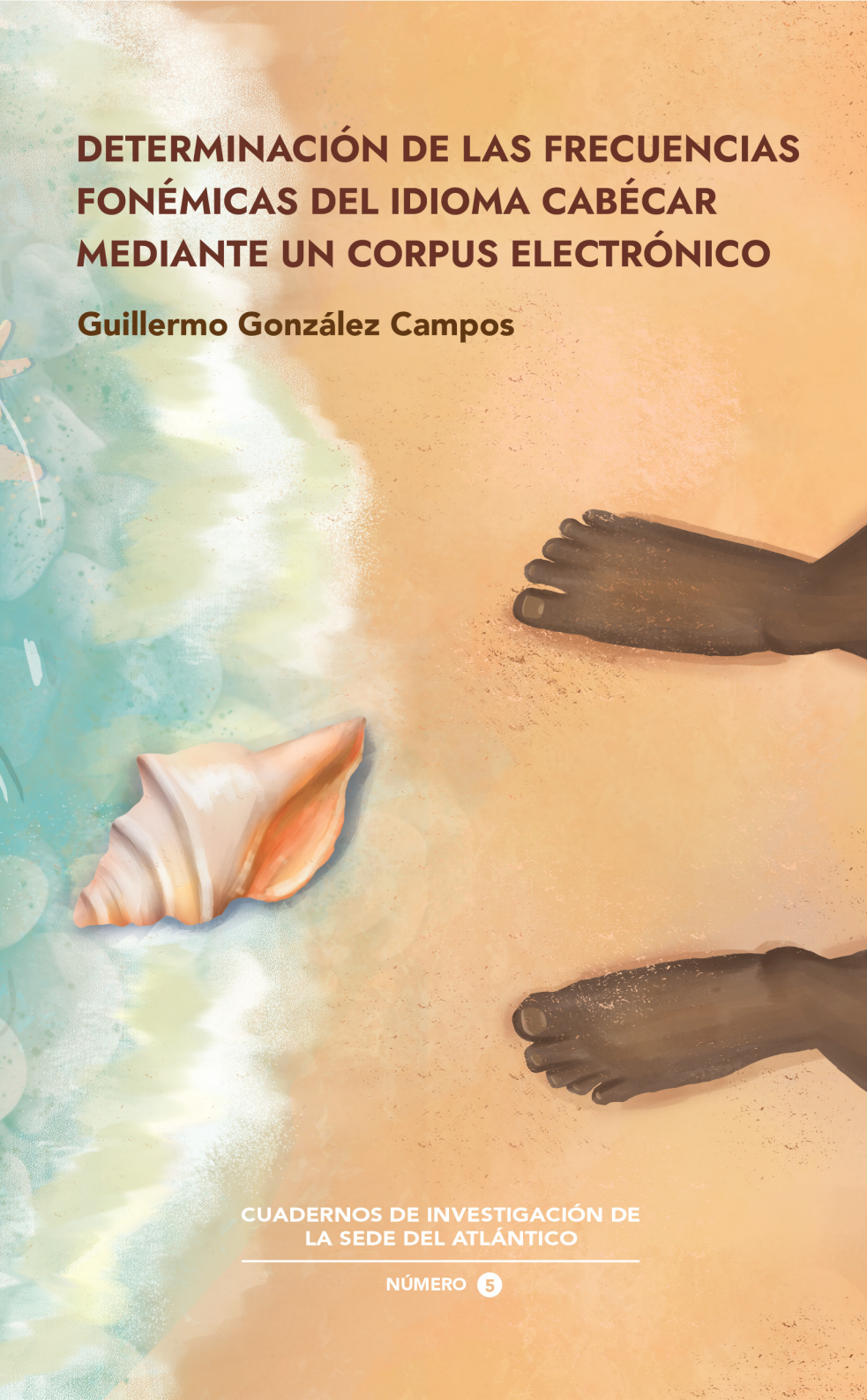
DETERMINACIÓN DE LAS FRECUENCIAS FONÉMICAS DEL IDIOMA CABÉCAR MEDIANTE UN CORPUS ELECTRÓNICO

Guillermo González Campos



CUADERNOS DE INVESTIGACIÓN DE LA SEDE DEL ATLÁNTICO

NÚMERO 5

UNIVERSIDAD DE COSTA RICA

SA Sede del Atlántico

DETERMINACIÓN DE LAS FRECUENCIAS FONÉMICAS DEL IDIOMA CABÉCAR MEDIANTE UN CORPUS ELECTRÓNICO

GUILLERMO GONZÁLEZ CAMPOS

CUADERNOS DE INVESTIGACIÓN DE LA SEDE DEL ATLÁNTICO

NÚMERO 5

CC.SIBDI.UCR - CIP/4101

Nombres:	González Campos, Guillermo, autor.
Título:	Determinación de las frecuencias fonémicas del idioma cabécar mediante un corpus electrónico / Guillermo González Campos.
Descripción:	Primera edición. | [Turrialba, Costa Rica] : Universidad de Costa Rica, Sede del Atlántico, 2024. | Cuadernos de investigación de la Sede del Atlántico ; número 5.
Identificadores:	ISBN 978-9930-9718-9-5 (blanda o rústica)
Materias:	LEMB: Cabécar (Lengua indígena) – Fonología. | Cabécar (Lengua indígena) – Fonética. | Cabécar (Lenguaindígena) – Frecuencia de palabras. | Lingüística matemática. | Lingüística computacional. | Lenguas indígenas – Costa Rica. | LCSH: Corpus (Lingüística).
Clasificación:	CDD 497.815 --ed. 23

La edición de esta obra fue aprobada por la Comisión Editorial de la Sede del Atlántico de la Universidad de Costa Rica.

Primera edición: 2024.

Diagramación y artes finales: Natalia Hernández Araya.
Corrección filológica: Guillermo González Campos.
Diseño de portada: Natalia Hernández Araya & Amets Muruzabal Ganuza.
Corrección de pruebas: Guillermo González Campos.

ÍNDICE

Resumen

Este texto presenta un estudio exploratorio que, por primera vez, ha establecido la frecuencia de fonemas del cabécar, lengua indígena costarricense. En concreto, este texto expone los principios conceptuales y el proceso metodológico de conformación de un corpus electrónico de datos lingüísticos de este idioma que sirvió de base empírica para la obtención de información estadística de tipo fonológico. Luego, presenta la frecuencia absoluta de aparición de cada uno de los fonemas de la lengua cabécar y lleva a cabo un análisis estadístico de dichos datos a la luz de los postulados y principios teóricos generados en el marco de investigaciones semejantes en otras lenguas del mundo.

Palabras clave: *lengua cabécar, estadística fonológica, frecuencia fonemática, lingüística de corpus, lingüística cuantitativa.*

Abstract

This text introduces an exploratory study which, for the first time, quantifies the frequency of phonemes of Cabécar, an aboriginal language of Costa Rica. This text sets out the conceptual principles and methodological process used in constituting a corpus of linguistic data of this language for getting information of phonological statistics. Then it presents the absolute frequency of occurrence of each of the phonemes of Cabécar. In addition, a statistical analysis of this data has been carried out based on theoretical postulates and principles generated by similar studies of other languages of the world.

Keywords: *Cabécar language, phonological statistics, phonemic frequency, corpus linguistics, quantitative linguistics.*

1. Introducción

A la fecha, nunca se ha llevado a cabo un estudio de estadística fonológica en ninguna lengua indígena de Costa Rica*. De hecho, ni siquiera se ha llevado a cabo uno en un corpus del español de Costa Rica. Esto se debe, sin lugar a duda, al hecho de que tanto la lingüística computacional como la lingüística de corpus constituyen campos emergentes en esta nación. Tal y como lo revela el estudio de Leoni de León (2010), que ofrece una panorámica general del desarrollo de dichas disciplinas en este país, lo único que ha habido hasta ahora en Costa Rica son esfuerzos puntuales y casos aislados que, inclusive hoy día, están en proceso de consolidación, lo anterior sin restarle mérito a algunos antecedentes de investigación y el desarrollo tanto académico como comercial de algunas herramientas informáticas, la mayoría de las cuales se encuentran ligadas sobre todo al ámbito de la lexicografía.[1]

En el contexto latinoamericano, Costa Rica no es un caso aislado. En general, dichas investigaciones son bastante escasas en lo que a lenguas aborígenes se refiere. Las lenguas europeas, en cambio, desde hace décadas cuentan con recuentos de este tipo. En el español, por ejemplo, el primer estudio de esta naturaleza fue llevado a cabo en 1939 por George K. Zipf y Francis M. Rogers. Hoy día, esta lengua cuenta con otras muchas investigaciones al respecto, incluidos algunos hechos con corpus de varios millones de palabras y otros referidos a variedades específicas del español. [2]

Esta carencia de estudios ha provocado que, en investigaciones de carácter general sobre las tendencias generales de las frecuencias fonémicas de las lenguas del mundo, los idiomas indoamericanos se encuentren escasamente representados. Por ejemplo, en la investigación de Tambovtsev y Martindale (2007), que tiene por objetivo examinar la frecuencia de fonemas en una gran cantidad de lenguas del mundo con el fin de

* Esta investigación se llevó a cabo en 2016. Posteriormente, apareció el trabajo de Krohn (2017).

1 Al respecto, no puede dejarse de mencionar algunos de los principales productos desarrollados en los últimos años. Entre ellos, sobresalen el *Calculador Léxico-estadístico de Frecuencia Dispersión y Uso* (CaLeFDU), documentado en Leoni de León (2012), la creación de un corpus electrónico anotado automáticamente, el *Corpus de Mensajes Presidenciales de Costa Rica* (CODIMEP-CR), cuyo desarrollo se explica en Jara Murillo (2013), y la puesta en marcha de la *Interfaz lexicográfica polivalente* (INLEXPO) para la gestión de diccionarios electrónicos (al respecto, véase Astorga, Cordero y Leoni 2013).

2 Para un recuento de los estudios que analizan la frecuencia estadística de los fonemas del español, consúltense los trabajos de Moreno Sandoval et al. (2006) y González y Mejía (2011).

determinar qué ecuación las describe de una mejor manera, no se incluye ninguna lengua indoamericana, con la excepción de una lengua esquimal.[3]

Este texto[4] se propone llenar parcialmente este vacío de estudios y a la vez contribuir con el desarrollo de la lingüística de corpus y la estadística fonológica en Costa Rica presentando un estudio de tipo empírico y exploratorio, el cual tuvo por objetivo identificar el número de ocasiones en que aparecen cada uno de los fonemas del cabécar dentro de la cadena de habla, con el fin de determinar sus rangos de frecuencia. Debido a esto, este artículo explora hechos de índole fonológica en datos lingüísticos reales, los cuales serán analizados desde un punto de vista cuantitativo.

Con el fin de llevar a cabo la investigación sobre una base observable, dicho estudio tuvo como punto de partida la constitución de un corpus electrónico de datos lingüísticos del cabécar, cuyo diseño, procesamiento y anotación se explicarán más adelante. Gracias a esto, la investigación se basó en datos empíricos primarios que se sometieron a un proceso de análisis estadístico automatizado mediante una aplicación informática. Además, los resultados obtenidos fueron interpretados a la luz de las asunciones teóricas propuestas con respecto al comportamiento general que manifiestan las lenguas del mundo a la hora de organizar las frecuencias de sus fonemas en los conjuntos textuales.

El cabécar es una de las cinco lenguas indígenas que aún se habla en el territorio costarricense. De todas, es el idioma con mayor número de hablantes. De acuerdo con el trabajo de Constenla (2005 y 2008), el cabécar pertenece a la estirpe lingüística chibchense, la cual, más comúnmente es conocida como familia de lenguas chibchas. Esta, a su vez, forma parte de un grupo mayor de idiomas conocido como

3 Sobre este tema, los autores señalaron lo siguiente: *"Our sample of 95 languages does not include languages from all language families. American and African languages are notably underrepresented. We were limited to languages for which we could find or compute tabulations of phoneme frequencies"* (Tambovtsev y Martindale 2007: 4). En otras investigaciones, Yuri Tambovtsev ha utilizado una selección mayor de lenguas americanas para las cuales ha calculado, de forma personal, las frecuencias de fonemas por medio de textos que ha logrado encontrar. En su libro de 2003 sobre la tipología de la cadena de sonido (obra que conocemos gracias a la reseña de Shipulina 2005), utilizó datos de 20 idiomas nativos de América. En un trabajo posterior (Tambovtsev 2009), logró reunir datos de 32 lenguas indígenas americanas. Cabe destacar que, en todo caso, se trata de datos reunidos sobre corpus textuales muy limitados y que los únicos idiomas centroamericanos incluidos en dichas investigaciones son dos lenguas mayas habladas en Guatemala.

4 Trabajo surgido en el seno del proyecto de investigación 510-B6-036 de la Universidad de Costa Rica. Deseo agradecer a los profesores Pascual Cantos Gómez, de la Universidad de Murcia, y Christian Lehmann, de la Universidad de Erfurt, los valiosos comentarios y sugerencias que hicieron a este texto, las cuales permitieron mejorarlo sustancialmente.

microfilo lenmichí, el cual comprende una serie de lenguas que se hablan o hablaron desde el sur de Honduras hasta el norte de Venezuela.

De acuerdo con Margery (1989), el cabécar presenta dos variedades principales que, a su vez, se subdividen en dos áreas dialectales: una norteña (que abarca Chirripó y el Valle de La Estrella) y otra sureña (que comprende Talmanca y Ujarrás). Aún no se ha realizado un estudio pormenorizado de las variaciones dialectales del cabécar; sin embargo, con base en observaciones propias hechas por el autor de este trabajo (González Campos 2011), puede afirmarse que los cambios fonéticos son los más notables y los que mejor permiten diferenciar una variedad de otra. Para efectos de este estudio, se tomó como punto de partida la propuesta de análisis del sistema fonológico del cabécar hecha por el autor de esta investigación en un trabajo anterior (González Campos 2011). De acuerdo con esta, el sistema fonológico del cabécar norteño incluye veintiséis fonemas, mientras que el cabécar sureño solo posee veintidós.

Es necesario que el cabécar cuente con un estudio de cuantificación de fonemas por las siguientes razones:

a) Es la base para futuras investigaciones y desarrollos de la lingüística computacional. Los datos de frecuencia de los fonemas pueden ser utilizados por diferentes estudios de procesamiento del lenguaje, tales como el reconocimiento automático de habla o el reconocimiento óptico de caracteres.
b) El establecimiento de estas frecuencias permitirá enriquecer los estudios del mismo tipo ya hechos con otras lenguas y verificar si los universales y tipologías establecidos (por ejemplo, la preeminencia de los segmentos coronales) se cumple en dicha lengua, contribuyendo al desarrollo de la lingüística como disciplina.
c) La conformación del corpus electrónico que comprenda todo el material lingüístico disponible sobre el cabécar facilitará de forma muy significativa la consulta de estos datos lingüísticos y, por ello, mejorará sustancialmente la manera como se hacen los estudios lingüísticos de esta lengua.

Como suele ser tradicional, este trabajo se estructura en tres apartados. Primeramente, se brinda una presentación de los contenidos conceptuales y metodológicos que dan sustento al estudio. A

continuación, se presentan y discuten los resultados obtenidos a lo largo del proceso investigativo. Al final, se enumeran las conclusiones obtenidas.

2. Consideraciones teóricas y metodológicas

El propósito de esta sección es exponer el paradigma investigativo que se utilizó para obtener los datos, así como las técnicas, procedimientos e instrumentos utilizados para su recolección y procesamiento durante la investigación que da sustento a este artículo. Esta descripción no solo abarcará una clara delimitación del tipo de estudio llevado cabo, sino que también incluirá una pormenorizada explicación de cómo se conformó el corpus electrónico que sirvió de base para el estudio estadístico, así como las herramientas utilizadas en el análisis y explotación de los datos.

2.1. Tipo de investigación

Los datos que dan sustento a este trabajo fueron obtenidos mediante la aplicación de un estudio de estadística fonológica y los métodos propios de la lingüística de corpus. Por lo tanto, resulta pertinente dar cuenta de algunos aspectos conceptuales relacionados con estas dos subdisciplinas de la lingüística cuyos campos de estudio se entrecruzan en esta investigación.

2.1.1. Sobre la lingüística cuantitativa y la estadística fonológica

Tal y como lo explica Těšitelová (1992: 11), la lingüística cuantitativa, a veces también llamada lingüística estadística, es una de las dos especialidades en que se divide la lingüística matemática, la cual es la disciplina encargada de estudiar el lenguaje utilizando los métodos propios de las así llamadas ciencias exactas. Se caracteriza, principalmente, por utilizar métodos cuantitativos como la estadística o el cálculo de probabilidades.[5] En palabras de Köhler y Rieger (1993: x), *"Quantitative Linguistics deals with the quantitative characterization of languages and text features in an exact mathematical form which*

5 Según esta misma autora, la otra subdisciplina de la lingüística matemática es la lingüística algebraica, la cual, en sus análisis, utiliza más bien métodos de tipo cualitativo (teoría de conjuntos, teoría de grafos, etc.).

allows for their further treatment by formal and numerical operations".

Debido a lo anterior, el ámbito conceptual de la lingüística cuantitativa es tan amplio como el de la lingüística general, pues, al igual que esta, pretende estudiar todas las propiedades del lenguaje que sean esenciales para la comprensión de su funcionamiento. Es decir, se abarca con ella todos los niveles tradicionales del análisis lingüístico (fonología, morfología, sintaxis, etc.). La única diferencia, entonces, radica en los procesos metodológicos. Dicha disparidad de enfoque se aprecia con claridad en la explicación de los propósitos de esta disciplina que hace Marie Těšitelová, la cual, retomando un trabajo del lingüista checo Vilém Mathesius, expone de la siguiente manera su campo de estudio:

> *Quantitative linguistics looks for quantitative data, quantifies the phenomena of different language levels and models their relations realized in lower units, in the word, as well as in higher units, in the sentence, text, etc. to enable us a better understanding of their causal mechanism, to know the dynamism of the development of a language, their functioning in their formal as well as semantic aspects, to disclose the causes of the potentiality of the phenomena of language.* (Těšitelová 1992: 13).

Justamente por eso, autores como Reinhard Köhler y Gabriel Altmann (véase Köhler 2005: 2 y Köhler y Altmann 2011: 696-697) proponen que el objetivo de esta disciplina es descubrir fenómenos pertinentes y describirlos en forma sistemática para, en la medida de lo posible, encontrar leyes que expliquen los hechos observados. El propósito final es aplicar el conocimiento obtenido en áreas como la lingüística computacional, la enseñanza de idiomas, el procesamiento del lenguaje natural, la optimización del texto, y muchos otros.

Constituye, entonces, un campo emergente que ha ido, poco a poco, haciéndose un espacio dentro de la ciencia lingüística misma, pues cabe destacar que, a lo largo de más de cincuenta años, ha enfrentado la reticencia de los sectores más conservadores. Un claro ejemplo de ello es la analogía que, según Köhler (2005: 2), puede hacerse entre la lingüística cuantitativa y la lingüística computacional. Actualmente, esta última es considerada una disciplina "separada", un "tema aparte", que incluso en algunas universidades se encuentra en un departamento distinto de la "lingüística general". Ello no ocurre en otras ciencias. No existe, por

ejemplo, una "física computacional", porque el uso de computadoras en física es tan obvio que no se considera un procedimiento fuera del canon.

Existe consenso en que la lingüística cuantitativa es un fenómeno propio del siglo XX, aunque, claro está, no deja de tener antecedentes en los siglos inmediatamente anteriores.[6] Těšitelová (1992: 15) menciona el caso del pedagogo checo Jan Amos Komenský (1592- 1670), quien, en su obra Janua linguarum reserata (1631), demostró que conocer la frecuencia de aparición de las palabras en una determinada lengua puede ser útil para aprenderla como un segundo idioma. Reinhard Köhler, por su parte, menciona diversos trabajos hechos en el siglo XIX que, sin lugar a dudas, constituyen precedentes de los métodos y objetivos de la lingüística cuantitativa:

> *Erste Zählungen von Einheiten der Sprache oder von Texten wurden schon im vorigen Jahrhundert vorgenommen. In Deutschland waren es wahrscheinlich Förstemann (1846, 1852) und Drobisch (1866), in Russland Bunjakovskij (1847), in Frankreich Bourdon (1892), in Italien Mariotti (1880) und in den USA wohl Sherman (1888), die als erste zu dieser Methode als Mittel der sprachwissenschaftlichen Beschreibung gegriffen haben.* (Köhler 2005: 3).

Sin embargo, en general, se considera que el nacimiento oficial de la disciplina ocurre en 1913, cuando el matemático Andréi Márkov (1856-1922) publica en el Bulletin de l'Académie impériale des sciences de Saint-Petersbourg su estudio sobre la probabilidad de aparición de las distintas secuencias de letras en la novela en verso Evgenij Onegin (Eugenio Oneguin), de Alexander Pushkin. En este trabajo, dicho autor presentó las "cadenas de Márkov", las cuales constituyen la primera propuesta de tipo teórico, luego de muchos años de estudios meramente descriptivos o censales. Como se sabe, las "cadenas de Márkov", de amplia aplicación en diversos campos, ofrecen un modelo para calcular las probabilidades de transición de un elemento dentro de una secuencia de unidades

6 Pawłowski (2008) sugiere que es factible encontrar antecedentes tanto de la lingüística cuantitativa como de la lingüística de corpus en el trabajo hecho por los bibliotecarios alejandrinos durante el periodo helenístico. En particular, sugiere que procesos como la *esticometría* (medición del número de líneas que contenía un rollo o volumen) y la *colometría* (división de un verso en determinadas secciones o *κῶλα*) constituyen "precursores" del actual trabajo realizado por dichas disciplinas. A menos que mejore sus argumentos, dichas aseveraciones no parecen tener un sustento lo suficientemente adecuado como para ser aceptadas.

dependiendo solamente del elemento inmediatamente anterior.

Más adelante, el principal hito de la disciplina llegaría con el trabajo de George Kingsley Zipf (1902-1950), quien propuso un primer modelo matemático para la explicación del rango de frecuencia de aparición de las palabras en un corpus lingüístico; justamente, lo que hoy se conoce con el nombre de *ley de Zipf.* Según esta ley, en un corpus compuesto por lenguaje natural, la frecuencia de cualquier palabra es inversamente proporcional a su rango de frecuencia. De esta forma, la palabra más frecuente se produce aproximadamente el doble de veces que la segunda palabra más frecuente, tres veces más que la tercera y así sucesivamente. [7]

Posteriormente, tal y como explican Köhler y Rieger (1993: ix), la lingüística cuantitativa tendría un desarrollo vertiginoso que la llevaría a un periodo de "florecimiento" durante la década de los cincuenta, sobre todo en los países de Europa del Este. Dicho auge estuvo ligado a figuras como Paul Menzerath (1883–1954), fonetista alemán quien descubrió la ley que lleva su nombre (luego reformulada por Altmann), la cual establece que a mayor constructo lingüístico, menores serán sus partes constituyentes; Gustav Herdan (1897–1968), pionero en el uso de la estadística para testear leyes lingüísticas; Wilhelm Fucks (1902–1990), físico alemán que fue uno de los iniciadores de la lingüística cuantitativa en su país; Charles Muller (1909-2015), fundador de la estadística léxica en Francia cuyos modelos de estudio del vocabulario tuvieron amplia influencia; Raijmund G. Piotrowski (1922-2009), autor de la ley que lleva su nombre sobre el cambio de lenguas, y Juhan Tuldava (1922-2003), estoniano creador de modelos matemáticos para el estudio de fenómenos textuales; entre otros muchos.

Como puede verse, muchos de ellos pertenecían a otras disciplinas diferentes de la lingüística y llegaron por sus investigaciones a interesarse en el tema. En años recientes, las investigaciones en este campo han venido desarrollándose por especialistas dedicados de forma exclusiva a este ámbito. En los últimos años, entre la gran cantidad de investigadores dedicados a ello sobresalen Gabriel Altmann, Karl-Heinz Best, Michail V.

7 Desde un punto de vista matemático, dicha ley se puede enunciar de la siguiente forma: $f = k/r$, donde f es la frecuencia de una palabra en el corpus, r es su posición en el ranking de frecuencias y k es una constante próxima a uno que depende del corpus. Zipf, en su libro *Human behaviour and the principle of least effort* (1949), propone que la razón de dicho fenómeno se encuentra en la ley del mínimo esfuerzo, pues, al hablar, siempre es más fácil usar una palabra conocida que una menos conocida. El polaco Benoît Mandelbrot, en su obra *Information Theory and Psycholinguistics* (1966) reformuló la *ley de Zipf*, al considerar en la fórmula el costo de la comunicación de una palabra, en términos de la cantidad de letras y el espacio que las separa.

Arapov, Jurij K. Orlov, Werner Lehfeldt, Reinhard Köhler y muchos otros más.[8]

Tal y como explica Těšitelová (1992), la lingüística cuantitativa tiene subespecialidades según el dominio lingüístico que, de forma particular, estudie. Así pues, existen la estadística léxica, la estadística gramatical, la estadística semántica, estadística tipológica, etc. Uno de los campos de estudio más típicos de la lingüística cuantitativa es la estadística fonológica o fonoestadística. Esta se encarga de estudiar fenómenos como la frecuencia, distribución y relación de las unidades fonológicas (fonemas, sílabas, etc.) en un determinado conjunto de datos lingüísticos.

Para Altmann (2005: 191), los principales objetivos de esta disciplina se encaminan a generar hipótesis sobre tres aspectos principales: la frecuencia de fonemas o sílabas que conforman las palabras de un determinado idioma, la distribución combinatoria y posicional de los fonemas en las palabras[9] y las formas canónicas que construyen las clases de fonemas (como consonantes y vocales).[10] No obstante, como lo señala Simons (1977), sus técnicas de investigación también pueden utilizarse para cuantificar las diferencias fonológicas entre dos tipos de habla o, incluso, entre dos lenguas diferentes, con la finalidad de examinar la cercanía entre estas y clasificarlas de forma tanto diacrónica como tipológica.[11]

Para efectos de este trabajo, interesan de forma particular el conteo de fonemas en distribuciones de frecuencias, fenómeno que, según Strauss, Fan y Altmann (2008), constituye uno de los problemas típicos de la lingüística cuantitativa a nivel fonológico. Como es bien sabido, en cualquier idioma, el total de unidades fonológicas, como los fonemas o

8 Para un desarrollo específico y bien detallado de la historia de la lingüística cuantitativa, véanse los diversos estudios presentes en el volumen editado por Köhler, Altmann y Piotrowski (2005). De particular importancia son los incluidos en la primera parte del libro, pues constituyen reseñas de investigaciones hechas en diversos países de Europa del Este y Asia.

9 El trabajo más influyente sobre esta línea de trabajo es, sin lugar a dudas, el artículo publicado por Harary y Paper (1957) titulado "Toward a General Calculus of Phonemic Distribution", el cual describe la coocurrencia de los fonemas en términos de la teoría de grafos. Dicho trabajo parte de la idea de que, si f representa un determinado fonema, el conjunto de fonemas que lo precedan en un corpus determinado será P*(f)* y el conjunto de los que lo sucedan es S(*f*). Por ejemplo, en un enunciado como *"no les preguntes"*, P*(e)* consiste en *{l,r,t}*, mientras que S*(e)* es *{s,g}*. Lo anterior permite hacer diversos cálculos, como el grado de asociatividad de un fonema, es decir, la capacidad de un fonema de conectarse con otros fonemas predecesores o sucesores. Para más detalles sobre este tipo de estudios, consúltese el trabajo de Lehfeldt (2005).

10 Para más detalles, sobre este tipo de análisis, véase Altmann (2005).

11 Dos ejemplos de este tipo de estudios son Tambovtsev (2008), que explora posibles relaciones entre el ainu, un idioma genéticamente aislado hablado en Japón, y diversas familias y lenguas del mundo, y Tambovtsev (2010b), el cual propone una nueva ubicación del polaco dentro del grupo de lenguas eslavas a partir de la fonoestadística de dichos idiomas.

las sílabas, es un conjunto relativamente limitado.[12] Dichas unidades, entonces, se combinan, de acuerdo con ciertas reglas, para crear el léxico de la lengua. En dicho proceso, algunas formas son más recurrentes que otras. Este tipo de estudios, entonces, busca determinar, desde un punto estadístico, la frecuencia de aparición de dichos segmentos.

Básicamente, puede decirse que existen dos maneras diferentes de estimar la frecuencia de los fonemas de una lengua, las cuales Stefan Frisch explica de la siguiente manera:

> *Frequency is the rate of occurrence of a phonological unit, and is unrelated to acoustic frequency. But there are still many possible frequencies, depending on what is taken to be the domain over which occurrences are counted. In studies of language using corpora of language usage, frequency is usually the frequency of occurrence in the corpus. This type of frequency is referred to as token frequency or usage frequency. In English, for example, the token frequency of the phonemes /ð/ and /v/ is relatively high, due to their presence in frequently used words like the and that, and of and very. […] Abstracting away from repeated usages of a word, phonological patterns can also be examined on the basis of the number of times the pattern is used across different words. This frequency is referred to as type frequency or lexical frequency. The type frequency of the phonemes /ð/ and /v/ in English is relatively low, as they are used in relatively few words.* (Frisch 2011: 2138).

Como lo advierten Strauss, Altmann y Best (2006), este tipo de cálculos, al igual que los estudios de estadística léxica, datan del s. XIX. Inicialmente, se trataba más bien de cálculos de frecuencia de letras, los cuales son de mucha utilidad para estenógrafos, impresores, creadores de fuentes, etc. Hoy día, gracias a los avances llevados por la fonología, se tiene claro que se trata de fenómenos diferentes, pues las letras o grafemas no siempre equivalen a los fonemas de una lengua; consecuentemente, sus cálculos van a diferir. En palabras de Altmann (2008: 152), "letter frequency is not identical with

12 Para el caso de los fonemas, Crystal (1997: 170) indica que al rotokas, lengua hablada en Nueva Guinea, es la que posee el sistema fonológico más pequeño, pues solo dispone de seis consonantes y cinco vocales. En el otro extremo, se encuentran algunas lenguas joisanas, como el !xũ, las cuales, gracias a los clics o chasquidos, pueden llegar a tener más de noventa fonemas.

phoneme frequency, grapheme frequency or «letters + punctuation marks frequency» or «graphemes + punctuation marks frequency»".

Debido a esto, los cálculos directamente fonológicos no tuvieron su adecuado impulso sino hasta el siglo XX. De particular importancia para su desarrollo, fue el trabajo de los lingüistas del Círculo Lingüístico de Praga tales como Vilém Mathesius, Josef Vachek y Bohumil Trnka. Un miembro de este grupo, Nikolái Trubetskói, ofreció una primera panorámica general del tema en su obra póstuma Principios de fonología (1939). En un capítulo titulado justamente "De la estadística fonológica", Trubetskói señala que la estadística puede utilizarse en fonología con dos propósitos. Uno de ellos es saber cuántas veces se presenta una unidad fonológica en una lengua dada, el otro es determinar el rendimiento funcional de determinados segmentos.

De esa época datan, además, los primeros intentos de generar un modelo matemático que explique el rango de las frecuencias de los distintos fonemas. Dichos intentos se debieron a George Udny Yule y George Kingsley Zipf. Como se indicó antes, la frecuencia de ocurrencia de las palabras en un determinado corpus puede determinarse por medio de la ley de Zipf, anteriormente mencionada. Sin embargo, no ha sido posible encontrar una función matemática equivalente que explique la distribución de las frecuencias de los fonemas. De acuerdo con Strauss, Altmann y Best (2006), las investigaciones comprueban que la frecuencia de los fonemas constituye una función decreciente, pero las fórmulas utilizadas para derivarla son diversas. Estos autores mencionan métodos propuestos por diversos investigadores; entre ellos, Juhan Tuldava, George U. Yule, George K. Zipf, Vriddhachalam K. Balasubrahmanyan y Sundaresan Naranan, Gabriel Altmann y Peter Grzybek, los cuales constituyen buenas aproximaciones al fenómeno, aunque no constituyen fórmulas que, de momento, puedan considerarse definitivas.[13]

13 Para un enfoque más detallado sobre esta cuestión, véase justamente el artículo de Strauss, Altmann y Best (2006), el cual contiene una bibliografía muy completa sobre el tema en cuestión. Véanse también los trabajos de Martindale *et al.* (1996) y Tambovtsev y Martindale (2007), los cuales sugieren que mejor ecuación para describir la distribución de los fonemas es la de Yule.

2.1.2. Sobre la lingüística de corpus

Es un hecho ampliamente reconocido que la lingüística cuantitativa se encuentra "*ligada metodológicamente con la lingüística del corpus"* (Terrádez Gurrea 2001: 39). Esto ya lo hacía ver Geoffrey Leech hace más de dos décadas, el cual explicaba las razones que motivaban esta conexión de la siguiente manera:

> t*he revival of corpus linguistics in the 1980s has demonstrated a close connection between "corpus linguistics" and "quantitative linguistics". The connection is made in both directions: (a) If we have a large computer corpus, one of the most obvious things we can do with it is to derive frequencies. For example, we can calculate relative transitional frequencies, and thus to set up a simple statistical model of how, at one level, language works (a Markov process model). (b) If one wishes to set upv a quantitative model of language performance, the most obvious requirement is a corpus from which quantitative counts can be made: we cannot rely on the Chomskyan native speaker's intuition for probabilities.* (Leech 1992: 110).

La lingüística de corpus constituye, entonces, una herramienta de primera línea para llevar a cabo estudios de lingüística cuantitativa; pero, ¿qué es exactamente la lingüística de corpus? Como suele señalarse, no es tarea sencilla ofrecer una respuesta concreta a esta pregunta. Para algunos es una disciplina, para otros es una metodología, hay quien lo considera un "paradigma" e incluso quienes proponen que es un "enfoque" (*approach*) o "teoría".[14] En términos generales, puede decirse que la lingüística de corpus se dedica a la constitución y explotación de recursos lingüísticos escritos y orales mediante el uso de un corpus electrónico. Obviamente, para poder llevar a cabo dicha tarea, ha desarrollado un complejo conjunto de principios teóricos y metodológicos que, en cierta medida, sería injusto titular como una mera y simple "metodología" de

14 Para más detalles sobre esta variedad de opiniones, véase Taylor (2008), quien hace un elente recorrido sobre la amplia bibliografía que existe alrededor de esta polémica.

trabajo. Aun así, muchos estudios suelen definirla de esta forma. Por citar un caso, Pintzuk (2011: 231), al conceptualizar la lingüística de corpus, señala lo siguiente: "*This term refers to linguistic research that uses corpus data as the primary object of study. The term, therefore, describes a methodology rather than a field of linguistics".* Similar es la definición que aporta Parodi (2010: 14), el cual manifiesta que esta *"constituye un conjunto o colección de principios metodológicos para estudiar cualquier dominio lingüístico y se caracteriza por brindar sustento a la investigación de la lengua en uso a partir de corpus lingüísticos con sustrato* [sic] *en tecnología computacional y programas informáticos ad hoc".*

En todo caso, a pesar de la diversidad de criterios con respecto a cómo definirla, no cabe la menor duda de que, como bien señala el autor anterior, existe cierta claridad sobre los principios y características que debe reunir un estudio para ser ubicado dentro del ámbito de trabajo de la lingüística de corpus. Los cuatro rasgos que se consideran típicos son los que se detallan a continuación:

1. *La LC [lingüística de corpus] es empírica, ya que se analizan patrones de uso lingüístico real en textos naturales*
2. *La LC utiliza una amplia y organizada colección de textos naturales como base del análisis, entendida como un corpus*
3. *La LC hace uso de los computadores para procesamientos y análisis, con base en técnicas automáticas e interactivas*
4. *La LC depende tanto de técnicas analíticas de tipo cuantitativo como cualitativo.* (Parodi 2010: 38)

Como puede verse, independientemente de la definición que se adopte, existe claridad sobre el hecho de que la lingüística de corpus básicamente dirige su campo de acción a la compilación y análisis de corpus mediante el uso de herramientas de tipo informático (Kennedy 1998: 1). Cabe, entonces, preguntarse qué exactamente es un corpus, noción sobre la que tampoco existe consenso.[15] Una definición clásica es la brindada por Sinclair (2005: 16), quien propone delimitar dicho concepto de la siguiente forma: "*A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as*

15 Sobre las discrepancias a la hora de definir corpus, véase Parodi (2010: 20-25).

far as possible, a language or language variety as a source of data for linguistic research." Como puede verse los puntos centrales de la definición son que el corpus se elabora con datos lingüísticos reales (o "lenguaje natural"), que este se codifica de forma electrónica mediante el uso de computadores y que debe tener algún tipo de diseño para asegurar su utilidad según los objetivos y el tipo de investigación lingüística por realizar.

Tal y como explican McEnery y Wilson (2001: 103-130), los corpus tienen una gran cantidad de aplicaciones dentro de los estudios lingüísticos. Pueden utilizarse prácticamente en todas las subdisciplinas de la lingüística.[16] Entre estas, pueden citarse la lexicografía y los estudios léxicos y semánticos, la pragmática y el análisis del discurso, la sociolingüística, la dialectología y los estudios sobre variación y cambio lingüístico e, incluso la lingüística diacrónica.[17] Su importancia radica en que aportan una base empírica a las investigaciones, lo cual permite sustentar con "datos reales" las hipótesis en lugar de tener que recurrir a percepciones de tipo individual. A ello cabría agregar que, gracias al soporte electrónico en que se encuentran codificados, los corpus facilitan enormemente el manejo de los datos lingüísticos.

Obviamente, no siempre los estudios lingüísticos se basan en un corpus electrónico. Se puede entender, por lo tanto, que existen dos tipos de investigaciones lingüísticas en la actualidad: aquellas que utilizan un corpus electrónico y aquellas que no lo hacen.[18] Los estudios que utilizan corpus, a su vez, han sido divididos por Tognini-Bonelli (2001) en dos grandes grupos: los enfoques corpus-based y los corpus-driven. Los primeros utilizan los corpus de forma auxiliar, es decir, como fuente para obtener datos que ejemplifiquen o justifiquen sus hipótesis y descripciones acerca de algún determinado fenómeno lingüístico. Los

16 Tal y como lo mencionan Rafel i Fontanals y Soler i Bou (2003: 46), los corpus también son un recurso fundamental para el desarrollo de diversos productos lingüísticos y tecnologías computacionales. Es el caso de los diccionarios electrónicos, los sistemas de conversión de texto en voz y las aplicaciones de reconocimiento automático del habla, entre otros muchos.

17 Para un tratamiento a profundidad sobre la utilidad y aplicación de corpus en las diferentes subdisciplinas de la lingüística, véanse los dos volúmenes del manual editado por Lüdeling y Kytö (2008).

18 En lingüística, se le llama corpus no solo a textos o partes de textos recopilados según un criterio y codificados electrónicamente para un determinado fin; de forma más general, también se le llama corpus a cualquier tipo de datos lingüísticos (sílabas, palabras, oraciones, textos, etc.) reunidos para realizar una determinada investigación (al respecto, véase Alcaraz y Martínez 1997: 151).

segundos, por su parte, se constituyen sobre la base de los datos mismos; debido a esto, las teorías acerca del lenguaje que proponen las obtienen, a través de procedimientos deductivos, de las evidencias dadas por el corpus. Tony McEnery y Andrew Hardie lo explican de la siguiente forma:

> *Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, typically one established in the current literature, in order to validate it, refute it or refine it. The definition of corpus linguistics as a method underpins this approach to the use of corpus data in linguistics. Corpus-driven linguistics rejects the characterization of corpus linguistics as a method and claims instead that the corpus itself should be the sole source of our hypotheses about language. It is thus claimed that the corpus itself embodies its own theory of language.* (McEnery y Hardie 2012: 6).

De acuerdo con estos mismos autores (McEnery y Hardie 2012: 29), los corpus suelen contener tres tipos de datos adicionales: los metadatos, que brindan información sobre el texto mismo (por ejemplo, nombre de quién dijo o escribió el texto o quién lo recopiló, fecha, género discursivo, etc.); el marcaje textual, que se refiere al sistema de codificación del texto en formato electrónico (este tipo de información permite, por citar un caso, identificar los turnos de los hablantes en una conversación), y la anotación lingüística, que consiste en agregar "etiquetas" que contienen datos de índole lingüístico y gramatical como el tipo de frase o sintagma constituido por un grupo de palabras.

Así pues, un corpus puede consistir en una simple recopilación de datos lingüísticos hecha con criterios que aseguren su representatividad, sin añadir ningún otro tipo de información. Sin embargo, suele ser una práctica común y muy útil agregarles a los textos que conforman el corpus algún tipo de anotación o etiquetado. Obviamente, la información que se le agregue puede ser de diverso tipo, según los propósitos que se ha planteado la investigación. Puede agregarse, por ejemplo, información sobre los rasgos prosódicos si se trata de un corpus de lenguaje oral. Existen también etiquetados de las funciones pragmáticas para hacer estudios sobre el uso de determinadas expresiones. Sin embargo, el etiquetado más usual es la asignación del tipo de palabra (*part-of-speech*

tagging), la cual consiste en asignar a cada ítem del corpus la unidad o categoría léxica correspondiente (verbo, sustantivo, preposición, etc.).

Ahora bien, ¿qué tipo de análisis suelen hacerse sobre los corpus? Según Evison (2010), los procedimientos más comunes que se suelen llevar a cabo son básicamente tres: la generación de listas de frecuencia, las cuales consisten en una lista completa de las palabras presentes en el corpus ordenadas según el rango de aparición (como se sabe, la gran mayoría de las palabras aparece solamente una vez); la identificación de las palabras clave (*key words*) de un texto, que no son necesariamente las palabras más frecuentes, sino aquellas que identifican estadísticamente un corpus de otro, y la creación de concordancias (también llamadas *key word in context* o KWIC), las cuales consisten en líneas de palabras generadas automá-ticamente que muestran el contexto lingüístico de un determinado ítem (letra, sílaba, palabra, frase, etc.), es decir, el conjunto de secuencias lingüísticas que la suceden y la preceden. A ellos, Parodi (2010: 44-45) agrega, como último procedimiento típico, la búsqueda de colocaciones o unidades fraseológicas, las cuales son definidas por él como la "*aparición simultánea de dos o más palabras en un segmento de texto*".[19]

Como puede verse, la lingüística de corpus es un ámbito de trabajo con amplias posibilidades que a lo largo de las últimas décadas ha venido desarrollándose y posicionándose de forma vertiginosa dentro de los estudios lingüísticos. Ha abierto, además, nuevos métodos de indagación y nuevos enfoques para el estudio de viejos problemas. Vale la pena, entonces, adentrarse en las posibilidades que ofrece.

2.2. Recolección, procesamiento y análisis de los datos

2.2.1. Conformación y características del corpus

Los datos empíricos utilizados en esta investigación proceden de un corpus electrónico de datos lingüísticos del cabécar, creado por medio del programa computacional denominado *Fieldworks Language Explorer (FLEx)*, el cual tiene por objetivo ofrecer una plataforma para

19 Se han referido aquí los procedimientos más comunes que utiliza la lingüística de corpus, los cuales están disponibles en todos los paquetes informáticos comúnmente utilizados por los investigadores. Claro está, que existen muchas otras posibilidades de indagación, las cuales, como explica Parodi (2010: 45) son "*tan diversas y versátiles como el programa con que se cuente lo permita*".

la documentación, organización y análisis de información de índole lingüística.[20] Actualmente, dicho programa informático constituye uno de los recursos más útiles en lo que respecta a herramientas de trabajo para los lingüistas, pues posee una gran cantidad de funcionalidades; por ejemplo, permite la gestión y creación de diccionarios, realizar el análisis interlineal de un texto, analizar morfológicamente las palabras y anotarlas (*tagging*), así como la creación de listas de frecuencia y concordancias, entre otras posibilidades.[21]

El corpus incluye todos los materiales disponibles sobre la lengua cabécar que ha sido posible ubicar y digitalizar. En este sentido, se trata de lo que Torruella y Llisterri (1999: 55) denominan un corpus oportunista, pues se construye sin utilizar un criterio de selección, sino atendiendo únicamente a la disponibilidad del material lingüístico. Tal y como explican estos autores, esto se ha hecho debido a la escasez de documentación lingüística del idioma cabécar y la imposibilidad material de generarla.[22] Como es bien sabido, la mayoría de las lenguas indígenas de América se encuentran muy mal documentadas y el cabécar no es una excepción. Por ello, se consideró pertinente recoger todo lo que se ha producido con respecto a este idioma, digitalizarlo e incluirlo en el corpus electrónico.

El uso de corpus oportunistas para el estudio de lenguas minoritarias, en peligro de extinción o ya extintas por completo, ha sido abordado por McEnery y Hardie (2012: 11-13). Dichos autores señalan que solo las lenguas oficiales de las naciones suelen tener el aparato estatal y financiero necesario para generar suficientes materiales lingüísticos como para construir con corpus según las expectativas de "más grande, mejor". Muchas lenguas, como el cabécar, tienen pocos hablantes que producen pocos materiales textuales. Otras ya no poseen hablantes, como el latín o el sánscrito. En dichos casos, no queda más que trabajar con lo que se tenga a mano, pues, como ellos mismos indican: "*A researcher must at times be guided by pragmatism.*

20 Dicha aplicación informática es de libre distribución y se encuentra disponible en la siguiente dirección web: http://fieldworks.sil.org/flex/.

21 Para más detalles sobre el *Fieldworks Language Explorer*, véanse las valoraciones hechas por Butler y van Volkinburg (2007) y Rogers (2010).

22 Sobre este aspecto, téngase presente, a modo de ejemplo, que la transcripción de la historia publicada por Obando y González (2015), cuya duración es de 36 minutos (en total está constituida por 4.290 palabras), demoró un año completo. De esta forma, mientras no exista un equipo nativo bien capacitado en la transcripción de su propia lengua, la recolección de un corpus diseñado *ad hoc* seguramente consumiría años de trabajo.

[…] *it should be noted, and accepted, that the corpora that we use and construct must sometimes be determined by pragmatic considerations*".

Ciertamente, se comprende las limitaciones de este tipo de recopilación, especialmente en cuanto a representatividad se refiere. Sin embargo, el tema de la investigación acá propuesta (la frecuencia de aparición de los fonemas de la lengua cabécar) hace pensar que dicha limitación no incidirá de forma significativa en los resultados. Para dichas estimaciones, no existen, hasta donde sabemos, especificaciones determinadas, al menos de tipo externo, con la excepción de la inclusión de textos poéticos que, por razones estilísticas, podrían presentar desviaciones significativas en la frecuencia de ciertos fonemas. A parte de esto, las únicas recomendaciones hechas con respecto al recuento de fonemas tienen que ver con el tamaño de la muestra. En su análisis cuantitativo de los grafemas del checo, Marie Königová (citada por Těšitelová 1992: 156) hace las siguientes recomendaciones:

- Para los grafemas más frecuentes, es suficiente un corpus de 7 000 a 8 000 grafemas.
- Para los grafemas de frecuencia media, se requiere un corpus de al menos 20 000 grafemas.
- Para los grafemas de baja frecuencia, el corpus debe tener entre 150 mil y 170 mil grafemas.
- Finalmente, para grafemas sumamente extraños se requiere un corpus de unos 4 millones de grafemas.

Tal y como se explicará con mayor detalle más adelante, el corpus de esta investigación posee un total de 164.011 fonemas, lo cual lo ubica dentro de un rango apto para recoger incluso elementos de baja frecuencia.

Con respecto al etiquetamiento, es importante mencionar que el corpus en cuestión es del tipo que se denomina *etiquetado o anotado*. De esta forma, al "texto plano" se le han añadido las siguientes especificaciones: análisis morfológico, especificación del lexema, glosa del lexema, categoría gramatical y glosa de la palabra, tal y como se puede apreciar en la siguiente imagen:

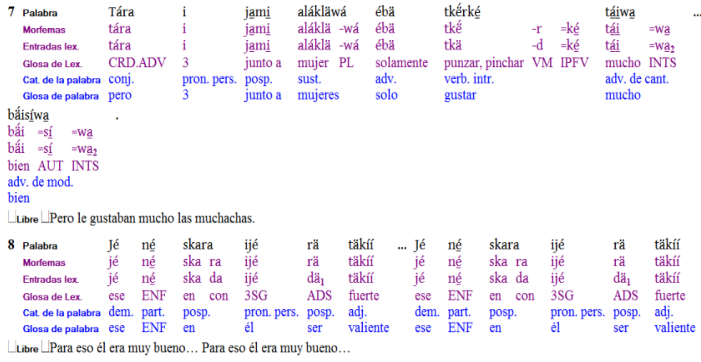


Imagen 1. Muestra del modelo de etiquetamiento utilizado en el corpus electrónico.

Obviamente, ello se ha hecho con el fin de enriquecer el corpus y poder darle, en un futuro, otras aplicaciones, pues la investigación aquí propuesta no requiere de dicha información.

Tal y como se dijo antes, el corpus contiene casi todo el material lingüístico sobre la lengua en cuestión que ha sido factible reunir y digitalizar, el cual está organizado en dos subcorpus, según el origen dialectal de los datos lingüísticos. Tal y como explican Biber y Jones (2009: 1301), el uso de un subcorpus resulta pertinente cuando se desea hacer estimaciones separadas en secciones de un corpus general con el fin de llevar a cabo luego una comparación. Es el caso de los corpus que tratan por separado el discurso oral y el discurso escrito. En nuestro caso, resulta pertinente debido a que, como se explicó antes, el cabécar sureño posee cuatro fonemas menos que el cabécar norteño, lo cual obliga a hacer cálculos separados para cada variedad lingüística.

Así pues, el corpus utilizado en esta investigación consta de dos subcorpus. El primero de ellos, que se denominará "A", comprende los materiales pertenecientes a la variedad dialectal hablada en la región de Chirripó, la cual como se dijo pertenece al cabécar norteño.[23]

23 El cabécar norteño también está compuesto por la variedad lingüística hablada en el Valle de La Estrella. De esta variedad, sin embargo, no existen suficientes materiales lingüísticos. Solamente, hay un libro publicado que contiene algunas historias y tradiciones. Se trata de la recopilación de textos hecha por Morales (2013), la cual, desdichadamente, está tan mal escrita desde un punto de vista ortográfico, que para poder ser utilizada requiere de una revisión y cotejo sumamente exhaustivos. Por ello, no fue incluido en esta investigación.

En específico los componentes de este subcorpus son los siguientes:

1. Corpus de oraciones obtenidas por educción (*elicited sentences*). Incluye datos obtenidos de dos fuentes diferentes:
    1.1. Conjunto de datos lingüísticos recopilados para la elaborar el *Diccionario Escolar del Cabécar de Chirripó*, el cual fue reunido por el autor de esta investigación. Se compone de 3.265 enunciados lingüísticos.
    1.2. Conjunto de datos lingüísticos recopilados en el marco del proyecto *Recopilación y análisis de material lingüístico* para el *estudio de la morfosintaxis del cabécar*, el cual fue reunido por el autor de esta investigación. Lo conforman 528 oraciones que ilustran diferentes aspectos gramaticales de la lengua.
2. Tres historias pertenecientes al género narrativo recopiladas por el autor de esta investigación:
    2.1. Historia de los Niños Huracanes (*Sẽrikẽ pákẽ*), recopilada de forma escrita.
    2.2. Historia del clan Kätsúibawák (*Kätsúibawák pákẽ*), relato oral de 36 minutos. Publicada por Obando y González (2015).
    2.3. Historia de Yẽbulẽ (*Yẽbulẽ pákẽ*), relato oral de 8 minutos.
3. Tres textos pertenecientes a la modalidad discursiva explicativa:
    3.1. Bukulú (‘*Primer embarazo*’), texto escrito que narra el régimen de vida que deben seguir los hombres y las mujeres cuando van a tener su primer hijo.
    3.2. El origen de los clanes, texto oral que explica cómo se establecieron los clanes del pueblo cabécar.
    3.3. La cultura cabécar de Chirripó (*Séjé duchíwák kq̱sẹ́nẹ́wq́*), texto publicado por Barquero Reyes et al. (2008) que presenta el modo de vida del pueblo cabécar.
4. Recopilación de diálogos destinados a lav enseñanza del cabécar como segunda lengua:
    4.1. Diálogos recogidos en el disco compacto *Curso de*

lengua y cosmovisión cabécar, aplicación multimedia. Sobre este material, consúltese el trabajo de Brenes Granados (2007), autor de dicha obra.

4.2.Diálogos recopilados por el autor de esta investigación para el desarrollo de un curso de cabécar como segunda lengua.

5. Un pequeño conjunto de textos misceláneos que recoge contenidos tomados de carteles, brochures, anuncios, etc., recopilados por el autor de esta investigación.
6. Una sección que reúne las listas de palabras, la cual, de momento, solo está compuesta por una enumeración de los topónimos que se encuentran en la región indígena de Chirripó.

Como puede verse, se ha abarcado prácticamente todo el material disponible. Únicamente se descartó dos obras. La primera de ellas es la colección de historias publicadas por Fernández Torres (2011). Estas no fueron incluidas debido a que fueron escritas por una persona que habla el dialecto sureño utilizando las formas de escritura de la variedad norteña. Esto ha provocado una cantidad enorme de inconsistencias. Por ejemplo, las consonantes aspiradas se transcriben de forma muy vacilante y no pocas palabras se escriben de dos o tres formas debido a esto. También se descartó utilizar la traducción del Nuevo Testamento al cabécar publicada recién-temente. Esta decisión se debió también a la inconsistencia ortográfica con que está escrito, así como a la enorme cantidad de palabras no pertenecientes al cabécar que este texto posee (se trata, sobre todo, de nombres propios de Palestina), los cuales evidentemente pueden inclinar el conteo de fonemas hacia un resultado no representativo de la realidad lingüística del cabécar.

Dado lo anterior, todo este material que compone el subcorpus del cabécar de Chirripó en total posee un volumen de 31.196 palabras, las cuales se distribuyen tal y como lo muestra la siguiente tabla:

	Cantidad de formas (*types*)	Cantidad de palabras (*tokens*)	%
1. Oraciones educidas	**3 469**	**22 049**	**69,08%**
1.1. Oraciones para el diccionario	3 336	18 880	59,16%
1.2. Oraciones para estudio de la morfosintaxis	530	3 169	9,93%
2. Narraciones	**832**	**5 274**	**16,52%**
2.1. *Sẽrikẽ pákẽ*	129	243	0,76%
2.2. *Kätsúibawák pákẽ*	672	4 290	13,44%
2.3. *Yẽbulẽ pákẽ*	244	741	2,32%
3. Textos explicativos	**548**	**1 872**	**5,87%**
3.1. *Bukulú*	237	520	1,63%
3.2. *El origen de los clanes*	155	668	2,09%
3.3. *La cultura cabécar de Chirripó*	270	684	2,14%
4. Diálogos	**551**	**1 832**	**5,74%**
4.1. Diálogos del curso multimedia	508	1 532	4,80%
4.2. Diálogos recopilados para enseñar cabécar	104	300	0,94%
5. Textos misceláneos	**260**	**596**	**1,87%**
6. Listas de palabras	**173**	**293**	**0,92%**
6.1. Topónimos de Chirripó	173	293	0,92%
TOTAL	**4 255**	**31 916**	100%

Tabla 1. Componentes del subcorpus A de cabécar de Chirripó.

El segundo subcorpus, que se identificará aquí como "B", reúne todos los materiales existentes sobre el cabécar sureño, el cual, como se dijo, se habla en Talamanca y Ujarrás. A diferencia del anterior, donde predomina material recopilado por el autor de esta investigación, en este caso la mayoría son textos tomados de publicaciones; por ello, la mayoría de los textos son de tipo narrativo. Este subcorpus se compone entonces de los siguientes componentes:

1.Corpus de oraciones obtenidas por educción (*elicited sentences*). Incluye datos obtenidos de dos fuentes diferentes:

1.1.Un conjunto de 506 oraciones recopiladas por David Bourland (1974) en Ujarrás para estudiar la morfosintaxis del cabécar.

1.2. Conjunto de datos lingüísticos recopilados en el marco del proyecto *Recopilación y análisis de material lingüístico para el estudio de la morfosintaxis del cabécar*, el cual fue reunido por el autor de esta investigación. Lo conforman 90 oraciones que ilustran diferentes aspectos

gramaticales de la lengua.

2. Un conjunto de textos de tipo narrativo publicados por diversos autores:
    - 2.1.Todos los textos recopilados y publicados por Stone (1961) en Ujarrás.
    - 2.2.Cuatro pequeñas narraciones orales transcritas y publicadas por Margery (1986).
    - 2.3.Las cinco historias publicadas por Varas y Fernández (1989), textos pertenecientes al registro escrito.
    - 2.4. El texto sobre *Sibö*, principal personaje de la religión cabécar, el cual fue transcrito y publicado por Margery (1995).
    - 2.5.Una pequeña historia recopilada y transcrita por Quesada y Lehmann (2007).
    - 2.6.Tres pequeñas narraciones utilizadas en la enseñanza escolar en Ujarrás recopiladas por Lehmann (2010) e incluidos en su página web.
    - 2.7.Una historia inédita de Severiano Fernández, texto perteneciente al registro escrito.
3. Una recopilación de textos pertenecientes a la modalidad discursiva explicativa reunidos por el autor de esta investigación:
    - 3.1.Tres textos pertenecientes al registro escrito producidos por estudiantes cabécares en el marco de un curso universitario dictado por el autor de esta investigación.
    - 3.2.Dos textos de tipo oral recopilados, transcritos y analizados por el autor de esta investigación.
    - 3.3.El texto publicado por Calderón Saravia (1996).

Como puede apreciarse, el material es considerablemente menor con respecto al reunido para el primer subcorpus. De hecho, en total está conformado por 14.039 cuya distribución se presenta en la siguiente tabla:

	Cantidad de formas (*types*)	Cantidad de palabras (*tokens*)	%
1. Oraciones educidas	**571**	**2 891**	**20,59%**
1.1. Oraciones de Bourland (1974)	464	2 498	17,79%
1.2. Oraciones para estudio de la morfosintaxis	161	393	2,80%
2. Narraciones	**1 981**	**7 495**	**53,39%**
2.1. Textos de Stone (1961)	516	1 596	11,37%
2.2. Textos de Margery (1986)	200	531	3,78%
2.3. Textos de Fernández y Varas (1989)	946	2 575	18,34%
2.4. Texto de Margery (1995)	302	1 116	7,95%
2.5. Historia de Quesada y Lehmann (2007)	101	185	1,32%
2.6. Narraciones de Lehmann (2010)	196	387	2,76%
2.7. Narración inédita de S. Fernández	353	1 105	7,87%
3. Textos explicativos	**1 313**	**3 653**	**26,02%**
3.1. Textos explicativos escritos	625	1 297	9,24%
3.2. Textos explicativos orales	180	588	4,19%
3.3. Texto de Calderón (1996)	656	1 768	12,59%
TOTAL	**3 366**	**14 039**	100%

Tabla 2. Componentes del subcorpus B de cabécar sureño.

Pudiera suponerse que ambos conjuntos son sumamente disímiles y, por ello, resultan inválidos para un análisis estadístico. Sin embargo, una comparación de la frecuencia léxica de ambos subcorpus nos permite considerar que esta idea no es correcta, pues ambos presentan similitudes evidentes. Así, un conteo de palabras (que no es el tema de esta investigación, pero que permite verificar la similitud entre ambas secciones del corpus) revela que, tal y como suele suceder en las lenguas del mundo, las posposiciones y pronombres, es decir las palabras funcionales, son las de mayor ocurrencia.

Aunque con porcentajes de aparición diferentes, en ambas secciones del corpus el pronombre *i* 'tercera persona' es el elemento léxico que más aparece, mientras que el segundo es la posposición *te/të*, partícula que marca el caso ergativo. Luego aparecen la posposición *dä/dö*, elemento morfológico que aparece en las oraciones copulativas; *jé/é*, un pronombre demostrativo con funciones gramaticales específicas; los pronombres *yís* 'yo' y *sá* 'nosotros inclusivo' y la partícula de negación *ká*.

El único elemento léxico en que difieren ambos subcorpus es la aparición, en el conjunto del cabécar sureño, de un sustantivo entre las primeras diez palabras más frecuentes. Se trata del nombre propio Sibö, el principal personaje de la religión cabécar. Ello, sin lugar a dudas, se debe a la gran cantidad de textos narrativos (más del 50 %) que conforman el material disponible para este dialecto. Obviamente,

al ser este el principal personaje, es evidente que juega un papel muy relevante en las historias y, por ello, aparece muchas más veces.

2.2.2. Fases de la investigación

La investigación llevada a cabo constó de tres etapas, cuyos pormenores se explican a continuación:

a) Digitalización de los textos y conformación del corpus electrónico

La primera parte de la investigación consistió en obtener los datos del cabécar por medio de consultas bibliográficas. Una vez recopilados todos los materiales, se procedió con la digitación de los textos, pues, con la excepción de los materiales recopilados por el autor de la investigación, todos los demás debían digitalizarse.

Para ello, se optó por transcribirlos utilizando un procesador de textos y la herramienta informática denominada Teclado Chibcha, la cual fue desarrollada por Flores Solórzano (2010) para poder escribir con facilidad las marcas diacríticas utilizadas por las lenguas indígenas de Costa Rica. Esta etapa fue la más ardua y la que consumió más tiempo, sobre todo, porque muchos textos tenían diversas inconsistencias ortográficas, las cuales debieron ser subsanadas. La utilización de dicha herramienta informática luego implicó un proceso adicional en la transformación del texto a su representación fonológica, pues el Teclado Chibcha apila los diacríticos y codifica cada uno de ellos como un carácter separado de Unicode. De esta forma, un grafema como *é* constituye, desde un punto de vista informático, la unión de tres caracteres diferentes: *e* + ¨ + ´.

Una vez que todos los textos estuvieron digitados, se procedió con la creación del corpus electrónico mediante el programa informático *FieldWorks Language Explorer* (FLEx). Se crearon, como se explicó antes, dos bases de datos separadas o subcorpus, una para cada dialecto del cabécar, con el objetivo de generar una estadística comparativa entre las diferentes variedades lingüísticas de este idioma.

b) Diseño e implementación del modelo de transcripción fonológica automático

Una vez digitados los textos y conformado el corpus, se procedió con el diseño de una herramienta que trasladara los datos textuales del cabécar, que en ese momento estarían escritos según la ortografía tradicional, a la escritura fonológica, es decir, a los signos gráficos propios del Alfabeto Fonético Internacional (IPA, por su nombre en inglés). Según Ríos Mestre (1999), este tipo de aplicaciones se denomina *transcriptor* o *fonetizador* y su diseño depende más que nada de la relación que existe entre los grafemas y los fonemas de la lengua que se transcribe. En este sentido, el cabécar constituye una lengua de fácil procesamiento, pues como se verá, casi todos los grafemas se corresponden con un único fonema.

Bisani y Ney (2008) hacen un recuento de las técnicas utilizadas para la conversión de grafemas a fonemas. De acuerdo con ellos, existen tres tipos principales: transcripción por diccionario (*dictionary look-up*), transcripción por reglas (*rule-based*) y transcripción basada en datos (*data-driven*). La primera es utilizada para lenguas como el inglés y consiste en asociar cada palabra a una pronunciación específica y luego hacer el cambio correspondiente en el corpus. Tiene el defecto de que requiere tener la pronunciación establecida de cada forma léxica en el corpus. En la última técnica, se utilizan modelos estadísticos con el fin de desarrollar un programa informático que, dados una cantidad de ejemplos suficientes, pueda predecir la pronunciación de las palabras por pura analogía. Tiene el problema de que aún no existe un algoritmo estandarizado para su implementación. El segundo modelo es el más tradicional y consiste en generar reglas de cambio que, como indica Ríos Mestre (1999), serán más eficaces y económicas en la medida en que la lengua utilice el principio fonémico de representación ortográfica. Este es justamente el caso del cabécar y, por ello, esta fue la técnica utilizada.

Al respecto, debe tenerse en cuenta que el alfabeto cabécar consta de 28 letras y varios dígrafos, casi todos ellos son monofonemáticos, es decir, representan siempre a un único fonema o alófono. En resumen, los valores fonológicos de los grafemas y los dígrafos del cabécar son los incluidos en la siguiente tabla:

LETRA O DÍGRAFO	VALOR FONOLÓGICO
a	Representa al fonema /a/.
ä	Representa al fonema /ɤ/. Representa al sonido [ə], alófono de /a/.[24]
a̲	Representa al fonema /ã/.
b	Representa al fonema /b/.
ch	Representa la unión de los sonidos /t/ y /ʃ/.
d	Representa al fonema /d/.
e	Representa al fonema /e/.
ë	Representa al fonema /ɪ/
e̲	Representa al fonema /ẽ/.
g	Representa al sonido [g], alófono de /k/.[25]
i	Representa al fonema /i/.
i̲	Representa al fonema /ĩ/.
j	Representa al fonema /h/.
k	Representa al fonema /k/.
kj	Representa al fonema /kʰ/.[26]
l	Representa al fonema /l/.
m	Representa al sonido [m], alófono de /b/
n	Representa al sonido [n], alófono de /d/. En posición de coda, representa al sonido [ŋ].[27]
ñ	Representa al sonido [ɲ], alófono de /ʤ/.
o	Representa al fonema /o/.
ö	Representa al fonema /ʊ/.
o̲	Representa al fonema /õ/.
p	Representa al fonema /p/.

24 La letra *ä* solo se utiliza en el cabécar norteño.
25 La letra *g* solo se utiliza en el cabécar sureño.
26 El dígrafo *kj* solo se utiliza en el cabécar norteño.
27 El estatus fonemático del sonido [ŋ] es controvertido en cabécar. Margery (1989) lo consideró un fonema, pero dicha consideración viola los universales fonológicos propuestos para las consonantes nasales. En nuestra propuesta de análisis fonológico (González Campos 2011: 14-15), se considera un sonido consonántico epentético alofónico.

pj	Representa al fonema /ph/.[28]
r	A inicio y final de palabra, representa al sonido [r], alófono de /d/. Entre dos vocales, representa al sonido [ɾ], alófono de /d/.
rr	Entre dos vocales, representa al sonido [r], alófono de /d/.
s	Representa al fonema /s/.
sh	Representa al fonema /ʃ/.
t	Representa al fonema /t/.
tj	Representa al fonema /th/.[29]
u	Representa al fonema /u/.
u̲	Representa al fonema /ũ/.
w	Representa al sonido [w], alófono de /u/. Entre dos vocales, representa [β̞], alófono de /b/.
y	Representa al fonema /ʤ/.

Tabla 3. Valores fonológicos de las letras y dígrafos del cabécar.

Como puede apreciarse, la ortografía del cabécar sigue en gran medida el principio fonémico, el cual propone una correspondencia única entre cada signo y cada uno de los fonemas de la lengua. Dicho principio solo se ve roto por dos situaciones. En unos pocos casos, un fonema se representa por dos o más signos, como /d/, que puede representarse por las grafías *d*, *n* y *r* o el dígrafo *rr*. Esta inconsistencia se debe a que el alfabeto cabécar busca parecer al máximo al abecedario español con el fin de facilitar la alfabetización de la población. En todo caso, esta violación al principio fonémico no ofrece gran problema, pues una sencilla regla de transformación la subsana con facilidad. El segundo tipo de inconsistencia es más difícil de solucionar. Se trata de tres signos difonemáticos, los cuales, como la *g* del español, poseen valores fonológicos diferentes que dependen de ciertas particularidades fonéticas. Se trata de las letras *ä*, *n* y *w*, que, como se verá a continuación, requirieron un tratamiento diferenciado a la hora de hacer la transcripción automática de la ortografía a la representación fonológica.

28 El dígrafo *pj* solo se utiliza en el cabécar norteño.
29 El dígrafo *tj* solo se utiliza en el cabécar norteño.

Debido a lo anterior, se procedió a diseñar un transcriptor basado en reglas, que convirtiera los grafemas y dígrafos del cabécar en una representación fonológica susceptible de ser cuantificada. La mayoría de las reglas consisten en simples cambios que deben hacerse sin importar el orden en que se hagan[30], las cuales puede apreciarse en la siguiente tabla:

REGLA	INPUT	OUTPUT
1	kj	q
2	pj	f
3	pj	θ
4	ch	tʃ
5	sh	ʃ
6	rr	d
7	r	d
8	g	k
9	m	b
10	ñ	y
11	a̱	ã
12	ë	ɪ
13	e̱	ẽ
14	i̱	ĩ
15	ö	ʊ
16	o̱	õ
17	u̱	ũ

Tabla 4. Reglas de cambio para la transcripción fonológica automática.

Como puede verse, las consonantes aspiradas debieron convertirse a un signo simple y no a su representación fonética debido a que, de lo

30 Además de las reglas de trasformación incluidas en la tabla, fue necesario implementar una regla de eliminación que quitara todas las tildes del corpus, pues este elemento no formó parte de los aspectos abordados en esta investigación. La eliminación resultó sencilla porque, como se dijo antes, la tilde en el Teclado Chibcha es un carácter más como cualquier otra letra. Así pues, bastó con una simple regla de transformación a cero para eliminarla.

contrario, hubieran sido leídas como secuencias de dos formas al llevar a cabo la cuantificación. También puede apreciarse que permanecieron sin cambios los siguientes grafemas: *a, b, d, e, i, j, k, l, o, p, s, t, u, y.*[31] El principal reto al implementar un modelo de reglas consiste en el tratamiento de tres grafemas: w y *ä* en el cabécar de Chirripó y *n* en ambas variedades de la lengua. El primer caso, se solucionó fácilmente, pues primero se identificaron las ocasiones en los que *w* representaba al fonema /b/, los cuales siempre ocurren cuando esta consonante está entre dos vocales *a* o dos vocales *ä*. Por medio de la utilidad incluida en FLEx que permite hacer concordancias, se identificaron los casos en los que esto ocurría y se procedió a hacer el cambio manual antes de aplicar la transformación de *w* en *u*. Con respecto a *n* se procedió de forma similar. Recuérdese que, en cabécar, [n] es alófono de /d/ en posición de ataque justo antes de una vocal nasal, mientras que en posición de coda representa [ŋ]. En este caso, se procedió de la misma forma, se ubicaron los casos en los que *n* aparecía en coda silábica (los cuales no superaron la decena) y, manualmente, se cambió el signo. Luego, se aplicó el cambio de *n* a *d*. Lo mismo se hizo en el último caso, aunque este es más complejo. La ä representa a [ə], alófono de /a/ en las sílabas débiles del cabécar. En este caso, se procedió de igual forma. Se identificaron los casos, que en esta ocasión fueron abundantes, y se cambió el signo manualmente por *a*. Luego de ello, se procedió al cambio por medio del transcriptor automático de *ä* en *ɤ*. Cabe señalar, finalmente, que las reglas de transformación se implementaron utilizando una hoja de cálculo de Excel mediante la función "sustituir", cuya sintaxis es la siguiente: "=SUSTITUIR(texto, texto_original, texto_nuevo, [núm_de_instancia])". Por ejemplo, la *sh* se cambió a *ʃ* mediante la siguiente fórmula: *=SUSTITUIR(A1;"sh";"ʃ").*[32] La lista a la que se aplicó esta función se obtuvo de la base de datos en FLEx. Esta aplicación informática tiene la posibilidad de exportar el repertorio completo de formas léxicas (*types*) con sus glosas y frecuencias a una hoja de cálculo. A continuación, se presenta una tabla

31 Evidentemente, pudo haberse también optado por transformar *y* en *ʤ*, por citar un caso entre varios, con el fin de tener una transcripción fonológica más apegada a las normas del IPA. Sin embargo, para efectos de la cuantificación, da lo mismo contar con uno u otro carácter. Consecuentemente, se consideró innecesario hacerlo.

32 También pudo haberse escrito un programa que utilizara un sencillo algoritmo de sustitución monalfabética, similar a los que se utilizan en encriptación. Pero ello hubiera implicado eliminar manualmente los términos españoles que hay dentro del corpus.

que muestra algunos ejemplos de cómo quedaron las formas léxicas antes y después del proceso de aplicación de reglas de transformación:

INPUT TRANSCRIPCIÓN ORTOGRÁFICA	OUTPUT TRANSCRIPCIÓN PARA EFECTOS DEL CONTEO
dɛ́jʊ̱lʊ̱tënɨ̱	dɪjũlũtɪdĩ
Kjɔ́lpanɛ̱́wák	qʊlpadẽuak
mɑ̱ñɑ́töbö	bãyãtʊbʊ
shkɑ́balöglö	ʃkãbalʊklʊ
yakéiwɑ̱klä	yakeiuãklɤ

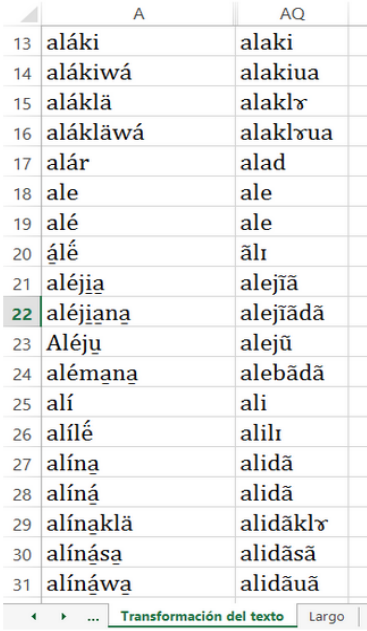
Tabla 5. Ejemplos del proceso de aplicación del transcriptor automático.

En aras de ilustrar el proceso que se siguió, se incluyen a continuación dos imágenes. En la primera se muestra la lista inicial de formas en FLEx, mientras que en la segunda se incluye un ejemplo de la transformación de los datos lingüísticos.

Palabras enteras

Forma	Glosas de palabras	Número en corpus	Los análisi…
Mostrar todos	Mostrar todos	Mostrar todo	Mostra
aláki	hembra	9	1
alákiwá	hembras	1	1
alákla̱	mujer	59	1
alákläwá	mujeres	22	1
alár	reumatismo	1	1
ale		4	0
alé		1	0
ɑ́lɛ́	gritar	1	1
Alejandro	NOMBRE	6	1
aléjia	hace rato	2	1
aléjiana	hace poco	3	1
Aléju		1	0
alémana	cerca	2	1
Alí	NOMBRE	1	1
älí	maduro	3	1
Alice	NOMBRE	4	1
älílɛ́	cocinar	1	1
älína	cocinar	2	1
älínɑ́	cocinar	1	1
älínakla̱	cocinar	2	1
älínɑ́sa	cocinar	1	1
älínɑ́wa	madurar	2	1

Imagen 2. Muestra de la base de datos en FLEx que contiene la lista de palabras del corpus.



	A	AQ
13	aláki	alaki
14	alákiwá	alakiua
15	aláklä	alaklɤ
16	alákläwá	alaklɤua
17	alár	alad
18	ale	ale
19	alé	ale
20	ą́lḗ	ãlɪ
21	aléjią	alejĩã
22	aléjiąną	alejĩãdã
23	Aléjų	alejũ
24	alémąną	alebãdã
25	alí	ali
26	alílḗ	alilɪ
27	alíną	alidã
28	alíną́	alidã
29	alínąklä	alidãklɤ
30	alínąsą	alidãsã
31	alínąwą	alidãuã

... Transformación del texto Largo

Imagen 3. Muestra del proceso de transformación del texto mediante una hoja de cálculo de Excel.

Como puede apreciarse en las imágenes incluidas anteriormente, al hacerse el traslado de la base de datos a la hoja de cálculo, se eliminaron los nombres y términos españoles, convenientemente etiquetados de previo, en aras de que no interfirieran en el cálculo que se hizo posteriormente.

C) Cálculo de las frecuencias fonémicas

Una vez que cada una de las palabras (*types*) de ambos corpus fueron transcritas en signos fonéticos o un equivalente cuantificable, se procedió a determinar las frecuencias de los fonemas. Dado que los datos estaban contenidos en una hoja de cálculo de Excel, debió buscarse una forma de establecer la cantidad de veces que aparecía un fonema concreto en la cadena de texto contenida en la celda. Esto no dejó de ser un problema, pues esta aplicación informática no posee una función específica que

permita calcular, por ejemplo, "cuántas veces aparece la letra b en una determinada celda".

Al final, se decidió desarrollar un algoritmo de cuantificación con base en la función "largo". Como se sabe, dicha función de Excel permite determinar el número de caracteres que compone una determinada cadena de texto contenida en una celda en particular. Ahora bien, la función cuenta todos los caracteres que forman parte de la cadena de texto, sin discriminar a ninguno de ellos. Entonces, para poder separar un carácter específico, se utilizó nuevamente la función "sustituir".

El proceso utilizado fue el siguiente: primero, se aplicó la función "largo" para saber cuántos caracteres había en la celda; luego, con la función "sustituir" se procedió a eliminar el carácter que se deseaba contar (sustituyéndolo por "nada"); de inmediato, se volvió a contar la cadena de texto (ahora con el carácter eliminado) y, finalmente, por medio de una resta, se obtuvo la cantidad de veces que este aparecía. El último procedimiento utilizado consistió en multiplicar el resultado obtenido por el número de veces que la palabra aparece en el corpus. De esta forma, la fórmula utilizada fue la siguiente:

*"=(LARGO(A3)-LARGO(SUSTITUIR(A3;C2;"")))*B3".*

Un ejemplo de este procedimiento se puede apreciar en la siguiente

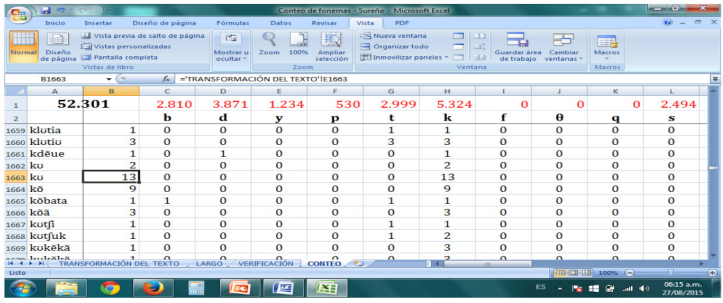


Imagen 4. Ejemplo de la hoja de cálculo de Excel utilizada para cuantificar los fonemas.

En la imagen incluida, puede verse que, por ejemplo, la palabra *klʊtia* ('bailar' en aspecto perfectivo) aparece una sola vez en el corpus, pues justamente la columna B contiene la frecuencia de aparición de la palabra contenida en la columna A. La fórmula logra identificar que, en esa palabra, la única consonante obstruyente que aparece es *k* y puede

notarse que se consigna en la columna correspondiente a esta letra un "1", que es el total de veces que aparece, en este caso, este fonema. El procedimiento es el mismo para todas las restantes celdas.

Hecho todo lo anterior, lo que resta para finalizar es determinar el total de apariciones del fonema, lo cual puede hacerse mediante una simple suma de todos los valores de la columna correspondiente. El dato final se colocó al inicio de la hoja de cálculo en la primera fila, tal y como se ve en la imagen incluida anteriormente.

3. Descripción y análisis de datos

Este apartado contiene una exposición de los resultados y hallazgos más relevantes obtenidos durante el proceso de investigación. Asimismo, proporciona una interpretación de estos a la luz de los conocimientos lingüísticos que se tiene con respecto al idioma cabécar, así como algunas consideraciones de tipo contrastivo y tipológico fundamentadas en estudios de cuantificación similares hechos con otras lenguas del mundo.

Se sigue en esta sección una organización que va de lo general a lo particular. Así pues, primero se discuten los resultados generales obtenidos en cada uno de los subcorpus antes definidos, para luego entrar en la descripción y explicación de algunos detalles de orden más específico.

3.1. Frecuencias generales de los fonemas

Una vez aplicado el proceso metodológico que se explicó en la sección anterior, se obtuvieron los siguientes datos. El subcorpus A, correspondiente al cabécar de Chirripó, se encuentra conformado por un total de 111.710 fonemas. Dado que este conjunto posee 31.225 palabras, se tiene que el promedio de fonemas por palabra es 3,58.

En la siguiente tabla, se puede observar la distribución de las frecuencias obtenida para los veintisiete fonemas de esta variedad de cabécar:

FONEMA x	RANGO x_i	FRECUENCIA ABSOLUTA n_i	FRECUENCIA ABSOLUTA ACUMULADA N_i	FRECUENCIA RELATIVA f_i	FRECUENCIA RELATIVA ACUMULADA F_i
/a/	1	12.393	12.393	11,09%	11,09%
/k/	2	10.242	22.635	9,17%	20,26%
/d/	3	8.942	31.577	8,00%	28,27%
/i/	4	8.558	40.135	7,66%	35,93%
/ã/	5	7.870	48.005	7,05%	42,97%
/t/	6	6.580	54.585	5,89%	48,86%
/u/	7	6.265	60.850	5,61%	54,47%
/b/	8	5.821	66.671	5,21%	59,68%
/ɤ/	9	5.795	72.466	5,19%	64,87%
/s/	10	5.355	77.821	4,79%	69,66%
/h/	11	4.909	82.730	4,39%	74,06%
/ɛ/	12	4.660	87.390	4,17%	78,23%
/ɭ/	13	4.237	91.627	3,79%	82,02%
/ĩ/	14	3.360	94.987	3,01%	85,03%
/ɪ/	15	3.321	98.308	2,97%	88,00%
/d͡ʒ/	16	3.124	101.432	2,80%	90,80%
/ɛ̃/	17	2.197	103.629	1,97%	92,77%
/ʃ/	18	1.963	105.592	1,76%	94,52%
/ũ/	19	1.556	107.148	1,39%	95,92%
/ʊ/	20	1.027	108.175	0,92%	96,84%
/ɔ/	21	999	109.174	0,89%	97,73%
/ɔ̃/	22	878	110.052	0,79%	98,52%
/p/	23	854	110.906	0,76%	99,28%
/kʰ/	24	574	111.480	0,51%	99,79%
/tʰ/	25	147	111.627	0,13%	99,93%
/pʰ/	26	65	111.692	0,06%	99,98%
[ŋ]	27	18	111.710	0,02%	100,00%
		111.710		100%	

Tabla 6. Frecuencia de los fonemas en el subcorpus de cabécar de Chirripó.

Como puede apreciarse, el fonema más frecuente es la vocal /a/. Dicho resultado, en lugar de sorprender, cae dentro de lo esperado. Como se sabe, sus particulares características articulatorias hacen de este sonido uno de los más frecuentes en las lenguas del mundo. De hecho, como lo recuerda Ladefoged y Ferrari Disner (2012: 178), se trata de "*the most common vowel in most languages*". Además, debe tenerse presente que,

en las sílabas débiles de esta lengua, el rasgo vocálico [- ALTO] solo puede ser ocupado por esta vocal, pues las vocales /ɛ/, /ɔ/ y /ɤ/ únicamente aparecen en las sílabas fuertes. Esta característica de la lengua seguramente también ha influido en el hecho de que la contraparte nasal de este sonido, el fonema /ã/, se encuentre en la posición número 5, pues recuérdese que debido a la expansión retrógrada de la nasalidad que presenta el cabécar, muchas veces la vocal de la sílaba débil adquiere la nasalidad presente en la sílaba fuerte.

Con respecto al sonido menos frecuente, tampoco sorprende que sea [ŋ], el cual, tal y como se mencionó anteriormente, solo aparece en posición de coda en sílabas fuertes y, debido a ello, posee un estatus fonológico cuestionado. Su presencia en el corpus es mínima; sin embargo, el examen de los pocos casos en los que aparece permiten postular que efectivamente no constituye un fonema, sino un alófono. En todos los casos registrados (con excepción de dos préstamos del español), la vocal que lo antecede siempre es una nasal. Además, siempre se encuentra en posición absoluta de frase o seguido de una consonante que posee el rasgo [- ANTERIOR]. Sin lugar a dudas, un sonido con un entorno tan limitado debe ser un alófono en lugar de un fonema.[33]

Fuera de este sonido nasal, la menor frecuencia de aparición les corresponde a las consonantes aspiradas, cuyo estatus fonológico también es controvertido.[34] En este caso, la escasez de muestras se debe a que se trata de sonidos emergentes en cabécar, la mayoría de ellos surgidos a partir de la fusión de una antigua secuencia C + /H/. No obstante, a diferencia del caso anterior, el proceso de fonologización se encuentra ya consolidado, pues no es posible restringir estos sonidos a un determinado entorno.

En todo caso, es importante señalar que esta baja frecuencia de aparición concuerda perfectamente con los datos observados en otras lenguas del mundo en las que un grupo de consonantes sordas se opone fonológicamente a un grupo de consonantes sordas aspiradas. En todas

33 Se demuestra aquí la utilidad de contar con un corpus que permita analizar los casos de aparición de los sonidos. Un análisis fonológico hecho únicamente mediante elicitación de palabras difícilmente logrará definir entornos para sonidos de aparición tan restringida. Más adelante, en un trabajo posterior, se abordará en detalle la situación de este sonido y se propondrá que se trata de un alófono de /ĩ/ que surgió a partir de la semivocal [j̃].

34 Al respecto, véase González Campos (2011: 13-14).

ellas, según Peust (2008), las aspiradas siempre son mucho menos frecuentes que las no aspiradas.

Siguen a estas consonantes, el fonema /p/ y las vocales posteriores, es decir, las que poseen los rasgos fonológicos de [+ REDONDEADO] y [- ANTERIOR], con la única excepción de /u/. Con respecto al primer caso, de acuerdo con lo reportado por Peust (2008), puede decirse que también cae dentro de lo observado en diversos estudios hechos anteriormente, los cuales señalan como tendencia general que las consonantes sordas tienden a tener una menor frecuencia de aparición que las sonoras y, dentro de la serie de las sordas, la /p/ es la que suele tener la frecuencia más baja o no existir del todo, como ocurre en árabe.[35] Lo mismo cabe acotar con respecto a la escasa presencia que tienen la mayoría de la vocales posteriores. Es un hecho ampliamente sabido que estas vocales tienden a ser menos comunes que las anteriores en los sistemas fonológicos del mundo y también a tener, con respecto a estas, un rango de aparición menor en los conteos de fonemas de una lengua.[36]

Todos estos sonidos, dentro del contexto del corpus, poseen una presencia sumamente limitada que alcanza apenas el 5,48 % del total de ocurrencias. Esta es otra tendencia general que el cabécar cumple con respecto a las lenguas del mundo. En general, hay un grupo de fonemas que tienen altas frecuencias de aparición, mientras que existen otros cuya presencia, como se ha dicho, es exigua.

Para apreciar mejor esto, se incluye a continuación un gráfico que contiene los datos de frecuencia relativa de la tabla 6.

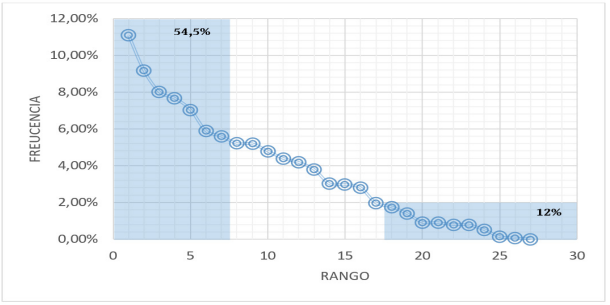


Gráfico 1. Frecuencia relativa de los fonemas en el subcorpus de cabécar de Chirripó.

35 Esto es una tendencia general, pues existen contraejemplos. En la serie de las sonoras, la que suele aparecer menos o no existir del todo es la /g/, sonido que no existe en cabécar como fonema.

36 Consúltese sobre este asunto a Papakyritsis y Granese (2013).

Se nota claramente que siete fonemas abarcan más de la mitad de las apariciones, lo cual, como se dijo, se ajusta a la perfección con la tendencia observada siempre en este tipo de cálculos. En orden decreciente, dichos fonemas son /a/, del que ya se habló, /k/, /d/, /i/, /ã/, /t/ y /u/. Con respecto a las consonantes, la presencia de la /k/ se debe a la gran cantidad de palabras que se construyen con dicho fonema, las cuales, a todas luces, resultan mayoritarias en la lengua.[37] Una explicación similar puede proponerse para /t/. La presencia de la /d/, por su parte, se debe a la versatilidad de este fonema, pues este es el único sonido oclusivo sonoro que puede aparecer en posición de coda silábica (en cuyo caso, se realiza como [r]).

Con respecto a las vocales, no cabe la menor duda que, como ya se dijo, el principal factor que influye en la frecuencia de los fonemas es la estructura de palabra del cabécar, la cual, como se explicó en el apartado 3.2., sigue el modelo métrico yámbico. Según este, las palabras bisilábicas se ajustan al esquema de "sílaba débil + sílaba fuerte" y existen rígidas restricciones para la aparición de las vocales en las sílabas débiles.[38] En estas, excepto que haya ocurrido armonía vocálica, solo pueden aparecer como núcleo de sílaba tres fonemas: /a/, /i/ y /u/. Nótese que son justamente estas vocales las que aparecen dentro del grupo de los fonemas con mayor número de apariciones. A ellas se une /ã/, la cual, como se propuso antes, seguramente debe su alto número de ocurrencias a la expansión regresiva de la nasalidad, típica del cabécar.

Los datos del subcorpus B, correspondientes al cabécar sureño, muestran una enorme similitud con lo ya visto para el subcorpus A. Este conjunto de datos lo componía un total de 52.301 fonemas correspondientes a 14.518 palabras. Por lo tanto, el promedio de fonemas por palabra es 3,60, prácticamente el mismo reportado en el corpus anterior.

De seguido, en una tabla, se sistematiza la distribución de las frecuencias

37 Para hacerse una idea de esto, téngase en cuenta que, en el diccionario de cabécar de Margery (1989), la letra K posee 1.564 entradas, casi el 20% del total de palabras incluidas.

38 Tradicionalmente, las sílabas cabécares se clasifican en dos tipos: las *fuertes,* que suelen aparecer al final de palabra, y las *débiles,* que solo se presentan en el interior de estas. Las sílabas débiles nunca son cerradas (es decir, no pueden tener coda); asimismo, desde un punto de vista prosódico, siempre presentan tono bajo y son menos intensas y más breves que las sílabas fuertes. De esta forma, dentro de la teoría inicialmente propuesta por Bruce Hayes, que clasifica los pies métricos en dos tipos (yámbico y trocaico), el cabécar se clasifica de forma abrumadora como una lengua con pies métricos yámbicos. Sobre esta teoría, véase Hyde (2011).

de aparición de los fonemas del cabécar sureño. Téngase en cuenta que, esta vez, solo se incluyen 23 fonemas, pues, tal y como se ha explicado con anterioridad, los dialectos meridionales carecen de los tres sonidos aspirados y la vocal media anterior no redondeada /ɤ/.

FONEMA x	RANGO x_i	FRECUENCIA ABSOLUTA n_i	FRECUENCIA ABSOLUTA ACUMULADA N_i	FRECUENCIA RELATIVA f_i	FRECUENCIA RELATIVA ACUMULADA F_i
/a/	1	6.052	6.052	11,57%	11,57%
/k/	2	5.324	11.376	10,18%	21,75%
/i/	3	4.234	15.610	8,10%	29,85%
/d/	4	3.871	19.481	7,40%	37,25%
/ã/	5	3.332	22.813	6,37%	43,62%
/ɛ/	6	3.097	25.910	5,92%	49,54%
/t/	7	2.999	28.909	5,73%	55,27%
/u/	8	2.940	31.849	5,62%	60,90%
/ʊ/	9	2.935	34.784	5,61%	66,51%
/b/	10	2.810	37.594	5,37%	71,88%
/s/	11	2.494	40.088	4,77%	76,65%
/h/	12	2.421	42.509	4,63%	81,28%
/ɻ/	13	1.789	44.298	3,42%	84,70%
/ɛ̃/	14	1.534	45.832	2,93%	87,63%
/ɪ/	15	1.515	47.347	2,90%	90,53%
/ĩ/	16	1.408	48.755	2,69%	93,22%
/dʒ/	17	1.234	49.989	2,36%	95,58%
/ũ/	18	619	50.608	1,18%	96,76%
/ʃ/	19	589	51.197	1,13%	97,89%
/p/	20	530	51.727	1,01%	98,90%
/ɔ/	21	282	52.009	0,54%	99,44%
/ɔ̃/	22	235	52.244	0,45%	99,89%
[ŋ]	23	57	52.301	0,11%	100,00%
		52.301		100,00%	

Tabla 7. Frecuencia de los fonemas en el subcorpus de cabécar sureño

Como se dijo, las coincidencias son evidentes. Se mantienen /a/ y /k/ como los fonemas con mayor frecuencia de aparición. Los siguen /d/ e /i/, aunque en cabécar sureño el segundo posee una cantidad ligeramente mayor de apariciones, por eso ocupa el tercer puesto. En todo caso, ello indica que se trata claramente de sonidos con una muy semejante frecuencia dentro del ámbito general de la lengua. El quinto puesto, en ambos dialectos, es ocupado por /ã/. Luego viene la única discrepancia de todo el conjunto. En el subcorpus B, la /ɛ/ ocupa el sexto puesto, pues

posee muchas más apariciones que en el subcorpus A, situación que se abordará con mayor detalle más adelante. Al ocupar la /ɛ/ este puesto, los sonidos /t/ y /u/ se han desplazado hacia las siguientes posiciones.

En cuanto a los sonidos más escasos, se mantienen las tendencias ya referidas. El sonido [ŋ] ocupa la posición más baja, aunque en este subcorpus su porcentaje de aparición es relativamente mayor. La /p/ se mantiene como el fonema consonántico de más baja frecuencia. En el caso de las vocales, /ɔ/ y /õ/ ocupan, al igual que en Chirripó, los últimos lugares. La única diferencia radica en la vocal /ʊ/. Esta, en lugar de aparecer en los últimos lugares, ocupa el lugar nueve dentro del rango de frecuencias. Esto se debe a la variación dialectal que explicábamos en el apartado 3.1 de este trabajo. Recuérdese que, en cabécar, la vocal /ɤ/ solo existe en las variedades norteñas de la lengua, pues en el sur, dicho sonido se transformó, de forma sistemática, en /ʊ/. De esta forma, las palabras que poseen /ɤ/ en el norte se pronuncian con /ʊ/ en el sur. Dada esta situación, es de esperar que la frecuencia de esta vocal, en el sur, sea mucho mayor que en el norte.

Otro punto de coincidencia radica en la tendencia, de por sí universal, de concentrar en unos pocos fonemas la mayor cantidad de ocurrencias. En el cabécar de Chirripó, los siete fonemas más frecuentes abarcan el 54,5 % del total de apariciones. En el sureño, los siete fonemas más frecuentes ocupan el 55,3 % de estas. Para apreciar esto de una mejor manera, se presenta a continuación un gráfico con los datos de frecuencia relativa incluidos en la tabla número 7.

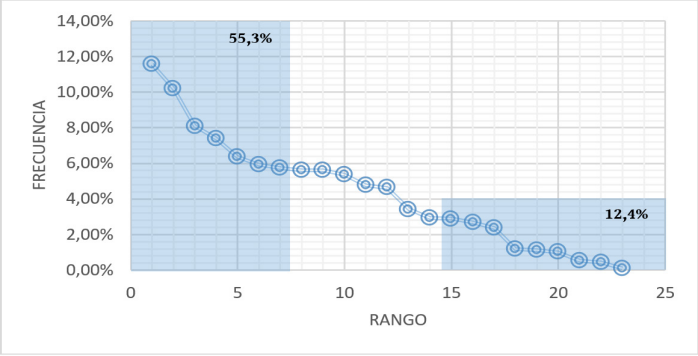


Gráfico 2. Frecuencia relativa de los fonemas en el subcorpus de cabécar sureño.

Se aprecia claramente, además, que se cumple con la tendencia de que la frecuencia de los fonemas constituye una función matemática decreciente, la cual, como se explicó en el marco teórico, se supone que es un universal lingüístico o una ley matemática dentro de la lingüística cuantitativa.

En aras de confrontar de una mejor manera los datos del cabécar de Chirripó y el cabécar sureño, es conveniente realizar una comparativa de las frecuencias relativas de cada fonema según los resultados obtenidos en cada uno de los subcorpus, la cual se ha sistematizado mediante la siguiente tabla:

FONEMA	SUBCORPUS A CABÉCAR DE CHIRRIPÓ FRECUENCIA RELATIVA	SUBCORPUS B CABÉCAR SUREÑO FRECUENCIA RELATIVA	Δ
/a/	11,09%	11,57%	0,48%
/k/	9,17%	10,18%	1,01%
/d/	8,00%	7,40%	-0,60%
/i/	7,66%	8,10%	0,43%
/ã/	7,05%	6,37%	-0,67%
/t/	5,89%	5,73%	-0,16%
/u/	5,61%	5,62%	0,01%
/b/	5,21%	5,37%	0,16%
/ɤ/	5,19%		-5,19%
/s/	4,79%	4,77%	-0,03%
/h/	4,39%	4,63%	0,23%
/ɛ/	4,17%	5,92%	1,75%
/ɻ/	3,79%	3,42%	-0,37%
/ĩ/	3,01%	2,69%	-0,32%
/ɪ/	2,97%	2,90%	-0,08%
/dʒ/	2,80%	2,36%	-0,44%
/ɛ̃/	1,97%	2,93%	0,97%
/ʃ/	1,76%	1,13%	-0,63%
/ũ/	1,39%	1,18%	-0,21%
/ʊ/	0,92%	5,61%	4,69%
/ɔ/	0,89%	0,54%	-0,36%
/ɔ̃/	0,79%	0,45%	-0,34%
/p/	0,76%	1,01%	0,25%
/kʰ/	0,51%		-0,51%
/tʰ/	0,13%		-0,13%
/pʰ/	0,06%		-0,06%
[ŋ]	0,02%	0,11%	0,09%

Tabla 8. Comparación entre las frecuencias relativas de los fonemas del cabécar

Se nota en la tabla anterior que las diferencias porcentuales son, en la mayoría de los casos, pequeñas. Hay solamente unos pocos casos que llaman la atención. Con respecto a los incrementos entre uno y otro subcorpus, el primero y más importante de todos es, sin lugar a dudas, el ya mencionado caso de /ʊ/, que presenta un incremento de 4,69 % entre una y otra variedad. Como ya se dicho, esto se debe a que dicha vocal en las variedades sureñas "suma", por así decirlo, las apariciones del fonema /ɤ/ en el norte.

El segundo incremento en importancia corresponde al fonema /ɛ/. En el subcorpus sureño aumenta 1,75 %, lo cual lo hace subir de rango, tal y como se acotó páginas atrás. En el norte este fonema ocupa la posición doce, mientras que en el sur se encuentra en el lugar seis. Resulta, en principio, difícil encontrar una explicación a este fenómeno. Una posible causa puede encontrarse en el hecho de que el subcorpus sureño muestra un mayor uso del pqronombre demostrativo *jé* [hɛ́]. Tal y como se indicó páginas atrás, en ambos subcorpus las cuatro palabras más frecuentes son las mismas. Ordenadas por orden de frecuencia de aparición, estas son *i* 'pronombre de tercera persona', *te*/*të* 'marca de ergatividad', *jé*/*é* 'pronombre demostrativo' y *dä*/*dö* 'partícula copulativa'. En el cabécar de Chirripó, existe, sin embargo, una diferencia con respecto al sur. En esta variedad, la partícula *dä*/*dö* ocupa el tercer puesto a expensas del pronombre *jé*, que queda relegado a la cuarta posición.[39] Podría pensarse que esa es una de las razones del incremento del fonema /ɛ/. En todo caso, en un futuro este asunto debe investigarse con mayor detalle, pues no deja de llamar la atención que el fonema /ɛ̃/ presenta también un incremento notorio, aunque no tan pronunciado, al comparar ambos subcorpus.[40]

El último incremento de la tabla 8 que llama la atención es el de /k/. Este, al igual que el de /ʊ/, se debe a la variación dialectal. Al no existir en sureño las aspiradas, todas las palabras que poseen /kh/ en el norte, presentan /k/ en el sur, lo cual provocó un aumento en el uso de este

39 Esto también se debe a que en Cabécar de Chirripó la partícula *dä*/*dö*, además de su uso original como cópula en las cláusulas ecuativas, se utiliza como marcador de foco, lo cual provoca que se utilice de forma más profusa.

40 Otro aspecto que podría estar relacionado con este fenómeno es el uso ortográfico. En el cabécar sureño, ocasionalmente se tiende a escribir la *schwa* [ə], un alófono del fonema /a/, como *e*. Por ejemplo, la posposición de esivo, cuya transcripción fonológica es /kàpí/, normalmente suele escribirse *gepí*. También suele suceder que no se le coloca la diéresis sobre la letra *ë*. En todo caso, es claro que, en un futuro cercano, debe buscarse una explicación para esta discordancia encontrada.

fonema. El aumento, en este caso, es mucho más notorio si se compara con las otras aspiradas debido a que el número de palabras que presenta este sonido es mucho mayor.

Con respecto a las disminuciones que se dan en el subcorpus B con respecto al A, no existe alguna que sea lo suficiente indicadora de alguna anomalía o variación significativa. La mayor es la que se da con el fonema /ʃ/, cuya frecuencia relativa alcanza un 1,76 % en Chirripó, mientras que el sur esta es de 1,13 %. En todo caso, la diferencia no parece ser, de algún modo, tan considerable como las comentadas anteriormente.

Para finalizar, se presenta a continuación un gráfico que compara las frecuencias de los fonemas obtenidas en cada uno de los dos subcorpus utilizados en esta investigación (en azul, las del subcorpus A del cabécar de Chirripó y en naranja, las de subcorpus B del cabécar sureño):

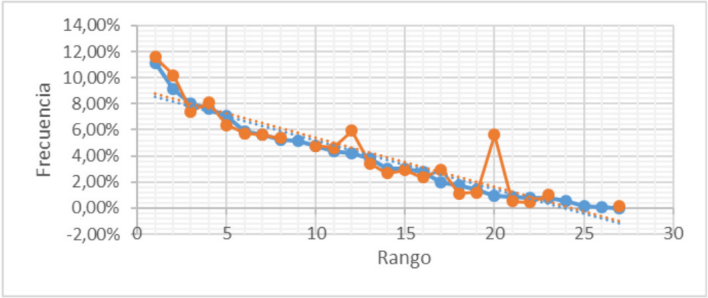


Gráfico 3. Comparación entre las frecuencias relativas de los fonemas del cabécar

En este gráfico, los puntos que sobresalen de forma notoria son aquellos que ocupan el puesto 12, que pertenece al fonema /ɛ/, y el puesto 20, correspondiente a /ʊ/. Ambos fenómenos ya han sido previamente comentados. Se incluye, mediante líneas punteadas, la regresión lineal correspondiente para ambos subcorpus, la cual, tal y como puede apreciarse, prácticamente coincide. Ello confirma la tesis de que las diferencias son mínimas entre uno y otro subconjunto.

3.2. Distribución de consonantes en relación con las vocales

Un aspecto importante por considerar en el análisis de las frecuencias fonémicas de una lengua es la relación que existe entre las consonantes

y las vocales. Existe un número denominado *coeficiente consonantal* que expresa la proporción matemática que existe entre las consonantes y las vocales en un determinado corpus, el cual se obtiene dividiendo el número total de consonantes presentes en este entre el número total de vocales.

Este coeficiente consonantal en lugar de ser universal, presenta una variación muy grande entre las lenguas del mundo. Debido a ello, Tambovtsev (1985) propuso que su cálculo y comparación entre diversos idiomas es un parámetro adecuado para establecer relaciones de orden tipológico y diacrónico y, por ende, es un medio para establecer una clasificación de estos. Esto se debe a que algunas lenguas, a causa de su parentesco genético o areal, suelen poseer coeficientes consonantales similares. Por ejemplo, de acuerdo con este autor, la familia de lenguas túrquicas se caracteriza por poseer un alto valor en este número, pues la presencia de consonantes en ellas es mucho mayor que las vocales. Por dar algunos números, en turco el coeficiente consonantal es 1,32, mientras que en kazajo es 1,33.

Con el fin de llevar a cabo dicho análisis, se presentan, a continuación, los valores absolutos de aparición de las consonantes y vocales en ambos subcorpus y el correspondiente cálculo del coeficiente consonantal.

	Subcorpus A Cabécar de Chirripó		Subcorpus B Cabécar sureño	
	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia absoluta n_i	Frecuencia relativa f_i
Consonantes	52.831	47,29%	24.118	46,11%
Vocales	58.879	52,71%	28.183	53,89%
Coeficiente consonantal	0,90		0,86	
TOTAL	111.710	100,00%	52.301	100,00%

Tabla 9. Frecuencia de las consonantes y vocales en cabécar.

Lo primero que llama la atención con respecto al coeficiente consonantal del cabécar es su bajo valor. Esto se debe que esta lengua, sin lugar a dudas, es un idioma en el que las vocales tienen una mayor frecuencia de aparición que las consonantes. Según lo reportado por Tambovtsev (2003), en Europa y Asia, es difícil encontrar valores tan bajos para dicho coeficiente.[41] Los más bajos son los del japonés (1,08), el nanai (1,02)

41 Quienes sí presentan altos índices de frecuencia vocálica son algunas lenguas austronesias

y el oroch (1,00).[42] Nótese que los valores del cabécar son ligeramente inferiores a estos, pues alcanza apenas 0,90 en el cabécar hablado en Chirripó y 0,86 en las variedades sureñas.

Sin embargo, esta alta concentración de vocales no es del todo extraña en lenguas americanas. Este mismo autor en un trabajo posterior (Tambovtsev 2009: 2), señala con respecto a la frecuencia de aparición de las vocales en lenguas de este continente los siguiente: "*In 32 American Indian languages taken for this study (Tab.1) the mean concentration of vowels is 44.06%. The least concentration of vowels is in Kadiweu — 35.73%. The major concentration of vowels is in Iquito — 58.84%*". Como puede verse, si bien el cabécar se encuentra muy por encima del promedio registrado, los datos recopilados no superan los del iquito, la lengua con mayor número de frecuencia de vocales. De hecho, en los datos calculados por este autor, hay dos lenguas americanas con porcentajes similares a los del cabécar. Se trata del cofán, una lengua amazónica aislada cuyas vocales constituyen un 53,04 % del total de sus sonidos, y el secoya, una lengua tucana hablada en Perú y Ecuador, la cual acumula una frecuencia de vocales de 51,43 %. Constituye un hecho lamentable que no se cuente con datos de cuantificación de fonemas de ninguna otra lengua perteneciente a la familia lingüística chibchense o de algún otro idioma centroamericano, pues ello permitiría comparar el cabécar con lenguas más cercanas tanto genética como arealmente.

Con respecto a este punto, otro aspecto que vale la pena comentar es lo relacionado con la eufonía, la cual se entiende como la suma de la frecuencia de las vocales y las consonantes sonoras de una lengua. Es un hecho bastante conocido que los sonidos sonoros tienden a ser más frecuentes que los sordos. En su análisis de la frecuencia de las oclusivas, Carsten Peust sostuvo que en la sonoridad suele ser mayor en las oclusivas anteriores. En palabras suyas, "voiced stops are generally easier to pronounce, and tend to be more frequent, at the fronted places of articulation" (Peust 2008: 120).

Esta idea fue desarrollada a su máxima expresión por Tambovtsev (2009), quien, tras analizar lenguas de todo el mundo, propuso que la eufonía es un rasgo lingüístico universal. Es decir, siempre los sonidos sonoros van a ser superiores a los sordos en una lengua. Según

como el hawaiano y el samoano.

42 Estas últimas son lenguas tunguses.

los datos recopilados por él, la eufonía más baja se da en una lengua caucásica denominada adigué, hablada en la república del mismo nombre dentro de la Federación Rusa. En esta lengua la eufonía alcanza un valor de 54,07 %.

Para poder corroborar dicha afirmación en el cabécar, se procedió a realizar el cálculo de las frecuencias de aparición de los sonidos sordos y sonoros en ambos subcorpus. Los resultados obtenidos se resumen en la siguiente tabla:

	Subcorpus A Cabécar de Chirripó		Subcorpus B Cabécar sureño	
	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia absoluta n_i	Frecuencia relativa f_i
Sonidos sordos	30.689	27,47%	14.357	27,45%
Sonidos sonoros	81.021	72,53%	37.944	72,55%
TOTAL	111.710	100,00%	52.301	100,00%

Tabla 10. Frecuencia de los sonidos sordos y sonoros en cabécar.

Como puede apreciarse, el principio enunciado por Tambovtsev (2009) se cumple perfectamente en el cabécar, pues la eufonía alcanza valores de 72,53 % en el primer subcorpus y 72,55 % en el segundo. De hecho, en comparación con las otras lenguas de América reportadas por este autor, el cabécar posee una de las eufonías más altas del continente, pues se acerca bastante a las señaladas en su trabajo como poseedoras de los mayores porcentajes de sonoridad, las cuales corresponden al guaraní (74,89 %), el siriano (75,02 %) y el iquito (76,85 %).[43]

3.3. Distribución de las consonantes según sus rasgos

Corresponde hacer ahora algunas apreciaciones sobre la distribución que presentan las frecuencias de las consonantes de acuerdo con sus rasgos articulatorios. Para empezar, se tomará en cuenta el modo de articulación, cuyos porcentajes de aparición se presentan a continuación:

43 Estas lenguas se hablan, respectivamente, en Paraguay, Colombia y Perú.

MODO DE ARTICULACIÓN		SUBCORPUS A CABÉCAR DE CHIRRIPÓ		SUBCORPUS B CABÉCAR SUREÑO	
		FRECUENCIA ABSOLUTA n_i	FRECUENCIA RELATIVA f_i	FRECUENCIA ABSOLUTA n_i	FRECUENCIA RELATIVA f_i
OBSTRUYENTES NO CONTINUAS	SONORAS	17.887	16,01%	7.915	15,13%
	SORDAS	18.462	16,53%	8.853	16,93%
OBSTRUYENTES CONTINUAS	SORDAS	12.227	10,95%	5.504	10,52%
NO OBSTRUYENTES	SONORAS	4.255	3,81%	1.846	3,53%
	TOTAL	52.831	47,29%	24.118	46,11%

Tabla 11. Frecuencia de las consonantes del cabécar según el modo de articulación.

A partir de los datos anteriores, puede decirse que, en general, el cabécar tiende a ajustarse a las tendencias generales observadas en otras lenguas del mundo. Para empezar las obstruyentes no continuas (u oclusivas) son, de forma notoria, mayoritarias en relación con las fricativas y otros sonidos consonánticos que poseen el rasgo [+ CONTINUO]. Esto se debe a que, como predijo Roman Jakobson[44], las fricativas son menos frecuentes en las lenguas del mundo tanto en el repertorio total de fonemas como en sus frecuencias de aparición. Además, debido a esto, en el habla infantil, las fricativas tienen a ser adquiridas mucho tiempo después que las oclusivas.[45]

Dentro de las oclusivas, las sordas y las sonoras tienden a tener cierto equilibrio, aunque las sordas presentan, de forma leve, frecuencias de aparición mayores, lo cual entra en contradicción con lo dicho antes en este trabajo sobre la preeminencia de los sonidos sonoros en las lenguas del mundo. Esto, en todo caso, no parece tener mucha importancia debido, sobre todo, a la ya mencionada cercanía porcentual entre las oclusivas sordas y sus contrapartes sonoras.

En otros aspectos, presenta el cabécar comportamientos mucho más disimiles según las tendencias observadas. Por ejemplo, de acuerdo con Peust (2008: 121), una tendencia que no es universal, pero

44 Un trabajo que resume, de forma adecuada y precisa, la teoría universalista de Roman Jakobson es Gómez Fernández (1993), quien además hace una revisión crítica de sus postulados a partir de los datos sobre la adquisición de fonemas en español recogidos por otros autores.

45 Al respecto, véase Li, Edwards y Beckman (2009). En otro estudio, Edwards, Beckman y Munson (2015) sostienen que la adquisición de un determinado sonido por parte de los infantes se encuentra relacionada directamente con la frecuencia que este presenta en la lengua. Los sonidos más frecuentes tienden a adquirirse más tempranamente que los de baja frecuencia. Este, sin embargo, es solo un factor entre otros que influyen en la adquisición fonológica. Para conocer una tesis contraria a esta, consúltese Yamaguchi (2008).

sí muy frecuente, es que las oclusivas sonoras se ajusten al siguiente orden de aparición d > b > g, mientras que el de las sordas sea t > k > p.[46] Esto se debe a que el segmento con el rasgo [+ coronal] suele ser el de mayor aparición. Como puede corroborarse observando las tablas 6 y 7, el cabécar cumple con la tendencia de las oclusivas sonoras, salvo por el hecho de que no posee un fonema /g/ (pero la posición de esta consonante es ocupada por el fonema africado /ʤ/, que diacrónicamente deriva justamente de *g).[47] Sin embargo, en lo que respecta a las oclusivas sordas, la tendencia no se cumple, pues el cabécar presenta el siguiente orden: k > t > p. Claramente, el fonema /k/ ha desplazado a /t/, lo cual no es de extrañar si se tiene en cuenta la alta cantidad de palabras que contienen este sonido, tal y como se explicó antes.

En lo que respecta a las obstruyentes continuas o fricativas, el cabécar presenta el siguiente alineamiento de fonemas: s > h > ʃ. Que /s/ sea el fonema más frecuente concuerda con tendencias generales vistas en otras lenguas, lo cual, no obstante, no constituye un universal, como lo reporta Peust (2008). Sin embargo, no deja de ser interesante que la frecuencia de /h/, si bien, es menor, se encuentra muy cercana a la de /s/ en ambos subcorpus. De hecho, en el cabécar sureño la diferencia entre uno y otro sonido es de solamente 0,14 %. Podría pensarse, en este sentido, en un equilibrio más que en una prevalencia de /s/, lo cual contradiría la tendencia general señalada.

Otro punto importante por considerar en la distribución de las consonantes es la frecuencia de estas de acuerdo con su punto de articulación. Dicha distribución, al igual que la proporción entre consonantes y vocales, es un aspecto que suele diferir considerablemente de una lengua a otra. Por mencionar un caso, Tambovtsev (2010a: 16) explica que el rango de variación de la frecuencia de las consonantes labiales es enorme en las lenguas del mundo: "*In the 128 languages which we took for our studies, the frequencies of occurrence of the labial consonants are spread in the range from 1.70% to 16.66%*". No obstante, también suele suceder que lenguas relacionadas tanto genética como arealmente presentan valores coincidentes. En ese mismo estudio, Yuri Tambovtsev demuestra, por ejemplo, que las consonantes labiales son

46 En este caso, el símbolo ">" significa 'más frecuente que'.

47 Al respecto, véase Constenla (2008).

mucho más frecuentes en las lenguas nigerocongolesas que en cualquier otra familia lingüística. Los rasgos articulatorios constituyen, por lo tanto, un medio para establecer relaciones entre los idiomas. De hecho, con base en esta situación, este autor desarrolló un método para medir la distancia fonotipológica entre dos o más idiomas, el cual compara la frecuencia de ocho grupos de consonantes por medio de una fórmula que combina la prueba estadística de la χ^2 con el cálculo de la distancia euclídea.[48]

En la tabla que se presenta a continuación, se han sistematizado las frecuencias de aparición de los grupos consonánticos del cabécar según su punto de articulación:

Punto de articulación	Subcorpus A Cabécar de Chirripó		Subcorpus B Cabécar sureño	
	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia absoluta n_i	Frecuencia relativa f_i
Labial	6.740	6,03%	3.340	6,38%
Dental	15.669	14,03%	6.870	13,14%
Alveolar	9.592	8,59%	4.283	8,19%
Palatal	5.087	4,55%	1.823	3,49%
Velar	10.834	9,70%	5.381	10,29%
Glotal	4.909	4,39%	2.421	4,63%
TOTAL	52.831	47,29%	24.118	46,11%

Tabla 12. Frecuencia de las consonantes del cabécar según el punto de articulación.

Los datos incluidos, como dijimos, corresponden a una tendencia propia de cada lengua en particular. Sin embargo, en ellos puede apreciarse una tendencia que se ha observado en otros estudios y se presupone universal: la preeminencia de los segmentos que presentan el rasgo articulatorio [+ CORONAL], el cual, como recuerda Yamaguchi (2008), se considera "no marcado" y, por lo tanto, el que acumule más frecuencias de aparición.[49] Ello, tal y como lo demuestra la tabla 12, se cumple a cabalidad en cabécar.

Al cumplir con esta característica, puede afirmarse que los datos recopilados sobre la frecuencia de fonemas del cabécar permiten

48 Pueden verse ejemplos de la aplicación de este método en Tambovtsev (2003, 2008 y 2010b).

49 La hipótesis de que los sonidos no marcados deben de ser los más frecuentes dentro de un determinado texto fue popularizada por Greenberg (1966: 14), quien retomó una idea de Trubetskói y afirmó justamente que "*in general the unmarked category has higher frequency than the marked*".

afirmar que esta lengua cumple con las dos características principales que se han postulado como tendencia general en dicho tipo de cálculos, a saber, la preeminencia de los segmentos con los rasgos [- CONTINUO] y [+ CORONAL] en el conteo final de las frecuencias de aparición.

3.4. Distribución de las vocales según sus rasgos

Al igual que las consonantes, es factible llevar a cabo un análisis de las vocales a partir de sus rasgos articulatorios y corroborar si los datos manifestados por el cabécar coinciden con las propuestas hechas con respecto a la distribución es este tipo de sonidos. La primera oposición que es importante abordar es la diferencia entre la cantidad de vocales orales presentes en el corpus en comparación con las vocales nasales, distinción fonológica que el cabécar manifiesta. Al respecto, no puede dejarse de mencionar la teoría propuesta hace cincuenta años por Joseph Greenberg (1966: 14-21), justo en los inicios de los estudios sobre los universales lingüísticos. En su trabajo pionero, este autor señaló que los sonidos no marcados son, por lo general, los más habituales en la cadena hablada. Estableció, además, que las vocales orales son siempre no marcadas frente a las orales, en aquellas lenguas en las cuales una distinción fonológica entre unas y otras. Consecuentemente, la frecuencia de aparición de las vocales orales en un texto (o corpus textual, para el caso de este trabajo) debe ser mayor que la de las vocales nasales.[50] De ello, él brindó algunos ejemplos.

A continuación, se presentan los datos de frecuencia de las vocales del cabécar organizados según la oposición [+/- NASAL].

50 Según Greenberg, esto se debe a que, desde un punto de vista diacrónico, las vocales nasales se originan de vocales orales nasalizadas por una consonante (obviamente nasal) ubicada luego de ellas, la cual, con el tiempo, se pierde.

Tipo de vocal	Subcorpus A Cabécar de Chirripó		Subcorpus B Cabécar sureño	
	frecuencia absoluta n_i	frecuencia relativa f_i	frecuencia absoluta n_i	frecuencia relativa f_i
Nasal	15.861	14,20%	9.425	13,63%
Oral	43.018	38,51%	18.758	40,26%
TOTAL	58.879	52,71%	28.183	53,89%

Tabla 13. Frecuencia de las vocales orales y nasales del cabécar.

Como es factible apreciar, el cabécar cumple muy bien con la hipótesis de Greenberg acerca de la distribución de las vocales nasales en relación con sus contrapartes orales. Las nasales poseen rangos de aparición bajos. De forma general, puede decirse que constituyen la cuarta parte del total de apariciones del total de vocales en el corpus. Es importante señalar que ello muy posiblemente se deba a la restricción de aparición de este tipo de sonidos en la lengua. Tal y como se ha mencionado varias veces en este trabajo, el cabécar suele presentar una métrica yámbica en sus palabras; debido a ello, hay dos tipos de sílabas: las fuertes y las débiles. En las sílabas débiles hay importantes restricciones fonológicas. De hecho, en esa posición, las únicas vocales fonológicamente distintivas son solo tres: /a/, /i/ y /u/. A nivel lexical, las nasales no pueden aparecer en sílabas débiles. La única excepción a esta regla la constituye el proceso de expansión retrógrada de la nasalidad.[51] En determinados casos (cuyos entornos varían dialectalmente), la nasalidad de la vocal que forma el núcleo de la sílaba fuerte se expande hacia las vocales de las sílabas débiles. Que dicho fenómeno es la razón de la frecuencia de las nasales en cabécar lo confirma la leve diferencia que existe entre un corpus y otro. En Chirripó, existen menos restricciones para la expansión de la nasalidad que en las variedades sureñas. En consecuencia, en este dialecto, la expansión de la nasalidad ocurre con mayor frecuencia, lo cual se puede apreciar en la tabla 13, en la cual se nota que el subcorpus

51 Todo ello confirma que, en cabécar, se cumple con la hipótesis de Greenberg acerca de las nasales, pues estas, claramente, constituyen el elemento fonológico "marcado". Ahora bien, lo que no se cumple es su hipótesis, aludida en la nota anterior, de que las nasales se originan siempre de nasalizaciones de vocales orales. Constenla (2008) dejó establecido, de forma muy firme, que las nasales de la familia lingüística del cabécar se remontan a su antepasado común, el protochibchense, cuya antigüedad es similar a la del protoindoeuropeo.

A presenta un ligero aumento en la aparición de las vocales nasales, producto sin lugar a duda de la diferenciación dialectal. Así pues, puede afirmarse que, siguiendo una tendencia aparentemente universal, las nasales en cabécar manifiestan una frecuencia de aparición baja con respecto a las vocales orales y ello se relaciona con las fuertes restricciones de aparición de este tipo de sonidos.

Otro aspecto que resulta importante valorar es la distribución estadística de las vocales según los rasgos articulatorios de localización y altura de la lengua, los cuales, como se sabe, suelen tomarse como parámetros para su clasificación. Para efectos de un cálculo de frecuencia de fonemas, el primero y más importante de estos rasgos es el grado de adelantamiento o retroceso de la lengua dentro de la boca, situación que produce tres tipos de vocales, a saber, las anteriores, las centrales y las posteriores. A continuación, se presentan los datos de distribución de los fonemas vocálicos del cabécar agrupados según dicho parámetro:

Localización	Subcorpus A Cabécar de Chirripó		Subcorpus B Cabécar sureño	
	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia absoluta n_i	Frecuencia relativa f_i
Anterior	22.096	19,78%	13.158	22,54%
Central	20.263	18,14%	9.384	17,94%
Posterior	16.520	14,79%	5.641	13,41%
TOTAL	58.879	52,71%	28.183	53,89%

Tabla 14. Frecuencia de las vocales del cabécar según la localización de la lengua.

Con respecto a los datos anteriores, resulta necesario hacer un comentario con respecto al hecho de que, tal y como se ha propuesto como tendencia universal, los sonidos vocálicos posteriores son minoritarios con respecto a los demás, pues constituyen justamente los elementos "marcados".[52] En la tabla 14, puede verse que estos, en ambos subcorpus, coinciden en ser los de menor frecuencia. Es más, si se observan las tablas 10 y 11, puede notarse que, tanto en Chirripó como en el cabécar sureño, las vocales de más baja aparición son /ɔ/ y su contraparte nasal /ɔ̃/, ambas vocales posteriores.

Esta situación se relaciona también con el orden que constituyen las vocales según la frecuencia de aparición, el cual suele ser

52 Al respecto, véase Yamaguchi (2008).

el siguiente: a > i > u.[53] Nótese que se sigue una secuencia de central, anterior y posterior. En el subcorpus A de cabécar de Chirripó, las tres vocales orales más frecuentes, de mayor a menor, son justamente esas, las cuales, como ya se ha dicho varias veces, son las únicas que en este idioma pueden aparecer en todo tipo de sílabas. En dicho conjunto, sus contrapartes nasales también se acomodan siguiendo este patrón, pues su orden de frecuencia es ã > ĩ > ɛ̃ > ũ > õ. En el subcorpus de sureño, hay algunas variaciones que se deben a la mayor frecuencia de las vocales anteriores /ɛ/ y /ɛ̃/, fenómeno del cual se habló antes. Debido a ellas, /ɛ/ supera a /u/ en las vocales orales, mientras que /ɛ̃/ sobrepasa por poco a /ĩ/ en la serie de las nasales. Aun así, en ninguno de los dos casos, se afecta el patrón señalado.

Resta, para finalizar, referirse a la distribución estadística de las vocales según el rasgo de altura. Los datos de dicho cálculo se han ordenado en la siguiente tabla:

Altura vocálica	Subcorpus A Cabécar de Chirripó		Subcorpus B Cabécar sureño	
	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia absoluta n_i	Frecuencia relativa f_i
Alta	19.739	17,67%	11.674	17,59%
Media	18.877	16,90%	7.125	18,35%
Baja	20.263	18,14%	9.384	17,94%
TOTAL	58.879	52,71%	28.183	53,89%

Tabla 15. Frecuencia de las vocales del cabécar según la altura de la lengua.

Sobre dicho rasgo, no se ha propuesto ninguna tendencia de tipo general. De hecho, es un aspecto que los estudiosos de las frecuencias fonémicas suelen dejar de lado, pues no se reportan estudios contrastivos al respecto. En todo caso, el cabécar, tal y como demuestra la tabla 15, presenta una distribución bastante equilibrada entre los tres rasgos. Cada uno de ellos, de forma general, tiene la tendencia a poseer la tercera parte de las apariciones en el ambos corpus. Solo se aprecia una pequeña disparidad en las vocales medias, las cuales son ligeramente superiores

53 Embarki (2013: 28) reporta que este es el orden de distribución de las tres vocales del árabe según su frecuencia de aparición en un corpus. Además, acota que "*this order probably follows a general tendency in the languages of the world*".

en número de apariciones en el cabécar sureño, efecto producido, a todas luces, por el fenómeno ya ampliamente mencionado de aumento de las vocales medias /ɛ/ y /ɛ̃/.

4. Conclusiones

Este trabajo presentó una investigación descriptiva de estadística fonológica en el cabécar, lengua indígena americana hablada en Costa Rica. Dicha investigación reunió un corpus electrónico con datos lingüísticos que abarcó casi todos los materiales disponibles sobre dicho idioma. Este corpus electrónico fue, por lo tanto, del así denominado tipo "oportunista" y se constituyó utilizando el programa computacional denominado *Fieldworks Language Explorer.* Debido a variaciones dialectales en el orden de lo fonológico, los datos se organizaron en dos subcorpus diferentes: uno con el cabécar de Chirripó y otro con el cabécar sureño. Hecho esto, se procedió a determinar las frecuencias estadísticas de los fonemas del cabécar de forma automatizada en cada subcorpus. Esto se hizo mediante una serie de fórmulas diseñadas *ad hoc* para esta investigación e implementadas en una hoja de cálculo de Excel. Así pues, de forma específica, puede decirse que esta investigación cumplió con los objetivos que se había propuesto.

A grandes rasgos y exceptuando las particularidades dialectales, ambos subcorpus coincidieron en la distribución de la frecuencia de los fonemas y la concomitancia de una serie de características planteadas en estudios anteriores como universales lingüísticos, leyes de la lingüística cuantitativa o tendencias generales de las lenguas del mundo. En concreto, dichos rasgos que se reafirman en los datos obtenidos del cabécar son los siguientes:

a. La frecuencia de los fonemas constituye una función matemática decreciente. Debido a esto, unos pocos fonemas concentran la mayor cantidad ocurrencias.
b. El fonema más frecuente es la vocal central /a/.
c. La frecuencia de aparición de los sonidos sonoros es mayor que la de los sonidos sordos.
d. Las obstruyentes no continuas (u oclusivas) son mayoritarias en relación con sus contrapartes continuas (o fricativas).

e. Los fonemas coronales presentan mayores frecuencias de aparición que los demás.
f. Dentro de la serie de oclusivas sordas, la /p/ es el sonido menos frecuente.
g. Las oclusivas sonoras se ajusten al siguiente patrón de aparición según su frecuencia en el corpus: d > b > g.
h. La frecuencia de las vocales nasales es mucho menor en relación con sus contrapartes orales.
i. Los sonidos vocálicos posteriores son minoritarios con respecto a los centrales y anteriores.

Dentro de la investigación, se detectó un hecho que no se ajusta a las tendencias generales señaladas por la bibliografía anterior. Se trata del patrón de aparición de las oclusivas sordas según su frecuencia, el cual debería ser t > k > p. En cabécar, /k/ es más frecuente que /t/. Además, se detectó, comparando un subcorpus con otro, un aumento notorio de las vocales anteriores medias /ɛ/ y /ɛ̃/ en las variedades sureñas de la lengua, cuya naturaleza debe investigarse con mayor profundidad en un futuro, pues de momento no es factible proponer una hipótesis segura que explique dicha situación.

Excepto por dicho aumento, el cabécar sureño presentó, como se dijo, distribuciones muy semejantes a las del cabécar de Chirripó. Solamente, se separó, como era de esperar, en la frecuencia del fonema /ʊ/, el cual, en las variedades meridionales de la lengua, tiene una mayor frecuencia debido a que subsume al fonema /ɤ/, cuya aparición solo se da en cabécar norteño. Por su parte, las consonantes oclusivas aspiradas, los otros fonemas que son exclusivos del cabécar norteño, no influyeron de forma significativa en las estadísticas de las frecuencias debido a su bajísima aparición.

Como en todo proceso investigativo, el estudio presentado en este artículo presentó algunas limitaciones que deberán solventarse en próximos proyectos. La principal fue la falta de materiales textuales suficientes y, sobre todo, más variados a los cuales recurrir a la hora de conformar el corpus. Por ejemplo, hubiera sido deseable contar con una mayor representación de géneros discursivos orales, como las conversaciones. En todo caso, debe ser una tarea futura dotar a hablantes nativos de este idioma de las destrezas ortográficas necesarias para que logren escribir su lengua de forma normalizada y puedan así

generar más documentación y materiales textuales. Debe promoverse, además, la recopilación de textos orales como diálogos o alocuciones que enriquezcan la variedad de textos disponibles.

Un caso particularmente lamentable es la ausencia en este estudio de datos propios de la región del Valle de La Estrella, la cual posee una variedad dialectal propia que debe necesariamente ser considerada en futuras investigaciones. No deja de ser una pena que no existan textos bien escritos en dicha variedad que puedan ser usados en investigaciones lingüísticas. Más adelante, hay que procurar documentar de forma profesional este dialecto del cabécar.

Relacionado con este asunto, se encuentra un problema más: la existencia de textos que no se pudieron utilizar en esta investigación debido a que se encuentran muy mal escritos desde un punto de vista ortográfico. Un claro ejemplo de esto, ya antes mencionado, es la reciente versión cabécar del Nuevo Testamento, que podría ser una fuente de información lingüística útil si estuviera redactado de forma coherente. Sería importante que se dedique tiempo a generar iniciativas que normalicen el registro escrito del cabécar. En este sentido, sería importante la publicación de un diccionario escolar tanto en papel como en línea e, incluso, la generación de un corrector ortográfico y su incorporación a algún procesador de textos. La conformación de corpus electrónicos es un hecho que puede ayudar en la creación de dichos productos.

Finalmente, cabe señalar algunas investigaciones futuras que pueden producirse a partir de este trabajo. A nivel interno, el corpus electrónico conformado debe aprovecharse en otros proyectos relacionados con el quehacer de la lingüística de corpus. En años venideros, sería conveniente emprender otro tipo de estudios, por ejemplo, los de estadística léxica. También sería importante determinar la frecuencia de las estructuras gramaticales fundamentales de la lengua. A nivel fonológico, puede diseñarse un algoritmo que permita identificar las sílabas de la lengua y estudiar las estructuras silábicas fundamentales, así como los fonemas que más aparecen en determinadas posiciones, como las codas.

A nivel externo, resultará de particular interés conformar diferentes corpus electrónicos de las restantes lenguas indígenas de Costa Rica, en particular, y Centroamérica, en general. Si ello se logra, se podrán proponer estudios similares que permitan comparar los datos aquí

logrados y otros más por obtener. Todo ello redundará, por supuesto, en un mejor conocimiento de las lenguas indígenas de la región y brindará a los lingüistas datos valiosos para ser considerados en trabajos tipológicos y areales.

Bibliografía

Alcaraz Varó, Enrique y María Antonia Martínez Linares (1997). *Diccionario de lingüística moderna*. Barcelona: Ariel.

Altmann, Gabriel (2005). "Phonic word structure". En *Quantitative Linguistik: Ein internationales Handbuch / Quantitative Linguistics: An International Handbook, editado por Reinhard Köhler, Gabriel Altmann y Rajmund G. Piotrowski, 191-208*. Berlín: Walter de Gruyter.

Altmann, Gabriel (2008). "Towards a theory of script". En *Analyses of Script*: Properties of Characters and Writing Systems, editado por Gabriel Altmann y Fan Fengxiang, 149-164. Berlín: Walter de Gruyter.

Astorga Campos, Julián, Sergio Cordero Monge y Jorge Antonio Leoni de León (2013). "INLEXPO: Una herramienta de gestión lexicográfica". *Actualizaciones en Comunicación Social* 1: 333–338.

Barquero Reyes, Rogelio, Vania Solano Laclé y Dalia Castillo Campos (2011). *Seje duchiiwak kasenewa*. *La cultura cabécar de Chirripó*. Turrialba: Universidad de Costa Rica, Vicerrectoría de Acción Social, Sede del Atlántico.

Biber, Douglas y James K. Jones (2009). Quantitative methods in corpus linguistics. En *Linguistics: An international handbook*. Vol. 2, editado por A. Lüdeling y M. Kytö, 1286-1304. Berlín: Mouton de Gruyter.

Bisani, Maximilian y Hermann Ney (2008). "Joint-Sequence Models for Grapheme-to-Phoneme Conversion". *Speech Communication* 50 (5): 434–51.

Bourland Hawley, David (1974). *A Generative-Transformational Grammar of an Idiolect of Cabécar*. San José, C.R.: Imprenta Semántica.

Brenes Granados, Cristian (2007). "Tecnologías de Información y Comunicación: el caso de las comunidades indígenas cabécares de Chirripó de Costa Rica". *Revista Electrónica Educare* 2: 177–92.

Butler, Lynnika y Heather van Volkinburg (2007). "Fieldworks Language Explorer from SIL International". *Language Documentation and*

Conservation 1 (1): 100–106.

Calderón Saravia, Ana Lucía (1996). *De cómo los bribris y los cabécares conservamos la herencia que nos dejó Sibö*. San José, C. R.: Proyecto Namasöl.

Constenla Umaña, Adolfo (2005). "¿Existe relación genealógica entre las lenguas misumalpas y las chibchenses?". *Estudios de Lingüística Chibcha* 24: 7-85.

Constenla Umaña, Adolfo (2008). "Estado actual de la subclasificación de las lenguas chibchenses y de la reconstrucción fonológica y gramatical del protochibchense". *Estudios de Lingüística Chibcha* 27: 117-135.

Crystal, David (1997). *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Edwards, Jan, Mary E. Beckman y Benjamin Munson (2015). "Frequency effects in phonological acquisition". *Journal of child language* 42 (02): 306-311.

Embarki, Mohamed (2013). "Phonetics". En *The Oxford Handbook of Arabic Linguistics*, editado por Jonathan Owens, 23-44. Londres: Oxford University Press.

Evison, Jane (2010). "What are the basics of analysing a corpus?". En T*he Routledge Handbook of Corpus Linguistics*, editado por Anne O'Keeffe y Michael McCarthy, 122-135. Nueva York: Routledge.

Fernández Torres, Severiano (2011). *Sibőte i-ka dieyá. El banquete de Sibö*. Historias y canciones del pueblo cabécar. Limón, C.R.: Fundación Naíri.

Flores Solórzano, Sofía (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): 155–61.

Gómez Fernández, Diego (1993). "La teoría universalista de Jakobson y el orden de adquisición de los fonemas de la lengua española". *Cauce: Revista de Filología y Su Didáctica* 16: 7–30.

González Campos, Guillermo (2011). "Dificultades para normalización ortográfica y problemas de escritura entre los cabécares de Chirripó". *Estudios de Lingüística Chibcha* 30: 7-35.

González Rátiva, María Claudia y Jorge Antonio Mejía Escobar (2011). "Frecuencia de fonemas en dos corpus de español de uso en

Colombia". En *La lengua, lugar de encuentro: actas del XVI Congreso Internacional de la ALFAL (Alcalá de Henares 6-9 de junio de 2011)*, editado por Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García, 105-115. Alcalá de Henares: Universidad de Alcalá, Servicio de Publicaciones.

Greenberg, Joseph H. (1966). *Language Universals with Special Reference to Feature Hierarchies*. The Hague: Mouton de Gruyter.

Harary, Frank y Herbert H. Paper (1957). "Toward a General Calculus of Phonemic Distribution". *Language* 33 (2): 143–69.

Hyde, Brett. 2011. "The Iambic-Trochaic Law". En T*he Blackwell Companion to Phonology*, editado por Marc van Oostendorp, Colin J. Ewen, Elizabeth V. Hume y Keren Rice, 1052–1077. Oxford y Nueva York: Wiley-Blackwell.

Jara Murillo, Carla (2013). "El treebank del español 'IPROCOLDI': componente anotado del corpus CODIMEP-CR". *Revista de Filología y Lingüística de la Universidad de Costa Rica* 39 (2): 143–71.

Kennedy, Graeme (1998). *An Introduction to Corpus Linguistics. Londres:* Longman.

Köhler, Reinhard (2005). "Gegenstand und Arbeitsweise der Quantitativen Linguistik". En *Quantitative Linguistik: Ein internationales Handbuch / Quantitative Linguistics: An International Handbook*, editado por Reinhard Köhler, Gabriel Altmann y Rajmund G. Piotrowski, 1-16. Berlín: Walter de Gruyter.

Köhler, Reinhard y Burghard B. Rieger (1993). "Preface". En *Contributions to Quantitative Linguistics. Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991*, editado por Reinhard Köhler y Burghard B. Rieger, ix-xii. Dordrecht; Boston; Londres: Kluwer Academic Publishers.

Köhler, Reinhard y Gabriel Altmann (2011). "Quantitative Linguistics". En *The Cambridge Encyclopedia of the Language Sciences*, editado por Patrick Colm Hogan, 695-697. Cambridge: Cambridge University Press.

Krohn, Haakon S. (2017). "Frecuencia de fonemas en las narraciones tradicionales en malecu". *Káñina* 41(2): 87-103.

Ladefoged, Peter y Sandra Ferrari Disner (2012). *Vowels and Consonants.* Oxford y Nueva York: Wiley-Blackwell.

Leech, Geoffrey (1992). "Corpora and theories of linguistic performance". En *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*, editado por Jan Svartvik, 105-122. Berlín y Nueva York: Mouton de Gruyter.

Lehfeldt, Werner (2005). "Phonemdistribution". En *Quantitative Linguistik: Ein internationales Handbuch / Quantitative Linguistics: An International Handbook*, editado por Reinhard Köhler, Gabriel Altmann y Rajmund G. Piotrowski, 181-190. Berlín: Walter de Gruyter.

Lehmann, Christian, recop. (2010). "Textos escolares". Disponible en línea: http://www.christianlehmann.eu/ling/sprachen/cabecar/textos/index.html.

Leoni de León, Jorge Antonio (2010). "Computational Linguistics in Costa Rica: An Overview". En *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, 40–45.

Leoni de León, Jorge Antonio (2012). "Calculador Léxico-estadístico de Frecuencia Dispersión y Uso (CaLeFDU)". *Káñina* 36 (2): 135–43.

Li, Fangfang, Jan Edwards y Mary E. Beckman (2009). "Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers". *Journal of Phonetics* 37 (1): 111–124.

Lüdeling, Anke y Merja Kytö (2008). *Corpus Linguistics: An International Handbook*. Berlín: Walter de Gruyter.

Margery Peña, Enrique (1986). "Cuatro leyendas cabécares". *Estudios de Lingüística Chibcha* 5: 45-57.

Margery Peña, Enrique (1989). *Diccionario cabécar-español español-cabécar*. San José, C.R.: Editorial de la Universidad de Costa Rica.

Margery Peña, Enrique (1995). "Sibö. Relato mitológico cabécar". *Estudios de Lingüística Chibcha* 14: 31-39.

McEnery, Tony y Andrew Hardie (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony y Andrew Wilson (2001). *Corpus Linguistics: An Introduction*. Edimburgo: Edinburgh University Press.

Morales, José Noel, recop. (2013). *Ditsä S̲iwa̲ (Ditsä Tjai wakwa S̲iwa̲*

/ Historias cabécares de Tayni). San José, C. R.: Ministerio de Cultura y Juventud, Dirección de Cultura.

Moreno Sandoval, Antonio; Doroteo Torre Toledano, Natalia Curto y Raúl de la Torre (2006). "Inventario de frecuencias fonémicas y silábicas del castellano espontáneo y escrito". En *IV Jornadas en Tecnología del Habla*, editado por L. Buera, E. Lleida, A. Miguel y A. Ortega, 77–81. Zaragoza: Universidad de Zaragoza.

Obando Martínez, Freddy y Guillermo González Campos (2015). *Historia del clan Kätsúibawák*. Turrialba: Universidad de Costa Rica, Sede del Atlántico.

Papakyritsis, Ioannis y Angela Granese (2013). "Cross-linguistic Study of Vowel Systems". En *Handbook of vowels and vowel disorders*, editado por Martin J. Ball y Fiona E. Gibbon, 186-206. Nueva York y Londres: Psychology Press.

Parodi, Giovanni (2010). *Lingüística de Corpus: de la teoría a la empiria*. Madrid: Iberoamericana.

Pawłowski, Adam (2008). "Prolegomena to the History of Quantitative Linguistics". *Glottotheory* 1: 48–54.

Peust, Carsten (2008). "On Consonant Frequency in Egyptian and Other Languages". *Lingua Aegyptia* 16: 105-134.

Pintzuk, Susan (2011). "Corpus Linguistics". En *The Cambridge Encyclopedia of the Language Sciences*, editado por Patrick Colm Hogan, 231-232. Cambridge: Cambridge University Press.

Quesada, Juan Diego y Christian Lehmann, recop. (2007). "Las piedras que se volvieron tigre", de Maribel Fernández Fernández. Disponible en línea: http://www.christianlehmann.eu/ling/sprachen/cabecar/textos/index.html.

Rafel i Fontanals, Joaquim y Joan Soler i Bou (2003). "El procesamiento de corpus: La lingüística empírica". En *Tecnologías del lenguaje*, 41-73. Barcelona: Universitat Oberta de Catalunya.

Ríos Mestre, Antonio (1999). "La transcripción fonética automática". *Estudios de lingüística del español* 4. Disponible en línea:http://www.raco.cat/index.php/Elies/article/view/194842.

Rogers, Chris (2010). "Fieldworks Language Explorer (FLEx) 3.0 from SIL International". *Language Documentation and Conservation* 4: 78–84.

Shipulina, Ludmila Alekseevna (2004). *Reseña de Tipologija funktsionirovanija fonem v zvukovoj tsepochke indoevropejskih, paleoaziatskih, uralo-altaiskih i drugih jazykov mira: kompaktnost' podgrupp, grupp, semej i drugih jazykovyh taksonov, de Yuri Tambovtsev. Acta Linguistica Hungarica* 52 (2-3): 328-338.

Simons, Gary F. (1977). "Phonostatistic Methods". *Workpapers in Papua New Guinea Languages* 21: 155–85.

Sinclair, John (2005). "Corpus and text – basic principles". En *Developing Linguistic Corpora: A Guide to Good Practice*, editado por Martin Wynne, 1-16. Oxford: Oxbow Books.

Stone, Doris (1961). *Las tribus talamanqueñas de Costa Rica*. San José, C. R.: Lehmann.

Strauss, Udo, Fengxiang Fan y Gabriel Altmann (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM-Verlag.

Strauss, Udo, Gabriel Altmann y Karl-Heinz Best (2006). "Phoneme frequency". Disponible en línea: http://lql.uni-trier.de/index.php/Phoneme_frequency.

Tambovtsev, Yuri (1985). "The Consonantal Coefficient in Selected Laguages". Canadian Journal of Linguistics 30 (2): 179-188.

Tambovtsev, Yuri (2003). "Phonological similarity between Basque and other world languages based on the frequency of occurrence of certain typological consonantal features". *The Prague Bulletin of Mathematical Linguistics* 79-80: 121-126.

Tambovtsev, Yuri (2008). "The Phono-Typological Distances between Ainu and Other World Languages as a Clue for Closeness of Languages". *Asian and African Studies* 17 (1): 40–62.

Tambovtsev, Yuri (2009). "Euphony in American Indian languages: A phonetic universal". *California Linguistic Notes* 34 (2): 1-21.

Tambovtsev, Yuri (2010a). "Labial consonant distribution in Niger-Congo: A study in typological distance". *California Linguistic Notes* 35 (2): 1–34.

Tambovtsev, Yuri (2010b). "Distances between Polish and Other Slavonic Languages: A Phono-Typological Comparison". *California Linguistic Notes* 35 (2): 1–14.

Tambovtsev, Yuri y Colin Martindale (2007). "Phoneme Frequencies Follow a Yule Distribution". *SKASE Journal of Theoretical Linguistics* 4 (2): 1–11.

Taylor, Charlotte (2008). "What Is Corpus Linguistics? What the Data Says". *ICAME Journal* 32: 179–200.

Terrádez Gurrea, Marcial (2001). *Frecuencias léxicas del español coloquial: análisis cuantitativo y cualitativo*. Valencia: Universitat de València.

Těšitelová, Marie (1992). *Quantitative Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Torruella, Joan y Joaquim Llisterri (1999). "Diseño de corpus textuales y orales". En *Filología e Informática. Nuevas tecnologías en los estudios filológicos*, editado por J. M. Blecua, G. Clavería, C. Sánchez y J. Torruella, 45–77. Barcelona: Editorial Milenio.

Trubetzkoy, Nikolai S. (1976). *Principios de fonología*. Madrid: Cincel.

Varas, Valeria y Severiano Fernández Torres (1989). *Historias cabécares*. Dos volúmenes. San José, C.R.: Editorial de la Universidad de Costa Rica.

Yamaguchi, Naomi (2008). "Markedness, Frequency: Can We Predict the Order of Acquisition of Consonants?". En *Proceedings of the 2nd Oxford Postgraduate Conference in Linguistics*, editado por Miltiadis Kokkonidis, 236-243. Oxford: Oxford University. Disponible en línea: http://www.ling-phil.ox.ac.uk/events/lingo/papers/Proceeings.pdf.

Este texto presenta un estudio exploratorio que, por primera vez, ha establecido la frecuencia de fonemas del cabécar, lengua indígena costarricense. En concreto, este texto expone los principios conceptuales y el proceso metodológico de conformación de un corpus electrónico de datos lingüísticos de este idioma que sirvió de base empírica para la obtención de información estadística de tipo fonológico. Luego, presenta la frecuencia absoluta de aparición de cada uno de los fonemas de la lengua cabécar y lleva a cabo un análisis estadístico de dichos datos a la luz de los postulados y principios teóricos generados en el marco de investigaciones semejantes en otras lenguas del mundo.

UNIVERSIDAD DE COSTA RICA
SA Sede del Atlántico