

Un corpus común para el bribri

Plan de trabajo para octubre de 2026 a mayo de 2027, y líneas que pueden continuar después

Ariel Arévalo Alvarado

Ciencias de la Computación e Informática, Universidad de Costa Rica

Versión de octubre de 2026

RESUMEN. En la última década, investigadores y proyectos de la Universidad de Costa Rica publicaron una cantidad considerable de material sobre el bribri: un corpus oral anotado de tres dialectos con su audio, diccionarios pictográficos con grabaciones, secuencias didácticas, libros y un diccionario general. Ese material está hoy en sitios y formatos distintos, y la anotación del corpus oral se publica solo como tablas en páginas web. Este documento propone un plan de 280 horas, del 15 de octubre de 2026 al 19 de mayo de 2027, con dos productos. El primero reúne el corpus oral y los materiales publicados en un formato abierto común, con un visor privado para consultarlos. El segundo usa ese corpus para medir qué parte del vocabulario del habla cubren las secuencias didácticas y los diccionarios. El plan no requiere trabajo de campo, anotación nueva ni presupuesto. La última sección describe las líneas que pueden continuar otros estudiantes.

- | | |
|--------------------------------------|------------------------|
| 1 Situación actual | 5 Qué se necesita |
| 2 Qué se propone | 6 Riesgos |
| 3 Calendario y horas | 7 Después de este plan |
| 4 Qué se entrega y cómo se comprueba | 8 Referencias |

1 Situación actual

TRES FUENTES concentran el material de este plan. El *Corpus oral pandialectal de la lengua bribri* reúne 35 textos transcritos de los dialectos de Amubri, Coroma y Salitre, con audio, análisis morfológico y traducción al español, y su proyecto contó con el apoyo de la Universidad.¹ DIPALICORI, el sitio de diversidad y patrimonio lingüístico de la Universidad, publica diccionarios pictográficos, una enciclopedia, secuencias didácticas, libros, juegos y escaneos de publicaciones antiguas.² El diccionario bribri-español de Haakon Krohn, profesor de la Universidad, ofrece más de seis mil entradas en bribri.³

Cada fuente conserva la forma en que se publicó (cuadro 1). El corpus oral existe como páginas web, una por texto, y su anotación se lee solo en esas tablas. En los diccionarios pictográficos, el significado de cada palabra es un dibujo: una ilustración por tema, con un círculo para cada área y una línea para cada punto que una palabra nombra.⁴ Sus grabaciones

están dentro de aplicaciones de escritorio. El texto de las secuencias didácticas se puede extraer de sus PDF, pero las marcas diacríticas del bribri salen duplicadas o separadas de su letra, y hay que normalizarlas antes de buscar una palabra. Las fuentes tampoco escriben una misma marca con el mismo carácter: la nasalidad aparece como dos caracteres distintos de Unicode, y el apóstrofo como tres. Esta variación ya se documentó para la escritura en teclado.⁵ Los escaneos más antiguos no tienen capa de texto.

Hoy no hay una versión común del conjunto, y quien quiera trabajar con él, para una secuencia, una herramienta o un estudio, convierte cada fuente por su cuenta. Existe un conjunto paralelo de unas seis mil oraciones en bribri y español, armado para traducción automática a partir de libros, diccionarios y del corpus oral.⁶ Pero no se encontró un corpus que reúna los materiales completos, con su audio y su procedencia, en un mismo esquema. Existe, aparte, una colección de documentación nueva de los dialectos de Amubre y Coroma, depositada en 2025 en el Endangered Languages Archive, con parte del acceso restringido hasta 2027.⁷ Esa colección no reúne el material publicado, y este plan no la toca. Un artículo de posición reciente sostiene, desde su título, que la documentación de lenguas necesita mejor infraestructura, no mejores modelos.⁸

Cuadro 1. Las fuentes, la forma en que se publican y su estado para este trabajo. Revisado en línea el 1 de octubre de 2026.

Fuente	Qué contiene	Cómo se publica	Estado para este trabajo
Corpus oral pandialectal	35 textos transcritos y un video, de tres dialectos y seis géneros, 13535 palabras	Una página HTML por texto, con un archivo WAV	El audio está en línea. La anotación se publica solo como tablas HTML.
Diccionarios pictográficos y enciclopedia	Vocabulario temático en bribri y español. Cada palabra señala una parte de una ilustración	PDF. El audio de las palabras y las ilustraciones van también dentro de aplicaciones de escritorio en Java	Texto extraíble. Las ilustraciones y las marcas que unen cada palabra con su parte son objetos separados en el PDF. En el texto extraído faltan la nasalidad y el tono escrito sobre la diéresis, aunque la página los muestra. El audio se extrae del paquete de cada aplicación.
Secuencias didácticas	Unidades para la enseñanza del bribri	PDF	Texto extraíble, con marcas diacríticas duplicadas o separadas. Requiere normalización.
Libro de narraciones Sa' sɪwə' dékə	Narraciones en bribri	PDF	Texto extraíble, con marcas diacríticas separadas de su letra. Requiere normalización.
Diccionario bribri-español	Más de seis mil entradas en bribri y su contraparte en español	Aplicación web que carga dos archivos XML	Datos estructurados y descargables.
Escaneos de revistas y del periódico de 1978	Artículos con ejemplos en bribri, y textos escritos en bribri	PDF de imágenes	Sin capa de texto. Fuera de este plan.
Portal de la lengua bribri	Textos bilingües, cursos y grabaciones	Google Sites y Google Drive	Fuera de este plan.

Las cifras del corpus oral son las que publica su propio sitio. El estado de las secuencias didácticas proviene de una extracción de prueba de dos de ellas, El de las ilustraciones y las marcas del diccionario de la casa tradicional proviene de comparar su texto extraído con la página impresa.

2 Qué se propone

EL PLAN TIENE DOS PARTES, y la segunda depende de la primera. La primera produce el corpus. La segunda es su primer uso, y responde una pregunta propia de la didáctica de segundas lenguas: qué vocabulario del habla enseñan los materiales que ya existen.

2.1 Reunir: el corpus, primera etapa

Cada fuente entra al repositorio por un convertidor propio y queda en tres formas: la copia original sin cambios, una versión en formato abierto y un archivo de metadatos con el titular de derechos, la licencia, la procedencia y el estado del permiso de publicación (figura 1). El formato abierto es CLDF, el formato tabular con esquema en que se distribuyen, por ejemplo, WALS y Glottolog.⁹ Cada texto y cada enunciado recibe un identificador estable. El texto se guarda tal como se publicó, y una capa normalizada escribe cada marca con un solo carácter,

según una tabla de correspondencias que se define en este trabajo y queda documentada en el repositorio, con la razón de cada decisión. Quien prefiera otra decisión parte de la copia original. Donde el texto extraído pierde marcas que la página muestra, sus metadatos lo indican hasta que se corrija contra la página. Así, cualquier capa posterior, sea una medición o una corrección, se refiere a ese identificador sin copiar el texto.

En este plan entran las fuentes de la Universidad: el corpus oral, los materiales de DIPALICORI con el audio de sus aplicaciones, y el diccionario de Krohn. El corpus no es solo texto. Del corpus oral entran el texto y el audio, y de los diccionarios pictográficos entran también sus ilustraciones. Cada entrada une la palabra en bribri, su traducción, su grabación y la ilustración, con el área o el punto que la palabra nombra. CLDF tiene desde 2021 una tabla propia para estos archivos de medios.¹⁰ Así, la definición de cada palabra, que en estos diccionarios es un dibujo, pasa también al formato abierto. Los metadatos de cada ilustración nombran a su ilustrador, que puede ser un titular de derechos distinto del autor del texto. A partir de las tablas del corpus oral se regenera un archivo ELAN por texto, con el tiempo de cada enunciado, para que se pueda abrir y corregir con la herramienta habitual de anotación. Un visor privado permite buscar una forma, ver los enunciados que la contienen, escuchar cada uno y ver la ilustración de cada entrada de diccionario. El repositorio y el visor quedan privados hasta que los autores de cada material acuerden su publicación.

Quedan fuera de este plan, y descritos en la sección 7: la publicación del sitio bajo un dominio de la Universidad, la alineación palabra por palabra, la capa de texto de los escaneos y las fuentes externas a la Universidad.

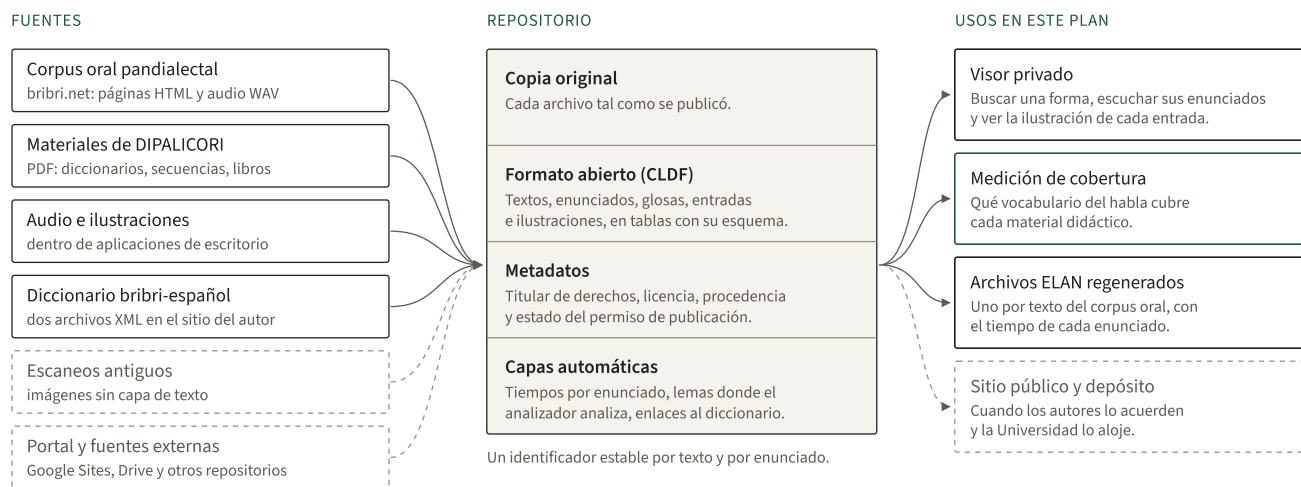


Figura 1. Del material publicado al repositorio y a sus usos. Las líneas continuas corresponden a este plan. Las punteadas, a etapas posteriores. El repositorio guarda cada material en cuatro capas, y cada capa se refiere a los mismos identificadores.

2.2 *Medir: lo que enseñan los materiales*

La investigación sobre vocabulario en segundas lenguas trabaja con la cobertura: la proporción de las palabras de un texto o de un discurso que conoce quien aprende.¹¹ Para el inglés y otras lenguas con corpus grandes, las listas de frecuencia que permiten calcularla existen desde hace décadas, y el paso de un corpus a listas de vocabulario para la enseñanza tiene un procedimiento publicado para el galés.¹² No se encontró una lista de frecuencias publicada para el bribri.

El corpus oral permite una primera aproximación. Con 13535 palabras es pequeño, y sus bandas de frecuencia resultan gruesas, pero es habla real de tres dialectos y seis géneros. La medición calcula, para cada secuencia didáctica y cada diccionario, qué proporción de las palabras del corpus cubre su vocabulario, por banda de frecuencia, por dialecto y por género. También produce dos listas: las palabras frecuentes en el habla que ningún material incluye, y las palabras de los materiales que no aparecen en el corpus.

La unidad de conteo es la forma escrita, después de normalizar la codificación de los diacríticos. Donde el analizador morfológico del propio corpus logra analizar una forma, el informe da también el lema, junto con la proporción de formas analizadas. El resultado es una página que se lee sin ayuda, pensada como insumo para la próxima edición de los materiales. Sus autores la ven antes que nadie.

3 Calendario y horas

EL PLAN CUBRE 28 SEMANAS DE TRABAJO y 280 horas, a razón de 10 horas por semana, del 15 de octubre de 2026 al 19 de mayo de 2027 (figura 2). Deja libres dos semanas de fin de año y la Semana Santa. Las horas de cada actividad son una estimación. La bitácora del TCU registra las horas cada semana, y cada reunión de seguimiento compara lo registrado con el plan. Si el trabajo se atrasa, el recorte cae en la medición, que puede limitarse a las secuencias didácticas. El corpus no se recorta, porque todo lo demás depende de él. El plan termina con margen antes del cierre del periodo de doce meses del TCU.

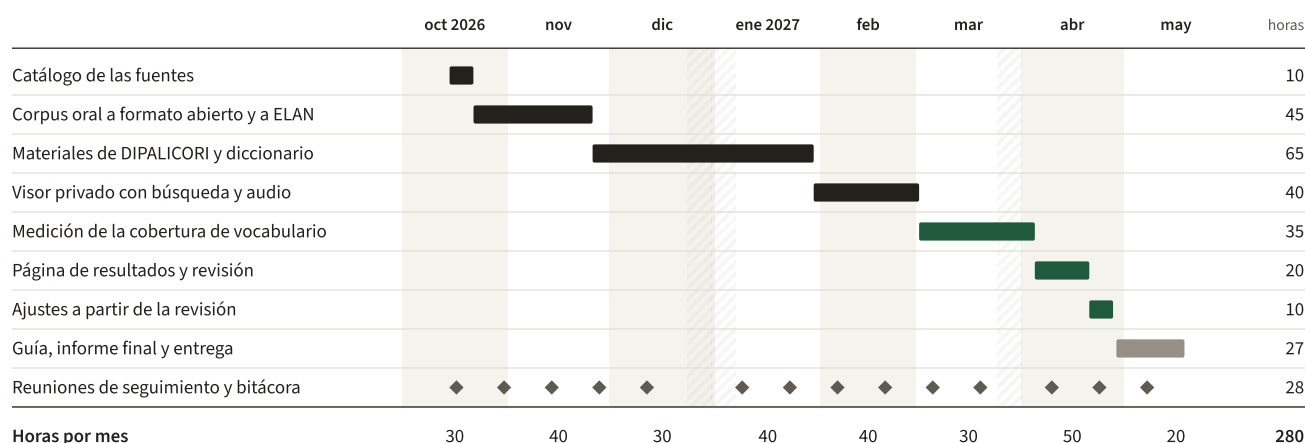


Figura 2. Calendario del 15 de octubre de 2026 al 19 de mayo de 2027. Las barras oscuras reúnen el corpus, las verdes miden los materiales y la gris cierra el trabajo. Las franjas rayadas marcan los recesos. Los rombos marcan las reuniones de seguimiento, cada dos semanas de trabajo. Las horas de reuniones y bitácora se reparten en todas las semanas.

4 Qué se entrega y cómo se comprueba

1. **El repositorio**, con el corpus oral, los materiales de DIPALICORI y el diccionario en formato abierto, con las ilustraciones de los diccionarios unidas a sus entradas, y cada material con su copia original y sus metadatos.

Se comprueba: el validador de CLDF lo acepta sin errores, una muestra de enunciados elegida en la reunión coincide con su fuente, y en una muestra de entradas cada palabra señala la parte correcta de su ilustración.

2. **Los archivos ELAN regenerados** del corpus oral, uno por texto.

Se comprueba: cada archivo abre en ELAN, y cada enunciado suena en su intervalo.

3. **El visor privado**, con búsqueda por forma, filtros por dialecto y género, audio por enunciado y la ilustración de cada entrada de diccionario.

Se comprueba: una demostración en la reunión de inicios de marzo.

4. **El informe de cobertura**, como página de resultados y como documento breve.

Se comprueba: revisión en la reunión de finales de abril, antes de cualquier difusión. Los ajustes que salgan de esa revisión entran antes de la entrega.

5. **Una guía para quien continúe**, y el informe final del TCU.

Se comprueba: una persona que no participó instala el repositorio y repite la medición siguiendo solo la guía.

5 Qué se necesita

- La aprobación del alcance y de este plan de trabajo, que el reglamento del TCU pide coordinar con el equipo responsable del proyecto.¹³
- Una reunión de seguimiento de 30 a 45 minutos cada dos semanas, presencial o virtual.

- Si existen, los archivos de anotación originales del corpus oral, o el contacto de quien los conserva. Con ellos, los archivos regenerados se comparan contra el original.
- Más adelante, y fuera de este plan, una decisión con los autores de cada material sobre la publicación del corpus.

El plan no pide nada a las comunidades. Todo el material ya está publicado, y el trabajo no requiere a nadie que lea bribri. No hay presupuesto: el trabajo corre en equipo propio, sin costo de operación.

6 Riesgos

Riesgo	Respuesta
El trabajo se atrasa.	La medición se limita a las secuencias didácticas. El corpus no se recorta. El plan termina con margen antes del cierre del periodo del TCU.
La extracción de un PDF pierde la estructura del documento.	La copia original se conserva, el formato abierto guarda lo que se extrae y los metadatos lo indican.
Los materiales usan ortografías distintas.	La normalización se limita a la codificación de los caracteres. Las diferencias de ortografía se informan y no se corrigen.
El corpus es pequeño para medir frecuencias.	El informe usa bandas gruesas, lo dice, y presenta el resultado como una primera aproximación.
La medición se lee como una evaluación de los materiales.	Sus autores ven los resultados primero, y el informe se plantea como insumo para la siguiente edición.
Aparecen los archivos de anotación originales.	Se usan en lugar de los regenerados, y la diferencia queda registrada.

7 Después de este plan

EL REPOSITORIO QUEDA como base para otros trabajos. Las líneas siguientes pueden tomarlas otros estudiantes de TCU, o continuar en un periodo posterior. Cada una tiene su propuesta completa, con su situación, entregables, plan, requisitos y riesgos.

El corpus completo unos tres meses

La publicación bajo un dominio de la Universidad con el acuerdo de los autores, la alineación palabra por palabra, la capa de texto de los escaneos, las fuentes externas y un depósito con identificador citable.

Lo que enseñan los materiales seis semanas

La medición de este plan, ampliada con una revisión de las formas dialectales que usa cada material y de cómo se repite el vocabulario a lo largo de una secuencia.

Un mapa de lo que existe dos meses

Un catálogo público de los recursos, conjuntos de datos, herramientas y artículos sobre el bribri, con un programa que revisa cada semana que sigan en línea.

Un corpus que se mantiene dos meses

Una plantilla para que cada secuencia didáctica nueva entre al corpus como datos, y un revisor automático que detecta problemas y deja preguntas para una persona.

El corpus en el aula un mes

Una vista del corpus para docentes, que funciona sin conexión, probada con dos docentes en una tarea real.

Lo que dice el sonido dos meses

Mediciones acústicas del tono del bribri en el habla del corpus, por dialecto. Es la línea de carácter más investigativo.

8 Referencias

1. Flores Solórzano, S. (2017). *Corpus oral pandialectal de la lengua bribri* [Corpus en línea]. bribri.net/
2. Universidad de Costa Rica. (s. f.). *Dipalicori: Diversidad y patrimonio lingüístico de Costa Rica* [Sitio web]. dipalicori.ucr.ac.cr/
3. Krohn, H. S. (2019–2025). *Diccionario bribri-español, español-bribri* [Diccionario en línea]. www.haakonkrohn.com/bribri/
4. Proyecto Diversidad Lingüística de Costa Rica. (2015). *Diccionario pictográfico de la casa tradicional bribri* [Diccionario]. Escuela de Filología, Lingüística y Literatura, Universidad de Costa Rica. dipalicori.ucr.ac.cr/wp-content/uploads/Diccionario_casa_tradicional_bribri.pdf
5. Flores Solórzano, S. (2010). Teclado chibcha: un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, 36(2), 155–161. <https://doi.org/10.15517/rfl.v36i2.1110>
6. Feldman, I., y Coto-Solano, R. (2020). Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. En *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3965–3976). International Committee on Computational Linguistics. doi.org/10.18653/v1/2020.coling-main.351
7. Herrera Miranda, R. (2025). *Documentation of Talamancan Bríbrí (Chibchan, Costa Rica)* [Colección bribri-herreramiranda-0778, proyecto ELDP IPF0426]. Endangered Languages Archive. hdl.handle.net/2196/cd0c64c6-951d-43d1-820c-abcd78bae63c
8. Gessler, L., Anastasopoulos, A., Auderset, S., Bodt, T., Chelliah, S., Christian, S., Fily, M., Herrera, S., Huber, E., Loáiciga, S., Meelen, M., Östling, R., Palmer, A., y Visser, E. (2026). The missing middle: Language documentation needs better infrastructure, not better models. En *Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9)* (pp. 136–147). Association for Computational Linguistics. doi.org/10.18653/v1/2026.computel-1.15
9. Forkel, R., List, J.-M., Greenhill, S. J., Rzymiski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., y Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use

in comparative linguistics. *Scientific Data*, 5, Artículo 180205. doi.org/10.1038/sdata.2018.205

10. Cross-Linguistic Data Formats. (2021). *CLDF Specification, version 1.1: MediaTable* [Especificación]. GitHub. github.com/cldf/cldf/tree/master/components/media
11. Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59–82. doi.org/10.3138/cmlr.63.1.59
12. Knight, D., Fitzpatrick, T., Morris, S., Tovey-Walsh, B., Prosser, H., y Davies, E. (2023). Corpus to curriculum: Developing word lists for adult learners of Welsh. *Applied Corpus Linguistics*, 3(2), Artículo 100052. doi.org/10.1016/j.acorp.2023.100052
13. Consejo Universitario, Universidad de Costa Rica. (2018). *Reglamento del Trabajo Comunal Universitario* (reforma integral, sesión 6206-06). www.cu.ucr.ac.cr/normativa/trabajo_comunal.pdf