and it is the one that discovered 23,400 `iecor-gold` members with no invariant.

SKEL · whole languages with no skeleton — UNDECLARED. It does not count languages without skeletons: it counts those not explained by any written decision. A language with more than two hundred forms and none with a skeleton either has its source declared out of scope, or is a declared lect — duplicate, misassigned, reconstructed — or is a hole. The first two are reported as decisions; the third fails.

LIC · table with no provenance, undeclared. The licensing invariant looks at tables that carry a `source_id` or `source` column, and that is its reach. For months the guarantee "no row without a licence leaves here" was stated without saying which tables it failed to reach, and 13.3 million rows sat outside its view (§9.6.3). This check discovers from the catalogue those tables that have neither provenance nor a link column through which to inherit it, and requires each to be declared with its reason; it fails if one appears that is not. There are thirteen today, and two of them — `concept`, from Concepticon, and `feature`, from CLTS — do hold third-party content: both are CC-BY-4.0 and their provenance is single for the whole table, so a `source_id` column would be a constant. The invariant's edge still exists; what changes is that it is written down, and that a new table cannot cross it in silence.

The difference from an ordinary check: these fail when the apparatus falls short, not when the data is wrong. It is the only defence against the class of error in §8.

7.5 · The five warnings, one by one

§7.3 says that a check that only warns is not read, and that sentence described what was happening here: `qa.py` reports five warnings on every run and none of them appeared in this work. Anyone cloning the repository and running `./verifica.sh` would see fourteen hundred genealogical cycles the paper did not mention. Here they are, one by one, with what each turned out to be on inspection.

ETY · 2-cycles (A→B and B→A) — 1,406 pairs, 105 lects, out of 12,378 edges. The warning calls them "impossible in inheritance" and for most of them they are not. `ancestry_edge` is an aggregate: it is built by lifting the edges of `form_etymology`, which run from word to word, up to the language level. Latin appearing as an ancestor of Portuguese and Portuguese as an ancestor of Latin does not say two languages beget each other: it says one Latin word went into Portuguese and another Portuguese word entered Latin, which happens, and in modern scientific Latin happens a great deal. 74% of the cycles — 1,042 — carry at least one edge marked as a loan, and those are exactly what two languages in contact should produce. The other 364 are **inheritance** ↔ **inheritance**, and those would indeed be impossible: scattered pairs — Bulgarian↔Romanian, Dutch↔Sicilian, Welsh↔Spanish — where a Wiktionary editor labelled as inheritance what was a loan. They are 2.9% of the edges and they come from the source, not from the computation (§10.1).

ETY · child OLDER than parent — 9 cases, and all of them the same one. All nine have Latin `la` as the child of something later: of Spanish from 1200, German from 1500, Old High German from 750, Medieval Latin from 700. The cause is not chronological but a modelling one: **la** is a single lect carrying Classical Latin's dates — up to 200 — and its entries include Medieval and Modern Latin, which did borrow from those languages. Lects for those stages exist (`la-eme`, `la-med`, `la-new`) and the words are not distributed between them. It is a declared defect of the lect division, not an impossibility in the data.

DUP · same (language, spelling) in >1 source — 181,404. These are not errors: they are the same word recorded by two projects. `ids+lexibank` (82,520) and `lexibank+nel` (78,824) dominate, and those are concept lists that overlap by design. Reconciling them — deciding when two rows are one node and when they are two — is open work, and until it is done the policy is to mark and not to filter (§7.2): both are kept, and whoever measures by source knows there is overlap.

DUP · same (language, spelling) same source >1 — 378,880. Of those, 100,193 differ by part of speech and are legitimate homographs, the kind of thing a corpus must keep apart. The remaining 278,687 share a part of speech, and there distinct Wiktionary entries for the same spelling — senses

with different etymologies — coexist with genuine duplicates. Telling one from the other calls for a criterion this work does not fix.

SKEL · skeleton longer than its spelling — 140, and the only one that uncovered a real defect. The check restricts itself to Latin script on purpose, so as not to count Arabic and Persian, where the skeleton is legitimately longer than what is written because the script notes neither short vowels nor gemination. Of the 140 Latin-script cases, some are abbreviations whose transcription spells out the whole word (`Dx` → `d·k·s`), but others are something else: **nord** yields **n·r·d·n·ɖ**, and it comes from **/nuːrd/[nuːɖ]**. Those are two transcriptions, the phonemic and the phonetic, concatenated into one skeleton. Measured across the whole database: 63 forms carry both **/…/** and **[…]**, all 63 have a skeleton, and in 48 the skeleton is literally its own first half repeated — `twelfth` comes out `t·w·l·f·θ·t·w·l·f·θ`. They are 0.002% of the skeletons and they are wrong. It is worth noting that the warning only catches them when the word is short: it was the check that led to them, not the other way round. The comma-separated variants in `lexibank` — `(j)et, jit` → `j·t` — are handled correctly, as was verified while measuring this.

What was done about that defect, and why not more. The origin is upstream, in `segments_raw`: the segmenter split on , which is how kaikki separates variants, but these two are separated by a space, and the cleaner stripped loose `/`, `[` and `]`, leaving the two strings glued together. The code is fixed — `ingest/segment_kaikki.py` now takes the first delimited transcription — and there is a unit test checking the property rather than the case: that the sequence cannot be its own first half repeated. No new form can enter this way.

The 63 rows already written are not repaired, and the reason is worth giving in numbers rather than as an apology. Repairing them touches 1,473 rows in the per-form tables, which is manageable; but it also forces the recomputation of six aggregate tables built by full passes — `morph_rule` (2,823,128), `morph_rule2` (2,180,499), `code_axes`, `code_specificity`, `skeleton_lineage` and `correspondence` — which feed the morphological decomposition and the irreducible nuclei, that is, already-published figures. Changing seven million derived rows to fix sixty-three forms of the 0.002% is not an improvement: it is a risk. It stands as a third declared debt with a ceiling (§7.6).

What the five have in common, and why they go together here: none is a failure of the computation, and four of the five are not failures at all — they are properties of integrating sources that overlap and do not always agree. The fifth is, and it is 63 forms. The difference between declaring them and not declaring them is that a reader can now judge that for themselves.

7.6 · Declared debt, with a ceiling

A broken invariant that cannot be fixed today has three possible fates: it is ignored — and forgotten — it is deleted — and the signal disappears — or it is declared with a ceiling.

There are three such. The first: 42,891 forms where `cv_template` is not the same length as `segments_raw`, annotated with its scope — it affects vowel accounting, not `cons_skeleton`, which is what the studies use — and with a ceiling of 45,000.

The second was declared on 6 September 2026: the 10,260 forms whose skeleton is glide-only and which therefore have no row in `skeleton_raw` (§5.1). Here the ceiling, 11,000, does a job worth pointing out: the divergence between the two layers is not repaired in this release, because aligning them would change 499,546 forms and three derived tables, and that decision carries a cost not taken in passing. What the ceiling guarantees is that the divergence does not grow in the meantime, which is the only thing that can be promised without taking it.

The third, from the same day: the 63 forms with two concatenated transcriptions of §7.5, with a ceiling of

1. It is the one that best shows what this figure is for, because its three parts separate cleanly. The

code

is fixed, so the problem cannot grow by the route it arrived through. The present rows are left alone, because repairing them would move seven million derived rows and published figures. And the ceiling covers what remains: that they enter by some other route we have not seen. A forward fix without a ceiling would be trusting that there is no other route.

They report while they do not grow; they bring the gate down if they exceed their ceiling. A known and bounded problem is not the same as an ignored one, and the difference is a number.

7.7 · A single gate

`verifica.sh` chains the five classes of check and returns a single exit code:

```
unit tests · PDF margins · internal references · schema contract · database invariants
```

It exists because each of them found real failures and none of them ran on its own. It has a `--rapido` mode without the database, in five seconds, so it can be used as a hook.

7.8 · Four rules the apparatus applies to itself

They came from concrete failures, and each is written in the code that applies it:

1. Discover from the catalogue, not from a hand-kept list. The licensing invariant walks the provenance columns by asking the schema. A new table is watched automatically; a hand-written list goes stale the day someone adds a table.
2. A check covering less than its name promises is worse than none, because it produces false confidence. The paper verifiers report their coverage: how many items they looked at out of how many there were.
3. Ask the parser, do not reimplement it. The coverage meta-check once reported 120% because it duplicated the parser's regular expressions instead of asking it for its own denominator.
4. A check that has never fired is not proven. The licensing invariant was validated by injecting a leak through each possible route — by withdrawn source name and by `redistributable = false` — and rolling back. Without that test, "reports zero" and "looks at nothing" are indistinguishable.

8 · Six errors, and why they had the same shape

This section is not a confession. It is the only material demonstrating that the apparatus of §7 is good for anything: an invariant that has never fired is not proven, and an apparatus described without the failures that originated it is a list of good intentions.

The six cases are real, they are in the repository history, and five of the six passed every check that existed at the time. That is the point of the section.

(Cases 1, 2, 3 and 6 affect the analysis built on the corpus, which since 2026-09-05 lives in a separate repository. They are included because they are what taught the discipline and because the pattern is the same.)

8.1 · The null that measured our own draw

Three papers published an enrichment figure with its confidence interval. The null was estimated by shuffling eighty thousand times, and the interval was presented as if it measured the uncertainty of the data.

It measured that of the draw. This was discovered by removing six Turkic forms and seeing the null move by 14% — a sensitivity impossible if the interval described the population. Replaced by the closed form `(Σ_c n_c² − Σ_cl n_cl²) /(N² − Σ_l n_l²)`, the published figures changed: 225×→218×, 77×→78×, 39×→37×.

What did not detect it: nothing. There was nothing to. The interval was wide, and a wide interval looks like prudence.

What detects it now: the null is exact and has no interval to misread. And there is a test validating the closed form against brute-force shuffling, because an unvalidated closed form is exactly what published a wrong figure for three papers.

8.2 · The appendix spliced three times, with 8/8 checks green

Six papers carried their appendix pasted in three times. All six passed the verifier's eight checks.

The check verified that the appendix existed. It did. Three times.

What detects it now: the splicer cuts at the first appendix heading and fails if exactly one does not remain at the end. The difference between `≥1` and `== 1` is the whole error.

8.3 · Five of six scripts giving a different result each run

The query loading a family's data had no `ORDER BY`. PostgreSQL promises no order without one, and gave none: the same family loaded differently in each process. Everything downstream inherited that order, and `Counter.most_common(1)` broke ties by insertion order.

Published result: the branch-pair substitution table said `b p` at 0.5× or `b n` at 1.9× depending on the run.

What did not detect it: the figures were plausible both times. Nobody runs the same script twice to compare.

What detects it now: a unit test that reads the query and requires the `ORDER BY` to end in a unique column. Ordering by language and concept leaves the ties loose again.

8.4 · Two columns with the same name and different meanings

`form.source_id` is the form's source. `lect.source_id` is that of the language record. The code loaded the second into a field called `fuente` and used it as if it were the first.

Published consequences: an entire section of the Indo-European paper said IDS lost 65.6% of its forms when it loses 100%; and a hygiene check attributed the distribution of Turkic infinitives to the wrong source — `ids` at 85%, not `nel` at 6.4%.

What detects it now: the schema contract declares, for each column, what it is confused with; there is a test requiring both `source_id` columns to be marked as mutually confusable, and another requiring the field loading the lect's source to be named `fuente_del_lecto` and not `fuente`. The error was born from a name.

8.5 · The licensing invariant reporting zero with 2,140 rows inside

When the non-redistributable sources were withdrawn, the sweep walked every table with a `source_id` column and deleted 5,971 rows. The licensing invariant reported zero rows in quarantine.

There were 2,140. The `cognate_set` table stores provenance in a column called `source`, without the suffix, and neither the sweep nor the invariant looked at it.

What detected it: not the licensing invariant, but another check — the cognacy meta-check, which complained about a network with no declared invariant. A licensing failure came to light through a cognate coherence check.

What detects it now: the invariant looks at both ways of naming the column, and additionally compares by declared name rather than only by `JOIN` against `source.redistributable` — because once the source row is deleted the `JOIN` finds nothing even if references remain, and would report zero precisely when there is a leak. It was validated by injecting a leak through each route and rolling back.

8.6 · The meta-check reporting 120% coverage

The paper verifier reports how many items it looked at out of how many there were. It reported 120%.

The denominator was computed by us, duplicating the parser's regular expressions instead of asking it for its own. The two versions diverged and the ratio went above one.

What detects it now: the denominator is returned by the parser itself. Ask the parser, do not reimplement it.

8.7 · The common shape, which is the thesis

All six are the same error:

> A check that covered less than its name promised, and reported conformity exactly when there was a failure.

The appendix check promised "the appendix is right" and verified that it existed. The licensing invariant promised "there is no quarantine" and looked at one of the two columns where there can be. The meta-check promised "coverage" and divided by a number it invented. In all three, green was more dangerous than the absence of a check, because a green check closes the question.

Hence the four rules of §7.8 and, above all, the meta-checks: checks whose object is not the data but the reach of the apparatus. They are the only ones that can detect this class of failure, because the failure is not in the data.

8.8 · A transferable recommendation

For anyone publishing a data resource with checks:

1. Write down what each check does NOT cover, alongside what it does. If it cannot be written, the check is not understood.
2. Make them fail, not warn. A warning in a hundred-line log does not exist.
3. Discover from the catalogue, not from a hand-maintained list.

4. Prove they fire. Introduce the failure, check that the check sees it, roll back. A check that has never fired is not proven — it merely has not failed yet, which is not the same.
5. Measure the apparatus's own coverage, and make that fail too.

8.9 · A closing note: the apparatus working

On 2026-09-05, separating the corpus from the analysis programme that uses it, five architecture tests failed at once. They asserted the previous structure: that `familia.py` was in the corpus package, that the verification gate chained the paper verifier, that `analysis/` was split into three particular subdirectories.

None was a failure. They were five descriptions of the architecture that had stopped being true, and the suite said so before anyone published anything about a structure that no longer existed. That is exactly what they are for.

9 · Licensing, distribution and availability

9.1 · The licence of the whole, and who imposes it

CC-BY-SA-4.0. It is not a choice: it is the lowest common denominator the sources impose.

licence	sources	forms
CC-BY-4.0	7 — CLTS, Concepticon, Glottolog, IDS, IE-CoR, Lexibank, NorthEuraLex	2,325,336 (61.4%)
CC-BY-SA-3.0	6 — the Kaikki family	1,462,960 (38.6%)

CC-BY integrates cleanly because BY is compatible towards BY-SA. But 38.6% of the forms come from Kaikki, which is CC-BY-SA, and ShareAlike is inherited: any derived work — including the computed layers of §5 and the catalogues the studies produce — carries the same licence.

It is the honest path and also the simple one: one licence for the whole thing, with no exceptions to remember.

9.2 · Attribution is not a promise, it is a column

CC-BY and CC-BY-SA require attribution. In this corpus the mechanism is structural: every form carries its **`source_id`**, and there is a `fail`-severity invariant that brings the gate down if any lacks it.

Zero forms without provenance, out of 3,788,294.

9.3 · The content, with no licence restrictions

"Without restrictions" here refers to the LICENCE, not to access, and it is worth separating before going on, because §9.5 says the repository is private and the two sound contradictory. They are not: the content carries no clause preventing its redistribution, and who can clone it today is a different question.

Until 2026-09-05 there were two databases: a research one with everything and a public one produced by a script filtering on export. Maintaining two truths forces you to remember which one you are looking at, and one of the two ends up stale. Now there is one, and it is the distributable one.

Three sources were withdrawn. The one that mattered is Pokorny (CC-BY-NC-ND): the `NC` would be saved by a free service, but the `ND` forbids sharing derivative work — and we do not copy the dictionary, we scrape, parse and link it, which is deriving. The other two, de Vaan and Kroonen, were declared and had never been ingested.

The cost, measured before deleting, by recomputing every claim with and without the source:

rows withdrawn	5,971
forms losing their path to Proto-Indo-European	5,508 (3.32%)
cognate sets left without reconstruction	2,140
forms left with no etymology at all	906

By branch the blow is uneven: Anatolian −30.5%, Italic −10.3%, Slavic −6.3%, Germanic −2.1%. Anatolian loses nearly a third of its links to PIE. That is the price of being able to distribute, and it is paid knowingly.

A `fail` invariant checks that no row from a withdrawn source re-enters, and it is validated through both routes by which it could (§7.8).

9.4 · The repository is the distribution

Raw data is not shipped: the repository is cloned and rebuilt. CLDF and SQL dumps are generated as pipeline outputs, not as the hosted product.

It has three good consequences: the compliance surface shrinks — CC licences activate on *sharing*, and whoever rebuilds obtains each source from its origin under its origin's licence; provenance stops being a claim and becomes executable; and versioning is that of version control, which serves both for citing and for working.

With one deliberate exception: the derived layer is archived separately. The 3.09 M skeletons are our computation and are what people will want to cite and reuse; requiring them to stand up a database and run hours of rebuilding would stop most readers. The studies' catalogues already work this way: they occupy 8.6 MB and live beside the paper explaining them.

9.5 · Access today: on request

The repository is private, and it is worth saying why, because the reason is simpler than it looks and should not have another attributed to it. The work is in progress and access is managed rather than left open; that is all. It is not a reservation about the data — the licence already permits redistributing it — nor a consequence of the reproducibility question in §9.6.1, which is not settled by opening or closing a repository.

An earlier version of this section did tie it to that question: it said a public clone would promise a reproducibility that did not exist. That reasoning does not hold once §9.6.1 was reframed, because `kaikki`'s version will never be pinnable and the repository would have stayed private forever for a reason that was not its own.

Access is requested by writing to alex@pixelspace.com, saying who you are and what for. Whoever receives it has full read access and can fork the repository to their own account, so they can work on it without depending on anyone accepting their changes.

The licence of the content does not change because of this: CC-BY-SA-4.0, and ShareAlike is inherited by any derived work. What is restricted is access to the repository, not the rights over what it contains.

Once the source versions are pinned, this restriction stops making sense and the repository becomes public. It is a state, not a policy.

9.6 · Three declared gaps

They are written here, not in a footnote, because they affect what can be claimed about reproducibility.

9.6.1 · No source has its version pinned · Reframed 2026-09-06. This section used to say that versions were unpinned and had to be pinned. On going to do it, it turned out that for the heaviest source that is no longer possible, and that the section described the problem badly.

What kaikki does. Its page publishes the date of the dump the extraction came from — today it reads *"extracted from the enwiktionary dump dated 2026-08-05"* — and it is regenerated at least once a week, overwriting. It does not keep earlier extractions and publishes no checksums. And the date lives on the page, not inside the file: this was checked by opening the downloaded JSONL, and every record carries `word`, `lang`, `lang_code`, `pos`, `senses` and `forms`, without a single field for date, version or source dump. Once the file is downloaded, the version is gone.

What that means here, measured. The 111 `kaikki` files this corpus came from were downloaded on nine different days, between 12 July and 8 August 2026. At weekly regeneration, that is four or five distinct extractions: Latin and Ancient Greek were fetched on 6 August, and Afrikaans and Aromanian on 14 July, three weeks and three dumps earlier. This corpus is not built on one version of kaikki but on several, mixed, and none of them can be recovered today. The same, to a lesser degree, holds for `lexibank` (two days a month apart) and `glottolog` (two days).

What was done instead. The version cannot be recorded, but which exact files there were can. `ingest/manifiesto_fuentes.py` walks the 148 raw files — 5.46 GB — and records for each its `sha256`, its size and its download date, in `docs/MANIFIESTO-FUENTES.md` and its JSON twin. That does not bring back the lost version; it turns *"we do not know what we had"* into *"we had THESE files, and anyone can check whether theirs are the same"*. Sources that do have a readable version carry it: `pyclts` 4.0.2 and `lingpy` 2.6.14. Concepticon is neither installed nor downloaded — its identifiers travel inside the CLDF of Lexibank, IDS, NorthEuraLex and IE-CoR — so its version is that of each source dataset.

And the disclaimer, which is the part that was missing. Kaikki is not ours. We do not control it, we do not version it, and we cannot promise what it will serve tomorrow. Anyone who wants to rebuild the corpus should download kaikki and use it: they will get that day's kaikki, run their tests and obtain their figures, and that is fine — it is their run, not ours. What this work guarantees is what is actually its own: that the pipeline is deterministic, that given the same input files it produces the same corpus, and that the input files are identified one by one by their fingerprint. What it does not guarantee — and no project integrating moving third-party data can — is that the third party will serve you the same files. Pretending otherwise would be the error; declaring it costs nothing and is exact.

9.6.2 · Two sources with licence "open" · CLOSED on 2026-09-06. `europarl` and `opensubtitles` were declared with licence "open", which is not a licence. It turned out neither is a data source: OpenSubtitles never contributed text — only counts — and Europarl's 22,281 sentences lived in a study table, `construction`, which was removed from the corpus. Both remain as measurement, and there is no third-party licence to comply with because no third-party content is retained. Detail in §12.1.

9.6.3 · 13.3 million rows with no provenance column · Tatoeba withdrawn · CLOSED 2026-09-06. This section previously held that `tat_sentence` (5,799,836 rows) and `tat_link` (7,516,808) were a blind spot in the licensing invariant because they lacked a `source_id`. That was the wrong reason. Tatoeba's licence is known and is CC BY 2.0 FR, which permits redistribution with attribution: the gap belonged to the apparatus rather than to the data, and adding a column would have closed it.

The reason they leave is of another order, and it concerns the object. This corpus is about forms, and

the chain that holds it together runs form → segment → skeleton → code; a sentence is not a form. Neither table carried a single foreign key towards `form`, `concept`, `lect` or `source`; no form declared `source_id='tatoeba'`; and no corpus process read them. They were an orphan layer of 13.3 M rows and close to a gigabyte, with nothing entering through it and nothing leaving towards it.

The only thing ever built from them was `construction` and `construction_align` (24,267 rows), the output of a qualic-fields study that had already left the corpus by way of §9.6.2. All four tables now live in the `qualicfields` database, which is the project whose object is indeed usage. The extractor that queries them never joined against `form` or `lect` at any point, so the separation takes nothing from it; the `qf_*` tables, which do sustain fourteen foreign keys into the corpus, stayed where they were.

The reach that is lost is worth stating in figures, because it is short. Tatoeba covered twelve languages —eng, ita, deu, fra, por, spa, nld, dan, swe, ron, nob and cat—, all from the same macrosystem, one of the 225 in the corpus, and within it the corner of western Europe. The usage context it supplied fell exactly where the corpus is already densest, and reached none of the other 224.

9.7 · What must be fixed BEFORE freezing the release accompanying this work

The preceding sections declare gaps, and declaring is not fixing. A reviewer may fairly reply: *"very well; fix it before publishing the resource."* This is the list, and this work is not submitted with any of the first four still open:

	what	why it blocks
1	Pin a version and checksum for every source · RESOLVED DIFFERENTLY 2026-09-06	Pinning `kaikki`'s version is no longer possible: the date lives on its page, not in the file, and earlier extractions are not kept. In its place there is a `sha256` fingerprint of the 148 raw files and the disclaimer in §9.6.1: the pipeline is ours and is deterministic; the third party is not, and nothing is promised on its behalf
2	Identify the licence of `europarl` and `opensubtitles` · DONE 2026-09-06	Neither was a data source; Europarl's text was removed and both remain as measurement (§12.1)
3	Give provenance to `tat_sentence` and `tat_link`, or remove them from the published database · DONE 2026-09-06	They left: 13.3 M rows of an object the corpus does not read. They live in the `qualicfields` database (§9.6.3)
4	Recover clicks in the skeleton · DONE 2026-09-06	488 of 491 forms keep it and none is left without a skeleton (§4.4)
5	Re-run the `skeleton_raw` pass · DONE 2026-09-06, and it did not close the gap	It ran and wrote 92 rows: 57,132 → 57,040. The gap was not a lagging pass. Its three real causes are now in §5.1, and one of them — 10,260 forms whose skeleton is glide-only — is declared as a divergence between layers, not repaired
6	Move the study-output tables to their own repository · RESOLVED DIFFERENTLY 2026-09-06	The premise was false: `codigos` reads NONE of the 25 movable ones and Meulemans reads 17. They are not output awaiting a move, they are what the query layer serves (§5.4). What really was output — `construction`, `construction_align` — already left. So did the 8 backup tables living inside the database: 1.5 M rows and 190 MB
7	Remove the 2 residual `wiktionary` forms · DONE 2026-09-06	Not deleted: one was the etymological parent of English *angst*, so its 4 edges were repointed to the `kaikki` twin before removal (§12.1)

The first four change what the work can claim. The last three do not, which is why they come after.

The four that block are closed as of 6 September 2026, though the first not in the way it was posed. It was posed as an inventory task — walk the eighteen sources recording versions — and turned out to be a question of scope: for the heaviest source the version no longer exists anywhere, and what needed fixing was not the inventory but the promise. That is in §9.6.1.

10 · Limits

§6.7 lists the questions the corpus cannot answer. This one collects the limits of the resource as such: what no amount of further work on what already exists will fix.

10.1 · It inherits whatever it receives

It is aggregation (§2), and therefore inherits its sources' errors without being able to correct them in general. When an error is detected it is declared — there is a Kam-Sui language classified as Uralic at source, written down as a declared fact (§3.6) — but detecting it required three independent checks to agree, and there is no reason to believe that was the only one.

An integrated corpus is no more reliable than its sources. What it can be is more auditable, which is a different thing and is what it aims at.

10.2 · Reproducibility reaches as far as what is ours

Two things are usually said together and need separating. The pipeline is ours and is deterministic: given the same input files it produces the same corpus, and those files are identified one by one by their `sha256` in `docs/MANIFIESTO-FUENTES.md`. Anyone can check whether theirs are the same.

The sources are not ours. `kaikki` — 38.6% of the forms — is regenerated at least once a week, overwriting, and keeps none of its earlier extractions; whoever clones and rebuilds will get the kaikki of the day they do it. That is not a defect to be resolved: it is what happens when moving third-party data is integrated, and this work makes no promises on a third party's behalf. Anyone who wants to use kaikki should use it, run their tests and obtain their figures — it will be their run, with their data, and it will be fine.

What does follow, and is worth saying, is this: a published figure must cite the frozen archive that produced it, not "the rebuild", because two rebuilds in different weeks are not the same. The limit exists and is declared; what does not exist is an obligation to take responsibility for it. See §9.6.1.

10.3 · The historical layer covers one macrosystem

Etymology, cognacy and proto-forms exist for Indo-European and nothing else. §6.4 gives the figures and what they imply for anyone working on the corpus.

What matters here is bounding what kind of limit this is. It is not that expert cognacy does not exist outside Indo-European — it does, and §6.4 names some sets: it is that it is not integrated here, and each integration costs mapping work that has not been done.

What is a limit of the resource rather than of this version is that a corpus can gather the judgements that exist but cannot produce those that do not: where a family has no published comparative tradition, no amount of aggregation substitutes for it.

10.4 · The normalisation loses by design

The consonant skeleton groups rhotics, nasals and dorsals into a canonical consonant by qualic value (§4.3). That is what makes two languages comparable and it is also a real loss of phonetic information.

It is mitigated by preserving `ipa_raw`, `segments_raw` and `skeleton_raw` intact, so the normalisation is an added and reversible layer. But anyone working on `cons_skeleton` works on a deliberately impoverished representation, and questions depending on the merged distinctions cannot be asked there.

10.5 · The apparatus has known blind spots

The licensing invariant's reach has an edge, and it is written down. Thirteen tables have neither a provenance column nor a link column through which to inherit one, and two of them — `concept`, from Concepticon, and `feature`, from CLTS — do hold third-party content (§3.3). There is no licensing problem, both are CC-BY-4.0, but the invariant does not reach there, and a meta-check fails if one more undeclared table appears (§7.4).

Three schema columns are declared and unpopulated (§6.5).

Three declared debts with a ceiling: 45,000 rows of template/segment mismatch, 10,260 forms where `skeleton` and `skeleton_raw` disagree about whether a glide is skeleton, and 63 forms whose skeleton carries two concatenated transcriptions — this one with the code already fixed (§7.6).

Five warnings that are not failures but are not nothing either, enumerated one by one in §7.5 with what each turned out to be: four are properties of integrating overlapping sources, and the fifth uncovered 63 forms whose skeleton carries two concatenated transcriptions.

All of them are written down because a check that does not cover something and does not say so produces false confidence, which is the thesis of §8. They will remain written until they are resolved.

10.6 · It is not a primary source, nor does it claim to be

It does not replace any of the resources in §2, does not document new languages and does not adjudicate linguistic questions. It offers an auditable union of other people's material, a computed layer on top, and the tools to check that both say what they say.

11 · How it is used and how to cite it

11.1 · Getting it

The repository is the distribution (§9.4), and access is requested (§9.5): write to alex@pixelspace.com saying who you are and what for. With access, it is cloned and rebuilt — and can be forked to your own account:

```
git clone https://github.com/toledoal/corpus-integrativo
cd corpus-integrativo
./db/setup_dev.sh               # the project cluster
./db/rebuild.sh --scope smoke   # one small family, minutes
./db/rebuild.sh --scope full    # everything, hours
./verifica.sh                   # the five classes of check
```

Rebuilding requires the sources to remain available, and comes with the caveat of §9.6.1: it will give you the kaikki of the day you run it, which is not the one that produced this work's figures. To check whether your files match ours there is `docs/MANIFIESTO-FUENTES.md`, with the `sha256` of all 148. If they do not match, nothing is broken: you have a different corpus, made by the same pipeline over a different input.

Anyone wanting only the derived layer — the consonant skeletons — needs none of this: it is archived separately, for the reason in §9.4.

11.2 · Two ways of reading it, and neither is the right one

The corpus has no canonical consumer, and that conditions how it is described (§6.1). The two that exist today ask different things of the same data:

A query layer for a word's history. It looks up a form and shows its sound, its sense, its cognates and its lineage. It is served by the 3.79 M forms, and the 1.24 M Kaikki forms with no assigned concept are not dead weight but its raw material. It displays reconstructions, because showing lineage is what it does.

A catalogue of form–meaning correspondences per macrosystem. It needs skeleton and concept together, because its unit is concept × language; there what is usable is 2.06 M. It excludes reconstructions, because its design is not to look at them.

Both figures are correct and both policies are correct. That is why the corpus declares facts about its lects and not policies (§3.6), and why this description gives the layers and their coverage rather than a single "usable corpus" figure.

11.3 · How to cite it

Every published number must cite an archived version, not the rebuild — for the reason in §9.6.1: two rebuilds run in different weeks do not start from the same input. The corpus citation goes together with those of the sources actually used: the `source` table carries each one's citation and licence, and is what to consult in composing the attribution for a particular piece of work.

The licence of the whole is CC-BY-SA-4.0 and it is inherited: any derived work, including catalogues built on it, carries the same.

11.4 · Contributing

What this corpus asks of anyone wanting to add something is what §7 described:

- A new source enters with its licence by name. "Open" is not a licence (§9.5).
- A decision is declared where it can be consulted, with its reason and its figure. An undeclared absence is indistinguishable from an oversight.
- A new check states what it does NOT cover, and is proven to fire by introducing the failure it should detect. A check that has never fired is not proven.
- **`verifica.sh`** passes before every change. It returns a single exit code and has a five-second mode without the database.

12 · Appendix · The eighteen sources

The text speaks of eighteen sources without ever listing all of them. Here they are, with what each contributes to the built corpus and its licence. The forms column is what each contributes to the **`form`** table: a source with zero forms is not an empty source — Glottolog contributes the classification of the lects and Concepticon the identity of the concepts, without contributing a single word.

#	id	what it is	licence	forms
1	`lexibank`	standardised wordlists in CLDF/CLTS	CC-BY-4.0	1,740,092
2	`kaikki`	English Wiktionary extracted with wiktextract	CC-BY-SA-3.0	1,457,643
3	`ids`	*Intercontinental Dictionary Series*, via CLDF	CC-BY-4.0	437,902
4	`nel`	NorthEuraLex, via CLDF	CC-BY-4.0	121,611
5	`iecor`	IE-CoR · expert cognacy judgements	CC-BY-4.0	25,731
6	`kaikki-descendants`	Wiktionary ”Descendants” sections	CC-BY-SA-3.0	5,315
7	`wiktionary`	earlier Wiktionary ingestion; today supplies only the record of 3 lects	CC-BY-SA-3.0	0 · its 2 forms were removed 2026-09-06
8	`kaikki-tree`	Wiktionary etymology chains	CC-BY-SA-3.0	0 · 383,234 etymologies
9	`kaikki-prose`	Wiktionary prose etymology	CC-BY-SA-3.0	0 · 37,994 etymologies
10	`kaikki-latin-etym`	Latin etymology templates	CC-BY-SA-3.0	0 · 2,007 etymologies
11	`glottolog`	glottocodes and classification	CC-BY-4.0	0 · identity of 4,735 lects
12	`concepticon`	concept catalogue	CC-BY-4.0	0 · identity of 3,205 concepts
13	`clts`	transcription standard and features	CC-BY-4.0	0 · features for 18,977,285 segments
14	`lingpy`	sound-class models (Dolgopolsky, SCA)	GPL-3.0	0 · 3,516 classes
15	`liv`	LIV², reconstructed Indo-European verbs	CC-BY-SA-4.0	0 · 143 proto-forms
16	`tatoeba`	community sentences and alignments	CC BY 2.0 FR	0 · withdrawn 2026-09-06: 13.3 M rows of running text, which is not the object of a database of forms (§9.6.3)
17	`europarl`	European Parliament proceedings, via OPUS	not applicable · measurement	0 · see §12.1
18	`opensubtitles`	aligned subtitles, via OPUS	not applicable · measurement	0 · see §12.1

12.1 · Two of the eighteen are not data sources

`europarl` and `opensubtitles` were declared with licence ”open”, which is not a licence. Looking at what they contributed showed the question was wrongly posed: they are not sources anything is ingested from.

OpenSubtitles never contributed text. Only frequency counts — facts, not protected expression —

and findings written by this project that cite it as evidence.

Europarl did contribute: 22,281 Parliament sentences, in the `construction` table. It was a study table belonging to the qualic-fields project — no family paper used it and the query layer did not query it — and it was removed from the corpus on 2026-09-06, along with `construction_align`. The backup was kept outside the repository.

Both remain as measurement: they are cited because measurements were made against them, not because anything of theirs is retained. And that resolves the licence without having to verify it, which was the gap in §9.6.2: no third-party content is redistributed, so there is no third-party licence to comply with.

`wiktionary` (#7) carried two forms from an earlier ingestion — German *Angst* and Spanish *angostura* — removed on 6 September 2026. The removal was not a deletion, and the detour is worth telling because it is the very error §8 describes. Both looked like innocent residue: no concept, no skeleton, no segments, and a twin in `kaikki` that had all three. But on deletion the database refused: **`de·Angst·N·001`** was the etymological parent of English *angst*, with four edges, and its `kaikki` twin had none. Deleting would have lost that etymology in silence. What was done instead was to repoint the four edges to the twin — the form that can actually carry a skeleton and a code — and only then remove the forms. The four affected cognate sets each lost one member, the duplicate: the twin was already inside all four. The source stays in the catalogue because it still supplies the record of three lects (German, Spanish, Old High German), and that is `lect` provenance, not `form` provenance (§3.3).

13 · References

13.1 · The other two works in this series

Both are built on this corpus, and none of the three stood up by citing the others — something this version corrects. They are the two consumers of §6.1, and their disagreement about which part of the corpus is usable is not a problem: it is that section's argument.

- Toledo Martínez, A. (2026). *Meulemans: a read-only query layer over a multi-layer lexical corpus of 225 macrosystems.* — It displays a word's lineage. It is read-only: it neither writes to nor computes anything here.
- Toledo Martínez, A. (2026). *The catalogue of consonantal codes: a comparative method without genealogy, cognacy or reconstruction.* — It reads the derived layer of §5 and forbids itself to look at the historical layer of §3.2. The `code` column this corpus publishes while asserting nothing about it (§5.1) is precisely the instrument that work studies.

13.2 · Sources and standards

On versions. The references below identify the work, not the specific version this corpus ingested, and none of them can carry a version number without lying. For `kaikki` that is final: the version is not in the file and earlier extractions are not kept, so there is no number to cite (§9.6.1). For the rest it is recoverable, and whoever needs it has a starting point: `docs/MANIFIESTO-FUENTES.md` gives the `sha256`, size and download date of the 148 raw files, which identifies the ingested artefact even though it does not name it with a version.

- Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A. & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5, 180205.

- List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J. & Gray, R. D. (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9, 316.
- Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. *Glottolog*. Max Planck Institute for Evolutionary Anthropology, Leipzig. <https://glottolog.org>
- List, J.-M., Rzymski, C., Tjuka, A., Greenhill, S. J. et al. *Concepticon: A Resource for the Linking of Concept Lists*. Max Planck Institute for Evolutionary Anthropology. <https://concepticon.clld.org>
- List, J.-M., Anderson, C., Tresoldi, T. & Forkel, R. *CLTS: Cross-Linguistic Transcription Systems*. <https://clts.clld.org>
- Key, M. R. & Comrie, B. (eds.). *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology. <https://ids.clld.org>
- Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., Mühlenbernd, R., Wahle, J. & Jäger, G. (2020). NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation* 54, 273–301.
- Heggarty, P., Anderson, C., Scarborough, M. et al. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science* 381, eabg0818. (IE-CoR)
- Ylönen, T. (2022). Wiktextract: Wiktionary as Machine-Readable Structured Data. *Proceedings of LREC 2022*. <https://kaikki.org>
- Rix, H., Kümmel, M., Zehnder, T., Lipp, R. & Schirmer, B. (2001). *Lexikon der indogermanischen Verben* (LIV²), 2nd ed. Wiesbaden: Reichert. Linked-data version from LiLa/CIRCSE.
- List, J.-M. & Forkel, R. *LingPy: A Python Library for Historical Linguistics*. <https://lingpy.org>
- Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of LREC 2012*. (Europarl and OpenSubtitles were obtained through it.)
- Tatoeba Project. <https://tatoeba.org>

And one reference no longer in the corpus, cited because §9.3 measures what removing it cost:

- Pokorny, J. (1959). *Indogermanisches etymologisches Wörterbuch*. Bern: Francke. Starostin (StarLing) digitisation. Withdrawn from the corpus on 2026-09-05 for its CC-BY-NC-ND licence.