Meulemans

Una capa de consulta de solo lectura sobre un corpus léxico multi-capa de 225 macrosistemas

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

Resumen

Un corpus de 3,79 millones de formas no se consulta: se le hacen preguntas. Y las que un lingüista hace de una palabra —de dónde viene, qué otras la acompañan, cómo suena en sus parientes— cruzan capas que en la base están separadas, de modo que responderlas exige escribir SQL contra ochenta y siete tablas y conocer sus trampas.

Meulemans es la capa que hace consultable el Corpus Integrativo: una API de solo lectura y una interfaz web, 3.006 líneas entre las dos, con siete vistas que responden siete preguntas distintas sobre el mismo dato. No añade datos ni los calcula: los enseña, con la fuente de cada uno a la vista.

Se describen tres decisiones que dieron forma al proyecto y que se defienden como transferibles: es de solo lectura, porque el aparato de verificación del corpus solo sirve si hay un único camino de escritura; no traduce los datos aunque la interfaz sea bilingüe, porque una glosa no es la versión inglesa de un concepto sino su identificador; y enseña la procedencia junto al dato, lo que además de cumplir las licencias CC hace visible la desigualdad del corpus en vez de disimularla.

Se dice también lo que la herramienta no es. Corre en local y no está desplegada; enseña un corpus cuya capa histórica cubre un macrosistema de 225, de modo que es un navegador etimológico indoeuropeo con un catálogo léxico muy ancho alrededor; y no valida lo que muestra, porque la validación es responsabilidad del corpus y una capa de consulta que «arreglara» lo que enseña introduciría una segunda verdad.

1 · El problema

Un corpus léxico de 3,79 millones de formas no se consulta: se le hacen preguntas. Y las preguntas que un lingüista hace de una palabra —¿de dónde viene? ¿qué otras palabras la acompañan? ¿cómo suena en sus parientes?— no caben en una tabla, porque cruzan capas que en la base están separadas: la forma, sus sonidos, sus sentidos, su etimología, sus cognados, su esqueleto consonántico.

Responderlas hoy exige escribir SQL contra ochenta y siete tablas y conocer sus trampas —empezando por las dos columnas `source_id` que significan cosas distintas—. Eso pone el corpus fuera del alcance de quien no lo construyó, que es la mayoría de la gente a la que podría servirle.

Meulemans es la capa que hace consultable el Corpus Integrativo. No añade datos ni los calcula: los enseña.

1.1 · El nombre

Honra a la Dra. Christiane S. Meulemans y al Dr. José Ángel Elías Fernández y Cuevas, de la Fundación Dr. J. Meulemans, creadores de la endolingüística. La herramienta no genera datos primarios:

integra fuentes existentes, cada una citada y con su licencia, para consultarlas.

1.2 · Qué es y qué no es

Es un navegador de la historia de una palabra: se busca una forma y se ve su sonido, su sentido, sus cognados, su colexificación y su linaje hasta donde llegue, con sus fuentes.

No es una herramienta de análisis. No calcula códigos, no construye catálogos, no propone reconstrucciones y no escribe en el corpus. Es de solo lectura por diseño, y §4.1 explica por qué esa restricción es la decisión de arquitectura más importante del proyecto.

No es tampoco un sustituto de la base. Quien vaya a hacer investigación sobre el corpus trabaja contra la base directamente; Meulemans sirve para mirar, comprobar y enseñar.

2 · Las siete vistas

Cada una responde a una pregunta distinta sobre el mismo dato. La lista no es un menú de funciones: es el conjunto de preguntas que el corpus puede contestar hoy.

vista	la pregunta que responde
Buscar	¿qué sé de esta palabra? Es la entrada: una forma, y de ahí su ficha completa
Comparar	¿cómo dicen esto otras lenguas? Un concepto a través de las lenguas que lo atestiguan
Coderivados	¿qué otras palabras comparten esqueleto con ésta, y para qué conceptos?
Genealogía	¿de dónde viene? El linaje de la forma, arista a arista, con la fuente de cada una
Racimos	¿qué familia de palabras forma? Agrupación por raíz y tema
Clases	¿qué sonidos usa esta lengua, y cómo se reparten? Inventario y matriz de clases
Mapa	¿dónde se habla? Distribución geográfica de lo que hay en pantalla

Y dos secundarias que no son accesorias: Acerca de, que dice qué es esto, y Legal, que lista las dieciocho fuentes con su licencia y su cita. La segunda no es cortesía: es el cumplimiento de las licencias CC bajo las que se distribuye el corpus, y por eso es una vista y no una nota al pie.

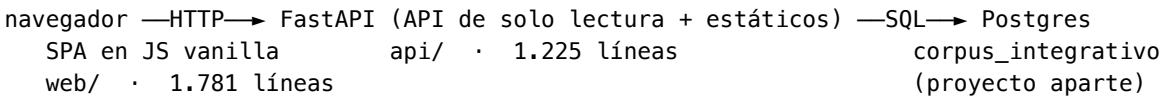
2.1 · La ficha de una forma

Es la vista central y merece describirse. Reúne, para una sola palabra:

- la grafía y la transcripción tal como las escribió la fuente, sin alterar;
- sus segmentos, uno a uno, con sus rasgos articulatorios;
- su esqueleto consonántico y su capa de clases;
- sus sentidos, con los conceptos a los que están mapeados;
- su linaje, hasta donde alcance la capa etimológica;
- sus cognados, cuando existen;
- y la fuente de cada una de esas cosas.

La última línea es el punto. Cada dato que la ficha enseña lleva de dónde salió, porque el corpus lleva procedencia por fila y sería un desperdicio no mostrarla.

3 · Arquitectura



pieza	tamaño	qué es
`api/queries.py`	783 líneas	las consultas, una por endpoint
`api/main.py`	253	el enrutado y los estáticos
`api/auth.py`	162	sesiones, contra una base propia y minúscula
`api/skeleton.py` · `db.py`	74	el esqueleto para mostrar, y el pool de conexiones
`web/js/`	1.781	la SPA: enrutador, render, i18n y una página por vista
`web/style.css`	399	todo el estilo

28 endpoints, todos `GET` salvo el de sesión.

3.1 · Sin paso de compilación

La SPA es JavaScript de módulos nativos servido tal cual. No hay bundler, no hay transpilación, no hay `node_modules`. Se edita un fichero y se recarga.

Es una decisión y tiene un coste —no hay comprobación de tipos ni árbol de dependencias— y una ganancia que en un proyecto de investigación pesa más: el código que se lee en el repositorio es exactamente el que corre en el navegador. No hay una capa de construcción entre lo que uno escribe y lo que ocurre, y por tanto no hay una clase entera de fallos donde el artefacto desplegado difiere de la fuente.

3.2 · Enrutado por hash, y todo el estado en la URL

Cada estado —la búsqueda, los filtros, la selección— va en el hash. De ahí salen dos cosas gratis: los enlaces profundos funcionan (se puede compartir exactamente lo que se está viendo) y el botón de atrás del navegador hace lo que debe.

Es lo que convierte la herramienta en algo citable: una afirmación sobre una palabra puede acompañarse del enlace que la enseña.

3.3 · Dos bases con requisitos opuestos

```
corpus_integrativo   SOLO LECTURA · grande · reemplazable
meulemans_ops        lectura/escritura · 7,9 MB · IRREEMPLAZABLE
```

La distinción importa porque induce el error caro: pagar respaldo y alta disponibilidad por gigabytes de datos que ya están duplicados en la máquina de origen y se pueden regenerar, y no respaldar los ocho megabytes de cuentas y registro de uso, que son lo único que no se puede recuperar.

4 · Tres reglas que dieron forma al proyecto

4.1 · Solo lectura, y no es una limitación

Meulemans no escribe en el corpus. Ni una tabla, ni una columna, ni un contador.

La razón no es prudencia sino separación de responsabilidades. El corpus tiene su propio aparato de verificación —invariantes que fallan, contrato de esquema, una puerta única— y ese aparato solo sirve si hay un único camino de escritura. Una capa de consulta que escribiera, aunque fuese un registro de accesos, abriría un segundo camino que los invariantes no vigilan.

De ahí que las cuentas y el uso vivan en una base aparte. No es duplicación: es mantener el corpus como un artefacto inmutable que se puede regenerar entero desde su pipeline sin perder nada.

4.2 · Los datos no se traducen

La interfaz es bilingüe, inglés por defecto. Y hay una regla que no se negocia: se traduce el cromo, nunca el dato.

Rótulos, explicaciones, estados y títulos de columna se traducen. No se traducen las palabras, la transcripción IPA, los esqueletos, los códigos, las glosas, los nombres de lengua, familia o rama, los identificadores de fuente ni las clases de Dolgopolsky o SCA.

Vienen de las fuentes o son neutrales, y traducirlos sería falsearlos. Una glosa inglesa de Concepticon no es «la versión inglesa» de un concepto: es el identificador de ese concepto, y traducirlo al español produciría un dato que ninguna fuente respalda.

4.3 · Cada dato lleva su fuente a la vista

El corpus lleva procedencia por fila porque las licencias CC exigen atribución (§9 del paper del corpus). Meulemans la muestra, y no como metadato escondido: la fuente aparece junto al dato.

Tiene un efecto secundario más útil que el cumplimiento legal: hace visible la desigualdad del corpus. Quien mira una ficha de una lengua indoeuropea ve etimología y cognados; quien mira una de otro macrosistema ve que esas secciones están vacías, y ve por qué. La carencia se enseña en vez de disimularse.

5 · Estado, y lo que costaría desplegarlo

Conviene decirlo sin rodeos: Meulemans corre hoy en local y no está desplegado. Un paper que describiera una herramienta pública sin serlo estaría mintiendo, así que esta sección dice dónde está y qué falta.

5.1 · Lo que hay

Funciona: las siete vistas, los 28 endpoints, el bilingüismo, los enlaces profundos y las sesiones. Se levanta con un comando contra la base local del corpus.

Lo que falta no es funcionalidad sino el pulido que separa una herramienta de trabajo de una herramienta pública: manejo de errores en los bordes, límites de tasa, y una revisión de qué se enseña a un visitante que no sabe qué es un esqueleto consonántico.

5.2 · Lo que se midió del despliegue

Las cifras son medidas sobre el clúster local, no estimadas:

la base que sirve	8,8 GB tras compactar (17 GB sin compactar)
las cuentas y el uso	7,9 MB — lo único irreemplazable
tablas que Meulemans consulta	`form`, `segment`, `skeleton`, `sense`, `racimo`, `form_etymology`, `concept`, `cognate_member`, `morph_rule`, `sound_class`

Los 8 GB de diferencia entre las dos cifras no son contenido: son espacio muerto. La tabla `form` ocupa 2.242 MB frente a 1.311 MB con exactamente las mismas 3.788.296 filas — hinchazón acumulada por las reconstrucciones. Es la clase de dato que solo aparece al preparar un despliegue y que conviene medir antes de pagar por él.

5.3 · Una corrección: el export público ya no hace falta por contenido

El documento de despliegue describe dos bases con contenidos distintos: la de investigación y una pública producida filtrando las fuentes no redistribuibles. Eso ha dejado de ser cierto.

Desde el 2026-09-05 la base principal del corpus es la distribuible: las fuentes no redistribuibles se retiraron de ella, con su coste medido y publicado. La diferencia de etimologías que aquel documento registraba —1.266.185 frente a 1.262.354— es exactamente esa retirada, y ya está aplicada.

De modo que el export público conserva una sola razón de ser, y no es la que decía: compactar. No filtra nada porque ya no hay nada que filtrar.

Queda anotado aquí porque es el tipo de documento que envejece en silencio: describía correctamente un estado del mundo que cambió en otro repositorio.

5.4 · Dónde está el código, y cómo se pide

El repositorio es privado y el acceso se da por solicitud, escribiendo a alex@pixelspace.com con quién se es y para qué. Quien lo recibe tiene lectura completa y puede bifurcarlo a su cuenta.

Es el mismo criterio con el que se dan las cuentas de la propia aplicación (§4.1) y por el mismo motivo: mientras la herramienta está madurando, cada acceso se da sabiendo quién lo pide, de modo que se pueda seguir cambiando la API sin romperle el trabajo a nadie sin avisar.

El corpus del que lee se distribuye bajo CC-BY-SA-4.0; el acceso al repositorio y la licencia del contenido son cosas distintas.

6 · Límites

6.1 · Enseña lo que el corpus tiene, con sus desigualdades

Meulemans no puede mostrar lo que no existe. La ficha de una palabra indoeuropea trae etimología, cognados y linaje; la de una palabra de otro macrosistema trae forma, sonido y sentido, y las demás secciones vacías — porque la capa histórica del corpus cubre un macrosistema de 225.

No es un fallo de la herramienta y no se disimula: se enseña, y se dice por qué. Pero quien llegue esperando un navegador etimológico universal encontrará uno indoeuropeo con un catálogo léxico muy ancho alrededor.

6.2 · No valida lo que enseña

La herramienta muestra lo que la base contiene, incluidas sus heurísticas. Dos columnas del corpus aparentan ser datos y son conjeturas —el detector de nombres propios por mayúscula inicial y la marca de préstamo heredada de cada fuente— y Meulemans no las corrige: como mucho puede señalarlas.

La validación es responsabilidad del corpus y de su aparato de invariantes, no de la capa de consulta. Una herramienta de visualización que «arreglara» lo que enseña sería peor: introduciría una segunda verdad.

6.3 · No es reproducible por sí sola

Depende del corpus, y el corpus tiene hoy un límite de reproducibilidad declarado: ninguna de sus dieciocho fuentes lleva versión fijada. Dos instalaciones de Meulemans contra dos reconstrucciones hechas en momentos distintos pueden enseñar cosas distintas, y hoy nada lo advierte.

Cuando el corpus fije versiones, Meulemans debería mostrar cuál está sirviendo. No lo hace todavía.

6.4 · Sin build, sin tipos

La decisión de §3.1 tiene su factura: no hay comprobación estática de nada. En 1.781 líneas de JavaScript es manejable; a partir de cierto tamaño dejaría de serlo, y la decisión habría que revisarla en vez de defenderla.

7 · Referencias

Meulemans no genera datos: los enseña. Eso hace que su aparato de referencias tenga dos partes que no conviene mezclar — de dónde sale lo que muestra, y sobre qué está construido lo que muestra.

7.1 · La procedencia de lo que se enseña

Ninguna de las palabras, glosas, etimologías ni cognados que aparecen en la interfaz es nuestra. Todas vienen de las dieciocho fuentes que integra el Corpus Integrativo, cada una con su cita y su licencia, y la lista completa no se reproduce aquí a propósito: duplicarla haría que las dos versiones se desincronizaran el día que entre o salga una fuente. Vive en la tabla `source` del corpus, que es la que se consulta para componer la atribución de un trabajo concreto, y se enumera en el §12 del paper del corpus.

Lo que sí conviene decir aquí es la consecuencia práctica: quien cite una pantalla de Meulemans está citando datos de terceros, y la atribución que le corresponde es la de esas fuentes, no la de esta capa.

7.2 · Los otros dos trabajos de esta serie

- Toledo Martínez, A. (2026). *Corpus Integrativo: una base léxica multi-capa de 225 macrosistemas, con procedencia por fila y su aparato de verificación.* — La base que Meulemans consulta. Meulemans es de solo lectura: no escribe ni modifica nada de lo que hay ahí.
- Toledo Martínez, A. (2026). *El catálogo de códigos consonánticos: un método comparativo sin genealogía, sin cognación y sin reconstrucción.* — El otro consumidor del mismo corpus, y el

contraste que explica por qué no hay una sola cifra de «corpus utilizable»: aquel trabajo se prohíbe mirar el linaje, y éste lo enseña porque enseñarlo es lo que hace.

7.3 · Lo que Meulemans usa como herramienta

- PostgreSQL Global Development Group. *PostgreSQL*. <https://www.postgresql.org>
- Ramírez, S. *FastAPI*. <https://fastapi.tiangolo.com>