Propuesta de Trabajo Comunal Universitario

Un corpus común para el bribri

Plan de trabajo de 280 horas para noviembre de 2026 a junio de 2027, y el trabajo que puede continuar después

Ariel Arévalo Alvarado
Ciencias de la Computación e Informática, Universidad de Costa Rica
Segunda versión, 28 de octubre de 2026 (qué cambió)

Resumen. DIPALICORI, el sitio de la Escuela de Filología, Lingüística y Literatura para las lenguas de Costa Rica, publica una cantidad considerable de material en bribri: diccionarios pictográficos con sus ilustraciones y grabaciones, una enciclopedia, secuencias didácticas y un libro de narraciones. Cada material conserva la forma en que se publicó, casi siempre un PDF, y las grabaciones de las palabras están dentro de aplicaciones de escritorio. Este documento propone un plan de 280 horas, del 12 de noviembre de 2026 al 16 de junio de 2027, con un solo producto: esos materiales reunidos en un formato abierto común, con un visor privado para consultarlos. El repositorio se construye de modo que otras fuentes puedan entrar después sin rehacerlo. El plan no requiere trabajo de campo, anotación nueva ni presupuesto. La sección 7, aparte del plan, describe el trabajo que puede continuar después del TCU.

1 Situación actual
2 Qué se propone
3 Calendario y horas
4 Qué se entrega y cómo se comprueba
5 Qué se necesita
6 Riesgos
7 Después del TCU
8 Referencias

1 Situación actual

DIPALICORI, el sitio de diversidad y patrimonio lingüístico de la Escuela de Filología, Lingüística y Literatura, reúne materiales en las lenguas de Costa Rica, en su mayoría producidos por los proyectos de acción social de la Escuela.[1] Para el bribri publica diccionarios pictográficos, una enciclopedia, secuencias didácticas, un libro de narraciones, láminas, juegos y escaneos de publicaciones antiguas. Este plan trabaja solo con esos materiales.

Cada material conserva la forma en que se publicó (cuadro 1). 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.[2] 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. Los materiales tampoco escriben una misma

marca con el mismo carácter: la nasalidad y el apóstrofo aparecen, según el material, como caracteres distintos de Unicode. Esta variación ya se documentó para la escritura en teclado.[3] Los escaneos más antiguos no tienen capa de texto.

Hoy no hay una versión común de estos materiales, y quien quiera trabajar con ellos, para una secuencia nueva, una herramienta o un estudio, convierte cada PDF 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 de un corpus oral,[4] pero no se encontró un corpus que reúna los materiales completos, con sus ilustraciones, 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.[5] 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.[6]

Cuadro 1. Los materiales de bribri en DIPALICORI, la forma en que se publican y su estado para este trabajo. Revisado en línea el 1 de octubre de 2026.

Material	Qué contiene	Cómo se publica	Estado para este trabajo
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' s̠iwą' dẽ́ką	Narraciones en bribri	PDF	Texto extraíble, con marcas diacríticas separadas de su letra. Requiere normalización.
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.

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. El catálogo del plan registra todos los materiales de bribri del sitio, también las láminas y los juegos.

2 Qué se propone

El plan tiene un producto: el corpus de los materiales de bribri de DIPALICORI, como primera etapa de un corpus común. La sección 2.1 describe el corpus. La 2.2 explica cómo el reposito-

rio queda listo para recibir otras fuentes, y la 2.3, cómo se toman las decisiones que requieren conocer la lengua.

2.1 *El corpus de DIPALICORI*

Cada material entra al repositorio 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.[7] Cada material, cada texto y cada entrada 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 queda documentada en el repositorio, con la razón de cada decisión (sección 2.3). 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.

El corpus no es solo texto. De los diccionarios pictográficos entran también sus ilustraciones, y de sus aplicaciones, la grabación de cada palabra. 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,[8] pero no define cómo señalar una parte de una imagen. Por eso el área o el punto de cada palabra se guarda como una anotación del modelo del W3C para anotaciones web, con un selector que delimita la región de la ilustración: un rectángulo o una forma en SVG.[9] Es el mismo modelo con que IIIF señala partes de una imagen.[10] Así, la definición de cada palabra, que en estos diccionarios es un dibujo, pasa también a un formato abierto que otras herramientas leen. Los metadatos de cada ilustración nombran a su ilustrador, que puede ser un titular de derechos distinto del autor del texto. Un visor privado permite buscar una forma, ver los textos y las entradas que la contienen, escuchar su grabación y ver la ilustración de cada entrada. El repositorio y el visor quedan privados hasta que los autores de cada material acuerden su publicación.

Entran los materiales de bribri que DIPALICORI haya publicado hasta el inicio del plan, el 12 de noviembre de 2026. Lo que el sitio publique después puede entrar por el mismo camino (sección 2.2). Quedan fuera de este plan, y descritos en la sección 7: los escaneos sin capa de texto, las fuentes que no están en DIPALICORI, la medición de lo que enseñan los materiales y la publicación del corpus bajo un dominio de la Universidad.

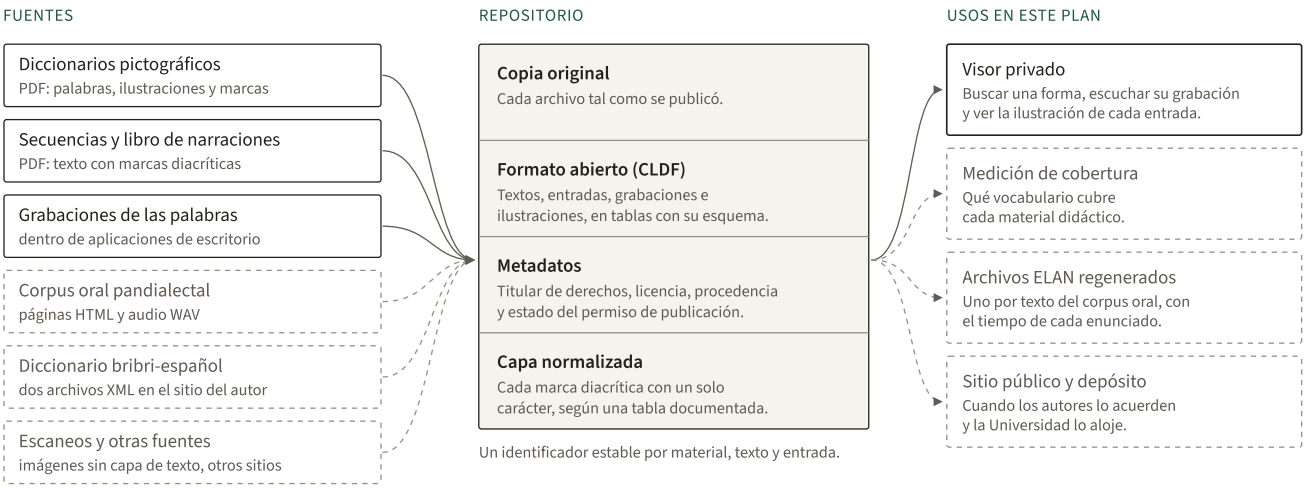


Figura 1. Del material publicado al repositorio y a sus usos. Las líneas continuas corresponden a este plan. Las punteadas, a etapas posteriores al TCU. El repositorio guarda cada material en cuatro capas, y cada capa se refiere a los mismos identificadores.

2.2 *Un repositorio que puede crecer*

Cada fuente se organiza como un conjunto de datos de cldfbench, la herramienta de CLDF para llevar datos heredados a ese formato.[11] Cada conjunto separa la copia original, lo que agregan quienes lo preparan y la salida en CLDF, y su programa produce la salida a partir de las otras dos. Cada tipo de material tiene un convertidor propio: uno para los diccionarios pictográficos con sus ilustraciones, uno para el texto de las secuencias y del libro, y uno para los paquetes de las aplicaciones. Todos cumplen el mismo contrato. Reciben los archivos originales y entregan las mismas tablas, los mismos metadatos y el mismo registro de lo que no pudieron extraer, y un validador común revisa cada resultado. Agregar un material de un tipo que ya existe requiere solo una entrada nueva en el catálogo.

Las fuentes que este plan no trata tienen ya un lugar. El catálogo las registra con su tipo, el estado de su permiso y el convertidor que necesitarían, por ejemplo páginas web con audio alineado por enunciado, un diccionario en XML o imágenes sin capa de texto. El plan no procesa esas fuentes ni escribe sus convertidores. Su objetivo es que entrar después no obligue a rehacer el repositorio. La conversión a CLDF de aplicaciones web heredadas tiene un método publicado, que sirve de referencia para ese contrato.[12]

2.3 *Las decisiones de lengua*

El autor de este plan no habla bribri. El plan está hecho para que la mayor parte del trabajo no lo requiera, y para que las decisiones que sí lo requieren no las tome él. Extraer el texto, las ilustraciones y las grabaciones, guardarlos con su procedencia y construir el visor son tareas de ingeniería. Los materiales son bilingües, y la traducción al español de cada palabra permite comprobar que una entrada quedó completa.

La regla del corpus es registrar, no interpretar. Cada material se conserva como se publicó, y cada cambio queda en una capa aparte, documentada y reversible. El trabajo no corrige la escritura, no traduce y no analiza las palabras.

Donde sí importa la lengua es en la escritura. Una misma marca aparece con caracteres distintos según el material, y en el texto extraído de un diccionario se pierden marcas que la página muestra: el dibujo dice kapǫ̀ y el texto extraído, kapö. Encontrar esas diferencias es mecánico, porque basta comparar el texto con la página. Decidir que dos formas son la misma letra, o la misma palabra, es una decisión de lengua.

Esas decisiones pasan por una persona de enlace, designada por la coordinación del proyecto, en el papel que en las secuencias didácticas cumple la asesoría lingüístico-cultural.[13] Las preguntas se reúnen en una lista y se resuelven en las reuniones de seguimiento o por escrito, y cada respuesta queda en la bitácora con su fecha. Mientras una pregunta espera, el dato se conserva como se publicó y queda marcado, de modo que el trabajo no se detiene.

3 Calendario y horas

El plan cubre 28 semanas de trabajo y 280 horas, a razón de 10 horas por semana, del 12 de noviembre de 2026 al 16 de junio 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 visor se reduce a la búsqueda y el audio, y los materiales que no alcancen a entrar quedan en el catálogo, con su convertidor listo. El repositorio 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.

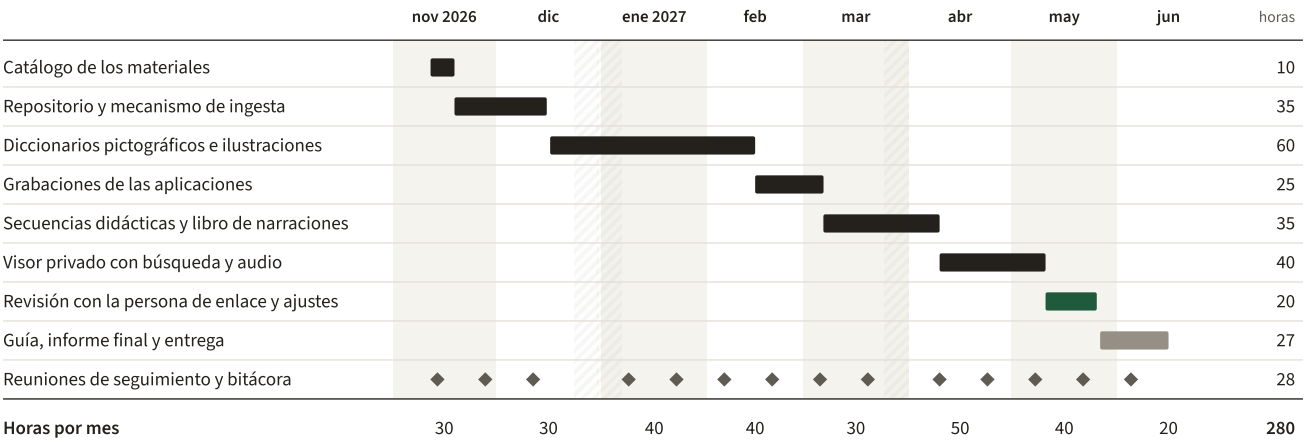


Figura 2. Calendario del 12 de noviembre de 2026 al 16 de junio de 2027. Las barras oscuras reúnen el corpus, la verde es la revisión con la persona de enlace y la gris cierra el trabajo. Las franjas rayadas marcan los recesos. Los rombos marcan las reuniones de seguimiento, cada dos semanas de trabajo. Las horas de reuniones y bitácora se reparten en todas las semanas.

4 Qué se entrega y cómo se comprueba

1. **El repositorio,** con los materiales de bribri de DIPALICORI en formato abierto, las ilustraciones unidas a sus entradas y las grabaciones de las aplicaciones, y cada material con su copia original y sus metadatos.
   Se comprueba: el validador de CLDF lo acepta sin errores, una muestra de textos y entradas elegida en la reunión coincide con su fuente, y en una muestra de entradas cada palabra señala la parte correcta de su ilustración.
2. **El mecanismo de ingesta,** con el contrato de los convertidores, el validador y el catálogo, que registra también las fuentes posteriores.
   Se comprueba: un material que se reserva a propósito entra al final con solo agregar su entrada al catálogo.
3. **El visor privado,** con búsqueda por forma, la grabación de cada palabra y la ilustración de cada entrada de diccionario.
   Se comprueba: una demostración en la reunión de mediados de mayo.
4. **La tabla de correspondencias y la lista de preguntas de lengua,** con las respuestas de la persona de enlace.
   Se comprueba: revisión de una muestra del corpus con la persona de enlace a finales de mayo. 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 agrega un material 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.[14]
- Una reunión de seguimiento de 30 a 45 minutos cada dos semanas, presencial o virtual.
- Una persona de enlace para las decisiones de lengua, designada por la coordinación del proyecto. Las preguntas se reúnen en una lista, para no ocupar más tiempo del necesario.
- 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 las decisiones de lengua se reúnen en una lista para la persona de enlace. No hay presupuesto: el trabajo corre en equipo propio, sin costo de operación.

6 Riesgos

Riesgo	Respuesta
El trabajo se atrasa.	El visor se reduce a la búsqueda y el audio. Los materiales que no alcancen a entrar quedan en el catálogo con su convertidor. 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.
El texto extraído pierde marcas que la página muestra.	Los metadatos lo indican. La corrección se hace contra la página, y los casos dudosos pasan a la persona de enlace.
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.
Una pregunta de lengua tarda en responderse.	El dato queda como se publicó y marcado, y el trabajo sigue.
El paquete de una aplicación no se puede leer.	Sus grabaciones quedan catalogadas con el paquete original, y el resto del material entra igual.
Preparar el repositorio para otras fuentes alarga el plan.	El plan escribe solo el contrato y el lugar de cada fuente en el catálogo, no sus convertidores.

Fuera del TCU

7 Después del TCU

Esta sección no forma parte del plan. Describe cómo puede continuar el trabajo en los proyectos de la Escuela, con otros estudiantes de TCU o en un periodo posterior. Todo entra por el mecanismo de la sección 2.2.

Las fuentes que quedan para después son el *Corpus oral pandialectal de la lengua bribri*, con 35 textos transcritos de tres dialectos, su audio y su análisis morfológico,[15] el diccionario bribri-español de Haakon Krohn, con más de seis mil entradas,[16] los escaneos sin capa de texto y otras fuentes en línea. Con el corpus oral en el repositorio se puede medir qué vocabulario del habla cubren los materiales didácticos. Una versión menor de esa medición compara los materiales de DIPALICORI entre sí: qué palabras de las secuencias definen los diccionarios, y qué formas se escriben distinto de un material a otro.

Cada una de las líneas siguientes tiene su propuesta completa, con su situación, entregables, plan, requisitos y riesgos.

El corpus completo unos tres meses
Las demás fuentes en el repositorio, 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 y un depósito con identificador citable.

Lo que enseñan los materiales seis semanas
La medición del vocabulario de los materiales contra el habla del corpus oral, 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. Universidad de Costa Rica. (s. f.). *Dipalicori: Diversidad y patrimonio lingüístico de Costa Rica* [Sitio web]. dipalicori.ucr.ac.cr/
2. 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
3. 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
4. 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
5. Herrera Miranda, R. (2025). *Documentation of Talamancan Bríbri (Chibchan, Costa Rica)* [Colección bribri-herreramiranda-0778, proyecto ELDP IPF0426]. Endangered Languages Archive. hdl.handle.net/2196/cd0c64c6-951d-43d1-820c-abcd78bae63c

6. 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

7. Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, 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

8. Cross-Linguistic Data Formats. (2021). *CLDF Specification, version 1.1: MediaTable* [Especificación]. GitHub. github.com/cldf/cldf/tree/master/components/media

9. Sanderson, R., Ciccarese, P. y Young, B. (Eds.). (2017, 23 de febrero). *Web Annotation Data Model* [Recomendación del W3C]. World Wide Web Consortium. www.w3.org/TR/annotation-model/

10. IIIF Consortium. (2020). *IIIF Presentation API 3.0* [Especificación]. International Image Interoperability Framework. iiif.io/api/presentation/3.0/

11. Forkel, R. y List, J.-M. (2020). CLDFBench: Give your cross-linguistic data a lift. En *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6995–7002). European Language Resources Association. aclanthology.org/2020.lrec-1.864/

12. Forkel, R., Swanson, D., y Moran, S. (2024). Converting legacy data to CLDF: A FAIR exit strategy for linguistic web apps. En *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3978–3982). ELRA; ICCL. aclanthology.org/2024.lrec-main.353/

13. Proyecto ED-3675, Coconstrucción de propuestas y recursos didácticos para la enseñanza-aprendizaje de la lengua bribri. (2025). *Comida y bebidas tradicionales* [Secuencia didáctica]. Escuela de Filología, Lingüística y Literatura, Universidad de Costa Rica. dipalicori.ucr.ac.cr/wp-content/uploads/Comidas-y-bebidas-tradicionales.-Secuencia-didactica.pdf

14. 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

15. Flores Solórzano, S. (2017). *Corpus oral pandialectal de la lengua bribri* [Corpus en línea]. bribri.net/

16. Krohn, H. S. (2019–2025). *Diccionario bribri–español, español–bribri* [Diccionario en línea]. www.haakonkrohn.com/bribri/