Corpus Integrativo

Una base léxica multi-capa de 225 macrosistemas, con procedencia por fila y su aparato de verificación

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

Resumen

Se presenta el Corpus Integrativo, una base de datos léxica que integra dieciocho fuentes en 3.788.294 formas de 4.735 doculectos sobre 225 macrosistemas, con segmentación fonológica, procedencia por fila y una capa calculada de 3.091.596 esqueletos consonánticos normalizados que ninguna fuente aporta.

La capa base es agregación: cada forma viene de otro sitio. Lo que se aporta es la unión hecha de modo que se pueda auditar de dónde salió cada cifra, una representación fonológica común calculada con las mismas reglas para los 225 macrosistemas, y las decisiones y huecos declarados con su coste medido.

Se describe también, y con el mismo detalle, el aparato de verificación que se publica con el corpus: 60 invariantes que fallan en vez de avisar —y los cinco que solo avisan, enumerados uno a uno—, un contrato de esquema que declara qué significa cada columna y con cuál se confunde, y meta-chequeos cuyo objeto no es el dato sino el alcance del propio aparato. Su justificación son seis errores reales de este proyecto —cinco de ellos pasaron todas las comprobaciones que había en su momento— que resultaron ser el mismo error seis veces: una comprobación que cubría menos de lo que su nombre prometía e informaba conformidad exactamente cuando había fallo.

La cobertura se da por capas y no como una cifra única, porque no existe: dos consumidores del mismo corpus —una capa de consulta de la historia de una palabra y un catálogo de correspondencias forma–sentido— necesitan capas distintas y llegan a cifras distintas, las dos correctas.

El corpus se distribuye bajo CC-BY-SA-4.0, licencia que impone el 38,6 % de formas procedentes de una fuente ShareAlike. El repositorio es la distribución: se clona y se reconstruye. El acceso se da por solicitud —se concede a quien lo pide, con permiso de bifurcar— porque el trabajo está en curso, no por ninguna reserva sobre los datos: lo que se gestiona es el acceso al repositorio, no los derechos sobre lo que contiene.

1 · El problema

Comparar palabras entre lenguas exige cuatro cosas a la vez: la forma, el sentido, una representación fonológica comparable y la procedencia de cada dato. Los recursos que existen dan cada uno una parte, y muy bien, porque cada uno se diseñó para una pregunta distinta (§2).

El resultado es que una pregunta que necesita dos de esas dimensiones a la vez no tiene dónde hacerse. Un catálogo de listas de palabras cubre 225 macrosistemas y no trae etimología; un volcado de diccionario trae toda la etimología de un macrosistema y no mapea sus entradas a conceptos; los juicios de cognación de experto existen sobre 170 conceptos y una familia.

Unir esas fuentes es la parte fácil. Unirlas de manera que después se pueda auditar de dónde salió cada cifra es la parte que decide si el resultado sirve, y es de lo que trata este trabajo.

Qué es este corpus, y qué no

Es una base de datos léxica multi-capa que integra dieciocho fuentes: 3.788.294 formas de 4.735 doculectos en 225 macrosistemas, con segmentación, procedencia por fila, y una capa calculada de 3.091.596 esqueletos consonánticos normalizados que ninguna fuente trae.

La capa base es agregación. Cada forma viene de otro sitio. No hay aquí una recolección nueva de datos primarios, y presentarlo de otro modo sería falso. Lo que se aporta son tres cosas, ninguna de ellas «más datos»:

1. Una unión auditable, no solo hecha: cada fila sabe de dónde vino, y eso es lo que permite rastrear después cualquier cifra hasta su fuente (§3).
2. Una representación fonológica común a los 225 macrosistemas, calculada con las mismas reglas para todos, que es lo que hace comparables lenguas documentadas por proyectos distintos (§4, §5).
3. Las decisiones y los huecos, declarados con su coste medido (§6, §9).

Y una cosa más, que quizá sea la más transferible

Un corpus integrado de dieciocho fuentes acumula errores que no se manifiestan como errores: no hay excepción, ni fila que falte, ni consulta que falle. Hay una cifra ligeramente distinta de la que debería ser, y nada la señala.

Este trabajo publica, con el corpus, el aparato construido para hacer visible esa clase de fallo: 53 invariantes que fallan en vez de avisar, un contrato de esquema que declara qué significa cada columna y con cuál se confunde, y —lo que creemos más útil— meta-chequeos, cuyo objeto no es el dato sino el alcance del propio aparato (§7).

La justificación de ese aparato es §8: seis errores reales de este proyecto, cinco de los cuales pasaron todas las comprobaciones que había en su momento, y que resultaron ser el mismo error seis veces.

Cómo leer lo que sigue

§2 sitúa el corpus entre los recursos existentes. §3 y §4 describen el modelo y cómo se calcula cada capa. §5 detalla lo calculado. §6 da la cobertura y dónde se acaba. §7 y §8 son el aparato de verificación y los errores que lo justifican. §9 cubre licencia y distribución, §10 los límites y §11 cómo citarlo y quién lo usa.

Quien vaya a usar el corpus debería empezar por §3.3 —la trampa de las dos columnas de procedencia— y por §6, que dice hasta dónde llega cada capa.

2 · Recursos relacionados

Conviene decirlo antes que nada: la capa base de este corpus es agregación. Cada forma viene de otro sitio, y los sitios son buenos. No hay aquí una recolección nueva de datos primarios, y presentarlo de otro modo sería falso.

Lo que sigue describe qué hace cada recurso —bien, y para lo que fue diseñado— y qué queda sin resolver cuando se necesitan juntos.

2.1 · La infraestructura de identidad

Tres catálogos que no aportan formas y sin los cuales nada de esto se podría unir:

Glottolog da el identificador estable de cada doculecto (el glotocódigo) y su clasificación. Es lo que permite saber que dos filas de fuentes distintas son la misma lengua — y es la base de la regla mecánica de duplicados de §3.6.

Concepticon da el identificador estable de cada concepto, enlazando las listas de elicitación que cada fuente usa. Sin él, «tres» de una lista y «three» de otra son dos cosas.

CLTS da el estándar de transcripción y los rasgos articulatorios, que en nuestro corpus cubren el 95,6 % de los segmentos (§6.5).

Y CLDF no es un dataset sino el formato que hace que todo lo anterior encaje. Es la razón de que agregar sea siquiera posible: sin un empaquetado común, integrar dieciocho fuentes sería dieciocho parsers y dieciocho conjuntos de supuestos tácitos.

2.2 · Las fuentes de forma, y en qué es profunda cada una

Medido sobre nuestra propia ingesta, que es lo que podemos afirmar con seguridad:

fuente	formas	lenguas	macrosistemas	conceptos	con esqueleto
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 es la anchura: 3.120 lenguas y los 225 macrosistemas, con listas de palabras elicitadas contra listas de conceptos y segmentadas de forma consistente. Es lo que hace posible cualquier trabajo comparativo de amplitud, y por eso los macrosistemas que están al 100 % de cobertura de esqueleto son exactamente los que vienen enteros de ahí. Lo que no da —ni pretende— es historia: es una colección de listas de palabras.

Kaikki / Wiktextract es la profundidad, en un solo macrosistema. 37 lenguas y un macrosistema, pero 1,46 millones de formas y prácticamente toda la capa etimológica: 1,24 de las 1,26 millones de aristas de etimología del corpus salen de ahí, más 72.023 protoformas y dos de las tres redes de cognación. Es un diccionario, y trae lo que un diccionario trae: definiciones en prosa, etimologías, descendientes. Lo que no trae es mapeo a conceptos —el 14,8 %, §6.2—, porque sus entradas no se elicitaron contra ninguna lista.

IDS es anchura de conceptos: 269 lenguas y 58 macrosistemas, todo mapeado. Su CLDF no publica segmentación, y de ahí sus cero esqueletos (§6.2).

NorthEuraLex aporta 107 lenguas del norte de Eurasia con IPA propio y consistente.

IE-CoR aporta lo más escaso que hay: juicios de cognación de experto, 152 lenguas indoeuropeas sobre 170 conceptos de léxico básico. Es la única cognación de oro del corpus, y por eso es el único sitio donde se puede comprobar contra una respuesta conocida un método diseñado para trabajar sin ella.

2.3 · Lo que no resuelve ninguno, y es el hueco

Léase la tabla por columnas y el patrón salta: cada recurso es profundo en una dimensión y llano en las demás, porque cada uno fue diseñado para una pregunta distinta.

- La anchura de lenguas y la profundidad histórica no coinciden: Lexibank cubre 225 macrosistemas sin etimología; Kaikki cubre uno con toda la etimología.
- La cobertura de conceptos y la de fonología no coinciden: IDS mapea el 100 % de sus formas a conceptos y no publica un solo segmento; Kaikki segmenta el 80 % y mapea el 15 %.
- La cognación de oro existe sobre 170 conceptos y un macrosistema.

Ninguno de estos es un defecto. Son recursos que hacen bien lo suyo. El hueco aparece cuando una pregunta necesita dos dimensiones a la vez, que es el caso de cualquier trabajo que quiera comparar formas por su significado en macrosistemas donde no hay reconstrucción: hace falta forma y sentido y una representación fonológica comparable, sobre las 225 familias y no sobre una.

2.4 · Qué añade este corpus, dicho estrechamente

Tres cosas, y ninguna es «más datos»:

1 · La unión, con procedencia por fila. Integrar dieciocho fuentes es fácil; integrarlas de modo que después se pueda auditar de dónde salió cada cifra no lo es, y es la diferencia entre un corpus utilizable y una mezcla. De ahí el modelo de §3 y el aparato de §7.

2 · Una capa derivada que no existe aguas arriba: 3.091.596 esqueletos consonánticos normalizados sobre 225 macrosistemas, calculados con las mismas reglas para todos (§4). Las fuentes traen segmentación; la consonante canónica por valor cuálico es cálculo nuestro, y es lo que hace comparables dos lenguas que realizan la misma consonante de forma distinta.

3 · Las decisiones y sus huecos, declarados y medidos. No «excluimos IDS» sino por qué, con la cifra de lo que cuesta; no «retiramos una fuente» sino qué afirmaciones dejan de sostenerse al hacerlo. §6 y §9.

Lo que este corpus NO es: una fuente primaria, un sustituto de ninguno de los recursos de §2.1 y §2.2, ni un dataset más ancho que Lexibank. Es la unión de varios con una capa calculada encima, y su valor está en que la unión sea auditable.

3 · Modelo de datos

87 tablas, organizadas en cuatro capas que se apilan: cada una se calcula desde la anterior y ninguna sobrescribe lo que recibió.

3.1 · El núcleo

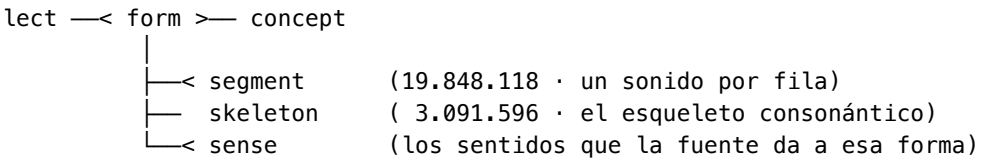

tabla	unidad	qué es
`lect`	un doculecto	una lengua *tal como la documenta una fuente*. No es una lengua en sentido glotológico: dos filas pueden ser el mismo idioma desde dos fuentes, y ésa es la razón de que exista una regla mecánica de duplicados (§3.4)

tabla	unidad	qué es
`concept`	un concepto	la glosa por la que se agrupan formas de lenguas distintas. Es la clave de todo trabajo comparativo
`form`	una forma	la palabra: grafía, IPA de la fuente, segmentación, procedencia
`segment`	un sonido	una fila por segmento IPA, con su posición y su anotación
`skeleton`	un esqueleto	las consonantes canónicas de la forma, calculadas desde `segment` (§4)

3.2 · La capa histórica

`form_etymology` (1.262.353 aristas forma→ancestro) · `ancestry_edge` (linaje entre lenguas) · `cognate_set` / `cognate_member` (157.327 conjuntos, 631.507 miembros) · `protoform_hypothesis` (72.166).

Como se dice en §6.4, esta capa existe solo para el indoeuropeo. Está en el modelo porque el modelo la contempla para cualquier macrosistema; está vacía en 224 de 225 porque ninguna de las fuentes hoy integradas aporta cognación fuera del indoeuropeo — que es una carencia de esta integración y no del campo (§6.4).

3.3 · Procedencia por fila, y la trampa que publicó dos errores

Toda forma lleva su fuente, y su identificador de origen. No es una preferencia de diseño: es el mecanismo de cumplimiento de las licencias CC, que exigen atribución, y hay un invariante que falla si alguna forma se queda sin ella.

Conviene precisar hasta dónde llega, porque «procedencia por fila» puede leerse como más de lo que a veces se ofrece. Aquí ofrece dos cosas: la fuente (`source_id`) y el identificador del registro de origen, que va dentro de `form.id` y conserva la estructura del conjunto del que salió:

```
lexibank:aaleykusunda-KusundaGM-above-1
   fuente : conjunto - doculecto - concepto - nº

ids:107-1-100-1
   fuente : lengua - capítulo - entrada - nº

nel:abk-1000_know-1
   fuente : iso - concepto - nº

kaikki:af:'n appeltjie te skil hê:verb
   fuente : lengua : entrada : categoría
```

De modo que desde una fila se puede volver al registro concreto del conjunto de origen, no solo al conjunto. Lo que no se puede todavía es volver a una *versión* concreta de ese conjunto, porque las versiones no están fijadas (§9.6.1): la trazabilidad llega al registro y se detiene antes de la versión.

Pero hay dos columnas `source_id`, y significan cosas distintas:

	unidad	significa
`form.source_id`	la forma	de qué fuente salió *esta palabra*. Es la que hay que usar para cualquier estadística por fuente
`lect.source_id`	el lecto	de qué fuente salió *el registro de la lengua*. No dice de dónde vienen sus formas

Una misma lengua reúne formas de varias fuentes, así que las dos reparticiones difieren, y mucho:

```
lect.source_id='wiktionary'  con formas de kaikki    133.120
lect.source_id='kaikki'      con formas de ids        45.404
lect.source_id='ids'         con formas de lexibank   19.206
```

El esquema permite escribir cualquiera de las dos y nada avisa de cuál se quiso. Confundirlas publicó dos errores: una tabla de pérdida por fuente que daba el 65,6 % donde era el 100 %, y un chequeo de higiene que atribuyó a la fuente equivocada el 85 % de los infinitivos turcos.

Este documento lo declara porque es un aviso para quien use el corpus, no solo un apunte interno: la trampa no está en nuestro código, está en el esquema, y cualquiera que escriba una consulta puede caer.

Y falta acotar una tercera cosa: qué tablas llevan procedencia. El título de este trabajo dice «procedencia por fila», y esa frase es cierta de la tabla `form` —de la que trata todo lo anterior— pero no de la base entera. Hoy hay trece tablas sin columna de procedencia ni columna de enlace por la que heredarla. Once son producción propia: agregados, estadísticas, tablas de salida de estudios; su procedencia somos nosotros y no hay nada que atribuir. Las otras dos sí conservan contenido de terceros: `concept` (3.205 filas de glosas y definiciones de Concepticon) y `feature` (33.912 filas de rasgos articulatorios de CLTS). Las dos son CC-BY-4.0, se citan en §12 y en las dos la procedencia es única para toda la tabla —cada fila de `concept` es de Concepticon—, de modo que una columna `source_id` sería una constante. No hay problema de licencia; lo que hay es un borde del invariante de §7, y llamarlo de otro modo sería describir una garantía que no existe.

Desde el 6 de septiembre de 2026 ese borde está escrito: las trece tablas se declaran una a una con su motivo en `ci/decisiones.py`, y un meta-chequeo las descubre del catálogo y falla si aparece una que no esté declarada (§7.4). La garantía sigue teniendo un límite; la diferencia es que el límite es auditable y que una tabla nueva sin procedencia no puede cruzarlo en silencio.

3.4 · El contrato de esquema

De ahí sale un artefacto que se publica con el corpus: `ci/esquema.py` declara, para cada columna que los estudios usan, cuatro cosas —cómo se escribe en SQL, de qué es propiedad (lecto, forma, concepto, esqueleto), qué significa en una frase, y con qué se confunde.

La cuarta es la que importa. Las dos `source_id` se declaran confundibles entre sí, y hay una prueba que falla si esa marca desaparece.

La tercera tiene una regla de uso: si el significado no se puede escribir en una frase, la columna no está entendida y no debería usarse en un análisis.

`analysis/tools/verifica_esquema.py` comprueba el contrato contra la base viva, para que no envejezca en silencio cuando una migración cambie una columna.

3.5 · Campos que parecen datos y son heurísticas

Dos columnas del esquema aparentan ser observaciones y son conjeturas. Se declaran juntas, con su nombre, porque nombrarlas es la mitad del trabajo:

`form.is_proper` no dice si la forma es un nombre propio: es un detector por mayúscula inicial. De las 99.698 formas marcadas en indoeuropeo, el 94 % simplemente empieza por mayúscula, y alemán e inglés son 82.000 de ellas. `Abzweig`, `Axt`, `Mahl` son nombres comunes alemanes. Filtrar por esta columna borraría el léxico común alemán.

`form.is_loan` es una marca de préstamo heredada de la fuente, con criterios distintos según cuál. No es comparable entre fuentes ni completa.

3.6 · Hechos declarados sobre los lectos

El corpus declara qué sabe de sus propios lectos, sin decidir qué hacer con ellos:

clase	qué afirma
`duplicado`	el mismo doculecto registrado dos veces desde fuentes distintas (mismo glotocódigo, solapamiento alto de formas). Se indica cuál conservar
`mal_asignado`	error de clasificación en el origen. Hay uno: una lengua kam-sui asignada a urálico, detectada por tres chequeos independientes —está a 6.446 km del centro de la familia, marca tono en dígitos cuando ninguna urálica lo hace, y sus formas son θaːm^{453} «tres» y ðɐm^{214} «agua»
`reconstruido`	el lecto no está atestiguado

La clase es un hecho; qué hacer con ella es política de cada consumidor. Un catálogo de correspondencias excluye las reconstrucciones porque su diseño es no mirarlas; una capa de consulta que enseña el linaje de una palabra hasta el proto-indoeuropeo las muestra, porque es lo que hace. Las dos son correctas, y por eso la decisión no puede vivir en el corpus.

4 · Ingesta y normalización

4.1 · Las dieciocho fuentes

Se ingieren por su formato nativo: CLDF para las bases del ecosistema (Lexibank, IDS, NorthEuraLex, IE-CoR, Concepticon, Glottolog, CLTS), volcado JSON para Kaikki, y raspado propio para las que no publican formato estructurado.

Cada una entra con su fila en `source`, que lleva su cita, su URL, su tipo y su licencia con nombre. «Open» no es una licencia, y dos fuentes estaban declaradas así — queda anotado en §9 como pendiente de identificar o retirar.

Dónde acaba lo que este corpus puede garantizar. La tabla `source` no lleva versión de fuente, y para `kaikki` —el 38 % de las formas— ya no puede llevarla: es un volcado rodante que se regenera desde Wiktionary sobrescribiendo, cuya fecha vive en la página y no en el fichero. Lo que sí hay es la huella `sha256` de los 148 ficheros crudos en `docs/MANIFIESTO-FUENTES.md`, de modo que la entrada del pipeline está identificada aunque su versión se haya perdido. El reparto de responsabilidad está en §9.6.1.

4.2 · La cadena de cálculo

```
form.ipa_raw          la transcripción tal como la escribió la fuente. NO se altera nunca
      ↓
form.segments_raw     la segmentación, de la fuente cuando la trae
      ↓
segment               una fila por sonido, con posición y anotación
      ↓
skeleton.cons_skeleton   las consonantes canónicas, separadas por ·
```

Ninguna capa sobrescribe la anterior. `ipa_raw` se conserva íntegro, y eso es lo que permitió recuperar el tono africano cuando se descubrió que la segmentación lo destruía (§6.6): el dato seguía ahí.

Y hay una regla que gobierna toda la cadena: el esqueleto se calcula desde **`segments_raw`**, nunca desde la grafía. Cuando una fuente no trae segmentación no hay esqueleto, y eso es lo que deja a IDS en cero (§6.2).

4.3 · Tres capas, y una función entre cada dos

El corpus distingue tres representaciones y cada paso entre ellas es una función reproducible, no una interpretación. Se dice así porque usar el corpus no debe exigir aceptar ninguna hipótesis teórica: quien solo quiera comparar formas puede quedarse en la segunda capa e ignorar la tercera.

```
segment          el sonido tal como lo segmenta la fuente
   │  f1 : tabla IPA -> consonante canónica
   v
cons_skeleton    las consonantes canónicas, separadas por ·
   │  f2 : tabla consonante -> símbolo de clase
   v
code             la secuencia de clases
```

Las dos tablas —`f1` en `ingest/recompute_skeleton.py`— son diccionarios explícitos, no procedimientos: dado un segmento, la salida está determinada y se puede reproducir sin ejecutarlas. Un invariante comprueba que las dos secuencias tengan la misma longitud, de modo que la correspondencia posición a posición está garantizada.

Sobre la tercera capa el corpus no afirma nada (§5.1): se publica como columna y su interpretación pertenece a otro trabajo. El nombre que esa literatura da al criterio de f_1 —«consonante canónica por valor cuálico»— no hace falta para usar el corpus: lo que hay que saber es qué agrupa la tabla, y eso está aquí abajo.

4.3bis · Qué es un esqueleto consonántico

La secuencia completa de consonantes de una forma, cada una sustituida por su consonante canónica según `f1`. `kərniajc` → `k·r·n·c`.

Colapsamos deliberadamente contrastes que en muchas lenguas son fonémicos. Conviene decirlo así y no de otra manera: las róticas —`ɾ ɹ ɻ ʀ ʁ ɽ`— se escriben `r`; las nasales `ŋ ɲ ɴ` se escriben `n`, y `ɱ` se escribe `m`; las dorsales fricativas `ɣ χ ç ʝ` se escriben `x`, y las glotales y faríngeas `ɦ ʔ ħ ʕ` se escriben `h`. La aspiración, la labialización y la palatalización se retiran.

No se afirma que la fonología trate esos segmentos como variantes: `/n/` y `/ŋ/` contrastan en muchísimas lenguas, y la aspiración y la palatalización son fonémicas en familias enteras. Lo que se hace es elegir una escala de comparación en la que esos contrastes no se representan, que es una decisión del instrumento y no una descripción de las lenguas.

Por qué esa escala. Sin colapsar, dos lenguas que dicen la misma palabra con distinta realización de la misma consonante no se agrupan, y la variación de superficie oculta la correspondencia. Con ello se paga un precio: cualquier pregunta que dependa de un contraste colapsado no se puede hacer sobre esta capa. El paper de método cuantifica ese precio para un caso concreto —fundir las sibilantes hace invisible el lambdacismo, que define la división primaria del turco.

Y por eso `ipa_raw`, `segments_raw` y `skeleton_raw` se conservan intactos: la normalización es una capa añadida y reversible, no una sustitución. Quien necesite el contraste lo tiene una capa más abajo.

4.4 · Cuatro decisiones que no son obvias

Se declaran porque cada una cambia lo que sale, y tres nacieron de errores encontrados.

Prenasalización ≠ nasalización. La nasalización vocálica (`ã`) es una modulación del segmento y se limpia. La prenasal de `ⁿt`, `ᵐb`, `ᵑg` es un segmento nasal y ocupa su posición. Estuvieron en el mismo saco, y la consecuencia era visible: `mainty` daba `m·t` mientras su cognado `metan` —sin prenasalizar—

daba `m·t·n`. La misma nasal, presente en una lengua y borrada en su hermana, por una decisión del instrumento. La aspiración, la labialización y la palatalización sí se siguen limpiando: la asimetría es deliberada, y está escrita donde se aplica.

Notación que no es segmento, y un defecto que sí lo es. La primera versión de este párrafo metía tres cosas distintas en una frase, y una de las tres estaba mal.

Los superíndices de tono y los paréntesis son notación, no segmentos, y no entran al esqueleto. Eso es correcto — y es exactamente lo que destruye el tono africano, donde el tono va como diacrítico sobre la vocal: para esas lenguas hay que leerlo de `ipa_raw`, que se conserva (§6.6).

Los clics son consonantes, y hasta el 2026-09-06 se descartaban. `ǀ ǁ ǂ ! ʘ` son consonantes nucleares en las familias que las usan, y estaban tratados de dos maneras equivocadas a la vez: `ǁ` vivía en el grupo de laterales del mapa de clases, de modo que **`ǁao`** daba **`l`** —fundiendo el clic lateral con la ele—, y los demás estaban en la lista de notación ignorada, de modo que desaparecían. En 39 formas el clic era la única consonante y la palabra se quedaba sin esqueleto.

Lo primero era peor que lo segundo, y conviene decir por qué: una ausencia se puede declarar y medir; un valor equivocado no, porque nada lo distingue de uno correcto. `l` es un esqueleto bien formado, no deja residuo y no dispara ningún invariante. Es la clase de error de la que trata §8.

Cómo se resolvió, y por qué no de otro modo. El clic entra al esqueleto como sí mismo: `ǀo` da `ǀ`, `ǁk'x'um` da `ǁ·k·x·m`. No se le asigna consonante canónica porque a qué corresponde un clic es una pregunta abierta que este trabajo no ha estudiado, y darle una sería escribir una hipótesis como si fuera dato — que es exactamente lo que hacía la versión anterior.

En la capa de clases lleva `[?]`, el símbolo que el corpus ya usa para «el elemento está en la palabra y el instrumento no puede mostrarlo», y que ya aplicaba a los glides por la misma razón. No es un hueco: es una posición ocupada por algo que aún no se clasifica.

Resultado: de 491 formas con clic, 488 lo conservan en el esqueleto y ninguna se queda sin esqueleto. Las tres restantes están en lenguas sin clics —anamgura, maleng—, donde el signo es un separador de la fuente y no una consonante.

Ocultamiento marcado, no borrado. Cuando el elemento está en la palabra y el instrumento no puede mostrarlo, la posición se conserva marcada con `[?]` en vez de desaparecer. `year` da `[?]·ʌ`, no `ʌ`. Es un suceso nuestro, no de la lengua, y por eso se escribe distinto de una ausencia real. Un invariante comprueba que ningún `[?]` residual llegue a `cons_skeleton`, que es la capa que usan los estudios.

La regla biclase existe en el código y NO está aplicada al dato. Está descrita aquí porque una versión anterior de este texto la presentaba como una decisión vigente, y no lo es. La regla contempla que algunos segmentos —las sibilantes desplazadas `ʃ ʒ ʂ ʐ ɕ ʑ`— aporten dos clases en vez de una, pero el corpus publicado se construyó con ella desactivada: `ʃ` da una sola entrada en el esqueleto y una sola en el código, y el invariante que exige `#clases = #consonantes` pasa a cero, que es la comprobación de que no está activa.

4.5 · Lo que NO se normaliza, y por qué

La grafía. `form.orthography` se muestra tal cual. Solo se retira el punto que algunas fuentes usan para separar segmentos —`b.i.r`, `w.s.w.r`: 296 formas turcas— y únicamente cuando separa caracteres sueltos, porque enseñar «b.i.r» en una tabla haría creer al lector que el turco dice eso.

Los conceptos entre fuentes. Cada fuente trae su glosa; no se unifican por juicio nuestro más allá del mapeo a Concepticon que la propia fuente declara. Unificar glosas es una decisión semántica y no cabe en la ingesta.

Nada por familia. No hay despiece morfológico automático. Recortar finales «que parecen sufijos» es la vía rápida a fabricar exactamente los códigos que uno quiere ver, así que las reglas de despiece son

declaradas a mano, por familia, con doble condición —ortografía y cola de esqueleto— y viven en el proyecto que las usa, no en el corpus.

5 · Las capas calculadas

Sobre el núcleo de §3 se calculan capas que ninguna fuente trae. Ésta es la aportación material del corpus, y conviene separar dos cosas que la base contiene mezcladas: las capas que el corpus ofrece a cualquiera y las tablas que un estudio concreto dejó ahí.

5.1 · El esqueleto consonántico y su capa de clases

tabla	filas	qué es
`skeleton_raw`	3.034.556	las consonantes antes de normalizar
`skeleton.cons_skeleton`	3.078.840	las consonantes canónicas por valor cuálico (§4.3), de 3.091.596 filas de `skeleton`: 12.756 formas no tienen ninguna consonante
`skeleton.code`	3.078.840	la secuencia de clases de esas consonantes
`skeleton_lineage`	461.686	agrupación de un mismo código a través de estadios y ramas

Que `skeleton_raw` se conserve aparte no es redundancia: es lo que permite deshacer la normalización y comprobar qué se agrupó. Una capa calculada sin su entrada es una afirmación sin manera de auditarla.

Y hay que explicar por qué tiene MENOS filas que la capa que precede, porque la intuición dice lo contrario. Son 3.034.556 de 3.091.596, el 98,2 %, y faltan 57.040. No hay ninguna en sentido contrario —cero filas de `skeleton_raw` sin su esqueleto—.

Una versión anterior de este apartado atribuía el hueco a que la pasada era aditiva y cubría solo lo que existía cuando corrió, y concluía que volver a correrla lo cerraba. Se volvió a correr el 6 de septiembre de 2026 y escribió 92 filas: el hueco pasó de 57.132 a 57.040. La explicación era falsa, y conviene decir en su lugar la verdadera, que son tres causas y suman exacto:

	formas	qué pasa
A	34.190 · 59,9 %	la forma no tiene **`segments_raw`**. La pasada lee esa columna, así que las excluye por construcción; su esqueleto se calculó por la vía ortográfica
B	12.590 · 22,1 %	la forma no tiene ninguna consonante: es toda vocal. No hay esqueleto crudo que escribir
C	10.260 · 18,0 %	el esqueleto consta solo de semiconsonantes (`j`, `w`, `ɥ`)

A y B no son defectos: son la definición de la capa. Una forma sin `segments_raw` no tiene un «antes de normalizar» que registrar, y una forma sin consonantes no tiene esqueleto en ninguna de las dos versiones. Ninguna pasada las va a cerrar, y decir que sí lo haría era prometer un arreglo imposible.

C sí es una inconsistencia, y se declara como tal. El 31 de agosto de 2026 se decidió que la semiconsonante no es una categoría de exclusión sino una posición del gradiente de saturación, y desde entonces `skeleton` la mete al esqueleto con el símbolo de elemento oculto (§4.4). La pasada de `skeleton_raw` no recibió ese cambio y sigue descartándola. Las dos capas discrepan sobre qué cuenta como esqueleto, y como `skeleton_raw` existe precisamente para auditar la normalización, la discrepancia le atribuye a la normalización una diferencia que en realidad es de inclusión de semiconsonantes.

Igualarlas cambiaría 499.546 formas —el 16,2 %— y obligaría a recalcular `morph_rule` (2.823.128 filas), `morph_rule2` (2.180.499) y `affix_code_lect`, que alimentan el despiece y los núcleos irreducibles. Decisión (2026-09-06): no se toca el dato en esta versión y la divergencia queda declarada, con un chequeo que la cuenta y falla si crece. Cambiar cifras publicadas para cerrar un hueco del 18 % de un 1,8 % no lo justifica; ocultarlo, tampoco.

Sobre la capa de clases, una advertencia que forma parte del dato. La columna `code` reduce cada consonante a una clase. Está en el corpus, se publica, y el corpus no afirma nada sobre ella: es un instrumento cuyo estatuto requiere un estudio que este trabajo no hace. Se ofrece como columna, no como resultado.

Y tiene un residual declarado: 303.434 esqueletos llevan **`?`** en **`code`** —clases sin clasificar— frente a cero en **`cons_skeleton`**. La asimetría es deliberada y es la que hay que mirar: la capa que los estudios usan está limpia; la que no se interpreta lleva sus huecos a la vista. Hay un invariante para cada una, con severidad distinta.

5.2 · Morfología

`morph` (1.307.991) y `morph_rule` (2.823.128) descomponen la forma en piezas con su papel —núcleo, afijo— y el código de cada pieza. `affix` (18.853) y `construction` (24.267) recogen el inventario resultante.

El origen importa y limita el alcance: la descomposición no se adivina por parecido ortográfico, se toma de la que Kaikki ya publica en la etimología —*«From angosto + -ura»*, *«sub- + marino»*—. Donde no hay descomposición declarada, el núcleo queda nulo, que es lo honesto. Recortar finales que *parecen* sufijos es la vía rápida a fabricar el resultado que uno busca.

Consecuencia: esta capa es tan ancha como la cobertura etimológica de Kaikki, es decir, indoeuropea (§6.4).

5.3 · Sentido y correspondencia

tabla	filas	qué es
`polyseme_link`	1.026.479	formas que la fuente da con más de un sentido
`colex`	231.313	pares de conceptos colexificados: expresados por la misma forma en una lengua
`correspondence`	111.778	correspondencias de sonido entre pares de lectos, con su entorno y su tipo
`sound_class`	3.516	el inventario de clases
`racimo` / `racimo_member`	5.531 / 29.152	agrupaciones por raíz y tema

`colex` es el material de las redes de colexificación; `correspondence` registra, para cada par de lectos, qué clase corresponde a cuál en qué entorno y cuántas veces, distinguiendo conservación de cambio.

5.4 · Lo que hay en la base y NO es una capa del corpus

Esto aparece al escribir esta sección y conviene dejarlo escrito en vez de disimularlo. La base contiene además tablas que son la salida de estudios concretos, no capas de uso general:

`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)

Se reconocen porque su esquema lleva la pregunta dentro: `code_axes` clasifica pares de redes en casillas como «resonancia pura»; `network_code_profile` guarda el valor modal de una red y su cobertura. Eso no es un dato sobre las lenguas: es el resultado de una manera concreta de mirarlas.

Desde la separación de 2026-09-05 esos estudios viven en un repositorio aparte, y este apartado dijo que estas tablas deberían seguirlos. Al ir a moverlas, el 6 de septiembre, resultó que la premisa era falsa y conviene corregirla aquí en vez de dejarla como deuda pendiente de una mudanza que no procede.

Se midió quién las consulta, distinguiendo quién las lee de quién las construye. De las veinticinco que no tienen claves ajenas y por tanto podrían moverse, **`codigos`** —el repositorio de los estudios— no lee ninguna, y Meulemans lee diecisiete: `morph_rule2`, `net_quality`, `network_code_profile`, las seis tablas `ncp_*`, `code_specificity`, `code_family_b`, `code_family_dmas`, `form_nucleus2`, `affix_code_lect`, `affix_allomorph` y `f6_bateria`. Las seis restantes tampoco están sueltas: `morph_rule`, `form_nucleus` y `affix_inventory` son la entrada con la que se construye `nucleus_inventory`.

De modo que no son salida de estudios a la espera de mudarse: son lo que sirve la capa de consulta, y su sitio es donde están, por la misma razón por la que las tablas `qf_*` se quedaron cuando Tatoeba salió (§9.6.3). Lo que sí se movió fue lo que de verdad era salida de estudio y nadie más leía: `construction` y `construction_align`, que salieron a la base `qualicfields`.

Queda en pie el criterio de abajo, y el reparo que lo motivaba: **`code_axes`** sigue siendo una tabla cuya definición es la hipótesis de un estudio, y que esté en la base del corpus es discutible aunque no haya adónde llevarla hoy. Lo que cambia es que eso ya no se llama deuda de una mudanza, sino una tensión entre el criterio y el hecho de que la capa de consulta se construyó encima.

El criterio para distinguirlas, por si sirve a quien construya algo parecido: una capa es del corpus si más de un consumidor la usaría sin discutir su definición. `skeleton` lo es —la usan tanto los catálogos como la capa de consulta—. `code_axes` no: su definición es la hipótesis de un estudio.

6 · Cobertura, y dónde se acaba

Esta sección describe las capas del corpus y hasta dónde llega cada una. No da una cifra de «corpus utilizable», porque no existe: lo utilizable depende de la pregunta, y dos consumidores del mismo corpus —una capa de consulta de historia de la palabra y un catálogo de correspondencias forma–sentido— necesitan capas distintas y llegarían a cifras distintas, las dos correctas. Lo que sí se da es cada capa con su cobertura, y las intersecciones que más varían entre macrosistemas.

6.1 · El corpus, por capas

lectos	4.735 (3.350 con glotocódigo)
macrosistemas	225
conceptos	3.205
formas	3.788.294
formas con concepto asignado	2.533.666 (66,9 %)
esqueletos consonánticos	3.091.596
segmentos	19.848.118
aristas de etimología	1.262.353
conjuntos de cognados	157.327 (631.507 miembros)
protoformas	72.166
fuentes	18, ninguna con licencia que impida redistribuir — con la salvedad de §9.6.2

No hay una sola cifra de «corpus utilizable», y conviene decirlo antes que nada, porque la tentación de dar una es fuerte y produce descripciones falsas. Lo utilizable depende de la pregunta:

- Un navegador de la historia de una palabra —como la capa de consulta Meulemans (§13.1)— necesita forma, sentido, etimología y cognados. Le sirven 3,79 M de formas, y una forma sin concepto asignado es materia prima perfectamente buena: se busca la palabra, se ve su linaje.
- Un catálogo de correspondencias forma–sentido —como el programa de códigos consonánticos (§13.1)— necesita esqueleto y concepto a la vez, porque su unidad es concepto × lengua. Ahí lo utilizable son 2.058.119 esqueletos, el 66,6 % de los que hay.
- Un estudio fonológico sobre inventarios no necesita concepto en absoluto, y le sirven los 3,09 M de esqueletos enteros.

Las tres cifras son correctas. La sección describe las capas y su cobertura, y quien use el corpus compone la suya.

6.2 · Dos fuentes grandes que fallan en direcciones opuestas

fuente	formas	con concepto		con esqueleto
`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

La asimetría explica la forma del corpus entero y conviene verla junta:

Kaikki trae la forma sin el sentido. Es un volcado de diccionario: entradas con definición en prosa, no elicitadas contra una lista de conceptos. Mapear una definición de Wiktionary a un concepto de Concepticon es un problema de emparejamiento semántico, y hoy resuelve el 14,8 %. Son 1,24 M de formas con esqueleto calculado y sin concepto: invisibles para un catálogo, plenamente utilizables para consultar una palabra.

IDS trae el sentido sin la forma. Sus 437.902 entradas vienen de una lista de conceptos, así que el 100 % está mapeado — pero su CLDF trae la columna `Segments` vacía en origen y sus formas son ortografía, no IPA. Sin segmentos no hay esqueleto: cero de 437.902. Convertir esa ortografía exigiría un perfil por lengua para 274 lenguas, y un perfil mal hecho no pierde la forma: fabrica un esqueleto falso que entra con aspecto de dato. Declarado fuera de alcance, con su motivo, en `docs/DECISION-IDS.md`.

6.3 · Por macrosistema

La penúltima columna es la intersección de dos capas —esqueleto y concepto—, no un juicio sobre lo que sirve: es la que necesita un catálogo de concepto × lengua, y se da porque es la que más varía entre macrosistemas.

macrosistema	lenguas	formas	esqueleto		esq. + concepto		conceptos
Indoeuropeo	185	1.554.812	1.270.368	81,7 %	323.476	20,8 %	2.785
Austronesio	586	233.351	218.840	93,8 %	218.840	93,8 %	1.825

macrosistema	lenguas	formas	esqueleto		esq. + concepto		conceptos
Najo-daguestánico	56	231.624	104.562	45,1 %	104.170	45,0 %	1.818
Sino-tibetano	270	200.399	200.399	100 %	200.339	100 %	2.737
Atlántico-congo	387	157.076	157.076	100 %	157.076	100 %	1.787
Urálico	28	101.065	82.952	82,1 %	81.142	80,3 %	1.929
Trans-neoguineano nuclear	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
Austroasiático	102	86.305	49.866	57,8 %	49.866	57,8 %	1.570
Turco	33	70.853	55.877	78,9 %	55.300	78,0 %	1.828
Afroasiático	112	69.237	64.460	93,1 %	64.323	92,9 %	2.034
Tupí	64	48.532	41.269	85,0 %	41.269	85,0 %	1.514
Pama-ñungano	155	45.855	45.855	100 %	45.855	100 %	1.160

Cola larga: 204 macrosistemas más, 657 lenguas, 392.008 formas, 328.051 esqueletos. Y 142 lectos con 222.204 formas sin macrosistema asignado.

El macrosistema mayor es el que menos se solapa, y es consecuencia directa de §6.2. El indoeuropeo tiene 1,55 M de formas y solo el 20,8 % lleva esqueleto y concepto a la vez, porque casi todo Kaikki es indoeuropeo. Los macrosistemas al 100 % —sino-tibetano, atlántico-congo, trans-neoguineano, pama-ñungano— lo están porque vienen enteros de Lexibank, que es material elicitado contra listas de conceptos.

De nuevo: eso no dice que el indoeuropeo sirva menos, dice para qué sirve cada parte. Sus 1,23 M de formas con esqueleto y sin concepto son exactamente donde vive la capa etimológica (§6.4), que no existe en ningún otro macrosistema. El tamaño de un macrosistema no predice qué se puede hacer con él, y tampoco lo predice una sola de sus capas.

6.4 · Etimología y cognación: indoeuropeo, y nada más

Ésta no es una carencia que el corpus deba disculpar: es la condición de partida del campo, y es la razón por la que todo lo construido encima está organizado como está.

macrosistema	formas	con etimología	en cognados
Indoeuropeo	1.554.812	424.787 (27,3 %)	279.010 (17,9 %)
los otros 224	—	0 (0,0 %)	0 (0,0 %)

Cero exacto, no «poco». Las tres redes de cognación del corpus lo confirman: las tres viven en un solo macrosistema.

red	conjuntos	miembros	lenguas	macrosistemas
`kaikki-cog`	82.664	356.354	36	1
`kaikki-etymology`	72.023	251.757	37	1
`iecor-gold`	2.640	23.400	152	1

El corpus es ancho en forma y estrechísimo en historia: 225 macrosistemas con esqueleto consonántico, uno con juicios de parentesco integrados.

Y hay que decir con precisión de qué es esto una afirmación, porque la formulación corta invita a leer más de lo que dice. Lo que se afirma es una carencia de esta integración, no del campo: fuera del indoeuropeo existen tradiciones comparativas maduras y conjuntos de cognación publicados —el *Austronesian Comparative Dictionary*, las bases de cognados urálicas, el trabajo comparativo bantú y sino-tibetano, entre otros— y ninguno de ellos está integrado aquí. La afirmación correcta es:

> Ninguno de los recursos históricos hoy integrados en este corpus ofrece cobertura de cognación comparable fuera del indoeuropeo.

Que es un enunciado sobre nuestras dieciocho fuentes y sobre el trabajo de integración pendiente, no sobre lo que la lingüística histórica ha hecho.

Lo que sí se puede hacer es construir para esa condición en lugar de contra ella, y es lo que explica la forma del trabajo que usa este corpus:

- Los catálogos se construyen por macrosistema y sin cognación, porque en 224 de 225 no la hay. Un método que la exija funciona en el 0,4 % de las lenguas del corpus.
- El indoeuropeo, precisamente por ser el único con juicios de oro, no es el objeto sino el banco de pruebas: es donde se puede medir cuánto pierde un método que trabaja sin ellos, comprobándolo contra la respuesta conocida.
- Y cada macrosistema se trabaja aparte, porque lo que hay que aprender es cuánto cuesta el instrumento en cada uno — no un promedio sobre 225 que no describiría a ninguno.

6.5 · Anotación fonológica: lo que el esquema declara y lo que hoy trae

columna	poblada	
`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	pendiente
rasgos articulatorios (`feature`)	18.977.285	95,6 %

Tres columnas están declaradas y pendientes de trabajar, no rotas: existen en el esquema porque el modelo las contempla, y no se han poblado todavía. La distinción importa y es la misma que gobierna §7: una columna vacía que se sabe vacía es una tarea; una que se cree llena es un error esperando a publicarse.

Los rasgos articulatorios son la capa mejor cubierta, con un 95,6 % de los segmentos.

6.6 · Tono: dos vías que no son intercambiables

El tono llega al corpus por dos caminos distintos, y confundirlos produce ceros falsos:

Como segmento propio (`prosodic='T'`), en las transcripciones del sudeste asiático: 339.543 segmentos, recuperables directamente.

Como diacrítico sobre la vocal, en las ortografías africanas: 41.953 formas solo en atlántico-congo. Ahí la segmentación lo destruye —`máma dáda` se convierte en `m a m a d a: d a`— y hace falta extraerlo del campo `ipa_raw`, decidiendo por lengua qué significa cada marca. Documentado en `../codigos/estudios/DECLARACION-TONO-AFRICANO.md`.

Y hay una trampa que conviene declarar porque ya produjo una lectura equivocada nuestra: la ausencia de tono no es observable. Gurung y Tamang son lenguas tonales de manual y en este corpus salen a 0,0 % de formas con tono, porque su fuente no lo transcribe. No se puede distinguir «esta lengua no tiene tono» de «esta fuente no lo escribió», y cualquier medida cuyo contraste dependa de esa ausencia estará midiendo la práctica del transcriptor.

6.7 · Lo que el corpus no puede responder

Se enumera aquí para que no haya que deducirlo de las tablas:

1. Nada que exija cognación fuera del indoeuropeo. §6.4.
2. Nada que exija esqueleto en las 274 lenguas de IDS. §6.2, y `docs/DECISION-IDS.md`.
3. Nada que exija saber que una lengua NO tiene tono. §6.6.
4. Nada que exija concepto en 1,24 M de formas de Kaikki, que tienen forma, esqueleto y procedencia pero no significado mapeado. Sirven para consultar una palabra y para preguntas sobre la forma; no para nada cuya unidad sea concepto × lengua. §6.2.
5. Nada prosódico: duración, acento y estructura silábica fiables. `is_stressed` y `syllable` están al 49,3 % y no cubren las familias tonales.
6. Nada morfológico por lengua. No hay segmentación morfológica; las reglas de despiece que existen son por familia, declaradas a mano y limitadas a formas de cita.

7 · Verificación

Un corpus integrado de dieciocho fuentes acumula errores que no se manifiestan como errores. No hay excepción, ni fila que falte, ni consulta que falle: hay una cifra ligeramente distinta de la que debería ser, y nada la señala. Esta sección describe el aparato que se publica con el corpus para hacer visible esa clase de fallo, y la sección siguiente da los casos que lo justifican.

7.1 · El principio: excepción declarada frente a hueco silencioso

Todo lo demás se sigue de aquí.

> Una ausencia sin declarar es indistinguible de un descuido.

Cuando una fuente no aporta esqueletos, cuando un lecto queda fuera, cuando una columna está vacía, hay dos mundos posibles: alguien lo decidió, o a alguien se le pasó. Desde fuera —y desde dentro, seis meses después— son idénticos. La única diferencia la produce escribirlo.

Por eso las decisiones viven en `ci/decisiones.py` y no dispersas en el código que las aplica, cada una con su motivo y su cifra; y por eso los invariantes las consultan: una fuente declarada fuera de alcance se reporta como decisión, y una fuente nueva con cero esqueletos sigue fallando.

7.2 · Tres capas, y ninguna sustituye a otra

capa	qué comprueba	cuántas
pruebas unitarias (`tests/unit/`)	la arquitectura: que el DSN se defina en un solo sitio, que nadie abra conexiones a mano, que ningún script calcule la raíz contando niveles, que la deuda declarada lleve techo	34
contrato de esquema (`ci/esquema.py`)	que cada columna publicada diga qué significa, en qué unidad está y con qué se confunde	11 columnas
invariantes de base (`tests/qa.py`)	el dato vivo: integridad referencial, cobertura de capas, coherencia de las redes, licencia	53

Las pruebas unitarias no tocan la base —hay un invariante que lo comprueba— y por eso corren en cinco segundos y sirven de gancho de pre-commit. Los invariantes sí, y tardan treinta y seis.

7.3 · Los invariantes: 61 en 20 clases

clase		qué vigila
`SKEL`	10	la capa de esqueletos: residuales, descuadres, lenguas y fuentes sin ninguno
`LIC COG`	6 c/u	procedencia y licencia · las redes de cognación, con la regla propia de cada una
`PROTO`	5	las protoformas
`INT ETY CLS RAW`	4 c/u	integridad referencial · linaje · clases de sonido · el hueco de `skeleton_raw`, separado por causa (§5.1)
`DUP VOW MARK FEAT POLY SUBS`	2 c/u	duplicados, vocales, anotación ajena, rasgos, polisemia, sustratos
`CPT CORR CRY COLX SPOT SEG`	1 c/u	conceptos, correspondencias, criptología, colexificación, muestreo, segmentación

Cada uno declara severidad: `fail` tumba la puerta, `warn` avisa, `info` describe. La proporción es deliberada, porque un chequeo que solo avisa no se lee, y por eso los cinco que avisan van enumerados uno a uno en §7.5 en vez de quedarse en la salida del programa.

En la ejecución que acompaña a esta versión: 39 en verde, 5 avisos, 3 deudas declaradas con techo, 14 informativas y ninguna en rojo.

Cuando uno pasa de `info` a `fail` se registra por qué. El de licencia lo hizo el 2026-09-05 al decidirse que el corpus se publica sin restricciones: antes podían existir filas no redistribuibles porque un script de exportación las quitaba al salir; ahora hay una sola base y es la distribuible, así que cualquier fila de una fuente retirada es una violación.

7.4 · Los meta-chequeos, que es lo que hace distinto a este aparato

Un invariante comprueba el dato. Un meta-chequeo comprueba que el aparato cubra lo que dice cubrir, y son tres:

`COG · red sin invariante declarado`. Cada red de cognación tiene su regla propia, porque no comparten criterio: una red por étimo exige que el miembro pertenezca a la familia que declara el conjunto, y la red `kaikki-cog` cruza familias por diseño, así que ahí basta con que el lecto exista. Generalizar una regla a todas producía falsos positivos. Pero separarlas abre un hueco: una red nueva

entra sin regla y nadie se entera. Este chequeo falla si alguna queda fuera, y es el que descubrió 23.400 miembros de `iecor-gold` sin invariante.

`SKEL · lenguas enteras sin esqueleto — SIN declarar`. No cuenta las lenguas sin esqueleto: cuenta las que no se explican por ninguna decisión escrita. Una lengua con más de doscientas formas y ninguna con esqueleto o bien tiene su fuente declarada fuera de alcance, o bien es un lecto declarado —duplicado, mal asignado, reconstruido—, o bien es un agujero. Los dos primeros se reportan como decisión; el tercero falla.

`LIC · tabla sin procedencia sin declarar`. El invariante de licencia mira las tablas que tienen columna `source_id` o `source`, y su alcance es ése. Durante meses la garantía «ninguna fila sin licencia sale de aquí» se enunció sin decir a qué tablas no llegaba, y llegó a haber 13,3 millones de filas fuera de su vista (§9.6.3). Este chequeo descubre del catálogo las tablas que no tienen procedencia ni columna de enlace por la que heredarla, y exige que cada una esté declarada con su motivo; falla si aparece una que no lo esté. Hoy son trece, y dos de ellas —`concept`, de Concepticon, y `feature`, de CLTS— sí conservan contenido de terceros: las dos son CC-BY-4.0 y su procedencia es única para toda la tabla, de modo que una columna `source_id` sería una constante. El borde del invariante sigue existiendo; lo que cambia es que está escrito y que una tabla nueva no puede cruzarlo en silencio.

La diferencia con un chequeo corriente: estos fallan cuando el aparato se queda corto, no cuando el dato está mal. Es la única defensa contra la clase de error de §8.

7.5 · Los cinco avisos, uno a uno

§7.3 dice que un chequeo que solo avisa no se lee, y esa frase describía lo que estaba pasando aquí: `qa.py` reporta cinco avisos en cada ejecución y ninguno aparecía en este trabajo. Quien clonase el repositorio y corriese `./verifica.sh` vería mil cuatrocientos ciclos genealógicos que el paper no mencionaba. Van uno a uno, con lo que resultó ser cada uno al mirarlo.

`ETY · 2-ciclos (A→B y B→A)` — 1.406 pares, 105 lectos, sobre 12.378 aristas. El aviso los llama «imposible en herencia» y para la mayoría no lo son. `ancestry_edge` es un agregado: se construye subiendo las aristas de `form_etymology`, que van de palabra a palabra, hasta el nivel de la lengua. Que el latín aparezca como ancestro del portugués y el portugués como ancestro del latín no dice que dos lenguas se engendren mutuamente: dice que una palabra latina fue al portugués y otra palabra portuguesa entró en el latín, cosa que ocurre y que en latín científico moderno ocurre mucho. El 74 % de los ciclos —1.042— lleva al menos una arista marcada como préstamo, y ésos son exactamente lo que se espera de dos lenguas en contacto. Los otros 364 son **`herencia`** ↔ **`herencia`**, y ésos sí serían imposibles: son pares dispersos —búlgaro↔rumano, neerlandés↔siciliano, galés↔español— donde un editor de Wiktionary etiquetó como herencia lo que era préstamo. Son el 2,9 % de las aristas y son de la fuente, no del cálculo (§10.1).

`ETY · child más VIEJO que parent` — 9 casos, y todos el mismo. Los nueve tienen al latín `la` como hija de algo posterior: del español desde 1200, del alemán desde 1500, del alto alemán antiguo desde 750, del latín medieval desde 700. La causa no es cronológica sino de modelo: **`la`** es un solo lecto con las fechas del latín clásico —hasta el 200— y sus entradas incluyen latín medieval y moderno, que sí tomó de esas lenguas. Los lectos para esos estadios existen (`la-eme`, `la-med`, `la-new`) y las palabras no están repartidas entre ellos. Es un defecto declarado del reparto por lectos, no una imposibilidad en los datos.

`DUP · misma (lengua, grafía) en >1 fuente` — 181.404. No son errores: son la misma palabra recogida por dos proyectos. Dominan `ids+lexibank` (82.520) y `lexibank+nel` (78.824), que son listas de conceptos que se solapan por diseño. Reconciliarlas —decidir cuándo dos filas son un nodo y cuándo dos— es trabajo abierto, y hasta hacerlo la política es marcar y no filtrar (§7.2): se conservan las dos y quien mida por fuente sabe que hay solape.

DUP · misma (lengua, grafía) misma fuente >1 — 378.880. De ésas, 100.193 se distinguen por categoría gramatical y son homógrafos legítimos, la clase de cosa que un corpus debe conservar separada. Las 278.687 restantes comparten categoría, y ahí conviven entradas distintas de Wiktionary para la misma grafía —sentidos con etimologías distintas— con duplicados verdaderos. Separar unas de otras pide un criterio que este trabajo no fija.

SKEL · esqueleto más largo que su grafía — 140, y el único que destapó un defecto real. El chequeo se restringe a grafía latina a propósito, para no contar el árabe y el persa, donde el esqueleto es legítimamente más largo que lo escrito porque la escritura no anota vocales breves ni la geminación. De los 140 latinos, unos son abreviaturas cuya transcripción deletrea la palabra entera (`Dx` → `d·k·s`), pero otros son otra cosa: **nord** da **`n·r·d·n·ɖ`**, y viene de **`/nu:rd/[nu:ɖ]`**. Son dos transcripciones, la fonémica y la fonética, concatenadas en un solo esqueleto. Medido en toda la base: 63 formas traen **`/…/`** y **`[…]`** a la vez, las 63 tienen esqueleto, y en 48 el esqueleto es literalmente su primera mitad repetida —`twelfth` sale `t·w·l·f·θ·t·w·l·f·θ`—. Son el 0,002 % de los esqueletos y están mal. Conviene notar que el aviso solo los caza cuando la palabra es corta: fue el chequeo el que llevó hasta ellos, no al revés. Las variantes separadas por coma de `lexibank` —`(j)et, jit` → `j·t`— sí se procesan bien, y se comprobó al medir esto.

Qué se hizo con ese defecto, y por qué no más. El origen está aguas arriba, en `segments_raw`: el segmentador cortaba por , que es como kaikki separa variantes, pero estas dos van separadas por un espacio, y el limpiador quitaba `/`, `[` y `]` sueltos dejando las dos cadenas pegadas. El código está corregido —`ingest/segment_kaikki.py` toma ahora la primera transcripción delimitada— y hay una prueba unitaria que comprueba la propiedad y no solo el caso: que la secuencia no pueda ser su primera mitad repetida. Ninguna forma nueva puede entrar así.

Las 63 filas ya escritas no se reparan, y conviene decir el porqué con números en vez de con una disculpa. Repararlas toca 1.473 filas en las tablas por forma, que es asumible; pero obliga además a recalcular seis tablas de agregado que se construyen por pasada completa —`morph_rule` (2.823.128), `morph_rule2` (2.180.499), `code_axes`, `code_specificity`, `skeleton_lineage` y `correspondence`— y que alimentan el despiece y los núcleos irreducibles, es decir cifras ya publicadas. Cambiar siete millones de filas derivadas para arreglar sesenta y tres formas del 0,002 % no es una mejora: es un riesgo. Queda como tercera deuda declarada con techo (§7.6).

Lo que los cinco tienen en común y por lo que van juntos aquí: ninguno es un fallo del cálculo, y cuatro de los cinco no son fallos en absoluto — son propiedades de integrar fuentes que se solapan y que no siempre coinciden. El quinto sí, y son 63 formas. La diferencia entre declararlos y no declararlos es que ahora un lector puede juzgar eso por su cuenta.

7.6 · Deuda declarada, con techo

Un invariante roto que no se puede arreglar hoy tiene tres destinos posibles: se ignora —y se olvida—, se borra —y desaparece la señal—, o se declara con un techo.

Hay tres así. El primero: 42.891 formas donde `cv_template` no tiene la misma longitud que `segments_raw`, anotado con su alcance —afecta a la contabilidad de vocales, no a `cons_skeleton`, que es lo que usan los estudios— y con un techo de 45.000.

El segundo se declaró el 6 de septiembre de 2026: las 10.260 formas cuyo esqueleto es solo semiconsonántico y que por eso no tienen fila en `skeleton_raw` (§5.1). Aquí el techo, 11.000, hace un trabajo que conviene señalar: la divergencia entre las dos capas no se arregla en esta versión porque igualarlas cambiaría 499.546 formas y tres tablas derivadas, y esa decisión tiene un coste que no se toma de pasada. Lo que el techo garantiza es que la divergencia no crezca mientras tanto, que es lo único que se puede prometer sin tomarla.

El tercero, del mismo día: las 63 formas con dos transcripciones concatenadas de §7.5, con techo de

70. Es el que mejor enseña para qué sirve esta figura, porque las tres partes se separan limpiamente. El código está arreglado, así que el problema no puede crecer por la vía por la que llegó. Las filas presentes no se tocan, porque repararlas movería siete millones de filas derivadas y cifras publicadas. Y el techo cubre la vía que queda: que entren por otro camino que no hayamos visto. Un arreglo hacia adelante sin techo sería confiar en que no hay otro camino.

Informan mientras no crezcan; tumban la puerta si superan su techo. Un problema conocido y acotado no es lo mismo que un problema ignorado, y la diferencia es un número.

7.7 · Una sola puerta

`verifica.sh` encadena las cinco clases de comprobación y devuelve un solo código de salida:

```
pruebas unitarias · márgenes sobre el PDF · referencias internas ·
contrato de esquema · invariantes de la base
```

Existe porque cada una encontró fallos reales y ninguna se ejecutaba sola. Tiene un modo `--rapido` sin base de datos, en cinco segundos, para poder usarse como gancho.

7.8 · Cuatro reglas que el aparato aplica a sí mismo

Salieron de fallos concretos, y cada una está escrita en el código que la aplica:

1. Descubrir del catálogo, no de una lista a mano. El invariante de licencia recorre las columnas de procedencia preguntándoselas al esquema. Una tabla nueva queda vigilada sola; una lista escrita a mano se desincroniza el día que alguien añade una tabla.
2. Un chequeo que cubre menos de lo que su nombre promete es peor que no tenerlo, porque produce confianza falsa. Los verificadores de papers reportan su cobertura: cuántos elementos miraron sobre cuántos había.
3. Preguntarle al parser, no reimplementarlo. El meta-chequeo de cobertura llegó a informar el 120 % porque duplicaba las expresiones del analizador en vez de pedirle su propio denominador.
4. Un chequeo que nunca ha disparado no está probado. El invariante de licencia se validó insertando una fuga por cada camino posible —por nombre de fuente retirada y por `redistributable = false`— y revirtiendo. Sin esa prueba, «informa cero» y «no mira nada» son indistinguibles.

8 · Seis errores, y por qué tenían la misma forma

Esta sección no es una confesión. Es el único material que demuestra que el aparato de §7 sirve para algo: un invariante que nunca ha disparado no está probado, y un aparato descrito sin los fallos que lo originaron es una lista de buenas intenciones.

Los seis casos son reales, están en el historial del repositorio, y cinco de los seis pasaron todas las comprobaciones que había en su momento. Ése es el punto de la sección.

(Los casos 1, 2, 3 y 6 afectan al análisis construido sobre el corpus, que desde 2026-09-05 vive en un repositorio aparte. Se incluyen porque son los que enseñaron la disciplina y porque el patrón es el mismo.)

8.1 · El nulo que medía nuestro propio sorteo

Tres papers publicaron una cifra de enriquecimiento con su intervalo de confianza. El nulo se estimaba barajando ochenta mil veces, y el intervalo se presentaba como si midiera la incertidumbre de los datos.

Medía la del sorteo. Se descubrió al quitar seis formas turcas y ver que el nulo se movía un 14 % — una sensibilidad imposible si el intervalo describiera la población. Sustituido por la forma cerrada `(Σ_c n_c² − Σ_cl n_cl²) /(N² − Σ_l n_l²)`, las cifras publicadas cambiaron: 225×→218×, 77×→78×, 39×→37×.

Qué no lo detectó: nada. No había con qué. El intervalo era ancho, y un intervalo ancho parece prudencia.

Qué lo detecta ahora: el nulo es exacto y no tiene intervalo que malinterpretar. Y hay una prueba que valida la forma cerrada contra el barajado por fuerza bruta, porque una forma cerrada sin validar es exactamente lo que publicó una cifra equivocada durante tres papers.

8.2 · El anexo empalmado tres veces, con 8/8 chequeos en verde

Seis papers llevaban su anexo pegado tres veces. Los seis pasaban los ocho chequeos del verificador.

El chequeo comprobaba que el anexo existiera. Existía. Tres veces.

Qué lo detecta ahora: el empalmador corta por la primera cabecera de anexo y falla si al terminar no queda exactamente uno. La diferencia entre `≥1` y `== 1` es todo el error.

8.3 · Cinco de seis scripts que daban un resultado distinto en cada corrida

La consulta que carga los datos de una familia no tenía `ORDER BY`. PostgreSQL no promete orden sin él, y no lo daba: la misma familia cargaba distinta en cada proceso. Todo lo de aguas abajo heredaba ese orden, y `Counter.most_common(1)` rompía los empates por orden de inserción.

Resultado publicado: la tabla de sustituciones por par de ramas decía `b p` a 0,5× o `b n` a 1,9× según la corrida.

Qué no lo detectó: las cifras eran plausibles las dos veces. Nadie corre el mismo script dos veces para comparar.

Qué lo detecta ahora: una prueba unitaria que lee la consulta y exige que el `ORDER BY` termine en una columna única. Ordenar por lengua y concepto deja los empates sueltos otra vez.

8.4 · Dos columnas con el mismo nombre y distinto significado

`form.source_id` es la fuente de la forma. `lect.source_id` es la de la ficha de la lengua. El código cargaba la segunda en un campo llamado `fuente` y la usaba como si fuera la primera.

Consecuencias publicadas: una sección entera del paper indoeuropeo daba que IDS perdía el 65,6 % de sus formas cuando pierde el 100 %; y un chequeo de higiene atribuía a la fuente equivocada el reparto de los infinitivos turcos —`ids` al 85 %, no `nel` al 6,4 %.

Qué lo detecta ahora: el contrato de esquema declara, para cada columna, con qué se confunde; hay una prueba que exige que las dos `source_id` estén marcadas como confundibles entre sí, y otra que exige que el campo que carga la del lecto se llame `fuente_del_lecto` y no `fuente`. El error nació de un nombre.

8.5 · El invariante de licencia que informaba cero teniendo 2.140 filas dentro

Al retirar del corpus las fuentes no redistribuibles, el barrido recorrió todas las tablas con columna `source_id` y borró 5.971 filas. El invariante de licencia informó cero filas en cuarentena.

Había 2.140. La tabla `cognate_set` guarda la procedencia en una columna llamada `source`, sin sufijo, y ni el barrido ni el invariante la miraban.

Qué lo detectó: no el invariante de licencia, sino otro chequeo — el meta-chequeo de redes de cognación, que se quejó de una red sin invariante declarado. Un fallo de licencia salió a la luz por una comprobación de coherencia de cognados.

Qué lo detecta ahora: el invariante mira las dos formas de nombrar la columna, y compara además por nombre declarado y no solo por `JOIN` contra `source.redistributable` — porque una vez borrada la fila de la fuente, el `JOIN` no encuentra nada aunque queden referencias, e informaría cero precisamente cuando hay fuga. Se validó insertando una fuga por cada camino y revirtiendo.

8.6 · El meta-chequeo que informaba el 120 % de cobertura

El verificador de papers reporta cuántos elementos miró sobre cuántos había. Informaba 120 %.

El denominador lo calculaba yo, duplicando las expresiones regulares del analizador en vez de pedirle las suyas. Las dos versiones divergieron y el cociente se fue por encima de uno.

Qué lo detecta ahora: el denominador lo devuelve el propio analizador. Preguntarle al parser, no reimplementarlo.

8.7 · La forma común, que es la tesis

Los seis son el mismo error:

> Una comprobación que cubría menos de lo que su nombre prometía, e informaba conformidad exactamente cuando había fallo.

El chequeo del anexo prometía «el anexo está bien» y comprobaba que existiera. El invariante de licencia prometía «no hay cuarentena» y miraba una de las dos columnas donde puede haberla. El meta-chequeo prometía «cobertura» y dividía por un número que se inventaba. En los tres casos, el verde era más peligroso que la ausencia de chequeo, porque un chequeo verde clausura la pregunta.

De ahí las cuatro reglas de §7.8 y, sobre todo, los meta-chequeos: comprobaciones cuyo objeto no es el dato sino el alcance del aparato. Son las únicas que pueden detectar esta clase de fallo, porque el fallo no está en el dato.

8.8 · Una recomendación transferible

Para cualquiera que publique un recurso de datos con comprobaciones:

1. Escribe qué NO cubre cada chequeo, junto a lo que cubre. Si no se puede escribir, el chequeo no está entendido.
2. Haz que fallen, no que avisen. Un aviso en un log de cien líneas no existe.
3. Descubre del catálogo, no de una lista mantenida a mano.

4. Prueba que disparan. Introduce el fallo, comprueba que el chequeo lo ve, revierte. Un chequeo que nunca ha disparado no está probado — solo no ha fallado todavía, que no es lo mismo.
5. Mide la cobertura del propio aparato, y haz que eso también falle.

8.9 · Una nota final: el aparato funcionando

El 2026-09-05, al separar el corpus del programa de análisis que lo usa, cinco pruebas de arquitectura fallaron a la vez. Afirmaban la estructura anterior: que `familia.py` estuviera en el paquete del corpus, que la puerta de verificación encadenara el verificador de papers, que `analysis/` estuviera repartido en tres subdirectorios concretos.

Ninguna era un fallo. Eran cinco descripciones de la arquitectura que habían dejado de ser ciertas, y la suite lo dijo antes de que nadie publicara nada sobre una estructura que ya no existía. Es exactamente para lo que están.

9 · Licencia, distribución y disponibilidad

9.1 · La licencia del conjunto, y quién la impone

CC-BY-SA-4.0. No es una elección: es el mínimo común denominador que imponen las fuentes.

licencia	fuentes	formas
CC-BY-4.0	7 — CLTS, Concepticon, Glottolog, IDS, IE-CoR, Lexibank, NorthEuraLex	2.325.336 (61,4 %)
CC-BY-SA-3.0	6 — la familia Kaikki	1.462.960 (38,6 %)

Las CC-BY se integran limpiamente porque BY es compatible hacia BY-SA. Pero el 38,6 % de las formas viene de Kaikki, que es CC-BY-SA, y ShareAlike se hereda: cualquier obra derivada —incluidas las capas calculadas de §5 y los catálogos que producen los estudios— sale con la misma licencia.

Es el camino honesto y también el simple: una sola licencia para todo el conjunto, sin excepciones que haya que recordar.

9.2 · La atribución no es una promesa, es una columna

CC-BY y CC-BY-SA exigen atribución. En este corpus el mecanismo es estructural: toda forma lleva su **`source_id`**, y hay un invariante con severidad `fail` que tumba la puerta si alguna se queda sin él.

Cero formas sin procedencia, de 3.788.294.

9.3 · El contenido, sin restricciones de licencia

«Sin restricciones» se refiere aquí a la LICENCIA, no al acceso, y conviene separarlo antes de seguir porque §9.5 dice que el repositorio es privado y las dos cosas suenan a contradicción. No lo son: el contenido no arrastra ninguna cláusula que impida redistribuirlo, y quién puede clonarlo hoy es otra pregunta.

Hasta el 2026-09-05 había dos bases: una de investigación con todo y una pública que un script producía filtrando al exportar. Mantener dos verdades obliga a acordarse de cuál se está mirando, y una de las dos acaba desactualizada. Ahora hay una, y es la distribuible.

Se retiraron tres fuentes. La que importaba es Pokorny (CC-BY-NC-ND): el `NC` lo salvaría un servicio gratuito, pero el `ND` prohíbe compartir obra derivada — y nosotros no copiamos el diccionario, lo raspamos, parseamos y enlazamos, que es derivar. Las otras dos, de Vaan y Kroonen, estaban declaradas y nunca se habían ingerido.

El coste, medido antes de borrar recomputando cada afirmación con la fuente y sin ella:

filas retiradas	5.971
formas que pierden su camino al proto-indoeuropeo	5.508 (3,32 %)
conjuntos de cognados sin reconstrucción	2.140
formas sin ninguna etimología	906

Por rama el golpe es desigual: anatolio −30,5 %, itálico −10,3 %, eslavo −6,3 %, germánico −2,1 %. El anatolio pierde casi un tercio de sus enlaces al PIE. Es el precio de poder distribuir, y se paga con conocimiento de causa.

Un invariante `fail` comprueba que ninguna fila de una fuente retirada vuelva a entrar, y está probado por los dos caminos por los que podría entrar (§7.8).

9.4 · El repositorio es la distribución

No se reparte dato en bruto: se clona el repositorio y se reconstruye. CLDF y volcado SQL se generan como salidas del pipeline, no como el producto hospedado.

Tiene tres consecuencias buenas: la superficie de cumplimiento se encoge —las licencias CC se activan al *compartir*, y quien reconstruye obtiene cada fuente de su origen bajo la licencia de su origen—; la procedencia deja de ser una afirmación y pasa a ser ejecutable; y el versionado es el del control de versiones, que sirve tanto para citar como para trabajar.

Con una excepción deliberada: la capa derivada se archiva aparte. Los 3,09 M de esqueletos son cálculo nuestro y son lo que la gente va a querer citar y reutilizar; obligar a levantar una base de datos y correr horas de reconstrucción para obtenerlos detendría a la mayoría de lectores. Los catálogos de los estudios ya funcionan así: ocupan 8,6 MB y viven junto al paper que los explica.

9.5 · El acceso, hoy: por solicitud

El repositorio es privado, y conviene decir por qué, porque la razón es más sencilla de lo que parece y no conviene atribuirle otra. El trabajo está en curso y el acceso se gestiona en vez de dejarse abierto; eso es todo. No es una reserva sobre los datos —la licencia ya permite redistribuirlos— ni una consecuencia de la cuestión de reproducibilidad de §9.6.1, que no se resuelve abriendo o cerrando un repositorio.

Una versión anterior de este apartado sí lo ataba a esa cuestión: decía que un clon público prometería una reproducibilidad inexistente. Ese razonamiento no se sostiene desde que §9.6.1 quedó reformulado, porque la versión de `kaikki` no se va a poder fijar nunca y el repositorio habría quedado privado para siempre por una razón que no era la suya.

Se pide escribiendo a alex@pixelspace.com, diciendo quién se es y para qué. Quien recibe acceso lo tiene de lectura completa y puede bifurcar el repositorio a su propia cuenta, con lo que puede trabajar sobre él sin depender de que nadie le acepte cambios.

La licencia del contenido no cambia por ello: CC-BY-SA-4.0, y ShareAlike se hereda en cualquier trabajo derivado. Lo que está restringido es el acceso al repositorio, no los derechos sobre lo que contiene.

Cuando las versiones de las fuentes estén fijadas, esta restricción deja de tener sentido y el repositorio pasa a ser público. No es una política, es un estado.

9.6 · Tres huecos declarados

Se escriben aquí, y no en una nota al pie, porque afectan a lo que se puede afirmar de la reproducibilidad.

9.6.1 · Ninguna fuente tiene fijada su versión · Reformulado el 2026-09-06. Este apartado decía que las versiones estaban sin fijar y que había que fijarlas. Al ir a hacerlo se vio que para la fuente que más pesa eso ya no es posible, y que el apartado describía mal el problema.

Lo que hace kaikki. Su página publica la fecha del volcado del que salió la extracción —hoy dice *«extracted from the enwiktionary dump dated 2026-08-05»*— y se regenera al menos una vez por semana, sobrescribiendo. No conserva las extracciones anteriores y no publica sumas de comprobación. Y la fecha vive en la página, no dentro del fichero: se comprobó abriendo los JSONL descargados, y cada registro trae `word`, `lang`, `lang_code`, `pos`, `senses` y `forms`, sin un solo campo de fecha, versión o volcado de origen. Descargado el fichero, la versión se ha perdido.

Lo que eso significa aquí, medido. Los 111 ficheros de `kaikki` de los que salió este corpus se descargaron en nueve días distintos, entre el 12 de julio y el 8 de agosto de 2026. Con una regeneración semanal, eso son cuatro o cinco extracciones distintas: el latín y el griego antiguo se bajaron el 6 de agosto, y el afrikáans y el aromanian el 14 de julio, tres semanas y tres volcados antes. Este corpus no está construido sobre una versión de kaikki, sino sobre varias mezcladas, y ninguna de ellas se puede recuperar hoy. Lo mismo, en menor grado, con `lexibank` (dos días separados por un mes) y `glottolog` (dos días).

Lo que se hizo en su lugar. No se puede registrar la versión, pero sí qué ficheros exactos había. `ingest/manifiesto_fuentes.py` recorre los 148 ficheros crudos —5,46 GB— y anota de cada uno su `sha256`, su tamaño y su fecha de descarga, en `docs/MANIFIESTO-FUENTES.md` y su gemelo en JSON. Eso no devuelve la versión perdida; convierte *«no sabemos qué teníamos»* en *«teníamos ESTOS ficheros, y cualquiera puede comprobar si los suyos son los mismos»*. Las fuentes que sí tienen versión legible la llevan anotada: `pyclts` 4.0.2 y `lingpy` 2.6.14. Concepticon no se instala ni se descarga —sus identificadores viajan dentro de los CLDF de Lexibank, IDS, NorthEuraLex e IE-CoR—, así que su versión es la de cada conjunto de origen.

Y el deslinde, que es la parte que faltaba. Kaikki no es nuestro. Nosotros no lo controlamos, no lo versionamos y no podemos prometer lo que sirve mañana. Quien quiera reconstruir el corpus que descargue kaikki y lo use: obtendrá el kaikki de ese día, hará sus pruebas y sacará sus cifras, y eso está bien — es su ejecución, no la nuestra. Lo que este trabajo garantiza es lo que sí es suyo: que el pipeline es determinista, que dados los mismos ficheros de entrada produce el mismo corpus, y que los ficheros de entrada están identificados uno a uno por su huella. Lo que no garantiza —y ningún proyecto que integre datos de terceros en movimiento puede garantizar— es que el tercero le sirva a usted los mismos ficheros. Fingir lo contrario sería el error; declararlo no cuesta nada y es lo exacto.

9.6.2 · Dos fuentes con licencia «open» · CERRADO el 2026-09-06. `europarl` y `opensubtitles` estaban declaradas con licencia «open», que no es una licencia. Resultó que ninguna es fuente de datos: OpenSubtitles nunca aportó texto —solo recuentos— y las 22.281 oraciones de Europarl vivían en una tabla de estudio, `construction`, que se retiró del corpus. Las dos quedan como medición y no hay licencia de terceros que cumplir, porque no se conserva contenido de terceros. Detalle en §12.1.

9.6.3 · 13,3 millones de filas sin columna de procedencia · Tatoeba retirada · CERRADO el 2026-09-06. Este apartado decía antes que `tat_sentence` (5.799.836 filas) y `tat_link` (7.516.808) eran un punto ciego del invariante de licencia por carecer de `source_id`. La razón estaba mal puesta. La licencia de Tatoeba se conoce y es CC BY 2.0 FR, que permite redistribuir con atribución: el hueco

era del aparato y no de los datos, y se habría cerrado añadiendo una columna.

La razón por la que salen es de otro orden, y es de objeto. Este corpus trata de formas, y la cadena que lo sostiene es forma → segmento → esqueleto → código; una frase no es una forma. Las dos tablas no tenían una sola clave ajena hacia `form`, `concept`, `lect` o `source`; ninguna forma declaraba `source_id='tatoeba'`; y ningún proceso del corpus las leía. Eran una capa huérfana de 13,3 M de filas y cerca de un gigabyte por la que no entraba nada y de la que no salía nada.

Lo único construido a partir de ellas fueron `construction` y `construction_align` (24.267 filas), salida de un estudio de campos cuálicos que ya había abandonado el corpus por la vía de §9.6.2. Las cuatro tablas viven ahora en la base `qualicfields`, que es el proyecto cuyo objeto sí es el uso. El extractor que las consulta no cruzaba en ningún punto contra `form` ni contra `lect`, de modo que la separación no le quita nada; las tablas `qf_*`, que sí sostienen catorce claves ajenas hacia el corpus, se quedaron donde estaban.

Conviene decir con cifras el alcance que se pierde, porque es corto. Tatoeba cubría doce lenguas —eng, ita, deu, fra, por, spa, nld, dan, swe, ron, nob y cat—, todas del mismo macrosistema, uno de los 225 del corpus, y dentro de él la esquina de Europa occidental. El contexto de uso que aportaba caía justo donde el corpus ya es más denso, y no alcanzaba a ninguno de los otros 224.

9.7 · Lo que hay que arreglar ANTES de congelar la versión que acompañe a este trabajo

Los apartados anteriores declaran huecos, y declararlos no es arreglarlos. Un revisor puede contestar con razón: *«muy bien; arréglalo antes de publicar el recurso»*. Esta es la lista, y este trabajo no se envía con ninguna de las cuatro primeras sin cerrar:

	qué	por qué bloquea
1	Fijar versión y suma de comprobación de cada fuente · RESUELTO DE OTRO MODO el 2026-09-06	Fijar la versión de `kaikki` ya no es posible: la fecha vive en su página y no en el fichero, y las extracciones anteriores no se conservan. En su lugar hay huella `sha256` de los 148 ficheros crudos y el deslinde de §9.6.1: el pipeline es nuestro y es determinista; el tercero no lo es y no se promete por él
2	Identificar la licencia de `europarl` y `opensubtitles` · HECHO el 2026-09-06	No eran fuentes de datos; el texto de Europarl se retiró y las dos quedan como medición (§12.1)
3	Dar procedencia a `tat_sentence` y `tat_link`, o sacarlas de la base publicada · HECHO el 2026-09-06	Salieron: 13,3 M de filas de un objeto que el corpus no lee. Viven en la base `qualicfields` (§9.6.3)
4	Recuperar los clics en el esqueleto · HECHO el 2026-09-06	488 de 491 formas lo conservan y ninguna se queda sin esqueleto (§4.4)
5	Correr de nuevo la pasada de `skeleton_raw` · HECHO el 2026-09-06, y no cerró el hueco	Se corrió y escribió 92 filas: 57.132 → 57.040. El hueco no era retraso de la pasada. Sus tres causas reales están ahora en §5.1, y una de ellas —10.260 formas de esqueleto solo semiconsonántico— queda declarada como divergencia entre capas, no arreglada
6	Mover a su repositorio las tablas que son salida de estudios · RESUELTO DE OTRO MODO el 2026-09-06	La premisa era falsa: `codigos` no lee NINGUNA de las 25 movibles y Meulemans lee 17. No son salida a la espera de mudarse, son lo que sirve la capa de consulta (§5.4). Lo que sí era salida —`construction`, `construction_align`— ya salió. Y salieron además las 8 copias de seguridad que vivían dentro de la base: 1,5 M de filas y 190 MB

	qué	por qué bloquea
7	Retirar las 2 formas residuales de `wiktionary` · HECHO el 2026-09-06	No se borraron: una era el padre etimológico del inglés *angst*, así que sus 4 aristas se repuntaron a la gemela de `kaikki` antes de retirarlas (§12.1)

Las cuatro primeras cambian lo que el trabajo puede afirmar. Las tres últimas no, y por eso van después.

Las cuatro que bloquean están cerradas al 6 de septiembre de 2026, aunque la primera no como se había planteado. Se planteó como una tarea de inventario —recorrer las dieciocho fuentes anotando versiones— y resultó ser una cuestión de alcance: para la fuente que más pesa la versión ya no existe en ninguna parte, y lo que había que arreglar no era el inventario sino la promesa. Está en §9.6.1.

10 · Límites

§6.7 enumera las preguntas que el corpus no puede responder. Ésta recoge los límites del recurso como tal: lo que no arregla ninguna cantidad de trabajo adicional sobre lo que ya hay.

10.1 · Hereda lo que reciba

Es agregación (§2), y por tanto hereda los errores de sus fuentes sin poder corregirlos en general. Cuando un error se detecta se declara —hay una lengua kam-sui clasificada como urálica en origen, y está escrita como hecho declarado (§3.6)— pero detectarlo requirió que tres chequeos independientes coincidieran, y no hay razón para creer que ése sea el único.

Un corpus integrado no es más fiable que sus fuentes. Lo que puede ser es más auditable, que es cosa distinta y es a lo que aspira.

10.2 · La reproducibilidad llega hasta donde llega lo nuestro

Hay que separar dos cosas que suelen decirse juntas. El pipeline es nuestro y es determinista: dados los mismos ficheros de entrada produce el mismo corpus, y esos ficheros están identificados uno a uno por su `sha256` en `docs/MANIFIESTO-FUENTES.md`. Cualquiera puede comprobar si los suyos son los mismos.

Las fuentes no son nuestras. `kaikki` —el 38,6 % de las formas— se regenera al menos una vez por semana sobrescribiendo, y no conserva sus extracciones anteriores; quien clone y reconstruya obtendrá el kaikki del día en que lo haga. Eso no es un defecto que haya que resolver: es lo que ocurre cuando se integran datos de terceros en movimiento, y este trabajo no promete por terceros. Quien quiera usar kaikki que lo use, haga sus pruebas y saque sus cifras — será su ejecución, con sus datos, y estará bien.

Lo que sí conviene decir es lo que se sigue de ahí: una cifra publicada debe citar el archivo congelado que la produjo, no «la reconstrucción», porque dos reconstrucciones en semanas distintas no son la misma. El límite existe y se declara; lo que no existe es la obligación de hacerse cargo de él. Ver §9.6.1.

10.3 · La capa histórica es de un solo macrosistema

Etimología, cognación y protoformas existen para el indoeuropeo y para nada más. §6.4 da las cifras y lo que implica para quien trabaje sobre el corpus.

Aquí interesa acotar de qué es un límite. No es que no exista cognación experta fuera del indoeuropeo —existe, y §6.4 nombra algunos conjuntos—: es que no está integrada aquí, y cada integración cuesta trabajo de mapeo que no se ha hecho.

Lo que sí es un límite del recurso, y no de esta versión, es que un corpus puede reunir los juicios que existen pero no producir los que no: donde una familia no tiene tradición comparativa publicada, ninguna agregación la sustituye.

10.4 · La normalización pierde por diseño

El esqueleto consonántico agrupa las róticas, las nasales y las dorsales en una consonante canónica por valor cuálico (§4.3). Eso es lo que hace comparables dos lenguas y es también una pérdida real de información fonética.

Se mitiga conservando `ipa_raw`, `segments_raw` y `skeleton_raw` intactos, de modo que la normalización sea una capa añadida y reversible. Pero quien trabaje sobre `cons_skeleton` trabaja sobre una representación deliberadamente empobrecida, y las preguntas que dependan de las distinciones agrupadas no se pueden hacer ahí.

10.5 · El aparato tiene puntos ciegos conocidos

El alcance del invariante de licencia tiene borde, y está escrito. Trece tablas no tienen columna de procedencia ni columna de enlace por la que heredarla, y dos de ellas —`concept`, de Concepticon, y `feature`, de CLTS— sí conservan contenido de terceros (§3.3). No hay problema de licencia, las dos son CC-BY-4.0, pero el invariante no llega ahí y un meta-chequeo falla si aparece una tabla más sin declarar (§7.4).

Tres columnas del esquema están declaradas y sin poblar (§6.5).

Tres deudas declaradas con techo: 45.000 filas de descuadre entre plantilla y segmentos, 10.260 formas donde `skeleton` y `skeleton_raw` discrepan sobre si la semiconsonante es esqueleto, y 63 formas cuyo esqueleto lleva dos transcripciones concatenadas — ésta con el código ya arreglado (§7.6).

Cinco avisos que no son fallos pero tampoco son nada, enumerados uno a uno en §7.5 con lo que resultó ser cada uno: cuatro son propiedades de integrar fuentes que se solapan, y el quinto destapó 63 formas cuyo esqueleto lleva dos transcripciones concatenadas.

Todos están escritos porque un chequeo que no cubre algo y no lo dice produce confianza falsa, que es la tesis de §8. Estarán escritos hasta que se resuelvan.

10.6 · No es una fuente primaria, ni pretende serlo

No sustituye a ninguno de los recursos de §2, no documenta lenguas nuevas y no arbitra cuestiones lingüísticas. Ofrece una unión auditable de material ajeno, una capa calculada encima, y las herramientas para comprobar que ambas cosas dicen lo que dicen.

11 · Cómo se usa y cómo se cita

11.1 · Obtenerlo

El repositorio es la distribución (§9.4), y el acceso se pide (§9.5): se escribe a alex@pixelspace.com diciendo quién se es y para qué. Con acceso, se clona y se reconstruye — y se puede bifurcar a la

cuenta propia:

```
git clone https://github.com/toledoal/corpus-integrativo
cd corpus-integrativo
./db/setup_dev.sh          # el clúster del proyecto
./db/rebuild.sh --scope smoke   # una familia pequeña, minutos
./db/rebuild.sh --scope full    # todo, horas
./verifica.sh              # las cinco clases de comprobación
```

La reconstrucción requiere que las fuentes sigan disponibles, y con la salvedad de §9.6.1: le dará el kaikki del día en que la corra, que no es el que produjo las cifras de este trabajo. Para comprobar si sus ficheros coinciden con los nuestros está `docs/MANIFIESTO-FUENTES.md`, con el `sha256` de los 148. Si no coinciden, no hay nada roto: tiene usted otro corpus, hecho con el mismo pipeline sobre otra entrada.

Quien solo quiera la capa derivada —los esqueletos consonánticos— no necesita nada de esto: va archivada aparte, por la razón de §9.4.

11.2 · Dos maneras de leerlo, y ninguna es la buena

El corpus no tiene un consumidor canónico, y eso condiciona cómo describirlo (§6.1). Los dos que existen hoy piden cosas distintas del mismo dato:

Una capa de consulta de la historia de una palabra. Busca una forma y muestra su sonido, su sentido, sus cognados y su linaje. Le sirven las 3,79 M de formas, y las 1,24 M de Kaikki sin concepto asignado no son peso muerto sino su materia prima. Muestra las reconstrucciones, porque enseñar el linaje es lo que hace.

Un catálogo de correspondencias forma–sentido por macrosistema. Necesita esqueleto y concepto a la vez, porque su unidad es concepto × lengua; ahí lo utilizable son 2,06 M. Excluye las reconstrucciones, porque su diseño es no mirarlas.

Las dos cifras son correctas y las dos políticas son correctas. Por eso el corpus declara hechos sobre sus lectos y no políticas (§3.6), y por eso esta descripción da las capas y su cobertura en vez de una cifra única de «corpus utilizable».

11.3 · Cómo citarlo

Todo número publicado debe citar una versión archivada, no la reconstrucción — por la razón de §9.6.1: dos reconstrucciones hechas en semanas distintas no parten de la misma entrada. La cita del corpus va acompañada de la de las fuentes que efectivamente se usaron: la tabla `source` lleva la cita y la licencia de cada una, y es la que hay que consultar para componer la atribución de un trabajo concreto.

La licencia del conjunto es CC-BY-SA-4.0 y se hereda: cualquier trabajo derivado, incluidos los catálogos que se construyan sobre él, sale con la misma.

11.4 · Contribuir

Lo que este corpus pide de quien quiera añadir algo es lo que se ha descrito en §7:

- Una fuente nueva entra con su licencia con nombre. «Open» no es una licencia (§9.5).

- Una decisión se declara donde se pueda consultar, con su motivo y su cifra. Una ausencia sin declarar es indistinguible de un descuido.
- Un chequeo nuevo dice qué NO cubre, y se prueba que dispara introduciendo el fallo que debe detectar. Un chequeo que nunca ha disparado no está probado.
- **`verifica.sh`** pasa antes de cada cambio. Devuelve un solo código de salida y tiene un modo de cinco segundos sin base de datos.

12 · Apéndice · Las dieciocho fuentes

El texto habla de dieciocho fuentes sin darlas nunca todas. Aquí están, con lo que aporta cada una al corpus construido y su licencia. La columna de formas es lo que aporta a la tabla **`form`**: una fuente con cero formas no es una fuente vacía — Glottolog aporta la clasificación de los lectos y Concepticon la identidad de los conceptos, sin aportar una sola palabra.

#	id	qué es	licencia	formas
1	`lexibank`	listas léxicas normalizadas en CLDF/CLTS	CC-BY-4.0	1.740.092
2	`kaikki`	Wiktionary inglés extraído con wiktextract	CC-BY-SA-3.0	1.457.643
3	`ids`	*Intercontinental Dictionary Series*, vía CLDF	CC-BY-4.0	437.902
4	`nel`	NorthEuraLex, vía CLDF	CC-BY-4.0	121.611
5	`iecor`	IE-CoR · juicios de cognación de experto	CC-BY-4.0	25.731
6	`kaikki-descendants`	secciones «Descendants» de Wiktionary	CC-BY-SA-3.0	5.315
7	`wiktionary`	ingesta anterior de Wiktionary; hoy aporta solo el registro de 3 lectos	CC-BY-SA-3.0	0 · sus 2 formas se retiraron el 2026-09-06
8	`kaikki-tree`	cadenas etimológicas de Wiktionary	CC-BY-SA-3.0	0 · 383.234 etimologías
9	`kaikki-prose`	etimología en prosa de Wiktionary	CC-BY-SA-3.0	0 · 37.994 etimologías
10	`kaikki-latin-etym`	plantillas de etimología latina	CC-BY-SA-3.0	0 · 2.007 etimologías
11	`glottolog`	glotocódigos y clasificación	CC-BY-4.0	0 · identidad de 4.735 lectos
12	`concepticon`	catálogo de conceptos	CC-BY-4.0	0 · identidad de 3.205 conceptos
13	`clts`	estándar de transcripción y rasgos	CC-BY-4.0	0 · rasgos de 18.977.285 segmentos
14	`lingpy`	modelos de clase de sonido (Dolgopolsky, SCA)	GPL-3.0	0 · 3.516 clases
15	`liv`	LIV², verbos indoeuropeos reconstruidos	CC-BY-SA-4.0	0 · 143 protoformas
16	`tatoeba`	frases y alineaciones de la comunidad Tatoeba	CC BY 2.0 FR	0 · retirada el 2026-09-06: 13,3 M de filas de texto corrido, que no es el objeto de una base de formas (§9.6.3)
17	`europarl`	actas del Parlamento Europeo, vía OPUS	no aplica · medición	0 · ver §12.1

#	id	qué es	licencia	formas
18	`opensubtitles`	subtítulos alineados, vía OPUS	no aplica · medición	0 · ver §12.1

12.1 · Dos de las dieciocho no son fuentes de datos

`europarl` y `opensubtitles` estaban declaradas con licencia «open», que no es una licencia. Al mirar qué aportaban resultó que la pregunta estaba mal planteada: no son fuentes de las que se ingiera nada.

OpenSubtitles nunca aportó texto. Solo recuentos de frecuencia —hechos, no expresión protegible— y hallazgos redactados por este proyecto que la citan como evidencia.

Europarl sí aportaba: 22.281 oraciones del Parlamento, en la tabla `construction`. Era una tabla de estudio del proyecto de campos cuálicos —ningún paper de familia la usaba y la capa de consulta no la consultaba— y se retiró del corpus el 2026-09-06, junto con `construction_align`. El respaldo quedó fuera del repositorio.

Las dos quedan como medición: se citan porque contra ellas se midió, no porque se conserve nada suyo. Y eso resuelve la licencia sin tener que verificarla, que era el hueco de §9.6.2: no se redistribuye contenido de terceros, así que no hay licencia de terceros que cumplir.

`wiktionary` (#7) llevaba dos formas de una ingesta anterior —el alemán *Angst* y el español *angostura*— que se retiraron el 6 de septiembre de 2026. La retirada no fue un borrado, y el rodeo merece contarse porque es el mismo error que §8 describe. Las dos parecían residuo inocente: sin concepto, sin esqueleto, sin segmentos, y con una gemela en `kaikki` que sí tenía las tres cosas. Pero al borrarlas la base se negó: **`de·Angst·N·001`** era el padre etimológico del inglés *angst*, con cuatro aristas, y su gemela de `kaikki` no tenía ninguna. Borrar habría perdido esa etimología en silencio. Lo que se hizo fue repuntar las cuatro aristas a la gemela —que es la forma que sí puede llevar esqueleto y código— y retirar después las formas. Los cuatro conjuntos de cognados afectados perdieron un miembro cada uno, que era el duplicado: la gemela ya estaba dentro de los cuatro. La fuente sigue en el catálogo porque aún aporta el registro de tres lectos (alemán, español, alto alemán antiguo), y ésa es procedencia de `lect`, no de `form` (§3.3).

13 · Referencias

13.1 · Los otros dos trabajos de esta serie

Los dos se construyen sobre este corpus y ninguno de los tres se sostenía citando a los otros, cosa que esta versión corrige. Son los dos consumidores de §6.1, y su desacuerdo sobre qué parte del corpus es utilizable no es un problema: es el argumento de esa sección.

- Toledo Martínez, A. (2026). *Meulemans: una capa de consulta de solo lectura sobre un corpus léxico multi-capa de 225 macrosistemas.* — Enseña el linaje de una palabra. Es de solo lectura: no escribe ni calcula nada aquí.
- Toledo Martínez, A. (2026). *El catálogo de códigos consonánticos: un método comparativo sin genealogía, sin cognación y sin reconstrucción.* — Lee la capa derivada de §5 y se prohíbe mirar la capa histórica de §3.2. La columna `code` que este corpus publica sin afirmar nada sobre ella (§5.1) es justamente el instrumento que aquel trabajo estudia.

13.2 · Fuentes y estándares

Sobre las versiones. Las referencias siguientes identifican la obra, no la versión concreta que este corpus ingirió, y ninguna puede llevar número de versión sin mentir. Para `kaikki` eso es definitivo: la versión no está en el fichero y las extracciones anteriores no se conservan, de modo que no existe número que citar (§9.6.1). Para las demás es recuperable, y quien lo necesite tiene por dónde: `docs/MANIFIESTO-FUENTES.md` da el `sha256`, el tamaño y la fecha de descarga de los 148 ficheros crudos, que identifica el artefacto ingerido aunque no lo nombre con una versión.

- 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), 2.ª ed. Wiesbaden: Reichert. Versión enlazada (LLOD) de 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 y OpenSubtitles se obtuvieron por esta vía.)
- Tatoeba Project. <https://tatoeba.org>

Y una referencia que ya no está en el corpus, citada porque §9.3 mide lo que costó retirarla:

- Pokorny, J. (1959). *Indogermanisches etymologisches Wörterbuch*. Berna: Francke. Digitalización de Starostin (StarLing). Retirada del corpus el 2026-09-05 por su licencia CC-BY-NC-ND.