Universidad Nacional de Educación a Distancia

Facultad de Filología
Departamento de Filologías Extranjeras y sus Lingüísticas

HACIA UN CORRECTOR ORTOGRÁFICO PARA LA NUEVA ORTOGRAFÍA DEL CHABACANO DE ZAMBOANGA

Marcelo Yuji HIMORO

Trabajo de Fin de Máster en las Tecnologías de la Información y la Comunicación en la Enseñanza y el Tratamiento de Lenguas.

Tutor: Dr. D. Antonio Pareja-Lora

Convocatoria de Septiembre
Curso 2018/2019

Si conversa tu con un gente na un lenguaje ele ta entende,
llega ’se hasta na de suyo pensamiento.
Si conversa tu con ele na de suyo lenguaje,
llega ’se hasta na de suyo corazon.

— **Nelson Mandela***

Bien misterioso el maga pasion humano [...]. Aquellos ya sale afectao no sabe que laya man con este explica, y aquellos no hay pa experiencia no puede gad con ese entende. Tiene quien ta pone na peligro el de ila vida para encanza el punta del un monte. No hay gad quien, pati sila mismo, ta puede hace claro si por que. Tiene alli ta arruina de ila cuerpo para gana el corazon de un tal fulano que no hay que ver kanila. Y tiene tambien alli ta destroza de ila cuerpo por causa na alegria de segui na maga entojo - o del tomahan. Tiene tambien alli maga decidido gana na juego de cen y ta queda no hay nada, acabar tiene pa tambien ta sacrifica todo cosas sila tiene na un ambicion que nunca sila ay puede realiza. Tiene ta pensa kay na otro lugar lang sila ay puede queda alegre, por eso ta dedica sila de ila vida para man viaje entero mundo. Acabar, tiene alli hende descansa hasta queda sila poderoso. Quiere ese decir, el maga diferente pasion igual de mucho con el maga gente.

— **Michael Ende***, Die unendliche Geschichte, 1979.

No hay cosa na todo pasada de mio que ya queda olvidao. El tristesa, el miedo, el huya, y el pagkatampa que no hay sirve. Ta resta el melancolia y el maga locuras que ya queda grande na sombra de esos. Ta resta el maga duda y el lasangan. El ataranto de tiene vez, de cuantos sueño. Acabar, ta resta el dolor de antes.

—— **Luisgé Martín***, El amor del revés, 2016.

Nunca yo ya puede queda como el maga otro. Nunca. Kay yo un Catòia. Pati kay yo un Mapolvos.

— **Joan Bodon***, Lo Libre de Catòia, 1966.

[...] vira-vira el Historia, un eterno de subi abaja. Nunca gayod ay dura hasta para cuando el un derecho que ya puede saca. Hende librao tambien el libertad contra na violencia, kay firme ese nuevo clase hechura ta dale mira. Cada nuevo ascenso ay enfrenta con el contralio estaba na humanidad. Pati el maga cosas que claro ya ay man duda ya tambien ole. Aquel ta hace ya kita de no hay nada con el libertad, y hende mas ta considera con ese el de con todo sagrado cosa kita ya puede saca, ay sale estaba na oscuridad del mundo del maga insticto un misterioso deseo para ataca con este [...].

— **Stefan Zweig***, Castellio gegen Calvin, 1936.

(* Ya hace Chavacano el autor de este trabajo. Responsabilidad de suyo todo el maga mali.)

AGRADECIMIENTOS

Quiere era yo rendi gracias con:

- el de mio *supervisor* por el de suyo bien largo paciencia, maga constructivo consejo pati maga palabra de suporta na tormento;
- mi familia y maga pariente, kay no hay sila man sumut conmigo suporta na emocional y na financial;
- el de mio maga amigo, kay si no hay ustedes conmigo firme manda sale na casa, segurao ya puede yo queda loco;
- Lucas Vinícius Avanço, *classmate* de mio na colegio, por causa del un seccion na de suyo masteral thesis que ya queda mio inspiracion para principia este trabajo.

Ta pedi yo gracias tambien con todo aquellos quien ya ayuda conmigo pati ta ayuda lang siempre na de mio aprendida del Chavacano, especialmente con:

- Alexandra-Jean Garcia David, el mio primer "maestra" de Chavacano;
- el grupo na Facebook "Zamboanga de Antes", particularmente con Jorge Seneca du Quillo, Nina Lacandalo-Nohay y Ronan Paul Dayot, por el bien grande ayuda sila ya dale conmigo por cuantos año, acabar pati con Ramon F. Covarrubias kay amo ele el quien ya introduci conmigo na grupo;
- Jerome Cadungog Herrera, na manada cosa yo ya puede aprende estaba con ele pati na de suyo bien grande contribucion na una y con todo importante parte de este trabajo.

Finalmente, dale lang era yo gracias con cuantos personas, quien na bien manada manera ya dale de ila contribucion para con este trabajo: Felino Manuel Santos, Chito Barrios, Vanessa Enriquez Madelo, Monica S. Macansantos, Dr. Norma Camins-Conti, Julia Teresita Enriquez, Aaron Misa, AimPinky (Pure Zamboangueños), Fr. Alberto Rossa, Aleii Ruales, Al-rass Amarillo (aka Pluma), Alshamir Bryan Barrera Aripuddin, Angelo Dean Bustillo, Anthony Val Acosta, Armee Jay

Cresmundo, Arnold Prio, Benjie A. Mahasol, Bobby Piedad, Charmaine Arcillas Lim, Charlie Villanueva, Christina Newhard, Daniel M. Taclap, Darren Bendanillo, Des Kan Sabess, Desena Ollie Ramos Fabiania, Dexter Ando, Don Ramon, Donex Magallon Bonifacio, EJ Natividad, Enrico Guido O. Canoy, Etienne Eunson, Floraime Oliveros Pantaleta, Gideon Cascolan, Gaspar A. Vibal, Herra Gend Fernandez, Isser Jorell L. Yao, Jhasmine Dangase, Jay-ar Ave, Jefferson Tuazon Sanico, Dr. Jervil Formilleza-Omaga, Dr. Jessie Grace Rubrico, Jinda Saint (aka mai2x), Joey Panganiban Dayagdag, John Leoner S. Tatil, John Nuluddin Kandero, Jomarie Navarro Alamhali, Joni Ginsberg, Jose Alfonso H. Lazaga, Jp Montuno, JR Natividad, Jullienne Fortich Tuazon Jaldon, Jurgen Sanes, Kenneth Joseph dela Zerna Martin, Mae-Ann Bustamente, Mavie Labor, Marion Pon Estrada, Mark Eingel N. Fernandez, Mark Lim Hao, Miller Lospendejos (aka zigzag), Mark Anthony Gregorio Tolentino, Nari Ramchand, Nickky Boy Auditor Ibarra, Oliver Bernardo, Patrick Jethro, Prince Aarol Fernandez, Dr. R. David Zorc, Randy Batiquin, Raybison James Enriquez, Refciejay Apiado Pagotaisidro (aka Twinkle), Rey Gaspar, Robert Barrera, Robert Nadera (aka Djbusted), Robert/Roland Villanueva Jr., Rodrigo Seiji Himoro, Roel M. Apolinario, Ronald Carpio (aka pondong), Ronel Rojas Buenaflor, Rogin Christ Eribal, Ruben Balagot, Ruffy Gerard Enopia Matarlo, ”Si, Chavacano También Yo“, UNiCA (Maldita), Dr. Walter Sauer, Weng Dela Peña, Zack Quijano, Ejay Reyes, Rose Zapanta Sanson, Anton Q. Darlucio, Christine Joan Albrecht, Sam Andico, Jayson Encarnacion, Kim Tayona, Mara Palengkera, Erik Bernardo, Jhay Andadi, Lando Barcode, Jhenno Mojica, Christopher Datol, Jim Shephard, Chiqui Basmayor, Ar-ar Solis Villablanca y Visita Zamboanga. Ta pedi gayod yo dispensacion pati gracias na maga yo ya puede olvida menta aqui.

Gracias tambien con todo maga ya dale de ila contestacion na de mio *survey* pati ya comparti con ese na de ila maga amigo y familia.

♣ ♣ ♣

Dedica yo este thesis na memoria del maga defunto escritor Antonio R. Enriquez y Francis C. Macansantos, kay no hay sila olvida con el Chavacano na de ila maga obra.

Dedica tambien era yo con este na todo maga musiquero de Zamboanga, especialmente con Comic Relief, Maldita (UNiCA), Nari Ramchand, Mirage, MushroomHead, Cheeze de Sal, Zero, nephelim, Lost Primoz, Chrysolite, DJ BUSTED, Zambo Top Dogz y Zigzag & Ivan, kay si no hay el de ila musica, no hay yo queda interesao aprende Chavacano. Idol gayod ustedes!

Acabar, dedica yo con este na maga Zamboangueño, na Zamboanga pati na *abroad*, quien dia mas dia no hay descansa conversa Chavacano.

RESUMEN

En la actualidad, el zamboangueño es la variedad de chabacano o criollo filipino de base española más hablado, contando con más de 400.000 hablantes nativos en todo el país en 2010, sin contar sus hablantes como segunda lengua. Desde 2012, se enseña como asignatura y sirve de lengua vehicular en los tres primeros años de la educación primaria en las escuelas públicas de la Ciudad de Zamboanga, en la región de Mindanao (Filipinas). En la primera parte de este trabajo, se muestra, a través de un análisis breve de algunas actitudes de los hablantes, que pese a su vitalidad, el zamboangueño puede estar amenazado, y proporcionamos muchas directrices para trabajos futuros a ese respecto. Mesuramos también, por medio de un cuestionario difundido en las redes sociales, el grado de familiaridad de los hablantes a la nueva ortografía del chabacano de Zamboanga. En la segunda parte, tomando esa grafía como referencia, procuramos analizar y clasificar errores de ortografía frecuentes tanto en contextos formales como informales en nuestro corpus y proponemos una aproximación para corregir *tokens* utilizando Traducción Automática Estadística de Caracteres. Los resultados obtenidos muestran que esta aproximación es sumamente adecuada y podría com-

ABSTRACT

Este tiempo, el Zamboangueño amo el de con todo ta conversa clase de Chavacano o ese ta llama "Philippine Creole Spanish" (PCS). Con ese ta conversa como lenguaje nativo mas de 400.000 persona na entero Filipinas na 2010, fuera maga *L2 speakers*. Principiando del año 2012, ya queda este *subject* y ta usa tambien como *medium of instruction* principia Grade 1 hasta na Grade 3 na maga escuela publico del Ciudad de Zamboanga. Na primer parte de este *research*, ta analiza kame el actitud del maga gente ta conversa Zamboangueño para dale kame mira kay, masquen bien vivo pa el lenguaje, baka *endangered* ese. Ta dale kame mucho tambien sugestion de trabajo futuro acerca con este. Por medio de un cuestionario que ya distribui na manada gente na *social media*, ta medi kame que laya de familiar el maga ta conversa Chavacano con ese ta llama "Zamboanga Chavacano Orthography". Na segunda parte, usando con este ortografia como referencia, ta analiza y ta clasifica kame el maga palabra hende amo el deletreada na formal pati na informal Zamboangueño que ta puede kame encontra na de amon corpus. Acabar, ta propone kame un solucion para corregi maga *tokens* con ese usando Character-Based Statistical Machine Translation. El maga resulta ta dale mira kay angay ese

binarse con las tecnologías de corrección ortográfica más utilizadas actualmente para obtener un mejor desempeño.

y posible man combinacion de este y el maga tecnologia que ahora recio ta usa na *spell checking* para mas bueno el *performance*.

Palabras clave: chabacano, zamboangueño, procesamiento del lenguaje natural, PLN, lenguas con pocos recursos, corrector ortográfico.

Keywords: Chavacano, Zamboangueño, natural language processing, NLP, Underresourced Languages, spell checker.

ABSTRACT

Zamboangueño is nowadays the most widely spoken Chabacano or Philippine Creole Spanish (PCS) variety, with over 400.000 native speakers in the Philippines in 2010, not including the numerous L2 speakers. Since 2012, it has been taught as a subject and serves as a medium of instruction from Grade 1 to 3 in the public schools of Zamboanga City, Mindanao, Philippines. In the first part of this research, we show through a brief analysis of some attitudes of the speakers that, despite its high vitality, Zamboangueño may in fact be endangered, while suggesting many directions for future works regarding this issue. By means of a questionnaire widely distributed to social media users, we also assess the speakers' level of familiarity with the so-called "Zamboanga Chavacano Orthography". In the second part, using that orthography as a reference, we aim to analyze and classify the most frequent spell errors found in both formal and informal Zamboangueño in our corpus and propose a Character-Based Statistical Machine Translation approach to correct tokens. The results show that this approach is suitable for the presented purposes and could well be combined with the current de facto spell checking technologies to achieve further performance.

Keywords: Chavacano, Zamboanga, Natural Language Processing, NLP, Underresourced Languages, spell checker.

ÍNDICE

Página

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ABREVIATURAS Y SIGLAS

CPH	Census of Population and Housing
CWZCC	Contemporary Written Zamboangueño Chabacano Corpus
CZ	chabacano zamboangueño
KWF	Komisyon sa Wikang Filipino
MTB-MLE	Mother Tongue Based - Multilingual Education
NIF	NLP Interchange Format
NLTK	Natural Language Toolkit
OCR	Contemporary Written Zamboangueño Chabacano Corpus
OWL	Web Ontology Language
PCS	Philippine Creole Spanish
PSA	Philippine Statistics Authority
RAE	Real Academia Española
RDF	Resource Description Framework
TA	traducción automática
TEI	Text Encoding Initiative
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

CAPÍTULO 1

INTRODUCCIÓN

Filipinas, u oficialmente la «República de Filipinas», es un archipiélago localizado en el sureste asiático, compuesto por 7.641 islas que se extienden sobre un área de 300.000 kilómetros cuadrados de territorio (Government of the Philippines, 2014). A lo largo de su historia, Filipinas ha sido colonizada por los españoles durante más de 300 años y después por los Estados Unidos de América durante 48 años, lo que ha dejado una profunda marca en las lenguas y culturas locales.

El chabacano, también llamado *Philippine Creole Spanish (PCS)*, es un conjunto de lenguas criollas de base española, resultado del contacto entre los colonizadores españoles y hablantes de diferentes pueblos locales, que no compartían una lengua común. Se conocen diferentes variedades del chabacano como, por ejemplo, el ternateño, el caviteño, el ermiteño, el cotabateño, el davaoeño y el zamboangueño. No existe consenso sobre la clasificación de estas variedades y hay diferentes teorías y un largo debate alrededor de su formación (M. Fernández, 2015; M. A. Fernández & Sippola, 2017; Grant, 2013; Lipski, 1992; Whinnom, 1956). Pese a su alto grado de inteligibilidad, la tendencia es, tanto en la literatura como por parte de sus hablantes, a tratar las diferentes variedades del chabacano de manera independiente.

Entre los organismos públicos del gobierno filipino, el tratamiento del chabacano tampoco es homogéneo. Según el mapeo lingüístico realizado entre los años de 2014 y 2015 por la Komisyon sa Wikang Filipino (KWF, 2015), organismo responsable de la política lingüística y la estandardización del tagalo y demás lenguas de Filipinas, en todo el país se hablan en la actualidad 135 lenguas, entre las cuales el chabacano figura de manera unitaria, sin hacer distinción a sus diferentes variedades. Sin embargo, en los censos gubernamentales de población y vivienda (*Census of Population and Housing*[1], CPH) realizados por la autoridad estadística filipina

[1] *Esta modalidad de censo se lleva a cabo a cada principio de década. Menos exhaustivos son los censos de población (Population Census o POPCEN), que se llevan a cabo entre dos CPH, normalmente a cada mitad de década.*

(*Philippine Statistics Authority*, PSA) se incluye cada una de esas variedades por separado.

En lo que concierne al número de hablantes total, no hay estadísticas fiables, ya que los CPH solo recogen el número de hablantes nativos por lengua. Procesando los datos de informes y documentos del CPH del 2000 y del 2010 disponibles en el sitio web de la PSA (National Statistics Office (NSO), 2003a, 2014a; Philippine Statistics Authority (PSA), 2014), se obtuvieron[2] los datos de las tablas 1 y 2.

En la tabla 1, podemos observar que el chabacano zamboangueño aparece en las posiciones 17 y 16 en el rango de lenguas maternas por número de hablantes nativos y por número de hogares respectivamente. Sin embargo, cabe recordar que estos números no incluyen los emigrantes filipinos de la diáspora. Además, debido al papel del chabacano zamboangueño como lengua franca de la ciudad de Zamboanga, si se tomaran en consideración también los hablantes no nativos, el zamboangueño probablemente subiría algunas posiciones en ese rango.

Tabla 1: Listado de las 17 lenguas autóctonas con mayor número de hablantes nativos y de hogares que la utilizan. (Estimación a partir de lo datos de National Statistics Office (NSO), 2003a, 2014a; Philippine Statistics Authority (PSA), 2014)

LENGUA	HABLANTES	HOGARES
Tagalog/Filipino[3]	33.391.225	7.569.251
Cebuano	22.218.692	4.838.843
Ilocano	7.363.799	1.638.317
Hiligaynon	6.178.719	1.339.177
Bicolano	4.236.487	873.595
Waray	2.616.230	555.118
Kapampangan	2.303.985	510.014
Maguindanao	1.341.165	264.337
Pangasinan	1.286.318	285.122
Tausug	1.159.364	206.278
Maranao	1.140.104	206.477
Kinaray-a	1.037.501	223.898
Capiznon	713.548	153.671
Masbateño	532.395	109.581
Aklanon	530.458	114.308

[2] Los números de hablantes se estimaron redondeando la multiplicación de la suma del número de hablantes en cada una de las regiones de Filipinas por el tamaño medio del hogar en esa región.

[3] El filipino es una forma estandardizada de tagalo. Aunque algunos consideren que la diferencia entre ambos es que el filipino incluye palabras de otras lenguas locales que no forman parte del tagalo, para la mayoría de las personas ambos son sinónimos y, por esta razón, los tratamos como tal en este trabajo.

Surigaonon	433.176	90.175
Chabacano (Zamboangueño)	427.268	90.246

Por su parte, la tabla 2 muestra el número de hablantes de cada variedad de chabacano en todo el país según los CPH del 2000 y del 2010. En el CPH del 2010, el caviteño y el ternateño figuran de manera unitaria bajo la etiqueta de «Caviteño-Chavacano», así que no disponemos de los datos individuales de esas variedades para el año en cuestión. Sin embargo, está claro que todas las variedades de chabacano, salvo el zamboangueño, presentan un descenso en el número de hablantes, mientras que el ermiteño ya no dispondría de hablantes nativos.

Tabla 2: Número de hablantes nativos y de hogares de las variedades vivas de chabacano según los CPH de 2000 y 2010. (Estimación a partir de lo datos de National Statistics Office (NSO), 2003a, 2014a; Philippine Statistics Authority (PSA), 2014)

	CPH 2000		CPH 2010	
VARIEDAD	**HOGARES**	**HABLANTES**	**HOGARES**	**HABLANTES**
Zamboangueño	69.041	356.165	90.243	405.798
Davaoeño	810	3.938	580	2.616
Cotabateño	980	5.016	573	2.741
Caviteño	796	3.959	1.177*	5.483*
Ternateño	495	2.416		

1.1 CHABACANO ZAMBOANGUEÑO

La variedad zamboangueña, a la cual a partir de aquí nos referiremos por la sigla «CZ», es más conocida por sus hablantes por el autoglotónimo de «chavacano». El CZ se habla sobre todo en la ciudad de Zamboanga, pero también en puntos de las islas de Basilan, Joló, Tawi-tawi y en otras regiones del país y en comunidades filipinas en el extranjero, como la de Kampung Air en Malasia (Grant, 2007; Lipski, 2003). El gentilicio que denomina a las personas de la ciudad de Zamboanga es «zamboangueño», pero también puede ser usado para referirse a un grupo étnico (la PSA lista «Chavacano-Zamboangeño» [sic] en los CPH). El CZ es, como se observa en las tablas 1 y 2, la variedad que disfruta de mayor vitalidad y la única que todavía observa un incremento en el número absoluto de hablantes nativos.

Con la implementación a nivel nacional del currículo K-12 y del programa MTB-MLE (*Mother Tongue Based - Multilingual Education*, o Educación Multilingüe en Lengua Materna), el CZ volvió a enseñarse en las escuelas públicas[4] de la

[4] Según ha reiterado en diversas ocasiones el periodista y ex-maestro Felino Manuel Santos, tanto en las redes sociales como en correspondencia privada con el autor: «En el 1964, el gobierno

Ciudad de Zamboanga y hoy funciona como lengua vehicular en los tres primeros años de primaria (Government of the Philippines, 2011). Es, por lo tanto, la única variedad de chabacano que se enseña actualmente en las escuelas. Esto ha llevado al ayuntamiento de la ciudad a invertir en la codificación y normativización[5] del CZ, plasmándose en la aprobación de una ortografía en el 2014, así como la publicación de una gramática básica, varias colecciones de textos, libros para niños, e incluso un diccionario normativo publicado a finales del 2018.

Antes limitados a temas religiosos, empiezan a publicarse cada vez más libros escritos en esta lengua, sobre todo en el campo literario. Además de las publicaciones del propio ayuntamiento ya mencionadas anteriormente, en los últimos años se han editado volúmenes de literatura local, como la selección de poemas «Balsa: poemas Chabacano» y el libro infantil «Cholo Chismoso», y también traducciones, como las dos de «El Principito» y la de «Si Amina y El Ciudad de Maga Flores». Pese a estas iniciativas, el número de publicaciones es aun escaso, teniendo en cuenta la vitalidad de la lengua.

El espacio del CZ en los medios de comunicación es también bastante limitado. En la televisión y en la radio, salvo en programas de debate, la mayor parte de la programación se limita a las noticias. De manera marginal, algunas radios transmiten radionovelas y canciones de artistas locales, pero sin regularidad. En lo que concierne a la producción musical, existe una escena musical local independiente, con grupos normalmente compuestos por jóvenes estudiantes. Gracias a los concursos musicales que se celebran cada año en el Festival Zamboanga Hermosa, ha habido un incremento en la cantidad de canciones nuevas en CZ.

En los apartados siguientes, destacamos algunos puntos importantes sobre la situación del CZ y algunos fenómenos relevantes que se están produciendo en la sociedad zamboangueña.

---

ordenó el uso del vernáculo, [que] en Zamboanga era el chabacano, como lengua vehicular en la escuela primaria. Era maestro de escuela pública en aquel entonces y participé de este proyecto. [...] Fui, por lo tanto, uno de los primeros maestros a utilizar el chabacano en las escuelas públicas.» («*In 1964, the government mandated the use of the vernacular, in Zamboanga it was Chavacano, as the medium of instruction in the elementary school. I was public school teacher then and participated in this project. [...] I was one of the early teachers who thus used Chavacano in the public schools.*»)

5 Es importante aclarar aquí la diferencia entre dos conceptos que se confunden con bastante frecuencia: el de la **normativización** y el de la **normalización**. Como explica Santarnarina (1995), la **normativización** se refiere a la elaboración de un modelo de reglas gramaticales y ortográficas y de un cuerpo léxico estable recogido en gramáticas y diccionarios respectivamente que sirven de modelo de referencia para los hablantes de la lengua; es decir, se refiere exclusivamente a un proceso de codificación de la lengua. La **normalización**, a su vez, es un proceso mucho más amplio, que consiste en una serie de medidas que conducen al uso de la lengua en todos los ámbitos de la vida y actividades de un pueblo.

1.2 LA SITUACIÓN DEL CZ

A pesar de no disponer de estadísticas oficiales o estudios que muestren una vista panorámica de la situación real del CZ, la lengua sigue siendo hablada por la mayoría de la población de la Ciudad de Zamboanga, sea como lengua materna o como segunda lengua. Sin embargo, aunque las cifras del CPH puedan parecer alentadoras, los relatos de los propios zamboangueños dejan entrever que esta situación no debería verse con tanto optimismo. En este apartado, se mencionan algunos factores que permitirían confirmar que el CZ se encuentra en realidad en una situación de fragilidad e, incluso, de amenaza.

1.2.1 *Diglosía, cambios demográficos y lingüísticos*

Desde hace algunas décadas, el CZ experimenta un proceso de minorización. Esto es, en parte, resultado de la presión del tagalo y del inglés (lenguas oficiales en todo país) en la educación y en los medios de comunicación y, gracias a los continuos cambios demográficos, también del cebuano (lengua de gran parte de las provincias vecinas). En términos prácticos, el CZ se enfrenta a una situación diglósica, o quizás sería más exacto hablar de una situación multiglósica, ya que se ve en competición con múltiples lenguas. Esto queda patente en las quejas constantes de hablantes mayores (ej. 1, 2, 3) en las redes sociales[6] acerca de una pérdida de espacio del CZ en ámbitos donde antes se utilizaba con normalidad.

(1) «Dol el lingua franca na pueblo hende ya Chavacano, dol mas mucho Tagalog pati Visaya.»
«Parece que la lengua franca en el centro de la ciudad ya no es el chabacano, me parece que el tagalo y el cebuano son mayoritarios.»

(2) «Masquen na KCC, maga saleslady Tagalog... 'De daw ta entende Chavacano...»
«Incluso en el KCC[7], las dependientas [te hablan en] tagalo... Dicen que no entienden el chabacano...»

(3) «Mio observaciones, especialmente na maga servidores ta conversa Tagalog: Por que hende na de aton dialecto? Na Ciudad de Zamboanga kita.»
«Mis observaciones, especialmente hacia los camareros que [te] hablan en tagalo: ¿Por qué no [habláis] en nuestro dialecto? Si estamos en la Ciudad de Zamboanga.»

---

6 Hemos cambiado la grafía, corregido la puntuación y anonimizado los nombres con el objetivo de guardar la identidad de las personas.

7 Se refiere al *KCC Mall de Zamboanga*, un centro comercial localizado en la ciudad.

Estos comentarios muestran que los hablantes son conscientes de la situación diglósica del CZ. Los comentarios siguientes relatan una de las consecuencias de la presión ejercida sobre todo por el tagalo sobre el chabacano: la lengua de socialización de los jóvenes, en muchos casos, ya no es el CZ.

(4) «Pero el nuevo generacion ahora hende ya ta conversa Chavacano, masquen de Zamboanga lang.»
«Pero la nueva generación ahora ya no habla en chabacano, aunque sea de Zamboanga.»

(5) «Cuando ya volve yo na de aton ciudad, ya espanta yo kay el maga dalagita y el maga solterito hende ya man ta conversa el de aton lenguaje. Puro Tagalog ya. Yo, masquen hende ya nace de mi maga anak na Zamboanga, ya enseña yo kanila conversa el mi lenguaje na Chavacano. Era ansina tambien sila aprecia de aton bonito lenguaje.»
«Al volver a nuestra ciudad me asusté, pues las chicas y chicos ya no hablan nuestra lengua; hablaban solo en tagalo. Aunque mis hijos no hayan nacido en Zamboanga, les enseñé a hablar mi lenguaje, el chabacano. Ojalá ellos también apreciaran de la misma manera nuestra bonita lengua.»

(6) «Enantes, a las doce mediadia, un grupo estudiante del Ateneo de Zamboanga alla na Terminal 3 ya llega. Ya observa yo puro Tagalog man ta conversa. Por que man?»
«Un poco más temprano, a las doce del mediodía, un grupo de estudiantes de Ateneo de Zamboanga llegaba [allá] al Terminal 3. He observado que hablaban únicamente tagalo. ¿Por qué será?»

Algunos comentarios (ej. 7, 8 y 9) van más allá, poniendo en relieve el prestigio del tagalo sobre el CZ, lo que llevaría a ciertos hablantes a pasarse a esa lengua, aun si su competencia en esa lengua es dudosa. También hay jóvenes (ej. 10 y 11) a los que les molesta esta situación.

(7) «Topao tu, [NOMBRE CENSURADO]. Maga ta'n pa sabiondo gayod sila man Tagalog, y despues hende man amo el maga palabra que ta usa sila. No sabe sila usa maga panlapi o prefix. Si ta oi yo ansina, ta correcta yo enseguidas. Y ta habla ya yo: "Por que man Tagalog pa y despues hende man amo? Man Chavacano kay taqui man kita na Zamboanga."»
«Tienes razón, [NOMBRE CENSURADO]. Hablan tagalo como si fueran unos sabelotodo, pero luego las palabras que usan no son correctas. No saben usar los prefijos. Cuando oigo a uno de esos, le corrijo enseguida. Y le digo: "¿Por qué hablas en tagalo si lo haces mal? Habla en chabacano, al fin y al cabo estamos en Zamboanga".»

(8) «Amo man... Conversa man Tagalog, pero quebrao man... Hay.»
«Así es... Hablan en tagalo, pero encima chapurreando... Ay.»

(9) «Ta pensa el los demas mas social si conversa Tagalog. Por eso mayoria de jovenes este tiempo no sabe mas conversa el dialecto Chavacano.»
«Los demás piensan que es más fino hablar en tagalo. Por eso, la mayoría de los jóvenes de hoy en día ya no saben hablar el dialecto (sic) chabacano.»

(10) «Chavacano is a beautiful dialect. No hate, but I wish instead of answering me in "baluktot tagalog" people will answer me in Chavacano na lang when I speak to them in Chavacano. Que bonito oi si ta conversa Chavacano.»
«El chabacano es un bonito dialecto (sic). Lo digo sin rencor, pero me gustaría que en lugar de contestarme en tagalo mal hablado la gente me contestara simplemente en chabacano cuando les hablo en chabacano. Qué bonito oír hablar en chabacano.»

(11) «Por que ba gad el maga gente na pueblo de Zamboanga, man cuento kanila Chavacano contesta con ikaw Tagalog?»
«¿Por qué demonios la gente en el centro de Zamboanga, si les hablas en chabacano, te contestan en tagalo?»

Pese a que la lengua es hablada también como segunda lengua por un número considerable de los habitantes de la ciudad de Zamboanga, algunos zamboangueños (ej. 12, 13, 14 y 15) sugieren que algunas personas nacidas fuera de la ciudad o algunos hijos de inmigrantes ya no estarían aprendiendo la lengua. Esto demuestra que la lengua empieza a perder su papel de lengua franca, aunque siga siendo el mayor símbolo de la ciudad.

(12) «Hace kita formal estudio porque el maga jovenes de Zamboanga, masquen ya nace aqui o mas de 10 años ya aqui na Zamboanga, no sabe pa conversa Chavacano.»
«Hagamos formal el estudio [de la lengua chabacana en las escuelas] porque los jóvenes de Zamboanga, sean los nacidos aquí o los que llevan más de 10 años aquí en Zamboanga, todavía no saben hablar en chabacano.»

(13) «Mismo, ya anda yo na pueblo, maga tendera ta'n Tagalog. Ya habla yo, hende yo ta entende Tagalog kay Zamboangeña yo. El uno tendera no sabe conversa Chavacano. Ya habla con el manager: para que taqui na Zamboanga si no sabe man Chavacano?»
«Yo también, fui al centro, las dependientas hablaban en tagalo. Les dije que no sé hablar tagalo porque soy zamboangueña. La dependienta no sabía hablar chabacano. Le dije al gerente: ¿Qué hace [la dependiente] en Zamboanga si no sabe hablar chabacano?»

(14) «Mucho maga negociante ta emplea con el maga empleados no sabe conversa chavacano. Maga clientes Chavacanos forzao conversa quebrao Tagalog.»
«Muchos dueños de negocio están contratando empleados que no saben hablar chabacano. Los clientes chabacanos estamos forzados a hablar en tagalo chapurreado.»

(15) «Tiene maga maestra no sabe conversa Chavacano. Debe na Abril hasta Mayo, tiene el DepEd programa para amola y enseña el Chavacano na maga maestra, especialmente con aquel hende de aqui kanaton, para efectivo el instruccion na clase.»
«Hay profesoras que no saben hablar chabacano. El Departamento de Educación debería tener un programa, entre abril y mayo[8]*, para afinar [los conocimientos] y enseñar el chabacano a las profesoras, especialmente las que no son de aquí, para que la enseñanza en clase sea efectiva.»*

Estos relatos, en cierta medida, corroboran las cifras de los últimos CPH. Aunque el número absoluto de hablantes nativos del CZ sigue en aumento, desde hace algunas décadas, la proporción de la población que lo tiene como lengua nativa sufre un decremento gradual. La tabla 3 nos permite observar ese decremento a lo largo de los últimos 40 años censados entre 1970 y 2010 (National Census and Statistics Office (NCSO), 1974, 1983; National Statistics Office (NSO), 1992, 2003b, 2014b).

Tabla 3: Porcentaje de hablantes nativos de CZ sobre el conjunto de la población de la Ciudad de Zamboanga según los CPH del 1970, 1980, 1990, 2000 y 2010. (National Census and Statistics Office (NCSO), 1974, 1983; National Statistics Office (NSO), 1992, 2003b, 2014b)

AÑO	%	HABLANTES	POBLACIÓN
1970	58,33 %	116.611	199.901
1980	53,15 %	182.701	343.722
1990	48.71 %	215.490	442.345
2000	46.57 %	280.252	601.794
2010	43.39 %	350.240	807.129

Las cifras y los comentarios aquí reproducidos, sin embargo, no son concluyentes. Se hacen necesarios estudios que verifiquen las consecuencias de los cambios sociales y demográficos sobre el uso del CZ en la Ciudad de Zamboanga para comprobar si de hecho hay una substitución lingüística en curso y a qué velocidad ésta se está produciendo. Resultaría igualmente interesante estudiar las actitudes de los hablantes, tanto nativos como de segunda lengua, hacia el CZ, y qué porcentaje de la población no tiene el CZ como lengua nativa y sabe hablarlo.

[8] Época de vacaciones escolares en Filipinas.

1.2.2 *Evolución e interferencias*

Algunos comentarios ponen en duda la competencia de los zamboangueños jóvenes y critican ciertos usos. La razón es que ciertos rasgos (ej. 16) y vocablos (ej. 17), antes asociados a hablantes no nativos y considerados incorrectos por las generaciones anteriores, ahora forman parte también del habla de los jóvenes nativos y de muchos mayores. En las generaciones más jóvenes, tiene lugar una «nivelación lingüística», es decir, la reducción de la variación lingüística interna del CZ, dando lugar a una mayor homogeneidad de la lengua. También se observa una aproximación a las lenguas del sustrato filipino (Lipski, 2013) en ciertos aspectos, además de muchas interferencias que afectan incluso a los profesionales que utilizan la lengua en los medios de comunicación de masas (ej. 18, 19 y 20).

(16) «Si Papa ta rabia gayod si ta oi ele na maga gente si ta usa el palabra ka que debe tu.
Ejemplo:
Si na pregunta "Onde **ka** anda?", "Cosa man **ka**?".»
«Papá se enoja mucho al oír a la gente usar la palabra "ka", que debería ser "tu".
Ejemplo:
A la pregunta "¿Adónde vas?", [contesta] "¿Qué es eso de 'ka'?"»

(17) «Cosa na Chavacano el **PINAKA**, **DAPAT**, **BAWAL**, **GAMIT**? Mayoria del maga jovenes este palabras de Tagalo ya ta usa... Tiene pa gale otro palabra: **BIGLA**...»
«¿Cómo se dicen en chabacano[9] *"PINAKA", "DAPAT", "BAWAL", "GAMIT"? La mayoría de los jóvenes ahora usan estas palabras del tagalo... Hay otra palabra más: "BIGLA"...»*

(18) «No sabe mas seguro el mas mucho del maga Zamboangueños si cosa el "tortuga", por eso el anunciador de ABS-CBN ta usa ya lang el palabra "pakiwan"... Makatriste man kay hasta el simple palabra ta perde ya lang lenguage Chavacano... Era ayuda estos na media para revivi ole este maga palabras...»
«A lo mejor la mayoría de los Zamboangueños no saben qué es una "tortuga", por eso el presentador de ABS-CBN[10] *simplemente usa la palabra "pakiwan"... Es triste que ahora incluso las palabras sencillas se estén perdiendo... Ojalá la gente de los medios ayudara a revivir otra vez estas palabras...»*

[9] En castellano, **pinaka~:** el más ~; **dapat:** debe; **bawal:** prohibido; **gamit:** cosa
[10] Nombre de una gran cadena de televisión en Filipinas.

(19) «El palabra "sucede" na TV Patrol Chavacano[11] ta habla sila "ocurra"!»
«La palabra "sucede" en TV Patrol Chavacano ¡la cambian por "ocurra"!»

(20) «Na radio ya habla "protecta". Cosa el recto: "protecta" o "protege"?»
«En la radio han dicho "protecta". Cuál es el correcto: ¿"protecta" o "protege"?»

Aunque es cierto que este tipo de interferencias es normal entre lenguas en contacto, hace falta estudios más profundos que señalen si el CZ avanza hacia la sustitución lingüística o no.

1.2.3 *Inseguridad entre los jóvenes*

Aunque, por un lado, los jóvenes son objeto de las críticas de los mayores por una supuesta falta de competencia adecuada en CZ, por otro, son a veces los mismos jóvenes los que expresan su inseguridad a la hora de utilizar la lengua. En las interacciones personales con el autor o en la recogida de datos de este y otro trabajo aun no publicado (Himoro, s.f.), algunos jóvenes (ej. 21-27) han manifestado sus inquietudes respecto a la manera de hablar y/o escribir la lengua. Reproducimos algunos de sus comentarios a continuación.

(21) «Ta pedi lang yo dispensa kay mio Chavacano ay hende mas gad igual de un real Chavacano igual del maga mayores. Mesclao ya por eso kame ta conversa.»
«Solo me gustaría pedirle disculpas porque mi chabacano no es igual al de un verdadero chabacano, como el de nuestros padres. Es que ahora lo hablamos mezclado.»

(22) «No sabe yo si justo 'quel mio contestada.»
«No sé si mis respuestas son correctas.»

(23) «Hende yo sure si amo ya 'quel.»
«No estoy segura de que [lo que contesté] sea correcto.»

(24) «Caso tiene word hende yo sabe na Chavacano. Hehe.»
«El problema es que hay palabras que no sé en chabacano. Jeje.»

(25) «Okay lang uy!!! Kay yo tambien not that good in chavacano!»
«Da igual, ¡¡¡eh!!! ¡Pues yo tampoco soy tan buena en chabacano!»

(26) «I am guilty that I was born and raised here in Zamboanga yet I do not know how to write in Chavacano. I really wanted to learn it because Chavacano is a very beautiful dialect.»
«Soy culpable por haber nacido y haber sido criado aquí en Zamboanga y no saber

[11] Telediario en CZ emitido por ABS-CBN

cómo escribir en chabacano. Me gustaría mucho aprenderlo porque el chabacano es un dialecto (sic) muy bonito.»

(27) «Hende recto el palabras de Chavacano que sabe yo.»
«Las palabras chabacanas que sé no son correctas.»

Los hablantes jóvenes parecen ser conscientes de las diferencias que hay entre su CZ y el de los mayores. Además del reproche de los de más edad, parte de la falta de confianza que tienen algunos hablantes jóvenes podría deberse a las burlas de las que son objeto, tema que en el apartado 1.2.4. Otra razón podría ser la falta de referencias adecuadas, como obras que sirvan de modelo de registros más formales de la lengua, o bien la falta de terminologías definidas para diferentes ámbitos que permitirían un uso más eficiente de la lengua.

Una codificación adecuada podría subsanar estos problemas y contribuir a contrarrestar la sensación de falta de competencia de los hablantes, o la idea errónea de que el CZ no tiene el mismo poder de expresión que otras lenguas como el tagalo o el inglés. Un estudio profundo de las actitudes de los hablantes hacia la lengua podría servir de guía para la implementación de una política lingüística concreta que contribuya verdaderamente a la preservación y el cultivo del CZ.

1.2.4 *Purismo y shaming en internet*

Si, por un lado, los medios de comunicación de masas tradicionales son a menudo vistos como villanos en la sustitución lingüística de las lenguas minoritarias, por otro, el surgimiento de la Web 2.0 permitió que cualquier persona produjera contenido en la lengua que le apetezca (Jones & Uribe-Jongbloed, 2013). No cabe hablar de igualdad de condiciones, ya que inevitablemente las lenguas mayoritarias siempre tendrán una gama mucho mayor de contenidos y recursos disponibles. No obstante, a diferencia de otros medios de comunicación, el usuario tiene libre control para elegir lo que desea consumir y, si así lo desea, también para crear contenido. La popularidad de perfiles de *memes* y de algunas celebridades, como *youtubers*, que utilizan el CZ muestra que hay espacio para las lenguas minoritarias y que apostar por la lengua materna puede ser una manera especial de conectar directamente con su audiencia.

En CZ, como en otras lenguas minoritarias y minorizadas, existen muchas personas que mantienen una postura purista o diferencialista frente a la(s) lengua(s) dominante(s). Al mismo tiempo que internet brinda más visibilidad al trabajo de los artistas y creadores de contenido, el anonimato que proporciona también hace que muchas personas se comporten de manera que no siempre es la más adecuada. Un comportamiento recurrente por parte de los usuarios suele ser el del *shaming* siempre

que identifican un error lingüístico u ortográfico: consiste en señalar públicamente el error, a veces con ánimo de educar, pero en ciertos casos también con ánimo de atacar, llegando al nivel de la grosería, del escarnio y de la humillación. Tal práctica es bastante común también en las lenguas más habladas (Badilla, 2014; Heisel, 2015; Neves, 2015).

A continuación reproducimos algunos comentarios extraídos de vídeos de grupos de música zamboangueños en YouTube. Algunos comentarios datan de hace más de 10 años, mientras que otros son un poco más recientes. Aunque algunas personas se ofrezcan para ayudar en privado (ej. 28), otros son más directos y señalan públicamente el error y la corrección (ej. 29, 30 y 31). Otros se burlan utilizando alguna forma de ironía (ej. 32 y 33) o simplemente son groseros (ej. 34, 35, 36 y 37), aun cuando intentan atenuar el tono de su mensaje con elogios antes o después de la crítica.

(28) «Si okay lang con vosotros, tiene yo ya hace lyrics del "Cuando". Mas o menos no hay man yo cosa ya cambia, el "conjugation" lang del palabra y maga "spelling". I have been an avid learner of the Chabacano language. Mas o menos ta hace yo manera para puede busca el maga propio palabra en Chabacano. Si quiere ustedes, ta pedi yo que man e-mail o man message conmigo aqui na YouTube, I'd be glad and try to help rewrite the spellings of each word... Contribution ya este para na lenguaje en Chabacano.»
«Si os parece bien, he hecho la [corrección de la] letra de "Cuando". No he cambiado gran cosa, solo la "conjugación" de las palabras y la ortografía. Soy un ávido aprendiz de la lengua chabacana. Me he centrado un poco en encontrar las palabras correctas en chabacano... Si lo queréis, os pido que me enviéis un correo o un mensaje aquí en YouTube, estaría encantado de intentar ayudar a revisar la ortografía de cada palabra... Esta es una contribución para la lengua chabacana.»

(29) «Sana yung correct spelling (in chavacano) dapat tama like "keda" dapat "queda", "resa" dapat "reza".»
«Si estuviera en grafía correcta (en chabacano), lo correcto sería que "keda" fuera "queda", "resa" debería ser "reza".»

(30) «Actualmente, este es debe deletrea como "Tiene Vez" con el letra zeta. :)»
«En realidad, esto debería escribirse "Tiene Vez", con la letra Z. :)»

(31) «Jendeh/hinde are accepted... They're all accepted... But "jendeh" is the right word.»
«"Jendeh"/"hinde" están aceptados... Están todos aceptados... Pero "jendeh" es la palabra correcta.»

(32) «Como clavos kalawang na de usted ojos el ortografia de este cancion? Jajajaja!!! Pero, ta mira yo ta entende man gaha los hispanicos el pronunciation del Chavacano...»
¿La ortografía de esta canción son como clavos oxidados en tus ojos? ¡¡¡Jajajaja!!! Pero creo que a lo mejor los hispanos entienden la pronunciación del chabacano...

(33) «Sana yung GRAMMAR ng lyrics nila itama kasi maganda ang song...»
«Estaría bien si corrigiera la gramática de la letra, ya que es una bella canción...»

(34) «I am from Zamboanga and I just can't believe that the one who posted the lyrics have not asked or consulted those who can speak Chavacano... Like the words "jalo", should have spelled "dehalo", "hinde" should have spelled "gendeh"... I like the song, really, and I am not an educator, it just so happen I got hold of Chavacano Bible and other stuff, hence I was able to know some misspelled Chavacano words... But then again, hope on their next video, they will post the correct ones ya...»
«Soy de Zamboanga y simplemente no puedo creer que el que publicó la letra no preguntara o consultara a los que saben hablar chabacano... Por ejemplo, la palabra "jalo", debería ser escrita "dejalo", "hinde" debería ser escrita "gendeh"... De verdad que me gusta la canción, y no soy un educador, solo resulta que tengo la Biblia chabacana y otras cosas, por eso sé que algunas palabras en chabacano están mal escritas... Pero de nuevo, espero que en el vídeo siguiente las publiquen correctamente...»

(35) «Next time kung man post cancion y palabra de Chavacano, favor tambien hacer claro y hacer buenamente el spelling. Pero vosotros cancion y tonos ay bien vale... Vaya con Dios, maga amigos.»
«La próxima vez, si publicáis [la] canción y [la] letra en chabacano, por favor corregidla y escribidla bien. Pero vuestra canción y tonos son muy guays... Id con Dios, amigos.»

(36) «Muy bonito el los canciones. Pero maga compañero, visia embuenamente el maga palabra. Hende mas amo el gramatico. El "ahora" ya queda "aura", el "no hay" ya queda "nuay", y hende mas amo el llevada del maga palabra. Pero bonito! Keep up the good work y vaya con Dios!»
«Muy bonitas las canciones. Pero compañeros, cuidad bien las palabras. La gramática no está bien. "Ahora" se convirtió en "aura", "no hay" se convirtió en "nuay" y el acento de las palabras no es correcto. ¡Pero lo hacéis bien! ¡Seguid así e id con Dios!»

(37) «Otro tambien pataka el deletreo/spelling. "Vira Ole" y hende "Bira Ole". Y na un lirico, "Yo ahora ta sufri" y hinde "Io aura ta supri"... Research-research anay.»

«Otra falta más en la ortografía. "Vira Ole" y no "Bira Ole". Y en una parte de la letra, "Yo ahora ta sufri" y no "Io aura ta supri"... Probad a investigar más antes (de publicar).»

Las reacciones de las personas criticadas, justamente por la manera agresiva en la que se dirigen a ellas, no siempre es positiva, como vemos en los ej. 38 y 39.

(38) «Para conmigo, no hay yo que ver na maga critico. Ansina gad el vida: masquen amo o mali ta hace, tiene gad maga critico... El lenguaje kita ta lleva ahora o cosa ya kita ya agranda palabra Chavacano, tormento ya kita 'se cambia... Uno pa ya engranda, maga gente 'qui na Zamboanga no hay proper education de lenguaje Chavacano... Ingles gale hende todo bueno y perfecto masquen tiene ya subject Ingles.»
«A mí me la suda [esa clase de] críticas. Así es la vida: lo hagas bien o mal, habrá críticas... La lengua que tenemos ahora o las palabras chabacanas que hemos creado, es complicado que las cambien a estas alturas... Además, la gente aquí en Zamboanga no tiene [acceso a una] educación adecuada en lengua chabacana... [Nuestro] inglés tampoco es del todo bueno ni perfecto, aunque tengamos la asignatura de inglés.»

(39) «Usted ya real, hace ya lang tambien 'te cancion!»
«Si tan bueno eres, ¡prueba tú hacer la canción!»

Pero no todos se lo toman a mal. Uno de los grupos musicales zamboangueños que se podrían considerar pioneros en YouTube mostró durante un corto periodo de tiempo una evolución bastante clara en la manera en la que escribía la letra de sus canciones, yendo de una grafía tagala a una grafía más castellana. Cuando fueron interrogados por la razón de esos cambios, confirmaron que se debió a las críticas recibidas en los comentarios (40):

(40) «Yeah... I mean, when we wrote the lyrics for that, we only wrote what we knew. And that was in "Tagalog". We don't really have proper knowledge of how Chavacano is written because it is not taught in our schools. We only know it because it is our native tongue. Someone told us to write it in formal Chavacano already and we need to be aware of our spelling. Because people now are listening to our songs.»
«Sí... Lo que quiero decir es que, al escribir las letras [para eso], escribimos lo que sabíamos. Y en ese caso, era en "tagalo". No tenemos conocimientos adecuados de cómo se escribe el chabacano porque éste no se enseña en las escuelas. Lo sabemos simplemente porque es nuestra lengua materna. Alguien nos dijo que lo escribiéramos en chabacano formal y que deberíamos poner atención a nuestra ortografía. Porque la gente ahora escucha nuestras canciones.»

La mayoría de estos comentarios datan de antes de la aprobación de una ortografía para el CZ. Puesto que no había publicaciones oficiales de ninguna autoridad de la lengua, las correcciones ofrecidas se basaban sobre todo en criterios e ideologías personales. Algunas personas algo más conscientes salían a contestar sobre la validez de esas correcciones (ej. 41 y 42)

(41) «[NOMBRE CENSURADO] El chabacano todavía no tiene estandardización y por eso es correcto escribir "keda" o "queda." Del punto de vista español son incorrectos, pero el español y el chabacano son idiomas distintos y [tienen] sus propias reglas...»

(42) «[NOMBRE CENSURADO] That's the thing: there is NO standard way of writing Chavacano. People write as they please. There is no right way there is no wrong way. Using Spanish as yardstick doesn't cut it, because Chavacano is a separate language now from Spanish.»
«[NOMBRE CENSURADO] Ese es el problema: NO hay una manera normativa de escribir en chabacano. La gente lo escribe como le da la gana. No hay manera correcta, no hay manera incorrecta. Usar el castellano como patrón no sirve, porque el chabacano es ahora una lengua separada del castellano.»

A veces, también hablantes de castellano manifestaban sus opiniones sobre el tema, no siempre encontrando el respaldo de los hablantes de CZ, como vemos en el ej. 43:

(43) «Por que sila ta reclama si que laya kita ta habla y ta escribi de aton lenguaje? No hay sila que ver kanaton.»
«¿Por qué se están quejando sobre cómo hablamos y escribimos nuestra lengua? Esos no tienen nada que ver con nosotros.»

También son víctimas de *shaming* algunos autores de *memes* en CZ, como podemos observar en los ej. 44 y 45, provenientes de páginas de Facebook.

(44) «Seguro hende original Chavacano el writer... El "hente" (gente) hende amo spelling.»
«El escritor tal vez no sea un verdadero chabacano... La escritura de "hente" (gente) es incorrecta.»

(45) «Hahaha. Real ba gayod 'se Chavacano. Cosa man "asa"? "Esperanza" gaha? Cosa ya gayod!»
«Jajaja. Este chabacano es la hostia. ¿Qué es "asa"? "Esperanza", ¿quizás? ¡Vaya!»

En algunos grupos de Facebook, aparecen asimismo algunas opiniones contrarias a la manera en la que la gente se corrigen los unos a los otros:

(46) «No mas lang ustedes rabia, pero tiene vez makaperde gana man post (dispensa kay no sabe yo na Chavacano el palabra "post") aqui kay mucho ta correcta el Chavacano y ta precura hace na Español. 'Cabar, ta rabia pa y ta hace huya na gente si medio hende amo el ya usa spelling (cosa el "spelling" na Chavacano?) o el palabra ya usa... Era si ta enseña man ustedes o ta correcta, no mas ya insulta o hace huya... Mucho maga miembro aqui tarda ya no hay na Zamboanga y hende ya Chavacano el de ila maga primer lenguaje. El Chabacano hende Español... El Chabacano un mismo lenguaje, tiene lang maga palabra daw palabra de Español. Gracias.»
«No os enfadéis, pero a veces se me van las ganas de escribir posts (perdón, no sé como se dice en chabacano la palabra "post") aquí porque muchos me corrigen el chabacano e intentan convertirlo en castellano. Al final, se enfadan y te ponen en ridículo si tu escritura (¿cómo se dice "escritura" en chabacano?) o la palabra que estás usando no son muy correctas... Me gustaría que, si desean enseñar, no insultaran o ridiculizaran a la gente... Muchos miembros aquí hace tiempo que no están en Zamboanga y el chabacano no es su lengua materna. El chabacano no es castellano... El chabacano es una lengua aparte, simplemente tiene palabras que son palabras del castellano. Gracias.»

El *shaming* termina por provocar miedo en algunos hablantes a la hora de utilizar el CZ en público, haciendo que estos se pongan a la defensiva y prevean ya las posibles críticas antes incluso de hacer su aportación, como vemos en los comentarios 47 y 48.

(47) «Perdoname si mali spelling de Chavacano words, no hay pa mother tongue de mio tiempo.»
«Perdonadme si la escritura de las palabras en chabacano está mal, no había [clases de] lengua materna en mi tiempo.»

(48) «Puede ba usted oi na mi cancion? Y basi puede tambien usted dale comento. Ta pedi yo dispensa si tiene man maga mali na mi maga letra o palabra aqui na mi cancion kay na mi generación bien poco ya lang el sabe conversa de aton Chabacano de antes y ta admiti yo kay uno ya yo na maga no sabe. Ya aprende lang yo con este usando mi maga orejas y no hay yo recurso como un libro de Chabacano. Pero uno tambien yo na maga persona que quiere preserva con el lenguaje Chabacano por medio de maga moderno canciones. Por favor dale usted consideracion. Gracias!»
«¿Puede usted escuchar mi canción? Y quizás también comentarla. Pido disculpas si hay faltas en mi letra o en las palabras aquí en mi canción, ya que de mi generación son muy pocos los que saben hablar nuestro chabacano antiguo y admito que soy uno de los que no lo sabe. Lo aprendí de oído y no tengo recursos como libros de

chabacano. Pero soy también una de esas personas que quiere preservar la lengua chabacana por medio de canciones modernas. Por favor, téngalo en consideración. ¡Gracias!»

Un creciente número de estudios sugiere que la retroalimentación correctiva en clases de idioma, aunque deseable hasta cierto punto, debe hacerse con cautela y moderación para no agobiar demasiado al estudiante ni socavar su confianza (Al-enzi, 2011; Martínez Agudo, 2008). Pese a la falta de estudios que evidencien los efectos del *shaming*, el paralelo con las clases de idioma nos lleva a creer que tal práctica podría ser daniña para los hablantes de la lengua, que en muchos casos podrían verse inhibidos y preferir escribir en inglés o en tagalo con tal de evitar las críticas. Aunque en algunos casos se haga con el ánimo de educar, corregir a los demás de manera no adecuada puede tener un efecto adverso y hacer que menos gente se anime a utilizar la lengua públicamente.

1.3 MEZCLA Y CAMBIO DE CÓDIGO

Un fenómeno particular de la sociedad filipina es el de la «mezcla de código» y el «cambio de código», generalizados incluso en los medios de comunicación de masas (Cook, 2018). Aunque no todos los autores distingan claramente entre los dos, una definición posible para «mezcla de código» es la transferencia de cualquier elemento lingüístico externo, tanto unidades léxicas como gramaticales, en una frase, mientras que el «cambio de código» consistiría en el cambio de una lengua a otra, pudiendo producirse a mitad de frase o entre frases (Grosjean, 1982 y Torres, 1989 citados en Kim, 2006). En muchos casos, resulta incluso difícil determinar en qué lengua se está hablando.

En Zamboanga, particularmente, las lenguas que entran en la mezcla pueden variar según la procedencia de los interlocutores que participan en la interacción. Algunos ejemplos extraídos de redes sociales ilustran el uso del CZ junto al tagalo (49, 56), al cebuano (50, 57), al inglés (49, 51, 52, 56) e incluso al tausug (54, 55) y al chino fukianés[12] (52, 53).

(49) «**Ang kahit gustong-gusto ko ng mag-reply**, pero ta'n pa caro anay yo para habla kay rabiao yo.»
«***Aunque me encantaría mucho contestar**, voy a hacerme un poco la difícil para decir que estoy enfadada.»*

[12] Originalmente, las palabras de origen china estaban escritas en alfabeto romano. Las hemos normalizado de acuerdo con el sistema POJ y agregado los respectivos caracteres chinos según las recomendaciones del Gobierno de Taiwán.

(50) «**Ok lang dinhi,** primo. Trabajo. Habla lang yo con **Daddy**.»
«***Aquí tirando**, primo. Trabajando. Voy a hablar con **papá**.»*

(51) «[NOMBRE CENSURADO] **Why** man? No hay 'le **work**? Manda con ele **man work**. Tsk.»
*«[NOMBRE CENSURADO] Pero, ¿**por qué**? ¿Él no tiene trabajo? Hazle **trabajar**. Joer.»*

(52) «阮欲**(gún beh) edit** luego :) 我**(góa) send** 予你**(hō–lí)** el **photos**.»
*«Ya edit**amos** :) **Te** mand**o** las **fotos**.»*

(53) «[NOMBRE CENSURADO] Gah, anda-anda otro lugar, 'cabar **無錢(bô chî^{n})**. Tsk... haha»
*«Vaya, siempre vas a otros lugares, y después **no tienes pasta**. Ejem... jaja»*

(54) «[NOMBRE CENSURADO] Gah. Bien moro man **ini**, bro. Haha. **Syu ini? Makaastul.**»
*«Ostras. **Esto** es muy moro, tío. Jaja. **¿Qué es eso? Qué asco.**»*

(55) «Ese yo quiere! **Magsukul**, 'Neng, ah este, Teacher [NOMBRE CENSURADO]. **Kaw na in mastal ko for Tausug 101.** :D»
*«Eso es lo que quería! **Gracias**, guapa, ah que va, Profesora [NOMBRE CENSURADO]. **Ahora eres mi profe de tausug básico.** :D»*

(56) «[NOMBRE CENSURADO] [NOMBRE CENSURADO] **Wala** gale, ha! **I told you a lot of times na, girl.** Evos lang 'se no hay plano anda. Haha.»
*«[NOMBRE CENSURADO] [NOMBRE CENSURADO] Así que **no**, ¿eh? **Ya te lo he dicho un montón de veces, maja.** Eres la única que no piensa ir. Jaja.»*

(57) «[NOMBRE CENSURADO] Taqui ya yo apartment ole. Ya vira **na pud** yo enantes para aqui. Anda kamo hospital?»
*«[NOMBRE CENSURADO] Ya estoy de nuevo en el apartamento. Acabo de volver **otra vez** aquí. ¿Vais al hospital?»*

Aunque la mezcla y el cambio de código puedan ocurrir, en mayor o menor medida, entre zamboangueños de cualquier edad y origen, estos suelen ser más frecuentes entre los más jóvenes. Ocurren en contextos informales, pero pueden ocurrir también en contextos semi formales. Ciertos sectores de la sociedad parecen desaprobar tal práctica, como se observa en los comentarios 58 y 59.

(58) «Si, por eso gane con mi maga sobrino y sobrina ta habla yo: "Si ta'n cuento masquen cosa lenguaje, el cosa ya principia amo hace acaba y no mezclao o halo-halo".»
«Si, exactamente por eso a mis sobrinos les digo: "Habléis la lengua que habléis, terminad con lo que habéis empezado y no [lo] mezcléis".»

(59) «Makaalmaria oi:
- Para onde **ka** anda?
- Anda yo compra **tinapay**!
- **Bukas** tiene **tayo** clase.
- **What** ba tu?
Y mucho pa.
Con mi maga anak, si ta conversa baliskad patuad kay ta segui na uso, ta habla yo kay hende yo ta entende... :)»
«*Me marea oír:*
*- ¿Adónde va**s**?*
*- ¡Voy a comprar **pan**!*
*- **Mañana** tenemos clase.*
*- ¿**Qué** eres?*
etc.
A mis hijos, si me hablan en ese guirigay por seguir la moda, les digo que no los entiendo... :)»

Por un lado, las razones detrás del cambio de código pueden ser muchas. Los trabajos acerca del *Taglish*, mezcla de tagalo y de inglés omnipresente en la vida cotidiana de los filipinos y también en los medios de comunicación de masas, señalan como principales razones para el cambio de código el prestigio (Sanchez, 2013), la falta de competencia lingüística, la precisión y la eficiencia en la comunicación, la identidad, la naturalidad, la creación de un efecto cómico, el secretismo y otros (Bautista, 2004; Goulet, 1971). Hacen falta trabajos que investiguen si las razones del cambio de código para el CZ son las mismas que para el *Taglish*.

Por otro lado, aunque en los registros formales escritos en CZ su ocurrencia es baja, en ciertos casos se utiliza la mezcla de código, junto a los préstamos, como recurso complementario para llenar los huecos terminológicos y estilísticos que todavía existen en CZ.

1.4 EL CZ ESCRITO

El único estudio que hemos podido localizar en la literatura que trata el tema de la práctica escrita en CZ es el de Tobar Delgado y Fernández (2019). En este estudio, son identificadas dos corrientes principales en las fuentes lexicográficas del CZ:

- La **corriente etimológica**, que defiende la conservación de la ortografía de la lengua de origen de la palabra. Esto se traduce en destacar el componente castellano del CZ (fuente de un 80 % del vocabulario) como marca de identidad,

lo que conlleva prescindir (1) de representar las innovaciones fonéticas del CZ y (2) de tener un sistema previsible y regular.

- La **corriente fonológica**, que defiende el uso de una grafía fonética, con lo que resulta en un sistema más intuitivo para los hablantes que desconocen el castellano, pero al mismo tiempo acerca el CZ al tagalo y al cebuano, lo que puede provocar el rechazo de algunos hablantes.

Sin embargo, cuando lidiamos con las prácticas del hablante corriente, nos topamos con un enorme horizonte de posibilidades. Aunque es cierto que muchos de ellos suelen decantarse, al menos en su intención, por una de las dos corrientes descritas, en la práctica sus grafías casi siempre podrían ubicarse en algún punto intermedio. El resultado son grafías mayoritariamente de una u otra corriente, pero de naturaleza mixta.

Dentro de la corriente etimológica, hay una posición que consiste en aplicar la etimología solo para las palabras de origen castellano, utilizando el alfabeto castellano para representar también el vocabulario local (ej. 60: ”kay“, de origen cebuana e ilonga, escrito ”cay“). No obstante, debido a las complicaciones de la grafía castellana y la evolución propia del CZ, en realidad esta acaba por convertirse en una grafía mixta, con aplicación de fonemas homófonos al azar, por hipercorrección o transcripción de la lengua hablada, según la mayor o menor consciencia que el hablante posea de la escritura castellana (ej. 60: «tamen», «aborido» y «centa» en lugar de «tambien», «aburrido» y «senta»).

(60) «Si talla **tamen** reclama **cay** caliente, **aborido**, pati no hay onde **centa**.»
«*Cuando está allá, se queja de que hace calor, está aburrido y no tiene donde sentarse.*»

Hay algunas palabras de alta frecuencia que, por influencia del uso del CZ en los medios y en carteles o panfletos, pueden aparecer escritas según su ortografía castellana en los escritos de aquellos que aplican una grafía más cercana a la del tagalo (ej. 61: ”corazon“).

(61) «Porke kuntigo yo ya skuhi? Awra mi **corazon** ta supri.»
(Maldita (2009). Porque. Transcripción no oficial encontrada en Twitter.)
«*¿Por qué te escogí? Ahora mi corazón está sufriendo.*»

Hay, además, casos que llevan el hablante a transferir grafías externas para el CZ, como pasa con los préstamos de origen castellano incorporados a las lenguas locales (ej. 62: «maskin» y «kontrabida» en lugar de «masquen» y «contravida»), palabras que tengan conatos en inglés (ej. 63: «official» en lugar de «oficial»), o casos de anglicismos castellanizados (ej. 63: «attende» [to attend] en lugar de «atende»).

(62) «**Maskin** onde tiene gayot **kontrabida**.»
«*Hay contrarios no importa donde.*»

(63) «Este debe **attende** el nuevo elegido **official** del Ciudad.»
«*De esto debe ocuparse el nuevo gobernante electo de la Ciudad.*»

En algunos casos, la misma palabra puede aparecer escrita de dos maneras diferentes en un mismo mensaje, como pasa con la palabra **tiene** en el ej. 64.

(64) «Ta pikura man yo **tiene** el kosa ka **chene**.»
«*Estoy intentando tener lo que tienes.*»

Esa gran variabilidad ortográfica es lo que exige un tratamiento diferenciado para la corrección ortográfica automática del CZ.

1.4.1 *Zamboanga Chavacano Orthography*

La ortografía del CZ, la «Zamboanga Chavacano Orthography», tiene sus orígenes en 2014, con la celebración del *Chavacano Orthography Congress*, evento en el cual se presentaron y discutieron los resultados de un trabajo de investigación de la Universidad Ateneo de Zamboanga (ADZU), financiado por el gobierno de la Ciudad de Zamboanga. El año siguiente se celebró la *Segunda Conferencia Nacional del Lenguaje Chabacano*, que tuvo como principal resultado la publicación del documento «*Revised Zamboanga Chavacano Orthography*» en el 2016 (Tobar Delgado & Fernández, 2019).

Como regla general, en este manual revisado, se define que, para las palabras cuya etimología pueda identificarse, sean seguidas las reglas de escritura de la lengua de origen; o sea, la ortografía pertenece a la corriente etimológica, mencionada en el apartado anterior. Sin embargo, el documento no entra en detalles sobre el tema, y algunos pasajes y ejemplos incluso llegan a contradecir lo dicho anteriormente.

En el 2018, el gobierno local publicó el diccionario «Chabacano/Chavacano Lexicography» en dos tomos. Durante la ejecución de este trabajo, desafortunadamente, no se ha tenido acceso al mismo, pero se espera que muchas de las dudas sobre cómo grafiar ciertas palabras se vean resueltas. Aquí, por lo tanto, nos guiamos únicamente por lo que indica el documento «*Revised Zamboanga Chavacano Orthography*».

Hasta la fecha desconocemos cualquier otra lengua criolla que, habiendo emprendido un proceso de normativización, haya elegido una ortografía semejante a la del CZ. Esto hace que sea de gran interés conocer las implicaciones que esa decisión pueda tener en las generaciones futuras con el paso del tiempo.

1.5 ESTRUCTURA DEL TRABAJO

En este capítulo, hemos presentado un panorama general del chabacano y sus variedades, centrándonos en el CZ, la variedad de interés para este trabajo, y hemos discutido algunos aspectos importantes de la situación del CZ y su problemática actual. En el capítulo 2, presentamos un estudio preliminar encaminado a verificar algunas hipótesis preliminares y a definir los objetivos de este trabajo. En el capítulo 3, presentamos algunos conceptos esenciales para la comprensión de este trabajo y las herramientas utilizadas en el mismo. En el capítulo 4, describimos brevemente la metodología empleada en este trabajo. En el capítulo 5, describimos las tareas y los experimentos realizados. En el capítulo 6, presentamos y discutimos los resultados de las tareas y de los experimentos realizados. En el capítulo 7, finalmente, presentamos nuestras conclusiones y algunas propuestas de trabajos futuros.

CAPÍTULO 2

JUSTIFICACIÓN Y OBJETIVOS

En este capítulo presentamos una hipótesis inicial, un estudio preliminar y un planteamiento de los objetivos de este trabajo.

2.1 HIPÓTESIS INICIALES

Como mencionamos anteriormente en la sección 1.1, el CZ tiene una ortografía desde 2014. Aunque, por un lado, esto supuso el fin a un largo y antiguo debate acerca de cómo debería escribirse la lengua, por otro lado, ha creado un nuevo problema, pues se llegó al acuerdo de que las palabras deben grafiarse siguiendo las reglas de escritura de la lengua de origen. Dado que no se trata de una grafía necesariamente fonética, la ortografía considerada como correcta para una palabra en muchas ocasiones no refleja ni la fonética ni la evolución particular del CZ.

Salvo raras excepciones, a diferencia de las generaciones anteriores, las más jóvenes no han tenido exposición al castellano. Con la supresión del requisito obligatorio de créditos de castellano en las universidades de Filipinas en el 1987, el contacto con el castellano hoy día se ve restringido a las personas que, sea por razones de trabajo (como es el caso de las personas que trabajan en uno de los centros de llamadas bilingües que operan desde Filipinas), sea por interés personal, deciden emprender su estudio.

Teniendo en cuenta la situación del CZ, inicialmente contemplamos las siguientes hipótesis:

1. La ortografía es demasiado complicada para el hablante medio si no se poseen conocimientos previos suficientes de castellano y de las lenguas del sustrato.

2. Tener alguna familiaridad con el castellano no basta para aplicar la ortografía de manera satisfactoria.

3. El hablante medio de CZ prefiere leer un texto en grafía no normativizada.

4. Un sector de la población demanda la creación de un corrector ortográfico de CZ.

Para comprobar esas hipótesis, llevamos a cabo el estudio preliminar descrito en el apartado 2.2.

2.2 MOTIVACIÓN: ENCUESTA Y ESTUDIO PRELIMINAR

Puesto que no hemos encontrado estudios relacionados con la ortografía o que contestaran nuestras suposiciones, nos propusimos elaborar un estudio preliminar por medio de un cuestionario en línea distribuido en las redes sociales. La población sujeto de este estudio son hablantes de CZ que residen tanto en Filipinas como en el extranjero, independientemente de si lo hablan como primera o segunda lengua. Se elaboraron dos versiones del cuestionario: una en chabacano y otra en inglés, ambas con contenidos idénticos. La versión chabacana fue validada y corregida por dos hablantes nativos. Antes de su distribución, el cuestionario fue evaluado con un número reducido de participantes.

Las preguntas fueron formuladas de acuerdo a nuestros objetivos: (1) evaluar la familiaridad de los hablantes con la ortografía del CZ; (2) su grado de preparación para aplicarla sin estudio previo; (3) el impacto del castellano en su aplicación; y (4) determinar si hay un grado de demanda suficiente para desarrollar un corrector ortográfico para el chabacano y en qué situaciones éste sería útil. Las preguntas eran de respuesta cerrada, aunque en muchos casos con un campo de respuesta libre «otro», para que el participante pudiera añadir una respuesta diferente de las proporcionadas. El cuestionario se dividió en cuatro apartados obligatorios:

- **Acerca de ti:** Este apartado recoge datos personales y sociales del participante, aunque no se solicitan datos demasiado invasivos, como los de naturaleza económica.

- **Acerca de tus usos del chabacano:** Este apartado se centra en el uso del chabacano en las diferentes esferas de la vida cotidiana, tanto en ambientes formales como informales, sean estos reales o virtuales.

- **Acerca del *Zamboanga Chavacano Orthography*:** Este es el apartado más importante del cuestionario. En la primera parte, presentamos la ortografía y, a partir de un texto corto, hacemos preguntas relativas a la familiaridad, la

legibilidad, y el interés del hablante respecto a la misma. En la segunda parte, evaluamos si se agradecería la existencia de un corrector ortográfico del CZ por la vía directa (pregunta explícita) y por la vía indirecta (desempeño en una tarea de identificación de errores ortográficos).

- **Acerca de este cuestionario:** El último apartado pregunta sobre cómo el participante se enteró del cuestionario y permite que éste deje su comentario en caso de considerarlo pertinente.

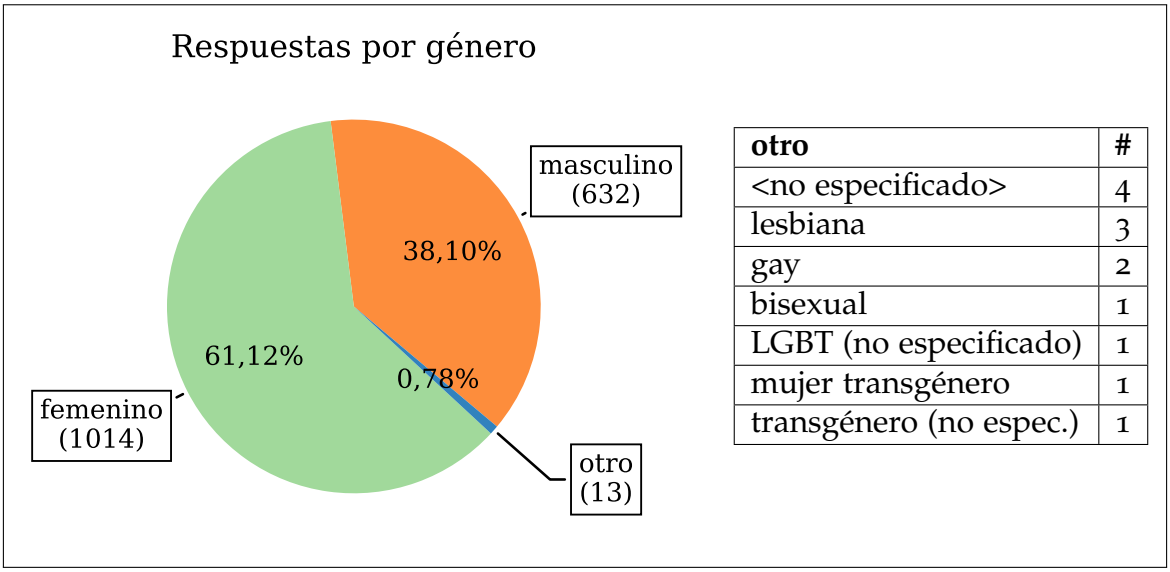
El cuestionario estuvo disponible en línea entre el 21 de mayo y el 25 de junio de 2018. El software utilizado fue LimeSurvey (LimeSurvey GmbH, 2018), una aplicación de código libre en lenguaje PHP y una base de datos MySQL/PostgreSQL/MSSQL, alojada en un servidor privado. Se distribuyó por Facebook, Twitter, Instagram y por reenvío directo de la información por los propios participantes a otras personas que también hablan la lengua. Algunas páginas y grupos de Facebook también aceptaron compartir nuestro cuestionario, lo que potenció su alcance y difusión. Además, como incentivo de participación, se ofreció el *ebook* de la versión en CZ del libro «What's in the pot?», distribuído de manera libre bajo una licencia *Creative Commons* de Reconocimiento.

La muestra contó finalmente con 1.659 participantes. Las dos versiones del cuestionario pueden consultarse en su integridad en los apéndices A y B. A continuación, analizamos brevemente los datos recogidos.

2.2.1 *Datos sociales*

El gráfico de la figura 1 muestra una prevalencia de participantes del sexo femenino (61,12 %), mientras la participación del sexo masculino sería de un 38,1 %, además de las personas que se consideran de otros géneros (un 0,78 %). A aquellos que eligieron la respuesta «otro», se les dio la posibilidad de detallar su género, aunque un 30,6 % prefirió no hacerlo. Hay que destacar que, entre los que detallaron su género, hay una clara confusión entre la noción de «orientación sexual» y la de «género».

En el gráfico de la figura 2, podemos observar la distribución de los participantes por edad y por género. Los porcentajes en el eje X, junto a cada grupo de edad, representan la proporción del grupo de edad dentro del total de participantes. Los porcentajes sobre las barras verticales del gráfico representan la proporción del grupo de edad dentro del total de participantes del género especificado. De esta manera, se puede observar que el 68,42 % de los participantes tiene entre 20 y 39 años. La baja participación de sujetos entre 10 y 19 años se debe sobre todo a la dificultad para llegar a esas personas. Respecto a los grupos de edad de 40 años o



otro	#
<no especificado>	4
lesbiana	3
gay	2
bisexual	1
LGBT (no especificado)	1
mujer transgénero	1
transgénero (no espec.)	1

Figura 1: Número de participantes del cuestionario por género

más, la barrera natural es sobre todo tecnológica, ya que esas personas tienden a utilizan menos internet. Asimismo, creemos que esta muestra, salvo para el grupo de edad de menos de 19 años, refleja bastante fielmente la distribución de la pirámide de población de Filipinas.

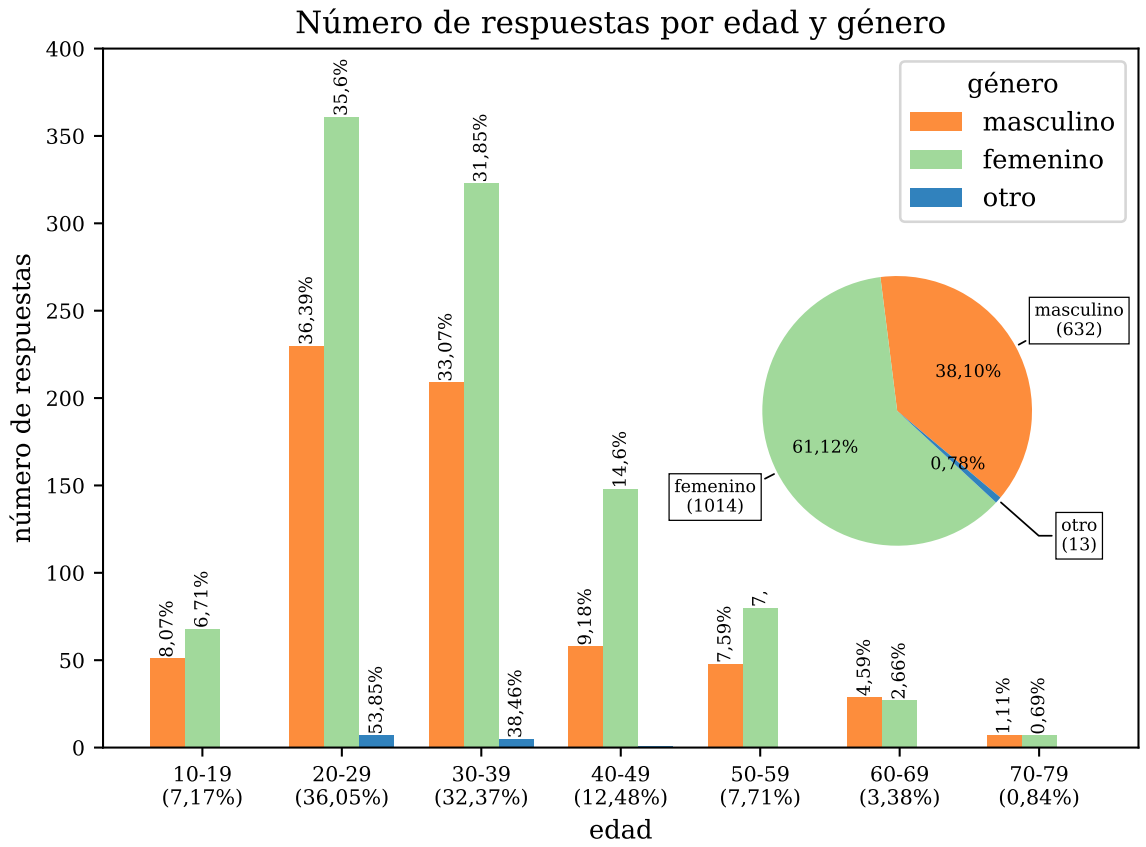
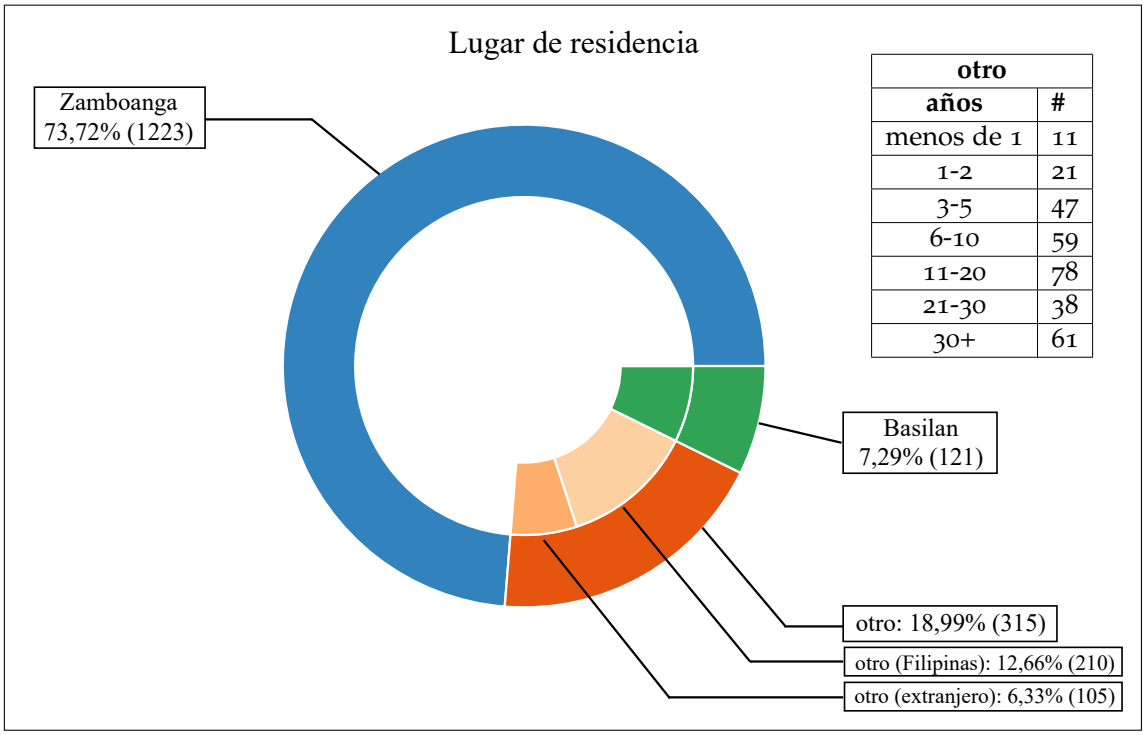


Figura 2: Distribución de los participantes del cuestionario por edad y por género

Respecto al lugar de residencia de los participantes, como se percibe en el gráfico de la figura 3, la gran mayoría afirma vivir en la Ciudad de Zamboanga (73,72 %), mientras el 7,29 % vive en Basilan. De los 18,99 % que afirmaron vivir fuera de esas dos localidades, el 66,67 % dice vivir en otros lugares de Filipinas y el 33,33 % en el extranjero. En la misma figura se incluye una tabla con el número de años que llevan viviendo fuera de las regiones chabacanófonas los participantes que eligieron la opción «otro». Esta tabla parece sugerir que muchos de los que se van lo hacen sin intención de volver durante un largo tiempo; esto es reflejo, tal vez, de la dificultad de encontrar trabajo en ciertas profesiones.



otro	
años	**#**
menos de 1	11
1-2	21
3-5	47
6-10	59
11-20	78
21-30	38
30+	61

Figura 3: Distribución de los participantes del cuestionario por lugar de residencia

Según el gráfico de la figura 4, la mayoría de los participantes tiene estudios superiores (90,72 %). Esto podría implicar que los resultados que encontremos en este estudio solo serían extrapolables a personas con un alto grado de instrucción. No obstante, este sesgo también podría deberse a un grado de cooperación en este tipo de estudios más bajo en los demás grupos, o bien a que la mayoría de seguidores de las páginas y grupos donde se compartió el cuestionario fueran personas con más años de estudios.

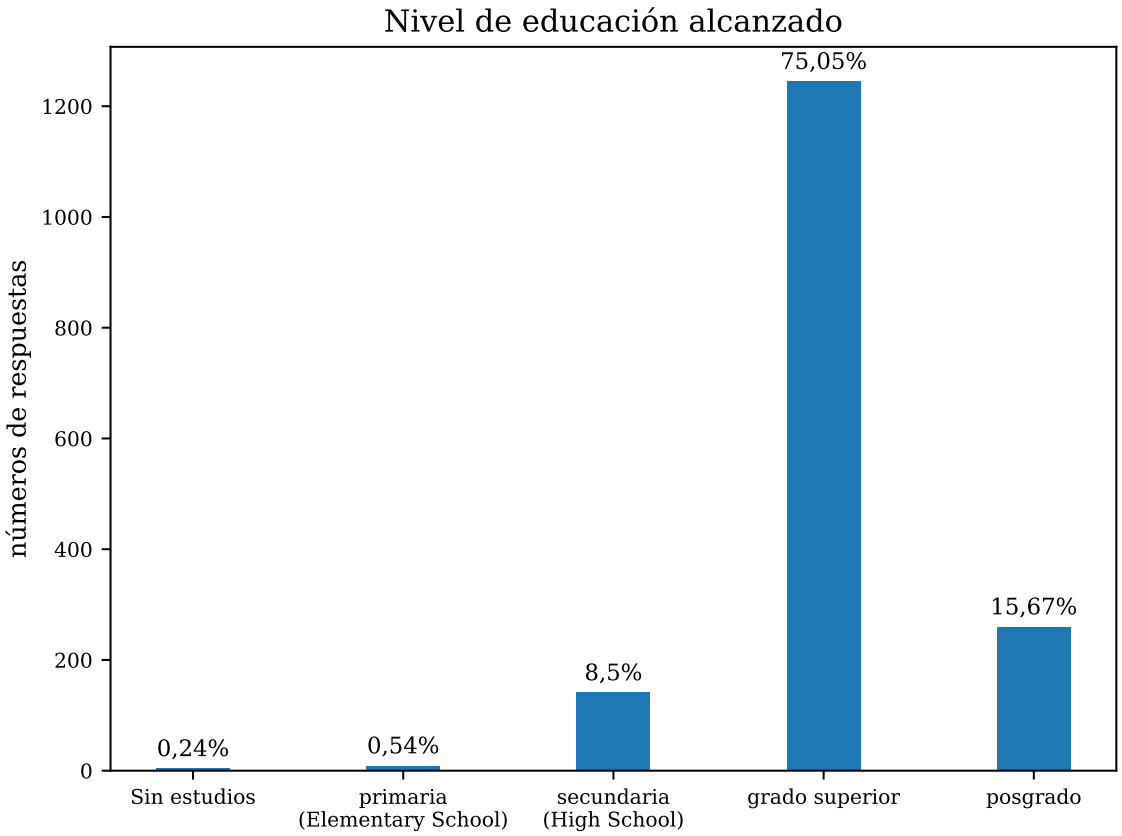
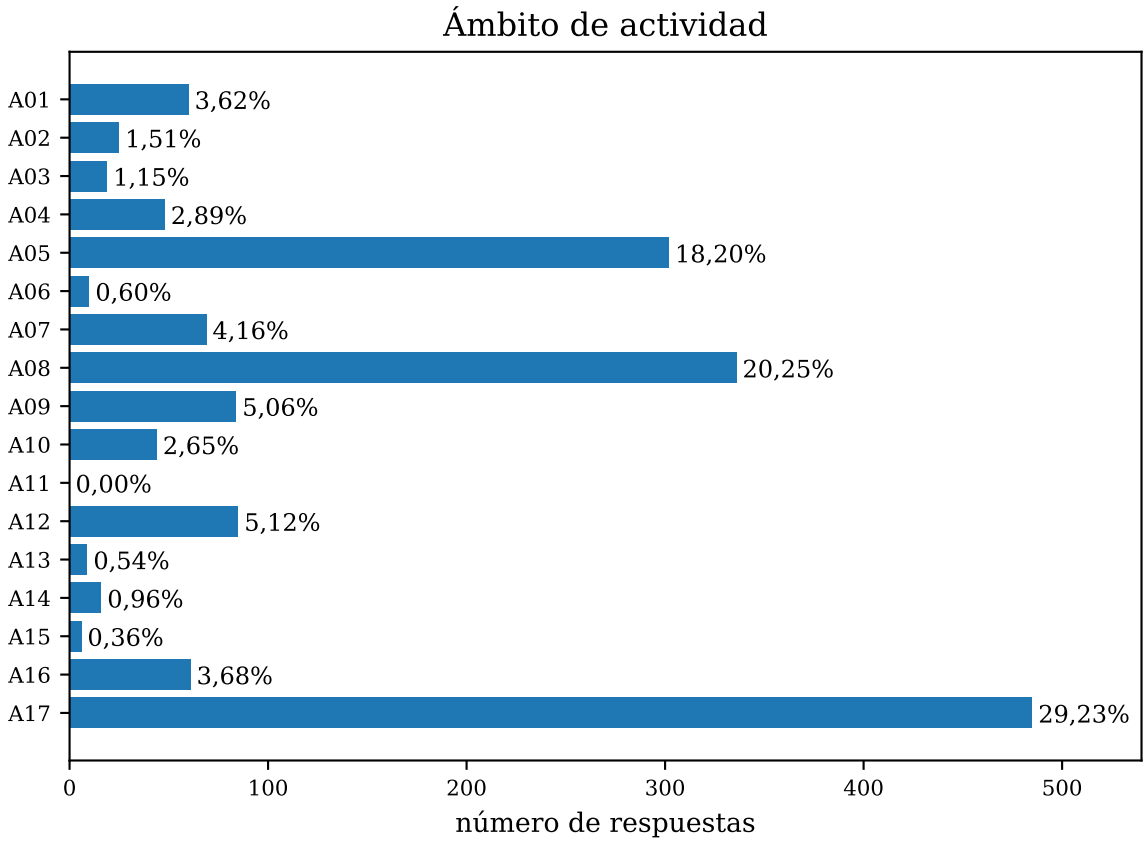


Figura 4: Distribución de participantes por nivel de educación alcanzado

Para evitar confusiones, en la pregunta relativa al nivel de educación alcanzado mantuvimos la nomenclatura del sistema antiguo de educación. El sistema K-12 empezó a implementarse en el año 2011 de manera simultánea en primaria y secundaria, comportando, además del cambio de nombre de «High school» a «Junior high school», la adición «Senior high school», compuesta por 2 años.

Como se observa en el gráfico 5, el 20,25 % de los participantes desempeña actividades en el sector de «Salud y asistencia social» y el 18,20 % en el sector de «Educación» (incluidos estudiantes). El 29,23 % marcó la opción «Otros servicios». La lista de ámbitos de actividad es la utilizada en los CPH.



Leyenda

A01	Actividades del servicio de alojamiento y alimentación
A02	Agricultura, silvicultura y pesca
A03	Arte, espectáculos y ocio
A04	Construcción
A05	Educación (incluidos estudiantes)
A06	Subministro de electricidad, gas, vapor y aire acondicionado
A07	Finanzas y seguros
A08	Salud y asistencia social
A09	Información y comunicación
A10	Industria manufacturera
A11	Industria extractiva
A12	Servicios profesionales, científicos y técnicos
A13	Sector inmobiliario
A14	Transporte y almacenamiento
A15	Actividades de suministro de agua, saneamiento, gestión de resíduos y descontaminación
A16	Comercio mayorista y minorista
A17	Otros servicios

Figura 5: Distribución de participantes por ámbito de actividad. (Clasificación según el National Statistics Office (NSO), 2014a.)

2.2.2 *Acerca del uso del chabacano*

Según el gráfico de la figura 6, la mayoría de los participantes (83,54 % y 75,35 % respectivamente) afirma que el inglés y el filipino eran lenguas vehiculares en el aula en su tiempo de primaria.

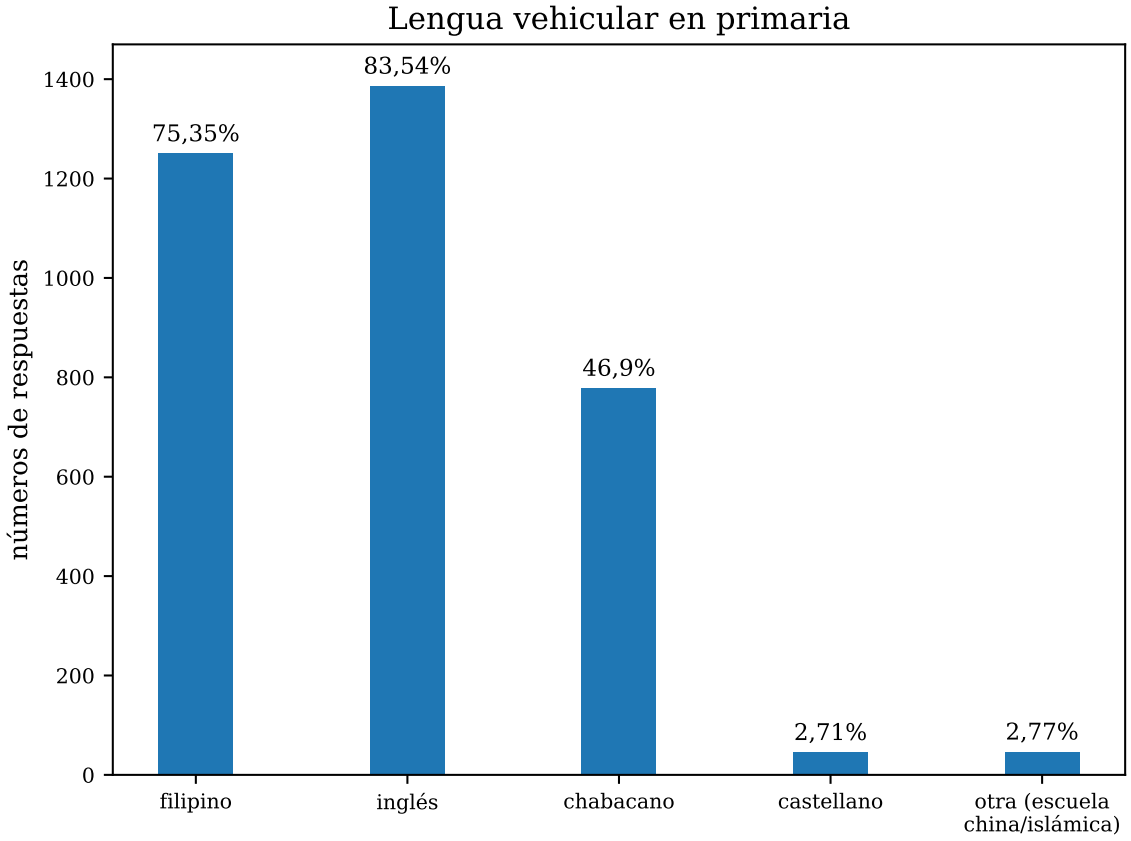


Figura 6: Lengua vehicular en primaria por número de respuestas

Lo que llama la atención a primera vista es que un 46,9 % afirma que el CZ también era lengua vehicular. Pese a que el CZ solo ha vuelto a conquistar el papel de lengua vehicular de la educación básica en tiempos recientes, según relatos de algunos zamboangueños (ej. 65, 66, 67, 68 y 69), éste se utilizaba (y probablemente sigue siendo utilizado más allá del tercer año de primaria) en las aulas de manera oficiosa por algunos profesores, sobre todo en las escuelas públicas de la ciudad, lo que explicaría el hecho de que tantos participantes hayan marcado esta lengua como una de las lenguas vehiculares.

(65) Yup, almost all of my teachers speak Chavacano.
Sí, casi todos mis profesores habla[ba]n chabacano [en el aula].

(66) Si na high school, tiene gad ansina.
En secundaria, sí que había de esos [que hablaban chabacano en el aula].

(67) Ahm, hm, dol de mi teachers antes kay ta'n Chavacano man si ta enseña, especially gad 'quel Elementary.
Ahm, hm, me da que mis profesores antes hablaban chabacano a la hora de enseñar, especialmente los de primaria.

(68) Yes. Especially if it's a public school. I studied in a private evangelical school and we were prohibited to speak Chabacano. So the teachers only spoke Tagalog and English. But in private catholic schools like [NOMBRE CENSURADO] and [NOMBRE CENSURADO], I know the teachers spoke Chabacano. When I was in a public science high school, the teachers spoke in Chabacano, except the Filipino teacher. But the English teacher too spoke in Chabacano.
Sí. Especialmente si es una escuela pública. Estudié en una escuela privada evangélica y nos prohibían hablar chabacano. Así que los profesores solo hablaban tagalo e inglés. Pero en escuelas privadas católicas como [NOMBRE CENSURADO] y [NOMBRE CENSURADO], sé que los profesores hablaban chabacano. Cuando estudiaba en una escuela secundaria pública de ciencias, los profesores hablaban en chabacano, salvo el profesor de filipino. Pero incluso el profesor de inglés hablaba en chabacano.

(69) Yes. I am from [NOMBRE CENSURADO]. Some teachers used Chavacano 'cause a lot of teachers are using it at home. But there was a point that time wherein the school needs to be placed on a strict English only policy for the accreditation purposes.
Sí. Soy de [NOMBRE CENSURADO]. Algunos profesores usaban el chabacano pues muchos profesores lo usan en casa. Pero hubo un momento en aquel tiempo en el que la escuela necesita[ba] establecer una política estricta de uso exclusivo del inglés a efectos de acreditación.

Los participantes que tuvieron la oportunidad de estudiar la asignatura de CZ en la escuela son los que al rellenar el cuestionario tenían 14 años o menos. Sin embargo, debido a la baja participación de este grupo de edad (tan solo 8 personas del total de 119 personas del grupo de edad 10-19), todavía no se puede hacer ningún tipo de afirmación sobre el impacto de la introducción de la lengua en los tres primeros años de primaria.

Como podemos observar en el gráfico de la figura 7, solo el 27,67 % de los participantes afirma tener algún tipo de estudios previos de castellano.

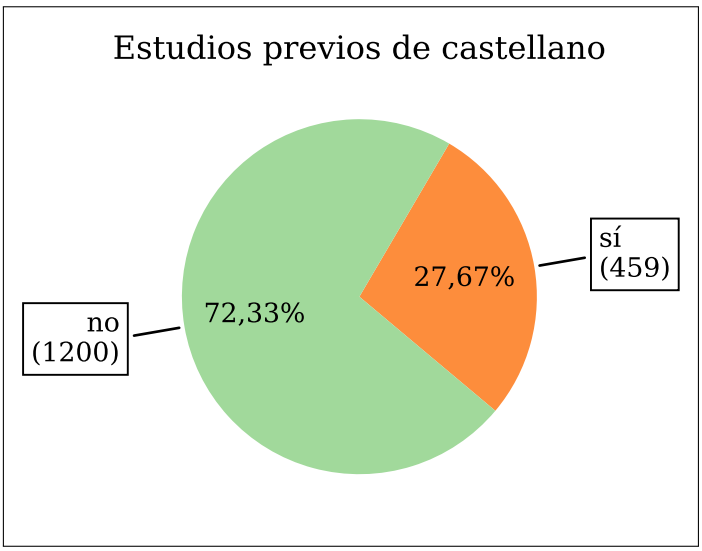


Figura 7: Distribución de participantes que poseen estudios previos de castellano

El gráfico de la figura 8 muestra el uso de las diferentes lenguas de los participantes en diferentes contextos comunicativos.

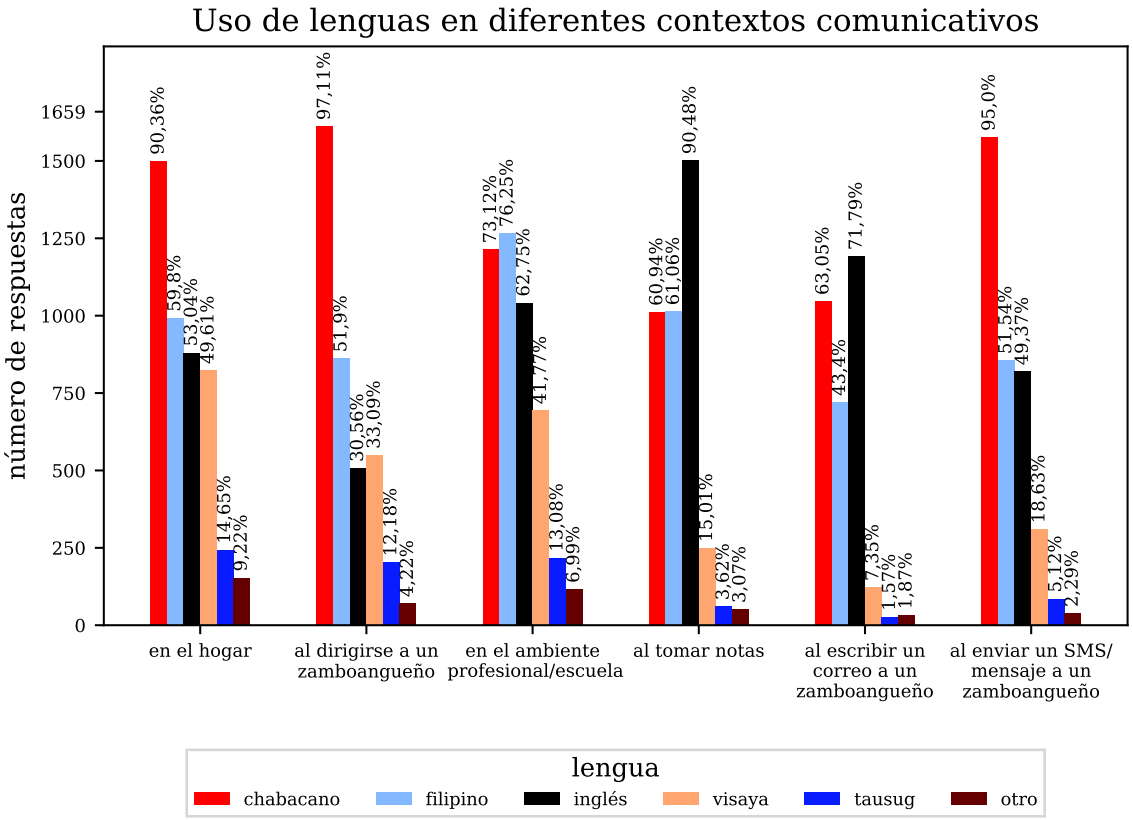


Figura 8: Uso de diferentes lenguas en diferentes contextos comunicativos

El 90,36 % de los participantes dice utilizar el CZ en casa, seguido por el tagalo, el inglés, el cebuano, el tausug y otras lenguas. El 97,11 % afirma utilizar el CZ para hablar con otros zamboangueños. En el ambiente laboral y académico, la

dominancia es del tagalo, aunque seguido de cerca por el CZ y el inglés. Para tomar notas, se prefiere con diferencia el inglés. Al escribir correos a un zamboangueño, la lengua preferida es el inglés, seguida del CZ. En los mensajes de texto y chats con un zamboangueño, hay una clara preferencia por el CZ. Se nota aquí que, en situaciones formales, se usa más el tagalo en la lengua oral y el inglés en la lengua escrita, aunque el CZ también puede hacer acto de presencia según la situación (escuela o ambiente de trabajo). En los contextos informales, se utiliza preferentemente el CZ.

A los que no marcaron el CZ como opción en ningún contexto escrito (112 participantes), se les preguntó la razón por la que prefieren escribir en otras lenguas. Como podemos observar en el gráfico de la figura 9, la mayoría de ellos afirma no hacerlo por no ser su lengua materna, seguido de la falta de reglas claras de ortografía y por no haberla estudiado en la escuela. Entre las personas que han elegido la opción «otro», la justificación fue la preferencia por otras lenguas, haber crecido fuera de Zamboanga y/o por vivir en el extranjero.

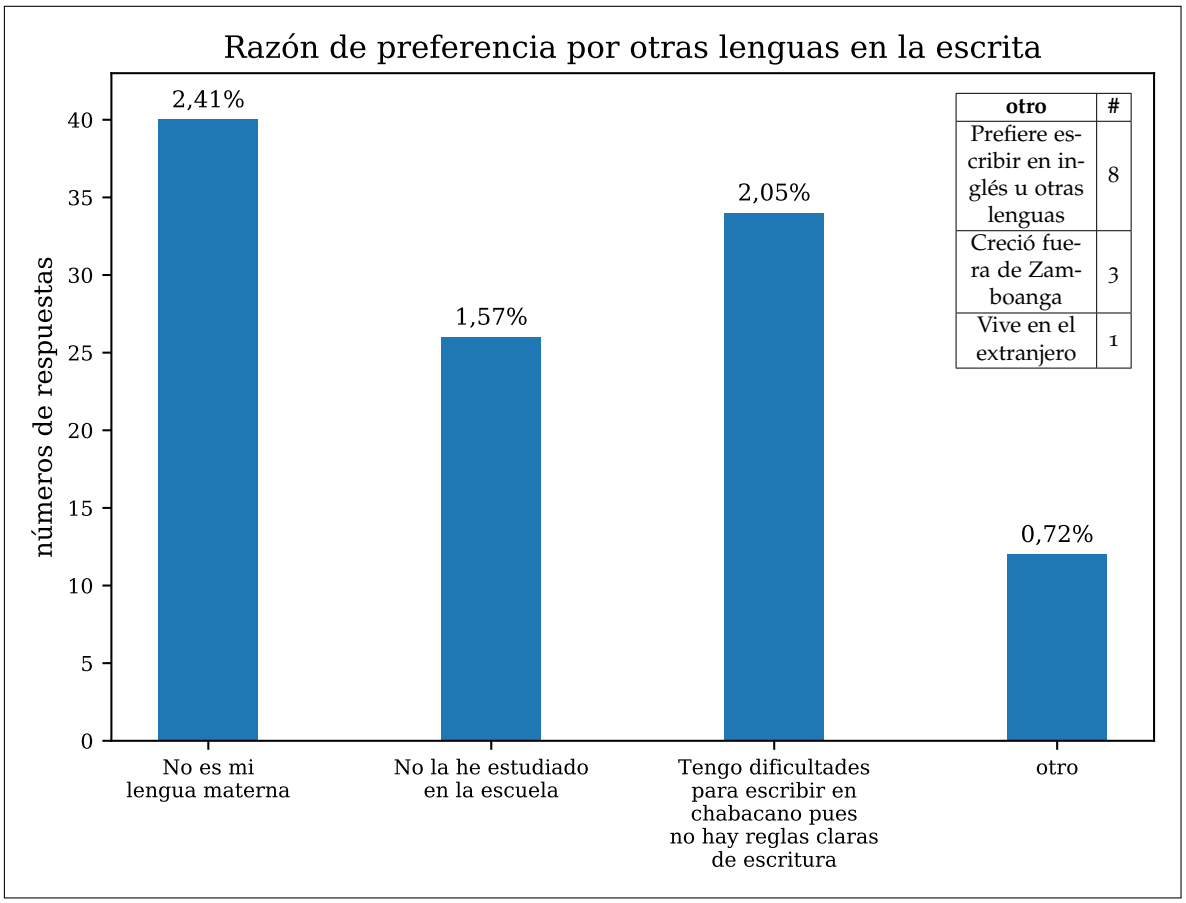


Figura 9: Principales razones por las que algunos hablantes no escriben en CZ

2.2.3 *Acerca de la ortografía*

Según el gráfico de la figura 10, un 38,4 % ha afirmado ser consciente de la existencia de la ortografía frente a un 61,6 % que afirma desconocerla. La tasa relativamente elevada de consciencia de la ortografía podría explicarse por el hecho de que el principio propuesto por la ortografía (grafiar las palabras según la lengua de origen) ya formara parte del repertorio de grafías utilizadas por los hablantes y no por una consciencia real de la existencia de la ortografía.

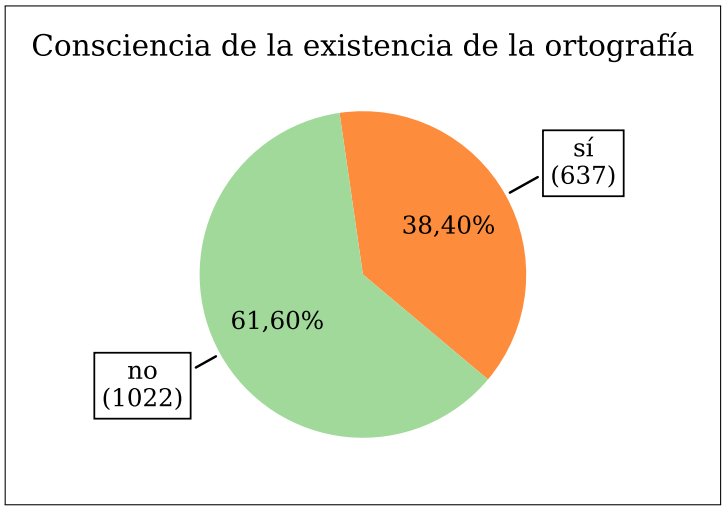


Figura 10: Distribución de participantes por su conocimiento de la ortografía

Como vemos en el gráfico de la figura 11, al ser presentado un mismo texto en la ortografía y en una grafía no normativa, solo el 27,12 % de los participantes afirmó que la ortografía es más legible que la grafía no normativa. Aunque coincide con el porcentaje de participantes que ha afirmado haber estudiado el castellano previamente (figura 7), los dos grupos solo se solapan en aproximadamente un 37 %.

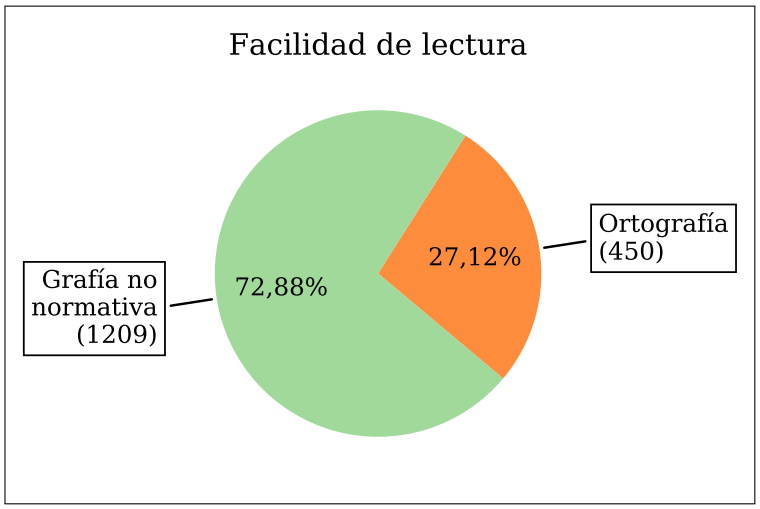


Figura 11: Facilidad de lectura según la grafía

Cabe destacar también que parte de los hablantes, al contestar este tipo de preguntas, se ven confrontados con principios ideológicos de defensa de la lengua: en su imaginario, el tagalo representa una amenaza a la lengua; y el castellano, la conexión con las raíces. Esto podría implicar también que la respuesta a esta pregunta no fuera del todo sincera en algunos casos.

Tomando en consideración la edad de los hablantes y la facilidad de lectura de una grafía sobre la otra, como observamos en el gráfico de la figura 12, la preferencia de la grafía no normativa sobre la ortografía permanece relativamente estable en todos los grupos de edad hasta los 49 años. A partir de los 50 años, la diferencia de preferencia sufre graduales decrementos hasta que haya una inversión.

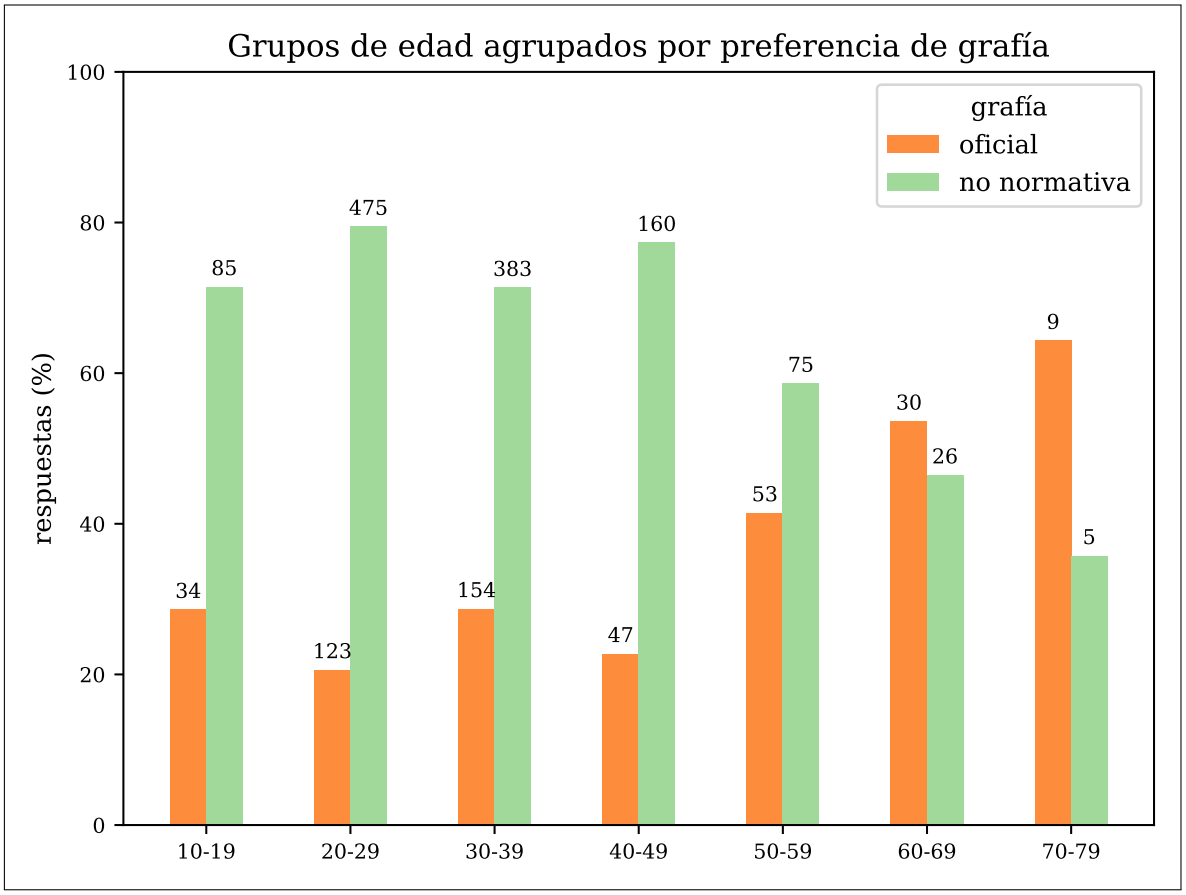


Figura 12: Facilidad de lectura según la grafía y la edad

Efectivamente, como muestra el gráfico de la figura 13, la inversión de preferencias ocurre a partir de los 52 años: poco más de la mitad de los participantes de ese grupo dice preferir la ortografía.

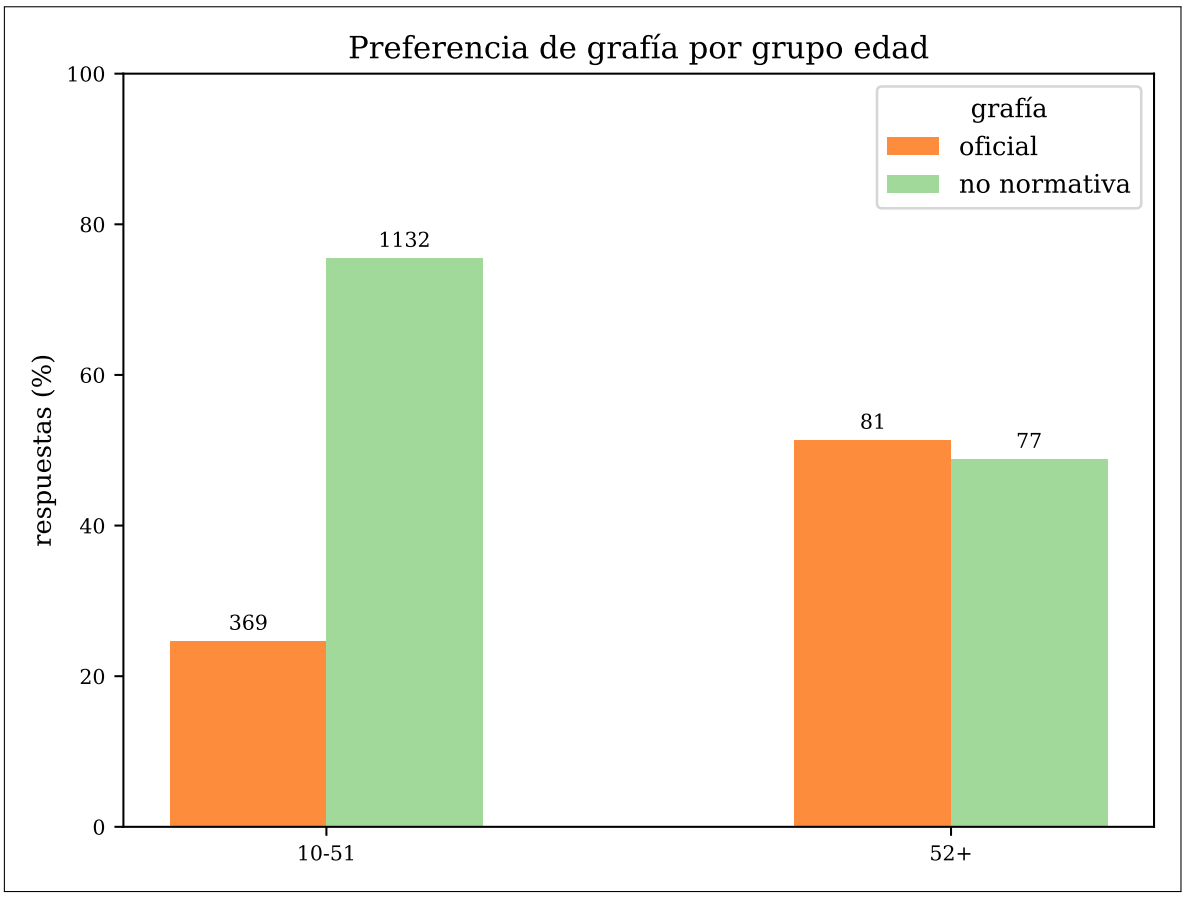


Figura 13: Punto de inflexión de la preferencia de grafía por edad

Al ser preguntados sobre su confianza al aplicar la ortografía, como podemos verificar en el gráfico de la figura 14, un número sorprendentemente alto de hablantes (58,58 %) se dice confiado, frente a una minoría (12,24 %) que afirma tener poca o ninguna confianza. El 34,12 % se muestra indeciso. Preguntados sobre el interés por aprender más acerca de la ortografía, la mayoría de los participantes (84,21 %) se han mostrado dispuestos. Solo un 3,56 % afirma tener poco u ningún interés por hacerlo, mientras un 12,24 % tiene moderado interés.

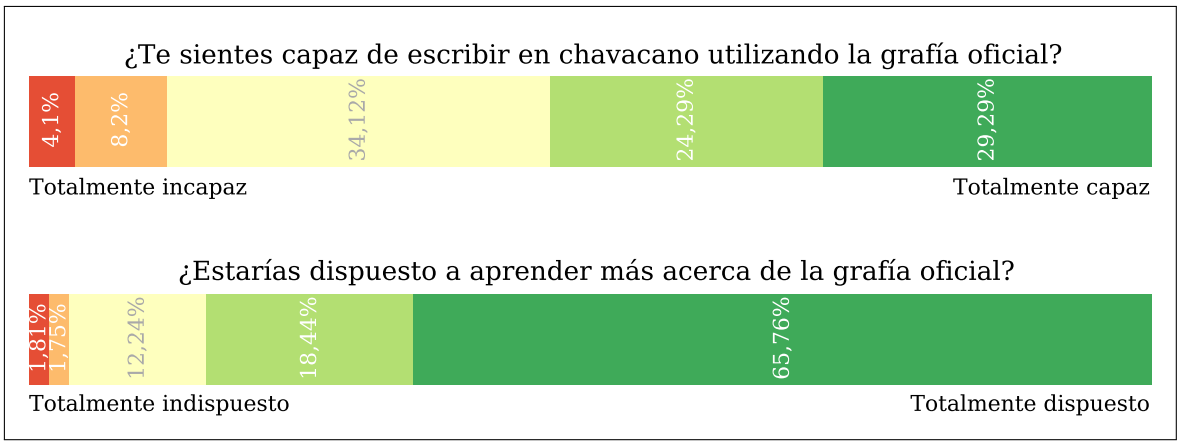


Figura 14: Grado de confianza al aplicar la ortografía e interés por aprender más al respecto

Según el gráfico de la figura 15, la gran mayoría de los participantes (97,53 %) afirma que es importante disponer de una ortografía para el CZ. Esto muestra que los esfuerzos del gobierno local por definir una ortografía para la lengua tienen el respaldo de los hablantes.

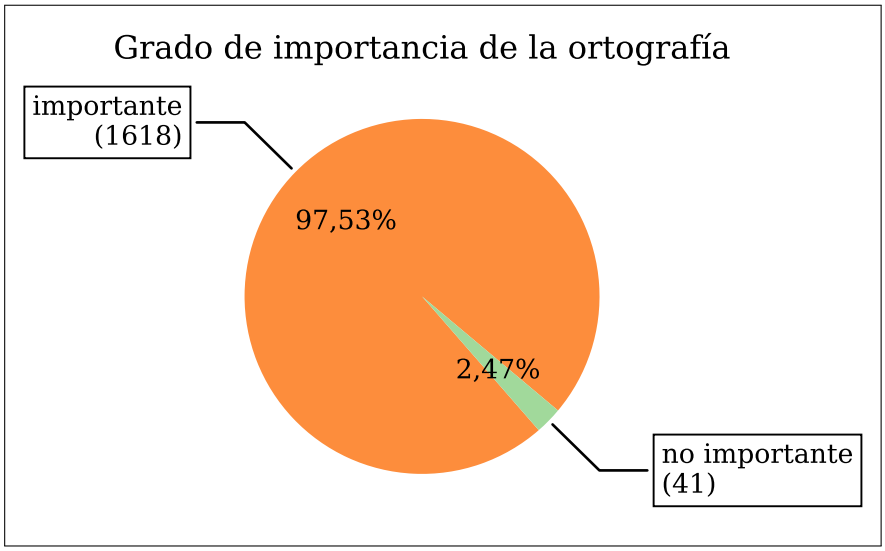


Figura 15: Distribución de participantes que cree que es importante disponer de una ortografía para el CZ

El gráfico de la figura 16 muestra que la gran mayoría de los participantes (96,20 %) afirma creer que sería útil disponer de un corrector ortográfico para el CZ.

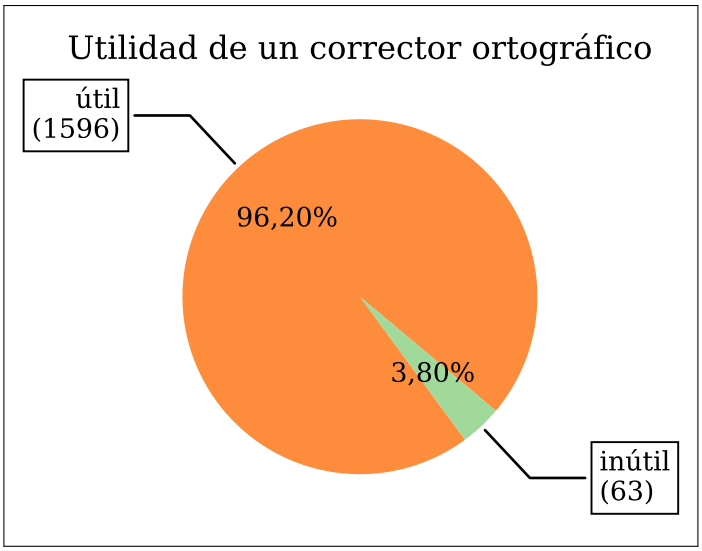


Figura 16: Distribución de participantes que creen que sería útil disponer de un corrector ortográfico para el CZ

El gráfico de la figura 17 muestra el desempeño de los participantes en una tarea de identificación de errores ortográficos: dado un texto en CZ con tres palabras destacadas y con errores ortográficos, se perdía que marcaran cuáles creían que estaban mal escritas. Solo el 5,67 % fue capaz de marcar todas ellas.

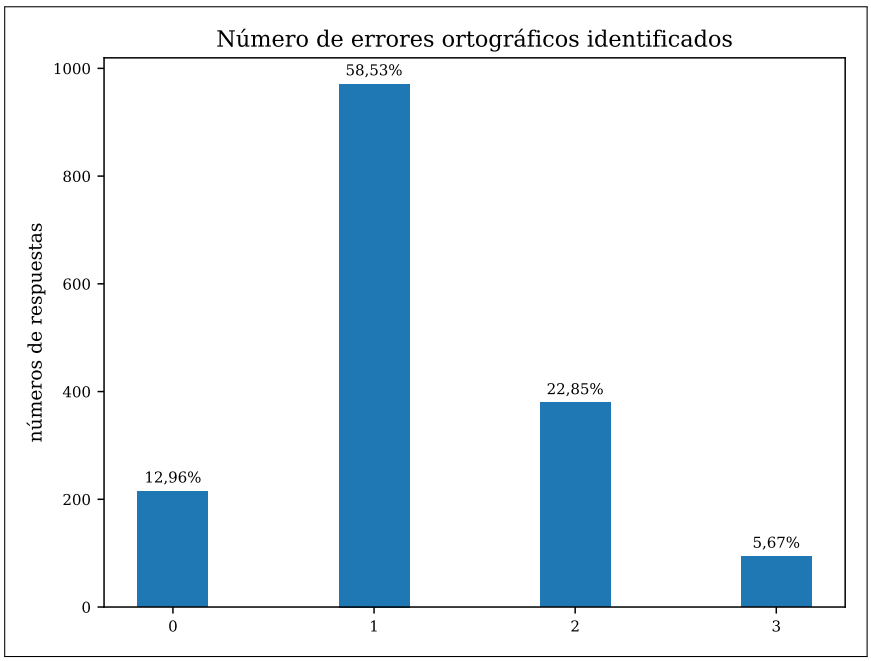


Figura 17: Desempeño de los participantes en la tarea de identificación de errores ortográficos

Asimismo, la figura 18 muestra el desempeño de los participantes en la tarea según su grado de confianza en la aplicación de la ortografía. De los 889 que decían tener confianza al aplicar la ortografía, solo un 7,5 % (67) fueron capaces de identificar las tres palabras mal escritas. Si consideramos solo dos de las respuestas, ese número sube al 22,16 %, lo que todavía se puede considerar un valor bastante bajo para un grupo de personas que se dicen confiadas.

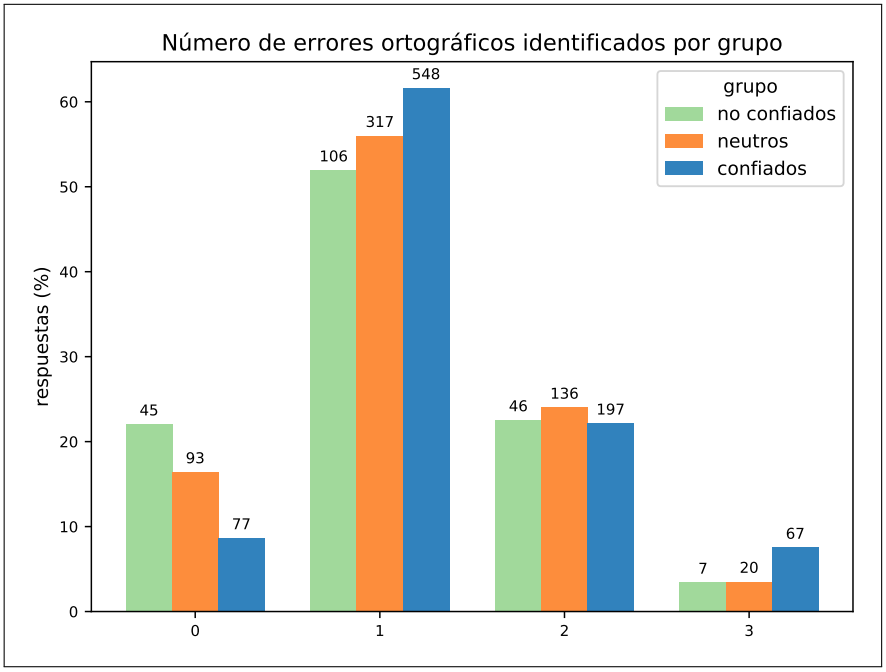


Figura 18: Desempeño de los participantes en la tarea de identificación de errores ortográficos por grado de confianza

Como muestra el gráfico de la figura 19, el haber estudiado castellano tampoco parece haber ayudado mucho en la identificación de los errores. Para ese grupo, obtuvimos solo un 9,15 % de respuestas totalmente correctas frente a un 4,33 % de las personas que no han estudiado el castellano. Esto implica que tener alguna familiaridad con el castellano no garantiza de por sí que esos hablantes tengan un mejor desempeño a la hora de aplicar la nueva ortografía. El estudio del castellano, tal como proponen algunos hablantes como solución para subsanar las deficiencias ortográficas de los hablantes de CZ, claramente no es la solución más conveniente, a menos que esas personas alcancen un nivel considerablemente alto de competencia en la lengua.

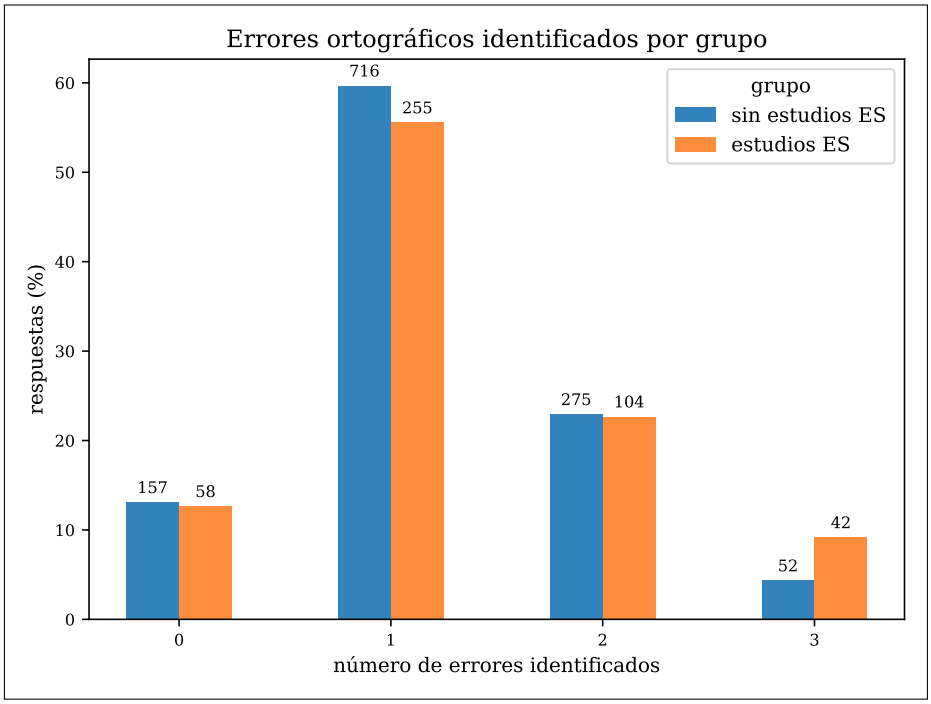


Figura 19: Desempeño de los participantes con estudios previos de castellano en la tarea de identificación de errores ortográficos

2.2.4 *Lengua de respuesta*

A la hora de distribuir el cuestionario, la versión compartida fue elegida de manera aleatoria. Sin embargo, el participante podía cambiar la lengua del cuestionario en cualquier momento, tanto de inglés a CZ como de CZ a inglés, si así lo deseaba. Entre los que decidieron cambiar de lengua, como vemos en el gráfico 20, un 10,41 % cambió de inglés a CZ, frente al 7,34 % que cambiaron del CZ a inglés. La mayoría (90,6 %) contestó al cuestionario en la lengua que lo recibió, fuera ésta el inglés o el CZ.

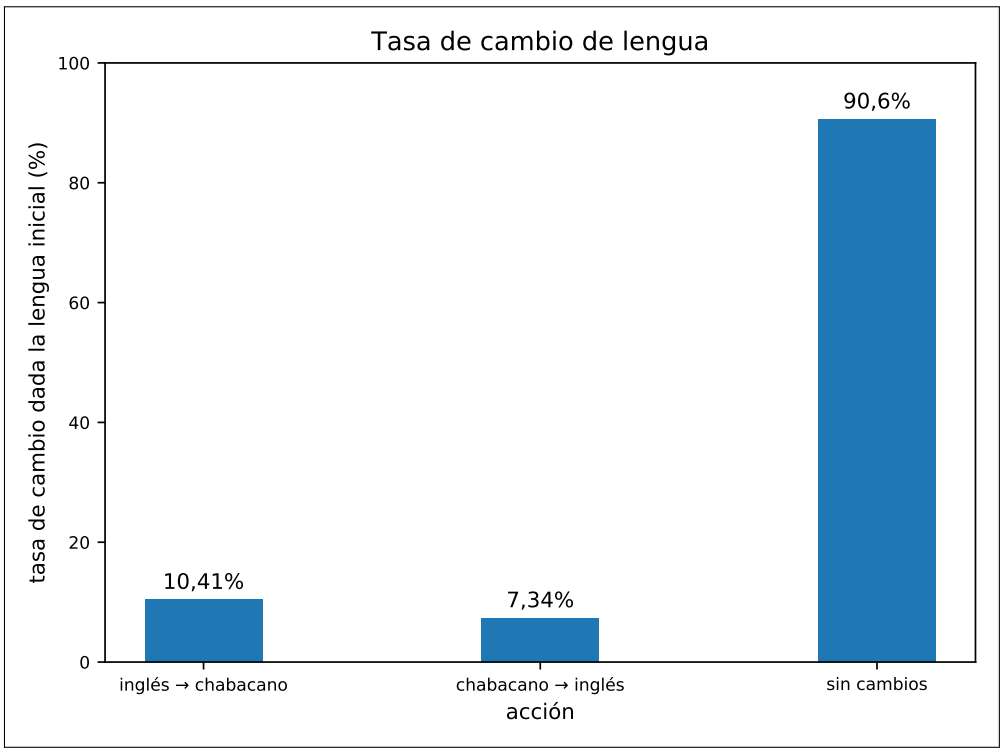


Figura 20: Tasa de cambio de lengua del cuestionario

Para terminar, en el gráfico de la figura 21 se muestra que la versión en inglés fue doblemente utilizada en comparación a la versión en chabacano (998 respuestas frente a 505). Como no tenemos conocimiento sobre cuál de las versiones fue compartida por los participantes y/o páginas y grupos de Facebook, es difícil establecer a ciencia cierta la causa. No obstante, es posible que se haya compartido más la versión inglesa del cuestionario. Algunas personas sugirieron un posible rechazo por parte de algunos participantes al recibir el cuestionario escrito en CZ.

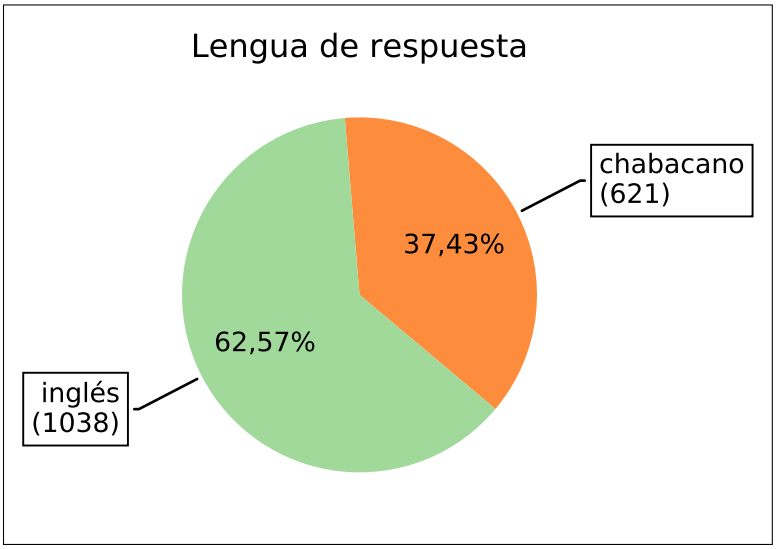


Figura 21: Distribución de participantes por la lengua de respuesta del cuestionario

2.2.5 *Conclusión*

Revisando las hipótesis iniciales (mencionadas en el apartado 2.1), concluimos que son verdaderas:

- ✔ La ortografía es demasiado complicada para el hablante medio si no se poseen conocimientos previos suficientes de castellano y de las lenguas del sustrato.
- ✔ Tener alguna familiaridad con el castellano no basta para aplicar la ortografía de manera satisfactoria.
- ✔ El hablante medio de CZ prefiere leer un texto en grafía no normativizada.
- ✔ Un sector de la población demanda la creación de un corrector ortográfico de CZ.

Además, las principales conclusiones adicionales que se sacan del estudio preliminar realizado es que:

- El chabacano se escribe sobre todo en contextos informales, en la comunicación interpersonal;
- La mayor parte de las personas desconoce la existencia de una ortografía para el CZ;
- La mayor parte de las personas se cree preparada para aplicar la ortografía del CZ, pero su desempeño no es satisfactorio sin previo estudio;
- La nueva grafía tiene el respaldo de los hablantes, que creen que es importante que el CZ tenga una ortografía y se muestran dispuestos a aprenderla.

Como limitación de este estudio, reconocemos que los resultados probablemente solo son extrapolables para personas con elevado grado de instrucción.

2.3 OBJETIVOS

En el mundo del Procesamiento del Lenguaje Natural (PLN), los esfuerzos en la construcción de recursos lingüísticos suelen concentrarse en un número extremamente reducido de lenguas. Las demás lenguas, muchas de ellas incluso oficiales en sus respectivos territorios, son conocidas como «lenguas con pocos recursos» (*under-resourced languages*) por vivir esa realidad de escasez de recursos de PLN, sea por su reducida población, sea por la falta de inversión en la investigación y en el desarrollo de los mismos. Dentro de ese grupo se encuentran no solo las lenguas minoritarias, sino también las minorizadas, como es el caso del CZ. El

desarrollo de aplicaciones de PLN puede ser de gran utilidad para la comunidad de hablantes de estos dos últimos grupos de lenguas, pues ayudaría a impulsar su uso y su inclusión en el siglo XXI.

Este trabajo pretende colaborar con el proceso de codificación del CZ anteriormente mencionado, de forma que se mitiguen de alguna manera los posibles efectos negativos que la implantación de la ortografía, de difícil aplicación para el hablante medio, pueda tener en su comunidad de hablantes. Se desea construir un prototipo que, además de corregir errores tipográficos sencillos, sea también capaz de detectar ciertos patrones erróneos de las diferentes grafías preexistentes y convertirlos a la ortografía del CZ, permitiendo así que cualquier hablante sea capaz de escribir su lengua sin cometer demasiadas faltas ortográficas, ni tener que preocuparse demasiado por la ortografía. Las correcciones sugeridas podrían también ayudar a aprender la ortografía al hablante que la desconozca, por medio de una exposición constante y su asimilación de manera paulatina.

Conscientes de la gran variabilidad de grafías existentes en CZ, proponemos una aproximación que sea capaz de capturar esa variabilidad y que además pueda ser combinada con las tecnologías más utilizadas en la actualidad.

2.4 PLAN DE TRABAJO

La investigación se llevó a cabo entre los meses de enero del 2018 y agosto del 2019, y como se ve reflejado en el cronograma de la figura 22, la duración de cada fase no ha sido homogénea. La tarea más larga fue la de la construcción del corpus, debido en especial a las dificultades para obtener permiso de uso de los textos protegidos por derechos de autor y, por lo tanto, a veces se realizó en paralelo con otras tareas. La revisión bibliográfica también se llevó a cabo junto con otras actividades, ya que ciertas decisiones del proyecto fueron siendo ajustadas a medida que avanzábamos en las demás tareas.

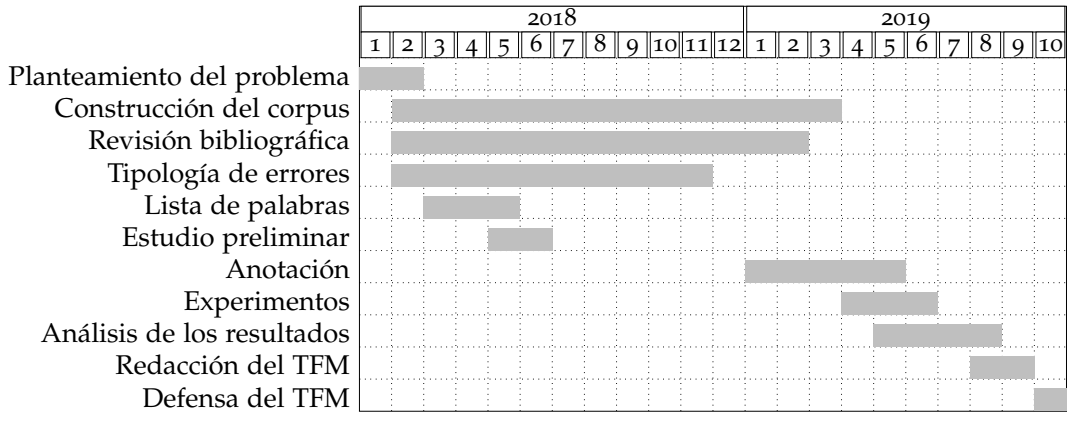


Figura 22: Cronograma del Trabajo de Fin de Máster.

CAPÍTULO 3

MARCO TEÓRICO

En este capítulo, describimos la base teórica fundamental para la comprensión de este trabajo. Este apartado va dividido en dos bloques, con sus respectivos apartados. El primer bloque (apartado 3.1) trata de los conceptos fundamentales y los antecedentes relevantes para este trabajo: el apartado 3.1.1 presenta algunas clasificaciones de los tipos de errores ortográficos desde el punto de vista lingüístico e informático; el apartado 3.1.2 trata de la clasificación de los textismos en diferentes lenguas; el apartado 3.1.3 presenta los lenguajes para representación de datos lingüísticos utilizados en este trabajo; y finalmente, el apartado 3.1.4 aborda una introducción al PLN y algunos conceptos relevantes. El segundo bloque (apartado 3.2, a su vez, describe las herramientas utilizadas en este trabajo.

3.1 ANTECEDENTES

3.1.1 *Errores ortográficos*

En lo que concierne a la clasificación de los tipos de errores ortográficos, es posible utilizar un enfoque lingüístico o uno informático. En este apartado, describimos algunas de sus tipologías asociadas, de interés para este trabajo.

3.1.1.1 *Enfoque lingüístico*

Muchas de las tipologías de errores ortográficos desarrolladas desde un punto de vista lingüístico provienen de estudios que analizan las faltas de ortografía cometidas por hablantes nativos de alguna lengua en la enseñanza básica. En tiempos recientes, sin embargo, parece haber un creciente interés por los errores ortográficos que comete el alumnado de segundas lenguas o los sujetos bilingües. Todos estos casos son de interés para el CZ, ya sea por la cualidad plurilingüe de la

sociedad zamboangueña, sea por la decisión de grafiar las palabras del CZ según la lengua de origen, aplicando reglas externas al sistema lingüístico.

3.1.1.1.1 Castellano

Como el castellano es la lengua lexificadora del CZ, es decir, la lengua de la que se origina la mayor parte del léxico, consideramos interesante analizar algunas tipologías de errores ortográficos cometidos en esa lengua.

Una clasificación de errores ortográficos encontrada en numerosas fuentes al respecto de la lengua castellana (Abregú Tueros, 2011; Rojas, 2009) es la que deriva del trabajo de Galí (citado en Cristóbal Rojo, 1982) para la enseñanza de la lengua catalana en la enseñanza básica:

- **Errores de ortografía natural:** Errores en palabras escritas tal y como se pronuncian.
- **Errores de ortografía arbitraria:** Errores en palabras cuya grafía no obedece ninguna regla específica y/o previsible.
- **Errores de ortografía reglada:** Errores en palabras cuya ortografía se rige por reglas. Originalmente, no figura en los trabajos de Galí, pero sí en sus derivaciones.

Otra clasificación más compleja e igualmente interesante es la propuesta por Balmaseda Neyra (2001), en la que se distinguen los siguientes tipos de errores:

- **Sustituciones:** Cambio de una letra por otra. Ej.: *pota, *vatalla, *docma, *cloriosa.
- **Confusión homonímica:** Sustitución de una palabra por otra de pronunciación semejante. Ej.: tu / tú, carabela / calabera, cayo / callo, caza / casa.
- **Omisiones:** Se produce por afonía, pronunciación defectuosa o percepción auditiva anormal. Ej.: *efica, *conocese (conocerse).
- **Condensaciones y segregaciones:** Problemas de unión o segmentación de palabras o sílabas. Ej.: *valla meses (bayameses)
- **Inserciones:** Inserción de letras, sílabas o tildes. Ej.: *protrector, *rostundo, *éxcito.
- **Duplicaciones:** Letra duplicada erróneamente. Ej.: *honrrar, *preveer.
- **Improvisaciones:** Escritura ad-hoc. Ej.: *haller (ayer).
- **Lapsus:** Errores por descuido.

3.1.1.1.2 Tagalo

El filipino, forma estandardizada de tagalo, tiene una grafía fonológica. En Octaviano, Go, Borra y Oco (2016), son listados los siguientes tipos de errores:

- **confusión en el uso de algunas consonantes (c/k)** - Ej.: *local (lokal, «local»).
- **confusión de grupos vocálicos (ia/ya, au/aw)** - Ej.: *industria (industriya, «indústria»), *audisyon (awdisyon, «audición»).
- **duplicación de caracteres** - Ej.: *opposisyon (oposisyon, «oposición»)
- **errores de segmentación** - Ej.: *parin (pa rin, «todavía» enfático)
- **errores de uso del guion** - Ej.: *ipinag-kaloob (ipinagkaloob, «prestado»), *pinaka malaki (pinakamalaki, «el más grande»).
- **sustitución de caracteres.** - Ej.: *kadalason (kadalasan, «a menudo»).

Todos los tipos de errores listados se producen también en CZ. Se atribuyen a la influencia del inglés la causa de los tres primeros.

3.1.1.1.3 Castellano como segunda lengua (filipinos)

Entre todas las tipologías analizadas, sin duda es la de Sánchez-Jiménez (2010), que clasifica los errores ortográficos cometidos por estudiantes filipinos de castellano, la que más despertó nuestro interés, ya que gran parte de los errores descritos coinciden con los que se producen en CZ. Esto se debe, ante todo, a los rasgos fonéticos compartidos entre el CZ y las lenguas del sustrato, y a la interferencia del inglés y del tagalo.

- Ortografía de la letra o grafemática:
  - Errores en el origen:
    - **Errores en el origen causados por la gramática** - Ej.: *tenado, *encontro, *hableme, lleguemos (llegamos);
    - **Errores en el origen causados por la interferencia** - Ej.: *bangko, *barocco, *adventurero, *ameliorar;
  - Errores contra el sistema:
    - **Errores contra la fonética natural vocálica** - Ej.: *miercules, *ordinador, *veintecuatro, *violonista;
    - **Errores contra la fonética natural consonántica** - Ej.: *Balderas, *Porge, *rab, *contemporareas;

        - **Errores en el uso de grafías complementarias** - Ej.: *matematika, *oscurros, *gymnasio;
        - **Errores por usar grafías castellanas con valores fonéticos ajenos** - Ej.: *hapones, *technicas, *mejo (medio);
        - **Errores por usar grafías impropias del castellano** - Ej.: *Espanya, *guapissima, *otso, *thesis;
    - Errores por arbitrariedad:
        - **Arbitrariedad por concurrencia de grafemas** - Ej.: *Unibersidad, *estranjera, *serveza;
        - **Arbitrariedad en el uso de la h** - Ej.: *Ola, *oras, *abido, *hojos;
    - **Errores por desatención** - Ej.: *briografias, *necho (noche),*distritrito, *apliar;
- Ortografía de la palabra o lexicológica:
    - **Acentuación** - Ej.: *¿Que?, *aqui, *tambien, *peliculas;
    - **Siglas y abreviaturas** - Ej.: *EU, *Phd, *RnB, *MRT;
    - **Mayúsculas y minúsculas** - Ej.: *Filipino, *Martes, *Mayo, *cervantes,*españa, *mediterraneo;
    - **Unión y separación de palabras** - Ej.: *veinte y cinco, *alas,*por que (causal: porque), *latino americanos;
    - **Signos diacríticos** - Ej.: *Linguistica, *espanol, *compania;
- Ortografía de la frase o sintagmática:
    - **Signos de puntuación** - Ej.: «*Yo, voy a vivir en hispano america.»; «*En el primer lugar»; «*he estudiado español.»; «*Pero no me gusta ir a la discoteca con ellos.»; «*Porque mis padres no se gusta.»;
    - **Signos de entonación** - Ej.: «*cómo estás?»; «*Hemos quedado de viaje en Caribe!»;
    - **Signos auxiliares** - Ej.: «Y tu mama tambien», «El Bola», Noviembre, «El Juego de la Verdad».

3.1.1.2 *Enfoque informático*

Es posible ver los errores ortográficos desde una óptica exclusivamente informática, sin tener en consideración la razón o cómo estos ocurren. Desde este punto de vista, un error ortográfico sería una deformación de la palabra original que, para ser corregido, requiere de la aplicación de una o más operaciones básicas:

- **inserción:** inserción errónea de un carácter;
- **eliminación:** eliminación errónea de un carácter;
- **transposición:** cambio de posición entre dos caracteres.

Algunas fuentes incluyen también la **sustitución**, pero esta puede ser obtenida con una **eliminación** y una **inserción**.

3.1.2 *Textismos*

Textismo es un término paraguas que abarca una gran cantidad de fenómenos, característicos de los mensajes intercambiados por móviles o por internet. Esta denominación es originaria de la palabra «text», término en inglés americano para referirse a los SMS, aunque los textismos sean anteriores a los móviles, y algunos de ellos, incluso a la era de internet.

Una clasificación bastante citada en la literatura es la del ensayo escrito por Craig (2003), a la cual agregamos algunos ejemplos ilustrativos en castellano:

- **sustituciones fonéticas:** Sustitución de uno o más fonemas por una serie de letras que producen el mismo sonido. Ej.: *salu2* (saludos), *ksi* (casi), *wapa* (guapa), *x* (por).
- **acrónimos:** Unión de la primera letra de las palabras más importantes de lo que se desea acortar. Ej.: *tqm* (te quiero mucho), *npi* (ni pu** idea).
- **abreviaturas:** Omisión de letras para acortar palabras. Ej.: *t* (te), *tb* (también), *nas* (buenas), *q/k* (que).
- **inanidades:** Neologismos, composiciones de varios tipos de jerga o transformaciones sin explicación. Ej.: *wolas* (hola), *nu* (no), *okis* (okay), *se* (sí).

La clasificación de Nocon, Cuevas, Magat, Suministrado y Cheng (2014) y de Nocon, Cuevas, Gopez y Suministrado (2014) categoriza los textismos del tagalo y del inglés de manera separada. Los textismos del filipino coinciden en gran parte con los existentes en CZ.

- **tagalo**:
  - **Omisión de la -a final** - Ej.: *n* (na), *s* (sa), *b* (ba);
  - **Esqueleto consonantal** - Ej.: *nmn* (naman), *bkt* (bakit), *dpt* (dapat);
  - **Estilo fonético** - Ej.: *i2* (ito), *xa/xya/cya* (siya), *d2* (dito);

  - **Reduplicaciones** - Ej.: *b3* (bababa), *pu2nta* (pupunta), *pinagsama2* (pinagsamasama);
  - **Omisión de la H o I** - Ej.: *kaiintay* (kahihintay), *lan* (ilan).

- **inglés**:
  - **Acortamiento** - Ej.: *morn* (morning);
  - **Contracciones** - Ej.: *wk* (week);
  - **Truncamiento** - Ej.: *goin* (going);
  - **Inicialismo** - Ej.: *WHO* (World Health Organization);
  - **Acrónimos** - Ej.: *AIDS* (Acute Immune Deficiency Syndrome);
  - **Letras/números homófonos** o **transformación ortográfica** - Ej.: *b4* (before), *U* (you);
  - **Error intencional** o **escrita no convencional** - Ej.: *nite* (night), *luv* (love);
  - **Estilización de acento** - Ej.: *wassup/wazzup* (what's up), *cos/coz/cuz* (cause);
  - **Emoticones** o **smileys** - Ej.: *:-)* (sonrisa), *:-(* (triste).

3.1.3 *Lenguajes para representación de datos lingüísticos*

En este apartado, describimos algunos lenguajes para representación de datos lingüísticos utilizados en este trabajo.

3.1.3.1 *XML*

XML (*eXtensible Markup Language*) es un lenguaje de etiquetado desarrollado por *World Wide Web Consortium* (W3C) para codificar datos de manera legible, tanto para las máquinas como para las personas. A pesar de simple, es bastante versátil y permite especificar diferentes tipos de datos.

Una manera sencilla de visualizar mentalmente un fichero XML es pensar en una estructura de árbol. En la figura 23, vemos que una «canción» (nodo raíz) tiene «título», «artista», «letrista» y «letra» (nodos hijos). La letra, a su vez, está formada por versos, que a su vez, se descomponen en líneas. Los nodos que no tienen hijos (título, artista, letrista, línea) son llamados de «nodos hoja», y es donde se encuentra la información asociada al nodo.

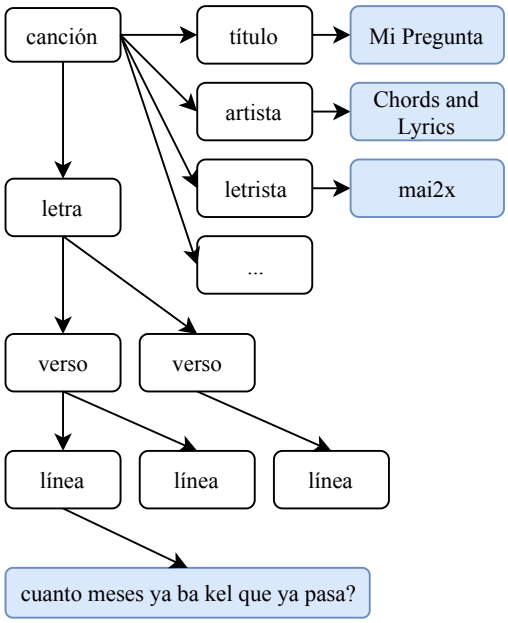


Figura 23: Ejemplo de archivo XML representado en forma de árbol.

3.1.3.2 *RDF*

RDF (*Resource Description Framework*) es un estándar para el intercambio de datos en la web desarrollado por W3C (*World Wide Web Consortium*). Se basa en el uso de tripletas para describir las relaciones entre dos recursos. Un recurso puede ser algo concreto o abstracto. Las tripletas siguen la estructura <*sujeto*> <*predicado*> <*objecto*>, siendo el sujeto y el objecto los dos recursos relacionados, y el predicado, la descripción de esa relación.

El ejemplo 24 ilustra ejemplos de relaciones que se pueden construir en RDF: la canción «Nuay Mas» (sujeto) tiene a «Comic Relief» (objecto) como «Artista» (predicado) y «Zack Quijano» (objecto) como «Escritor»/«Compositor» (predicado).

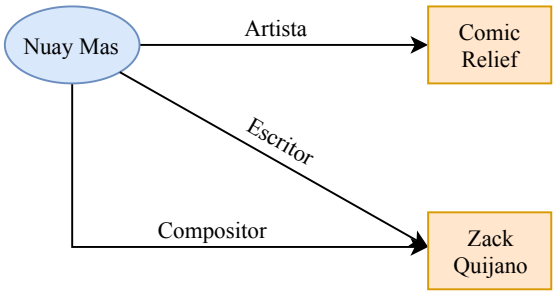


Figura 24: Ejemplos de relaciones que pueden representarse en RDF.

3.1.3.3 *Ontologías y OWL*

Una ontología, desde el punto de vista informático, es una especificación formal de una conceptualización, es decir, una representación del conocimiento de un dominio que queremos modelar. Los componentes básicos de una ontología son

(A. Gómez-Pérez, Fernandez-Lopez & Corcho, 2004; Asuncion Gómez-Pérez & Benjamins, 1999; Gruber, 1995):

- **Clases:** Son representaciones de conceptos, sean estos concretos o abstractos, del dominio que se desea formalizar. Se organizan en taxonomías.
- **Relaciones:** Son representaciones de los diferentes tipos de interacciones que pueden producirse entre los conceptos del dominio.
- **Funciones:** Son un tipo especial de relación en el que se calcula un elemento por medio de otros elementos de la ontología.
- **Instancias:** Son utilizadas para representar elementos dentro del dominio.
- **Axiomas:** Son utilizados para modelar las relaciones y condiciones que deben cumplir los elementos de la ontología.

El lenguaje estándar de W3C (*World Wide Web Consortium*) para la representación de ontologías es OWL (*Web Ontology Language*). Construido como una extensión del vocabulario RDF, OWL proporciona un conjunto de primitivas que permite modelar en una ontología los conceptos y sus relaciones, como disyunción, cardinalidad, igualdad, características de las propiedades, clases enumeradas, etc. (The World Wide Web Consortium (W3C), 2004, 2013).

3.1.3.4 *Corpus*

Un corpus es un conjunto de textos que sirven de base para el análisis de una lengua. El corpus puede estar formado por textos o transcripciones del lenguaje oral. Los textos pueden ser de un determinado campo o de diferentes temáticas, según los criterios y el uso que se pretenda dar al corpus construido. Hay también corpus bilingües o paralelos, o sea, que contienen los mismos textos o textos similares en dos lenguas diferentes.

Dos criterios importantes a la hora de construir un corpus son el de la representatividad y el del equilibrio. Para que los resultados de la investigación utilizando el corpus sean válidos, el corpus de trabajo debe tener una extensión suficientemente grande y una distribución que represente suficientemente el fenómeno que se desea estudiar. En el caso de lenguas con pocos recursos, estos criterios suelen ser extremadamente difíciles de alcanzar.

3.1.3.4.1 NIF

NIF (*NLP Interchange Format*) es un formato basado en RDF/OWL para anotación de textos que utiliza estándares W3C y que se adhiere al paradigma de los datos

enlazados (*Linked Data*). Las anotaciones en NIF son recursos RDF definidos de manera unívoca por identificadores URI (*Uniform Resource Identifier*). El objetivo de NIF es facilitar la interoperabilidad de diferentes herramientas de PLN (Soroa y col., 2017). NIF se puede utilizar para construir corpus anotados.

3.1.3.4.2 TEI-XML

TEI-XML es un formato XML definido por el consorcio *Text Encoding Initiative* (TEI) que responde a las necesidades de codificación de diferentes tipos de documentos de las ciencias humanas y sociales: manuscritos, diccionarios, poemas, prosa, transcripciones, etc. La ventaja de utilizar este formato es su usabilidad, ya que hay una gran disponibilidad de herramientas capaces de procesar ficheros en formato XML (Canadian Institute for Research in Computing and the Arts, 2010; Text Encoding Initiative (TEI), 2019), lo que hace que en los últimos tiempos se haya adoptado ampliamente en la creación de corpus.

3.1.4 *Procesamiento del Lenguaje Natural*

El Procesamiento del Lenguaje Natural (PLN) es un campo multidisciplinar de la informática, de la inteligencia artificial y de la lingüística computacional relacionado con la manipulación, la comprensión y la generación de lenguaje humano de manera automática por ordenadores. Tiene una relación muy estrecha con la lingüística computacional y las tecnologías del habla, hasta tal punto que dichos términos se confunden en ciertas ocasiones.

Entre las diversas áreas que forman parte del PLN, podemos citar: la sumarización automática, el análisis de discurso, la traducción automática, el análisis morfológico, el análisis sintáctico, el reconocimiento óptico de caracteres, la búsqueda de respuestas, el reconocimiento de entidades con nombre, la resolución de correferencias, la extracción de relaciones, la resolución del límite de la frase, el análisis de sentimientos, el reconocimiento y la sintetización del habla, la segmentación del habla, la clasificación de documentos, la alineación de textos, la desambigüación lingüística, la recuperación y la extracción de información, el procesamiento digital de voz, la corrección ortográfica, y otras.

En muchas de estas áreas se reconocen al menos dos enfoques principales: el basado en reglas y el estadístico. Los sistemas basados en reglas exigen la participación de un lingüista, que debe determinar las reglas que deben seguirse para la realización de una determinada tarea. Los sistemas estadísticos, a su vez, derivan automáticamente dichas reglas a partir de un modelo estadístico y una gran cantidad de datos.

3.1.4.1 *Traducción automática (TA) estadística*

La Traducción Automática Estadística (TA estadística) es un paradigma empírico de traducción automática. En otras palabras, a diferencia de la TA basada en reglas, utiliza poca o ninguna teoría lingüística para realizar la traducción de una lengua fuente a una lengua meta.

La idea subyacente en la TA estadística es la del modelo *Noisy Channel* (Brown y col., 1990): dada una frase en la lengua fuente (LF), hay que descodificarla, eliminando el ruido, para revelar la frase en la lengua meta (LM). Para ello, se utilizan un modelo de cómo se distorsiona el mensaje (modelo de traducción) y un modelo de los mensajes probables (modelo lingüístico) (figura 25).

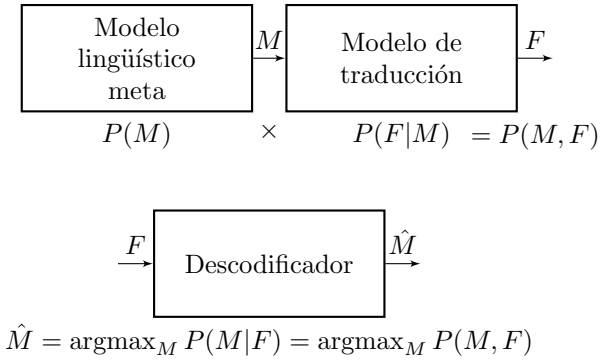


Figura 25: Un sistema de traducción automática estadística. Adaptado de Brown y col. (1990).

Para estimar estos modelos, son necesarios datos estadísticos extraídos de corpus paralelos, que se obtienen como muestra la figura 27: de los textos paralelos en una lengua fuente (LF) y una lengua meta (LM), se computan sus alineamientos, es decir, las equivalencias entre sus elementos (figura 26), obteniendo las «unidades de traducción» o «frases». De esa manera, para realizar la traducción de una frase F en la LF, el descodificador debe, dado el modelo lingüístico y el modelo de traducción, buscar la traducción más probable, $\hat{M}$ (figura 25).

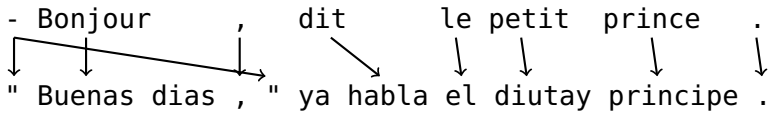


Figura 26: Ejemplo de alineamiento léxico. (Frase extraída de Saint-Exupéry, 1943, p. 77; Saint-Exupéry, 2018, p. 66).

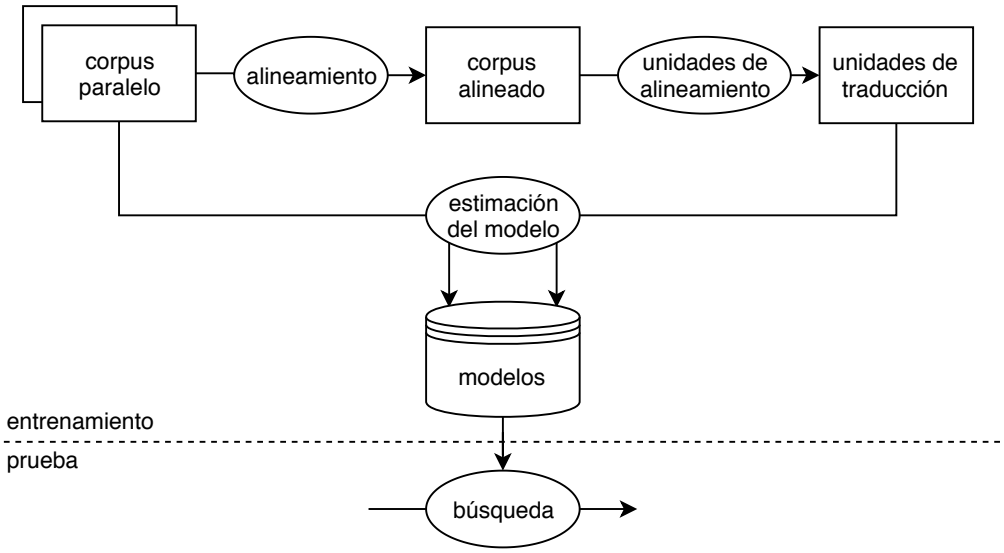


Figura 27: Flujo de un sistema de traducción automática estadística. Adaptado de Singla (2015).

Para obtener una traducción de calidad aceptable es necesario disponer de una cantidad razonable de textos bilingües.

3.1.4.1.1 TA estadística de caracteres

Una aproximación diferente, propuesta por Vilar, Peter y Ney (2007), es la de la TA de caracteres (*Character-Based Statistical Machine Translation*): consiste en tratar la traducción de una frase no como una secuencia de palabras, sino como una secuencia de caracteres. El alineamiento se hace, por lo tanto, entre los caracteres de la LF a la LM, como se puede ver en la figura 28.

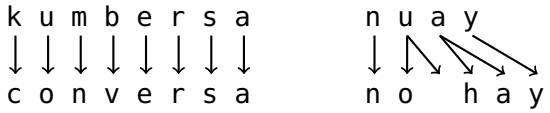


Figura 28: Ejemplo de alineamiento de caracteres.

Esta aproximación se aplicó sobre todo en la traducción entre lenguas cercanas con muchos cognados (Nakov & Tiedemann, 2012; Tiedemann, 2009; Vilar y col., 2007), en muchos casos, siendo usada como complemento para los modelos de traducción por palabras tradicionales. En la literatura también hay casos de aplicación en tareas más allá de la traducción, como la normalización de textos escritos en otras épocas (Korchagina, 2017; Schneider, Pettersson & Percillier, 2017) o de textos en variantes dialectales (Scherrer, Samardžić & Glaser, 2016), transliteraciones (Karimi, 2008; Tiedemann & Nabende, 2009), generación de cognados (Beinborn, Zesch & Gurevych, 2013) y corrección ortográfica de términos de búsqueda (Hasan, Heger & Mansour, 2015).

3.1.4.2 *Correctores ortográficos*

Un corrector ortográfico es un software capaz de identificar errores ortográficos en un texto y sugerir posibles correcciones para los mismos. Hay dos tipos de errores: los que generan palabras inexistentes (*non-word errors*) y los que generan palabras válidas, pero incorrectas en el contexto (*real word errors*) (Jurafsky & Martin, 2018).

Se dice que el primer corrector ortográfico, *SPELL*, surgió en el 1971 en el *Artificial Intelligence Laboratory* de la Universidad de Stanford. *SPELL* fue escrito en Assembly por Ralph E. Gorin y posteriormente mantenido por Wayne E. Matson en el 1974 y W. Bill Ackerman en el 1978. En el año 1983, *SPELL* fue portado al lenguaje C por Pace Willisson bajo el nombre de *ispell* (Earnest, 2016; Peterson, 1980; Willisson, 2015), dando origen a la familia de correctores ortográficos *SPELL* (figura 29).

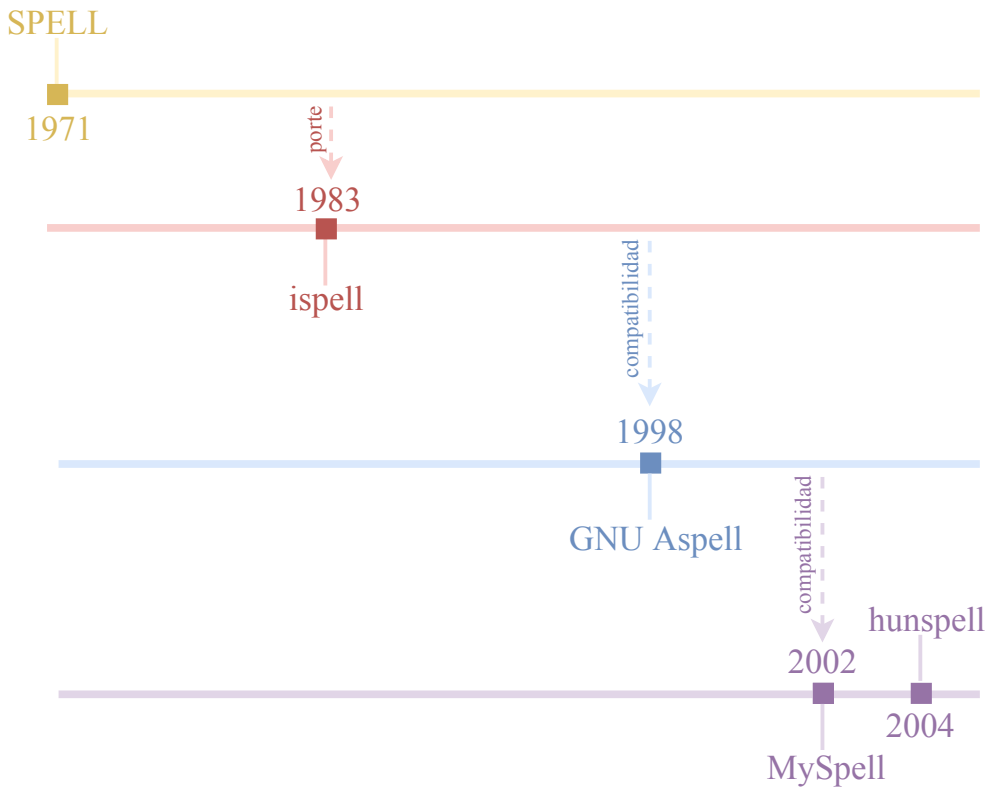


Figura 29: Línea del tiempo de los correctores ortográficos de la familia SPELL

Posteriormente, en 1998, aparece *GNU Aspell*, escrito por Kevin Atkinson y retrocompatible con *ispell* (Atkinson, 2019). *MySpell*, escrito por Kevin Hendricks en el 2002, surge como corrector ortográfico del paquete de ofimática *OpenOffice*, siendo retrocompatible parcialmente con *ispell* y *Aspell*. De *MySpell*, finalmente, nace *hunspell* (descrito en detalles en el apartado 3.2.4). *hunspell* es en la actualidad es el corrector ortográfico preferido por muchas tecnologías.

3.1.4.3 *Tokenización*

La tokenización es una tarea esencial de preprocesamiento en PLN que consiste en segmentar una cadena de caracteres en *tokens*. En el caso del castellano, los *tokens* normalmente están delimitados por espacios y signos de puntuación. El software que realiza la tokenización recibe el nombre de «tokenizador».

3.1.4.4 *Algoritmos fonéticos*

Los algoritmos fonéticos son utilizados para representar bajo un mismo código palabras que, aunque se escriban diferentes, tienen una pronunciación cercana según algún criterio definido. Los más conocidos de esos algoritmos son *Soundex* (1918) y *Metaphone* (Philips, 2000).

Por ejemplo, para los homófonos «their», «there» y «they're», el algoritmo *Metaphone* genera la misma salida **0R**. Lo mismo pasa con «your» y «you're», cuya salida es **YR**. Hay adaptaciones de ambos para otras lenguas, incluso el castellano.

3.2 HERRAMIENTAS UTILIZADAS

En este apartado, se describen las herramientas utilizadas en la ejecución de este trabajo.

3.2.1 *Python*

Python es un lenguaje de programación bastante utilizado en PLN. Es un lenguaje interpretado (no requiere que el código del programa sea previamente compilado antes de ser ejecutado), de tipado dinámico y con recolector de basura (Python Software Foundation, 2019). Es también bastante versátil, ya que permite diferentes paradigmas de programación: procedimental, funcional y orientado a objetos. Su popularidad de debe a la facilidad de aprendizaje y de uso, así como la enorme gama de bibliotecas disponibles.

3.2.2 *NLTK*

NLTK (*Natural Language Toolkit*) es la biblioteca más popular y más potente de PLN en *Python*. Incluye los algoritmos más populares de tokenización, etiquetado morfológico, lematización, análisis de sentimientos, clasificación de documentos y reconocimiento de entidades con nombres. Es un software libre y está disponible bajo la licencia *Apache License*, versión 2.0 (Bird, Klein & Loper, 2009).

3.2.3 *Moses*

Moses es un conjunto de herramientas de código libre que permite crear sistemas de TA estadística de manera sencilla (Koehn y col., 2007). El paquete contiene todas las herramientas necesarias para entrenar un modelo, siendo necesario tan solo disponer de datos paralelos en dos lenguas. Aunque permite utilizar diferentes alineadores y modelos de lenguaje, una combinación frecuente es la de utilizar *Moses* con el alineador *GIZA++* (Och & Ney, 2003) y el modelo de lenguaje *KenLM* (Heafield, 2011) o *SRILM* (Stolcke, 2002).

3.2.4 *hunspell*

hunspell es un corrector ortográfico y analizador morfológico de código abierto desarrollado por László Németh bajo las licencias LGPL[1]/GPL[2]/MPL[3]. Además de incluir soporte para el tratamiento de caracteres del conjunto UTF-8, permite codificar peculiaridades lingüísticas como composición y morfología complejas, afijos, reglas específicas para homónimos, reglas fonéticas, excepciones, sugerencias basadas en la distribución de las teclas del teclado, y muchas otras características (Németh, 2019).

Actualmente, es el corrector ortográfico estándar y se utiliza en muchas tecnologías populares, como los paquetes ofimáticos *LibreOffice* y *OpenOffice*, los navegadores *Mozilla Firefox* y *Google Chrome*, *Thunderbird*, en algunas distribuciones de *Linux*, y también en paquetes de software propietario, como *macOS*, *InDesign*, *memoQ*, *Opera* y *SDL Trados*.

Las dos operaciones más importantes que se pueden realizar con *hunspell* son la de verificar si una palabra es correcta y de generar candidatos de corrección para la misma. Por ejemplo, al introducir como entrada la palabra «dia» sin tilde, además de indicar que la palabra es inválida, *hunspell* proporcionará como candidatos de corrección (en orden): día, ida y di.

Una vez presentados los conceptos e ideas principales que fundamentan el desarrollo de este trabajo, pasamos ahora a discutir la metodología seguida en el mismo.

---

1 *GNU Lesser General Public License*
2 *GNU General Public License*
3 *Mozilla Public License*

CAPÍTULO 4

METODOLOGÍA

Como se ha discutido en el apartado 1.4, el CZ presenta una gran diversidad ortográfica, resultado tanto de la influencia de otras lenguas como de las diferentes corrientes ideológicas que orientan su escritura. La distancia que muchas veces separa la forma propuesta por la ortografía y la grafía del hablante corriente requiere un tratamiento especial para el caso del CZ, diferente del que se aplica normalmente en la corrección ortográfica de otras lenguas.

Juzgamos necesario, como paso inicial, estudiar las variaciones existentes en los escritos en CZ por medio de un corpus. Era por entonces ya conocida la existencia de un corpus del CZ, desarrollado en el contexto del *Chavacano Language Corpus Project* (CLCP), entre los años 2003 y 2004 por la Universidad Ateneo de Zamboanga (ADZU), bajo la coordinación del *Language Research Center* (LRC) de la editorial *Dunwoody Press*. Sin embargo, no nos fue concedido acceso al mismo, ya que McNeil Technologies, grupo del cual el laboratorio formaba parte, fue adquirida en 2010 por AECOM, y esta afirma no ser dueña de los datos. Por ello, se decidió emprender la construcción de un corpus propio para este trabajo de investigación.

Para estudiar los errores ortográficos del CZ, consultamos diferentes tipologías de errores existentes para otras lenguas (detalles en el apartado 3.1.4.2). Como parte de los errores que deseábamos corregir, estudiamos también algunas tipologías de textismos (incluidas en el apartado 3.1.2). Este estudio culminó en la elaboración de una tipología de errores ortográficos del CZ de naturaleza jerárquica.

Para solucionar el problema ortográfico del CZ, se propuso una solución de corrección ortográfica basada en TA estadística de caracteres. El modelo de TA estadística entrenado debía ser capaz de identificar patrones entre las diferentes grafías existentes y la ortografía y aprender a corregir tanto formas presentes en los datos de entrenamiento como palabras que encontrara por primera vez. Debía, además, no aplicar correcciones en palabras correctas en los textos originales.

El requisito principal del entrenamiento de un sistema de TA estadística es disponer de un corpus paralelo. En el caso de este trabajo, paralelo debe entenderse como material en la grafía original del autor y su respectiva conversión a la ortografía del CZ. La conversión de grafía y elaboración de los datos de entrenamiento fue realizada manualmente, con la ayuda de listas de palabras extraídas de diccionarios y un algoritmo fonético ad-hoc construido para el CZ.

Finalmente, para evaluar el funcionamiento de nuestra aproximación utilizando TA estadística de caracteres, implementamos un diccionario *hunspell*, que es la tecnología estándar actual para la corrección ortográfica. En el capítulo siguiente detallamos el desarrollo de cada una de las fases descritas.

CAPÍTULO 5

DESARROLLO

En este apartado, se discute el desarrollo de este trabajo de investigación. El diagrama de la figura 30 muestra el conjunto de las actividades realizadas y de qué manera éstas se interconectan e interrelacionan.

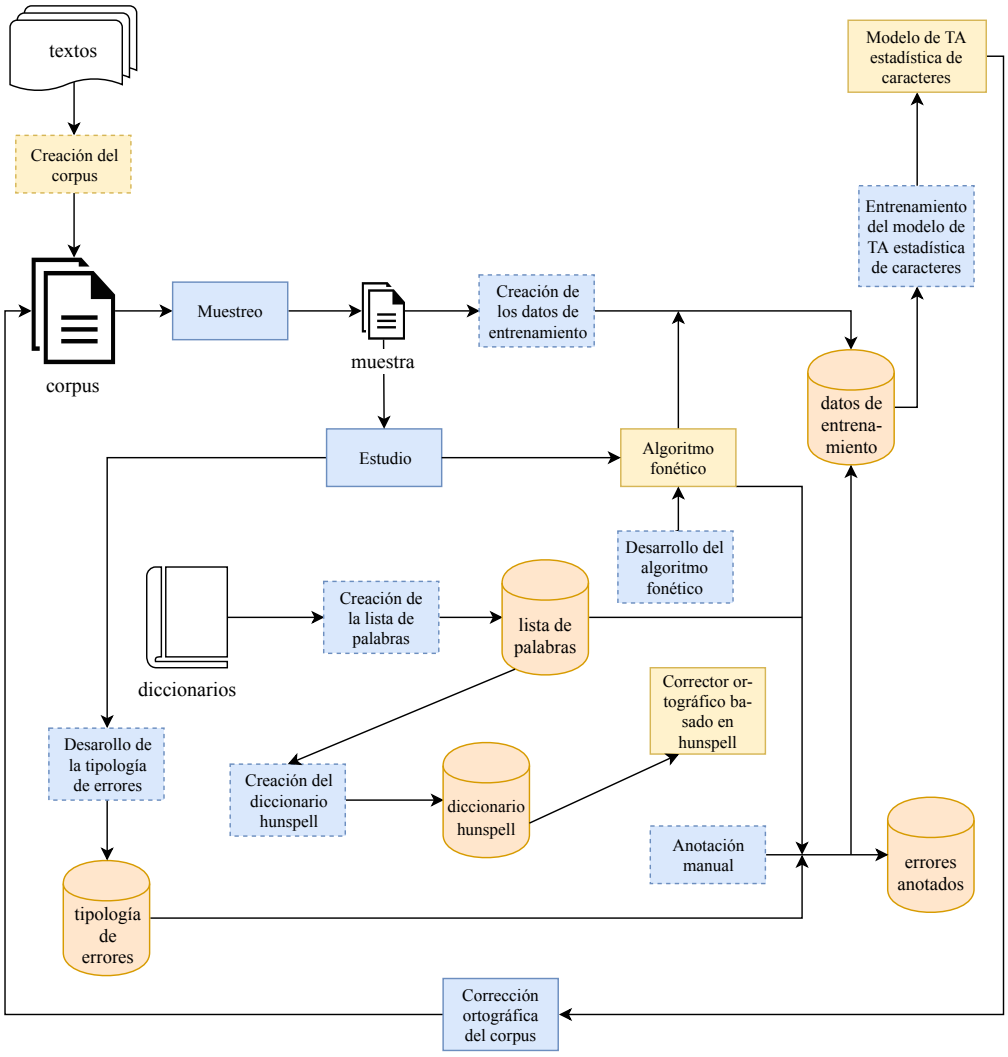


Figura 30: Tareas realizadas en el trabajo de investigación

5.1 CONSTRUCCIÓN DEL CONTEMPORARY WRITTEN ZAMBOANGUEÑO CHABACANO CORPUS

A partir de los resultados del estudio preliminar (véase el apartado 2), definimos los criterios de construcción del *Contemporary Written Zamboangueño Chabacano Corpus* (CWZCC).

- Puesto que nuestro objetivo era el de estudiar los errores ortográficos que ocurren en CZ, se buscó reunir material de diferentes fuentes escritas, incluyendo, además de todo el material del que se dispusiera en línea, texto de libros, revistas, periódicos, carteles, panfletos promocionales y cualquier otra fuente escrita en la lengua.
- Toda vez que la cantidad de material escrito en la lengua no era abundante y su inclusión en el corpus estaría sujeta a las licencias bajo las que este se publicó o a permisos de uso por parte de sus autores, decidimos maximizar el número de palabras en la medida de lo posible, aun cuando esto pudiera generar desequilibrios en el corpus.
- El estudio preliminar mostró que el uso del CZ se concentra en las esferas de comunicación interpersonal informal, por lo que los textos de este género deberían ocupar una parte importante del corpus.
- Puesto que había que tratar de capturar la diversidad de registros escritos y de variaciones ortográficas del CZ usado actualmente por los hablantes, prescindimos de incluir transcripciones de la lengua oral o escritos demasiado antiguos.
- El corpus resultante debía servir para el estudio de los errores ortográficos que ocurren en el CZ escrito y también como fuente de datos para el entrenamiento de herramientas de corrección ortográfica.

El formato original del corpus eran ficheros CSV para los datos provenientes de redes sociales y ficheros de texto (sin estructura pre definida) para los demás géneros. Después de entrenar un modelo de TA estadística de caracteres (véase el apartado 5.7), lo utilizamos para corregir el corpus y generar versiones del mismo en NIF y TEI-XML (véase los apartados 3.1.3.4.1 y 3.1.3.4.2 respectivamente). Los errores, en caso de haber sido clasificados anteriormente en la fase de entrenamiento, van acompañados de su respectiva clasificación según la tipología de errores (véase el apartado 5.2); de lo contrario, van marcados como no clasificados, indicando que la corrección fue generada por el modelo de TA estadística y necesita ser revisada manualmente.

Algunos materiales impresos que no se encontraban en formato digital tuvieron que ser escaneados y posteriormente digitalizados con la ayuda de un programa de reconocimiento óptico de caracteres (OCR). Debido a los frecuentes errores producidos, sobre todo en las regiones cercanas al lomo de los libros, el texto tuvo que ser revisado manualmente. Es posible, sin embargo, que todavía queden algunos errores.

En el apartado 5.1.1, describimos la composición del corpus y detallamos cada género que lo compone.

5.1.1 *Composición*

Entre los textos que se recogieron se identificaron 9 géneros distintos. En la tabla 5 pueden verse el número de palabras de cada género, así como el total de palabras del corpus CWZCC.

Tabla 5: Número de palabras de cada género del corpus CWZCC.

GÉNERO	Nº DE PALABRAS
Textos educativos	111.777
Ficción	104.130
Poesía	15.398
Canciones	41.351
Noticias	158.349
Religión	1.631.820
Autoayuda	14.426
Internet	5.896.517
Otros	64.432
Total	8.038.200

Se podría cuestionar la representatividad del CWZCC. Sin embargo, hay que considerar las limitaciones impuestas por la disponibilidad de material en CZ y la dificultad de obtener permiso para el uso de muchos de los textos. En el apartado 5.6 describimos nuestra estrategia para reducir el sesgo del modelo entrenado hacia registros de naturaleza informal, ya que son mayoritarios en el corpus. No obstante, dado que son estos los registros que decidimos priorizar, la composición del corpus está alineada con los objetivos definidos inicialmente.

5.1.1.1 *Textos educativos*

En este género, incluimos libros complementarios en la enseñanza del CZ a niños de primaria y libros de frases para extranjeros:

- Mangaser, V. D. & Abelardo, F. E. (2015). Filipino Children's Treasury Chavacano Filipino-Ingles Ilustrao. Ciudad de Quezón, Filipinas: Vibal Publishing House, Inc.
- Mayer, Audrey (comp.) (1979). Languages of the Southern Gateway: A Phrase Book of Chavacano, Sinama, Tausug, Yakan, and Including English and Pilipino. Filipinas: Department of Education, Culture and Sports and the Summer Institute of Linguistics.
- Rubrico, J. G. (2007). Conversa Kita Chavacano. Ciudad de Quezón, Filipinas: Language Links Foundation, Inc.
- Tatoeba (2019). Sentences in Chavacano. Recuperado de https://tatoeba.org/eng/sentences/search?query=&from=cbk&to=und
- Villaneza, R. J. & Abelardo, F. E. (2014). Mi Primer Diccionario-Caton Chavacano-Filipino-Ingles Illustrado. Ciudad de Quezón, Filipínas: Vibal Publishing House.

Además, se incluyen en este género varios conjuntos de diálogos de diferentes manuales de tagalo para extranjeros adaptados al CZ que el autor utilizó en clases con hablantes nativos para su aprendizaje privado.

Pese a nuestros esfuerzos, en razón de la dificultad de obtención de permiso de uso, materiales provenientes de los libros de texto de primaria editados por el Departamento de Educación no pudieron incorporarse al corpus.

5.1.1.2 *Ficción*

En este género incluimos algunos cuentos populares, obras de prosa y de teatro, de autores locales o traducidos. Por falta de permisos de uso adecuados por parte de los autores y/o editores responsables, algunas obras de teatro y material proveniente de las coletáneas de textos editados por el gobierno de la Ciudad de Zamboanga no pudieron ser incorporados al corpus.

A continuación, listamos las obras originalmente escritas en CZ incorporadas a este género:

AUTOR	TÍTULO
Antonio Reyes Enriquez	El Pirata Tagal
Cristhel Camille Del Fierro	Maltrato
Cristhel Camille Del Fierro	Unrato Lang Gale
Dianelle Bucoy	Mi Tata o Mi Nobyo
Donnalei Grace Ruales	Gracias Kontigo

AUTOR	TÍTULO
Ella Monette Galvez	Promesas No Hay Puede Hase
Michael Spencer Salvador	Dolor y Kambyo
Nina Lacandalo-Nohay	Pulumbato
Nina Lacandalo-Nohay	Chonggo y Tortuga
Rodelei L. Rodriguez	Kuwerpo y Sakripisyo
Rodelei L. Rodriguez	Ay! Pensaba Yo Amo Ya

Las obras traducidas incorporadas son:

- Alonzo, H., Warren, C., Rat Western (autores), & Anoya, J. P. (trad.) (2018). Cosa adentro na paso? Borrador no editado, Proyecto Coloring Colorao.
- Dingwall, C., McKimmie, S., Stoop, T. (autores), & Ramos Fabiania, D. O. (trad.) (2018). El grande plano del Diutay Subay. Borrador no editado, Proyecto Coloring Colorao.
- Giono, J. (autor), & Villano, R. C. (trad.) (2018). El gente ya sembra pono. Borrador no editado, Proyecto Coloring Colorao.
- Houareau, J., Cuzen, B., Lamoral-Rosmarin, C., & Agil, M. (trad.) (2018). Surpresa na compleanyo de Thato. Borrador no editado, Proyecto Coloring Colorao.
- Jobson, L., Breytenbach, J., Thesen A. (autores), & Agil, M. (trad.) (2018). Peskaw y Regalo. Borrador no editado, Proyecto Coloring Colorao.
- Lal, A. (autor), & Lozada, J. (trad.) (2018). Mi amigo. Borrador no editado, Proyecto Coloring Colorao.
- Newhard, C. (autor), & Pantaleta, F. O. (trad.) (2017). Si Amina y el Ciudad de maga Flores. Filipinas: Sari-Sari Storybooks.
- Pulu, P. (autor), & Lozada, J. (trad.) (2018). Mi awto. Borrador no editado, Proyecto Coloring Colorao.
- Rahalkar, S. (autor), & Lozada, J. (trad.) (2018). Perdido kabar ya puwede ingcontra. Borrador no editado, Proyecto Coloring Colorao.
- Reyes Enriquez, A. (autor y trad.) Subanons. Chapter 1. Recuperado de https://www.scribd.com/document/23599741/Calandracas-1
- Saint-Exupéry, A. de (autor), & Herrera, J. (trad.) (2018). El Diutay Principe. Manila, Filipinas: Jerome Herrera.

- Saint-Exupéry, A. de (autor), & De Los Reyes, A. (trad.) (2018). El Principe Niño. Neckarsteinach, Alemania: Edition Tintenfass.

5.1.1.3 *Poesía*

En este género, incluimos la totalidad de los 31 poemas encontrados en el libro:

- Macansantos, F. C. (2011). Balsa: poemas chabacano. Manila, Filipinas: National Commission for Culture and the Arts.

Incluimos también poemas de autores varios que pudimos encontrar en internet y algunas traducciones:

AUTOR	**TÍTULO**
Charmaine Arcillas Lim	Cosa ya lang ba el ya queda na de aton amor?
Charmaine Arcillas Lim	Para con acquel estangero
Charmaine Arcillas Lim	Quiere yo ama otra vez...
Rumi; Charmaine Arcillas Lim (trad.)	El Amiga del Alma
Rumi; Charmaine Arcillas Lim (trad.)	Este Sufrimiento
Rumi; Charmaine Arcillas Lim (trad.)	Go with muddy feet
Rumi; Charmaine Arcillas Lim (trad.)	Miedo
Rumi; Charmaine Arcillas Lim (trad.)	Paloma
Desconocido; Charmaine Arcillas Lim (trad.)	<Sin titulo>
Darren Bendanillo	Departir
Darren Bendanillo	El puno tambis
Darren Bendanillo	Isla del Santa Cruz
Darren Bendanillo	Ladron
Darren Bendanillo	Maga pajaro na pueblo del Zamboanga
Darren Bendanillo	Samal
Darren Bendanillo	Zamboanga
Jesthoni Acosta	Mi Nana
Jesthoni Acosta	Ya Ama
Mavie Labor	Pabor Daw???
Mavie Labor	Solamente tu!

5.1.1.4 *Canciones*

El género de letras de canciones es, sin duda, el más diverso del CWZCC. Excepto canciones de intérpretes y compositores ilustres de la ciudad de Zamboanga como Norma Camins-Conti, Major Chords y Titang Jaldon, nuestro objetivo fue la recompilación de canciones disponibles en YouTube y Facebook subtitulados o que tuviesen sus letras con el video (en la descripción o en los comentarios). No recogimos, por lo tanto, canciones que no tuvieran sus letras disponibles en internet, salvo cuando estas nos fueran ofrecidas por sus autores.

El mayor reto de la construcción de esta parte del corpus fue determinar la autoría de las canciones, ya que algunas de ellas no traían en el vídeo ni siquiera indicación del nombre del artista. Esto solo fue posible gracias a la ayuda de diferentes artistas de la escena musical local, que nos ofrecieron su ayuda y conocimiento, lo que nos permitió recuperar una parte de la historia de la canción zamboangueña de los últimos 10 años. Algunos pocos autores no pudieron ser localizados o no se obtuvo respuesta de su parte y, por ende, su material no se incorporó al corpus.

En la tabla 6 listamos las letras de canciones recompiladas, junto con sus respectivos letristas y uno o más posibles intérpretes. En ciertos casos, logramos recoger más de una transcripción de la letra de la misma canción, lo cual va indicado por el número de la columna «V.».

Tabla 6: Listado de las letras de canciones del género «Letras de canciones» del CWZCC.

Letrista	Título	Intérprete(s)	V.
Aaron Misa	La Bella de Zamboanga	Aaron Misa	1
Aaron Misa	Otro Mujer	Aaron Misa	1
Aaron Misa	Solamente si Duterte	Aaron Misa	1
Aaron Misa	Vamos	Gianella Marcos & Neil Alvarado	1
AimPinky	Ci Un Porhemplo Man To AI Esta Lejos	AimPinky	1
AimPinky	Habla tu permi	AimPinky	1
AimPinky	I Believe	AimPinky	1
AimPinky	Masquin umpoko de lastima	AimPinky	1
AimPinky	Otravez	AimPinky	1
AimPinky	Promete To	AimPinky	1
Al-rass Amarillo	Alegre	Al-rass Amarillo	1
Al-rass Amarillo	Vene Ya	Pluma & Tweena Campos	1

LETRISTA	TÍTULO	INTÉRPRETE(S)	V.
Alshamir Bryan Barrera Aripuddin	Eternamente	Keidi Shay Lim Napalcruz & Jadie Mae Natividad	1
Alshamir Bryan Barrera Aripuddin	Hasta na final	Keidi Shay Lim Napalcruz	2
Alshamir Bryan Barrera Aripuddin	Para Contigo	Keidi Shay & Rogin Christ	1
Alshamir Bryan Barrera Aripuddin	Para Contigo	Keidi Shay Lim Napalcruz & Rogin Christ Eribal	2
Alshamir Bryan Barrera Aripuddin	Tu Siempre (Mi Madre Tierra)	Alshamir Bryan Barrera Aripuddin	1
Angelo Dean Bustillo	Inspiracion	Watch Your Step Band	1
Angelo Dean Bustillo	Solo	Watch Your Step Band	1
Anna Liza Dela Zerna-Martin	Zamboanga	Khen Martin & BlindCurrent	2
Arnold Prio	Habla Tu	Arnold Prio	1
Benjie A. Mahasol	Viaje na Wow Zamboanga	Benjie A. Mahasol	1
Bobby Piedad	Verdad el mi amor	Bobby Piedad	1
Charlie Villanueva	Baila	Kathryn Daculan / rainplay	2
Charlie Villanueva	Chinita	Lando Bisaya	1
Daniel M. Taclap	Miyo ya lang tu	Facil Cielo	1
Des Kan Sabess	Buwan	R3mediaproduction	2
Dexter Ando	Despacito	Dexter Ando	2
Dexter Ando	Petate on the floor	Bobby Piedad	1
Donex Magallon Bonifacio Jr	Man unido kita	Donex	1
EJ Natividad	Bira Ole	cheeze de sal	1
EJ Natividad	Tiene Duenyo	cheeze de sal	1
Gideon Cascolan	De Ultramar	Francel Joy de Leon	1
Isser Jorell L. Yao	Algun Dia	Isser Yao	1
Isser Jorell L. Yao	Etu	Isser Yao & Joan Kuan	1
Isser Jorell L. Yao	Etu	Sarah Baul	2
Jay-ar Ave	Etu mi alegriya	Ji & Jr	1

LETRISTA	TÍTULO	INTÉRPRETE(S)	V.
Jay-ar Ave	Mi atencion	JJ & JR ft. UNIQUE BUDDIES	1
Jay-ar Ave	Tu pati mi tristesa	JJ & JR Ave	1
Jefferson Tuazon Sanico	Orgullo del Paiz	Skemberlu Band	1
Jervil Formilleza Omaga	Deficil	Camille Remulta	1
Jervil Formilleza Omaga	Fiesta Hermosa na Ciudad Latina	Xena Estrada, Aaron Fernandez, Binibining Beats & Miko Gwapito	1
Jervil Formilleza Omaga	Havana	Mara Palengkera	1
Jervil Formilleza Omaga	Pascua de Zamboanga	Armand Saavedra Cruz & Sweet Xena M. Estrada	1
Jervil Formilleza Omaga	Señorita	Armando Cruz, Jr	2
Jervil Formilleza Omaga	Telefono (Mi Amor, Bolbe Ya)	Sweet Xena M. Estrada	1
Joey Panganiban Dayagdag	Esta aqui	Ice Cream Bear	1
John Leoner S. Tatil	Para Contigo	nephelim	1
John Nuluddin Kandero	Fuerte El Tama	Cheapshot of BANGSAMORO	1
Jomarie Navarro Alamhali	Contigo lang	Nathalie Napalcruz & Kobe Tompong	1
Jomarie Navarro Alamhali	Mi Amor Zamboanga	Rommel Fernandez & Clarise Paras	1
JP Montuno	Sabe ba etu	Song Syndrome	1
JR Natividad	Contigo Lang	Mirage Zamboanga	1
JR Natividad	Etu Lang (Nuay mas Otro)	Mirage Zamboanga	2
JR Natividad	Pakilaya [Iyo Contigo ta Ama]	Mirage Zamboanga	1
JR Natividad	Puede Ya	Mirage Zamboanga	1
JR Natividad	Sabor de Alegria	Mirage Zamboanga	1
Julia "Titang" Enriquez Jaldon	Como un cancion	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	El Dalaga Zamboangueña	Titang Jaldon	2

LETRISTA	TÍTULO	INTÉRPRETE(S)	V.
Julia "Titang" Enriquez Jaldon	El Lenguaje Chavacano	Titang Jaldon	4
Julia "Titang" Enriquez Jaldon	El modo de comer	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	El Verdadero Zamboangueño	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	Moda del Circa	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	Por causa tuyo	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	Sueño de Zamboanga	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	Tu Amo Di Mi Sueno	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	Zamboanga de Antes	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	Zamboanga Derrotada	Titang Jaldon	1
Julia "Titang" Enriquez Jaldon	Zamboangueña de mi sueño	Titang Jaldon	2
Jullienne Fortich Tuazon	Estrellas	Julienne Tuazon	1
Kenneth Joseph Martin dela Zerna	babe	Khen Martin	1
Kenneth Joseph Martin dela Zerna	Miss	Khen Martin & BlindCurrent	1
Jinda Saint (aka mai2x)	Mi Pregunta	chords and lyrics	1
Major Chords	Aire de Zamboanga	Major Chords	1
Major Chords	Canciones de Bata	Major Chords	2
Major Chords	Canciones de Patente	Major Chords	2
Major Chords	Malo mucho vicio	Major Chords	1
Major Chords	Zamboanga Chavacano	Major Chords	2
Whey Guevara	Iyo ya lang era	Maldita	5
Whey Guevara	Martir	UNiCA	1
Whey Guevara	Porque	Maldita	8
Whey Guevara	Porque (solo estribillo)	Maldita	2
Mark Anthony G. Tolentino	Dejalo ya	Jubaira S. Garcia	2

LETRISTA	TÍTULO	INTÉRPRETE(S)	V.
Mark Anthony G. Tolentino	Solo en mi sueño	Jubaira S. Garcia & Jomarie Navarro Alamhali	1
Mark Anthony G. Tolentino	Yo si Paz	Mary Grace Delos Reyes	1
Mark Eingel N. Fernandez	right here waiting	Herra Mark & Rey	1
Mark Lim Hao	Kuntigo lang	Mark Lim Hao	1
Mark Lim Hao	Nececita lang yo sabe	Mark Lim Hao	1
Mark Lim Hao	Para kuntigo	Mark Lim Hao	1
Mark Lim Hao	Para lang kuntigo	Mark Lim Hao	1
Mark Lim Hao	rabiona	Mark Lim Hao	1
Mark Lim Hao	Senyorita	Mark Lim Hao	1
Mark Lim Hao	Un Zamboangueña	Mark Lim Hao	1
Nari Ramchand	Bulag El Amor	Nari Ramchand	1
Nari Ramchand	Tiene Ves	Nari Ramchand	1
Norma Camins-Conti	Ay Apaga la Luz Contigo Ya Aprende Yo	Prince Anthony Tangon & High Hatters Singers	2
Norma Camins-Conti	Desde'l Corazon	WMSU GRAND CHORALE	1
Norma Camins-Conti	Solo Tu	High Hatters Singers	1
Norma Camins-Conti	Himno de Zamboanga de Antes	Jaypee Mendoza	1
Norma Camins-Conti	La Rosa de Zamboanga	WMSU Grand Chorale & The High Hatters	2
Norma Camins-Conti	Mi Pais	-	1
Norma Camins-Conti	Mientras que yo ta vivi	Major Chords	3
Norma Camins-Conti	Paseo de Amigos	Major Chords	1
Norma Camins-Conti	Un Sitio de Amor y Paz	Norma Camins-Conti	1
Norma Camins-Conti	Un Sueño de Paz	Alshamir Bryan Barrera Aripuddin	1
Norma Camins-Conti	Vamos a Zamboanga	Major Chords	3
Oliver Bernardo	Hasta Ahora	Oliver Bernardo & Steel Gray Band	1
Patrick Jethro	Fantasia	Patrick Jethro	1
Pondong	Nohay ni rason	Pondong	1
Pondong	Promesa	Mushroomhead	3
Prince Aarol Fernandez	Mi amor	Prince Aarol Fernandez	1

LETRISTA	TÍTULO	INTÉRPRETE(S)	V.
QUEEN	Birahan	QUEEN	1
QUEEN	Mi Esperanza	QUEEN	2
Randy Batiquin	El Primer Mujer	Randy Batiquin	1
Randy Batiquin	Mi Barangay	Randy Batiquin	1
Randy Batiquin	Nuay dia que hinde yo cuntigo ta pensa	Randy Batiquin	1
Randy Batiquin	Porque ba	Randy Batiquin	1
Refciejay Apiado Pagotaisidro	Amor Que Tarda Entero Vida	Twinkle	1
Refciejay Apiado Pagotaisidro	Baila Cun Mi Tata Ole	Twinkle	2
Refciejay Apiado Pagotaisidro	Dimiyo Ya Lang Tu	Twinkle	1
Refciejay Apiado Pagotaisidro	Porque Aura Lang Tu	DJ Darel & Twinkle	1
Refciejay Apiado Pagotaisidro	Na Mi Sueño	Twinkle	1
Rey Into Gaspar	Buko	Jayvee Manalansan	1
Rey Into Gaspar	Chinito	Keidi Shay Lim Napalcruz	2
Rey Into Gaspar	Cay contigo ta ama	Rey Gaspar	1
Rey Into Gaspar	Etu Lang	Rey Gaspar	1
Rey Into Gaspar	Si io viejo ya y el pelo blanco ya	-	1
Rey Into Gaspar	Serca na cielo	Rey Gaspar	1
Rey Into Gaspar	Io para contigo y etu comigo	Rey Gaspar	1
Rey Into Gaspar	Despues de todo	Rey Gaspar	1
Robert Barrera	Este dia	SYMPHONIA	1
Robert Barrera	No este amor	WARD-9	1
Robert Djbust Nadera	El Chavacano	DJBUSTED	1
Robert Djbust Nadera	El Mensaje	DJBUSTED ft. Rosemar Fuentes	1
Roel M. Apolinario	Palabra	LOST PRIMOZ / Chrysolite band / Makahiya Grass Band	1
Roel M. Apolinario	Suenio	Chrysolite band	2

LETRISTA	TÍTULO	INTÉRPRETE(S)	V.
Rogin Christ Eribal	Tu y yo	Rogin Christ Eribal	2
Roland Villanueva Jr	Cambio	Zero	2
Roland Villanueva Jr	Deja	Zero	2
Roland Villanueva Jr	Mas Bueno Pa	Robert Villanueva	1
Roland Villanueva Jr	Para Contigo	Robert Villanueva	1
Ruben Balagot	El pregunta	Society9	1
Ruben Balagot	Junto iyo con tigo	Los Mariachi	1
Ruben Balagot	Langga	Los Mariachi	1
Ruben Balagot	Neneng	Los Mariachi	1
Ruben Balagot	Señorita	Los Mariachi	1
Ruben Balagot	Tomada blues	Los Mariachi	1
Ruffy Gerard Enopia	Llega El Mañana	Nicole Lim	1
Weng Dela Peña	Levanta Zamboanga	Weng Dela Peña	1
Zack Quijano	Cuando	Comic Relief	8
Zack Quijano	El Mujer	Comic Relief	2
Jeloh Bangcal	El Mujer II	Comic Relief	1
Whey Guevara	Etu Lang	Comic Relief	1
Zack Quijano	Jimmy Loco Baila	Comic Relief	1
Zack Quijano	Maria Ozawa	Comic Relief	1
Zack Quijano	Mi Estrella	Comic Relief	4
Mark De Leon	Necesita	Comic Relief	3
Jeloh Bangcal & Jeffrey Buhian	Nuay Etu, Nuay Manyana	Comic Relief	5
Zack Quijano	Nuay Mas	Comic Relief	5
Zack Quijano	Nytlyf	Comic Relief	1
Zack Quijano	REAL!!!	Comic Relief	1
Zack Quijano	Sharon O. Burgos	Comic Relief	1
Zack Quijano	The Zamboanga Drama	Comic Relief	1
Zack Quijano	Topeng	Comic Relief	2
Zambo Tog Dogz	Amable	Zambo Top Dogz, Dogz go to heaven, Klass Wreckordz	1
Zambo Tog Dogz	Chismosa	Zambo Tog Dogz	1
Zambo Tog Dogz	Jalo ya lang	Zambo Top Dogz, Dogz go to heaven, Klass Wreckordz	1
Zambo Top Dogz	Manyakul	Zambo Tog Dogz	1

LETRISTA	TÍTULO	INTÉRPRETE(S)	V.
Zambo Tog Dogz	Pedasito de papel	Zambo Top Dogz	1
Zambo Tog Dogz	Snowman	Zambo Top Dogz	1
Zambo Tog Dogz	Ulan de Golpe	Zambo Top Dogz	1
Miller Lospendejos (aka Zigzag)	cuntigo lang	Zigzag & Ivan	1
Miller Lospendejos (aka Zigzag)	Donde el amor	Zigzag & Ivan	1
Miller Lospendejos (aka Zigzag)	El Deverasan Amor	Lil L	1
Miller Lospendejos (aka Zigzag)	El mi amor	Ivan, Vivian Ft. ZigZag	1

5.1.1.5 *Noticias*

En este género, incluimos noticias de dos fuentes: RMN Zamboanga y algunos artículos proporcionados por el periodista Felino M. Santos o publicados en sus perfiles de redes sociales. Aunque muchas otras fuentes fueron localizadas, en razón de la falta de respuesta por parte de los propietarios del contenido, estos contenidos no pudieron ser incorporados al corpus. Tampoco pudimos incluir, por las mismas razones, material recogido de columnas periodísticas.

5.1.1.6 *Religión*

Los materiales religiosos tienen como característica el hecho de utilizar grafías más cercanas a la del castellano, además de haber servido durante mucho tiempo como referencia de escritura para algunos hablantes, quizás por ser el único material escrito en CZ que tenían a su alcance. En el corpus, además de los textos disponibles en el sitio web de los Testigos de Jehová, se incorporó la totalidad de las dos traducciones del Nuevo Testamento:

- Wycliffe Bible Translators, Inc. (1981). El Nuevo Testamento en Chavacano. Wycliffe Bible Translators, Inc. Recuperado de https://www.scriptureearth.org/data/cbk/PDF/00-WNTcbk-web.pdf
- Rivas, C. de & Tardio M. (1982). El Buen Noticia: el Nuevo Testamento Chabacano. Ciudad de Zamboanga, Filipinas: Claretian Publications.

5.1.1.7 *Autoayuda*

De libros de autoayuda, solo hemos encontrado una única fuente, pero juzgamos que su contenido era suficientemente distinto a los demás como para no encajar en las categorías ya existentes.

- Hubbard, L. R. (2018). ALEGRIA: El Camino para Un Sentido Comon Guia para na Mas Bueno Vida. Recuperado de: http://thewaytohappinessphilippines.org/the-way-to-happiness-bicolano-cebuano-and-chavacano/

5.1.1.8 *Internet*

En este género, además de recoger datos anónimos de perfiles de Twitter y blogs, recogimos datos de los hilos zamboangueños de los foros «PinoyExchange.com», «PhilBoxing» y «Cockfighting Sabong», publicaciones del foro «Learn Computer Basic and Advance», del grupo de Facebook «Zamboanga de Antes» y de las páginas «Habla Zamboangueño» y «Pure Zamboangueños». En el caso de las fuentes de Facebook, tratamos de transcribir el texto de las imágenes y de los memes.

Aunque procuramos anonimizar las etiquetas de nombres de personas utilizando la cadena de caracteres «NAMECENSORED», somos conscientes de que es imposible garantizar el anonimato total de los datos. Seguramente hubiera sido posible reunir más fuentes, pero entre las opciones que teníamos disponibles, procuramos elegir las que aportaran la mayor diversidad posible considerando la edad de los participantes, su entorno y los temas tratados.

Dado el volumen de datos con el que trabajábamos, la tarea de revisar todos los contenidos manualmente requeriría demasiado tiempo, así que procuramos eliminar de manera automática las publicaciones que aparentemente no contenían palabras en CZ. Es probable, sin embargo, que parte de las publicaciones en CZ se haya perdido y algunas publicaciones que no poseen ningún contenido en CZ hayan permanecido en el corpus.

5.1.1.9 *Otros*

Este género hace las veces de cajón de sastre para el resto de contenidos del corpus. Incluye, además de textos seleccionados de Wikipedia, textos extraídos de carteles, propagandas, panfletos, certificados y cartas.

5.2 DESARROLLO DE LA TIPOLOGÍA DE ERRORES

Para desarrollar la tipología de errores de este trabajo, seguimos un proceso iterativo que constaba de los siguientes pasos:

1. Extracción de pequeñas muestras del corpus con fragmentos de texto de diferentes fuentes;
2. Corrección manual de las muestras;
3. Compilación de una lista de palabras de la muestra corregida en su forma original y forma corregida;
4. **Primera iteración:** Clasificación de los errores encontrados utilizando como punto de partida las tipologías de errores identificadas en el apartado 3.1.1.1.
5. **Siguientes iteraciones:** Clasificación de los nuevos errores encontrados a partir de la clasificación desarrollada en la primera iteración.
   - Definición de nuevos tipos de errores cuando no es posible clasificarlos bajo las categorías ya existentes.
6. Volver al paso 1.

Después de repetir el proceso anterior varias veces, procuramos agrupar las categorías por similitud y nombrar categorías más generales que abarcaran dos o más de las categorías iniciales. Durante el proceso, más reestructuraciones fueron necesarias con tal de mantener una coherencia interna entre las categorías o los criterios de división: algunas categorías fueron unificadas, y otras, subdivididas en otras categorías menores. En la figura 31 puede verse un diagrama con el flujo de los pasos descritos.

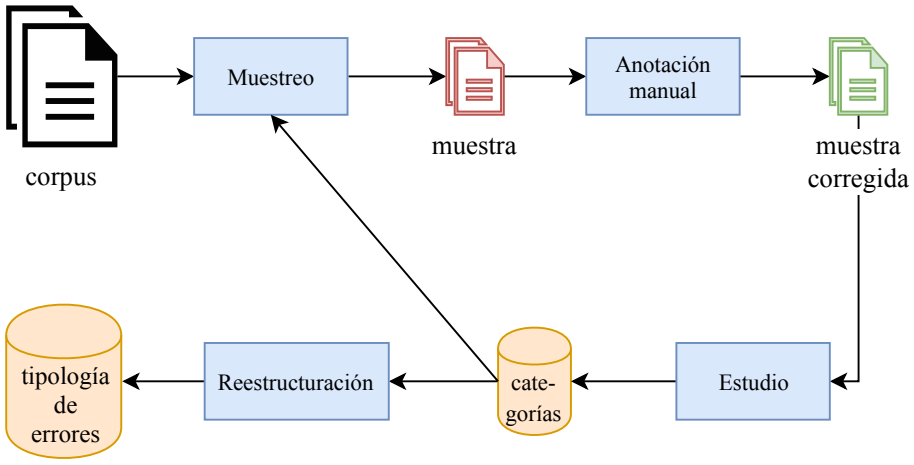


Figura 31: Tareas realizadas en el trabajo de investigación

La tipología de errores definitiva resultante se presenta y se discute en el apartado 6.1.

5.3 DESARROLLO DEL ALGORITMO FONÉTICO

Con el objetivo de acelerar el proceso de corrección de las palabras de los diccionarios (véase el apartado 5.4) y la generación de los datos de entrenamiento (véase el apartado 5.6), desarrollamos un algoritmo fonético para la corrección del CZ.

Para definir qué sustituciones serían pertinentes, estudiamos las diferencias entre grafemas y fonemas que existen en castellano y en tagalo. Posteriormente extendimos, a partir de las muestras utilizadas para crear la tipología de errores, algunas relaciones entre las dos lenguas y el inglés, además de agregar algunas variaciones frecuentes no normativas de grafemas del tagalo.

Sin que este algoritmo sea definitivo ni del todo exhaustivo, listamos a continuación las sustituciones realizadas. En su lectura, deberá tenerse en cuenta que se ha utilizado la notación siguiente:

- Los caracteres representados entre <> se conservan. Los notamos así para detallar un caso específico, en su contexto.
- <V> = a, e, i, o, u
- <SV> = y, w
- <V1> = a, o, u
- <V2> = e, i
- <C> = consonantes
- <C-> = consonantes, excepto y, w

▷ **Listado de sustituciones**

- **Eliminación de todas las tildes**
- **Letras dobles**
  - ff → f
  - pp → p
  - tt → t
  - dd → d
  - rr → r
  - mm → m

  - nn → n
  - cc<V1> → k

- **Dígrafos**
  - ph → p
  - rh → r
  - th → t
  - ck → k
  - mn → n
  - m<p> → n
  - m<b> → n
  - ct → kt
  - ksh/cc/ks/kt/sh/c/s + i/y/iy + <o> → si
  - ts/tch/ch/ts/t + <V> → ts
  - xc/sc + <V1> → sk
  - xc/sc + <V2> → s

- **Iniciales**
  - au-/aw-/ao-/u- → u
  - es-/is-/s- → s
  - an/en/in<C> → an
  - al-/ar- + <C-> → a
  - ex- + <V> → s

- **Finales**
  - -z/-s → -s
  - -t/-d → -t
  - -k/-g → -k
  - -ado/-ao/au/aw → aw
  - <V> + mg/ng → n

- **Consonantes**
  - v → b
  - qu/qw/quw/ku/kw/kuw/cu/cw/cuw + <V2>, q/k/c + <V1+SV> → k

  - z/c + <V2> → s
  - g/gh/gu/gw/guw/h/hu/hw/huw/ju/jw/juw + <V2>, j/dj/di/dy/diy + <V> → ∅
  - f → p
  - l/ll + i/y/iy + <V> → ly
  - ñ, n + i/y/iy/e + <V1> → ny
  - x/z + <V> → s
  - h → ∅

- **Vocales y diftongos**
  - ea/eia/eya/eiya/ia/iya → ya
  - ie/ye/iye → e
  - eo/eio/eyo/eiyo/io/iyo → yo
  - eu/eiu/eyu/eiyu → yu
  - ai/ae → ay
  - ei → ey
  - o/ow/u/uw + <a/e>, o/oh/ow/oy/u/uh/uw/uy/ou/ouh/ouw/ouy + <i>, u/uw + <o> → w
  - ou → ow
  - ao/au/aw → o
  - u → o
  - i → e

Algunos grafemas son especialmente problemáticos, ya que representan diferentes realizaciones en las tres lenguas en cuestión. En la tabla 7 listamos las realizaciones de los grafemas g, gu, gw, h, hu, hw, j, ju y dy. Las múltiples posibilidades de realización para un mismo fonema crean colisiones entre diferentes grafemas, lo que nos obliga a agruparlos todos bajo la misma sustitución. La consecuencia es que el algoritmo acaba generando el mismo código para palabras que se realizan de manera distinta.

Para un proceso de revisión manual, como es nuestro caso, los resultados son aceptables y el algoritmo tiene su utilidad, pero esta aproximación es poco recomendable para cualquier intento de corrección automática de palabras, pues podría encuentrarse reglas más específicas para tratar cada caso. Parte de las reglas resultantes tampoco se podría utilizar como reglas de sustitución de correctores ortográficos como *Aspell* o *hunspell* (véase los apartados 3.1.4.2 y 5.5).

Tabla 7: Realizaciones de los grafemas g, gu, gw, h, hu, hw, j, ju y dy en castellano (ES), tagalo (TL) e inglés (EN).

	e/i			a/o/u		
	ES	**TL**	**EN**	**ES**	**TL**	**EN**
g	x	g	dʒ/g	g	g	g
gu	g	-	g	gw	-	gw
gw	-	gw[1]	-	-	gw	-
h	muda	h	h	muda	h	h
hu	w/u	-	h/w	w	-	w
hw	-	hw[1]	-	-	hw	-
j	x	-	dʒ	x	-	dʒ
ju	xw	-[1]	dʒ	xw	-	dʒ
dy	-	dʒ	-	-	dʒ	-

5.4 CREACIÓN DE LA LISTA DE PALABRAS

Como uno de los requisitos para implementar un corrector ortográfico tradicional es disponer de una lista de palabras correctas, extraimos las palabras de CZ incluidas en los siguientes diccionarios/obras:

- Camins, B. S. (1999) Chabacano de Zamboanga Handbook and Chabacano-English-Spanish Dictionary (2da. ed.). Ciudad de Quezón, Filipinas: Claretian Publications.
- Miravite, R. M., Sanchez, U. C. N., Tardo, D. S., Viloria, S. J. B. & Delos Reyes, D. J. M. (2009). Chavacano Reader. Hyattsville, MD, USA: Dunwoody Press.
- Santos, R. A. (2010) Chavacano de Zamboanga: Compendio y Diccionario Chavacano-English/English-Chavacano. Ciudad de Zamboanga, Filipinas: Ateneo de Zamboanga University Press.
- Santos, F. M. (2011) Chavacano Handbook: El Español que se habla en Zamboanga, Usage and Dictionary. Ciudad de Zamboanga, Filipinas: Linus Multi-Media & Editorial Services.

[1] En tagalo existen las secuencias *guwa*, *guwe*... y *huwa*, *huwe*..., que se pronuncian como dos sílabas separadas, y a menudo se confunden con las secuencias unisilábicas. En CZ, además de secuencias como *jua*, *jue*..., también se pueden encontrar grafemas mixtos de castellano y tagalo: *jwa*, *jwe*...

- Maletsky, F. H. (2019) Chavacano - Chabacano: The original online Chavacano - English Dictionary. Recuperado de https://www.zamboanga.com/chavacano/chavacano_de_zamboanga_speak.htm

Para cada diccionario, fue utilizado el siguiente proceso iterativo:

1. Digitalización de las palabras del diccionario.
2. **Primera iteración:** Corrección manual de las palabras.
3. **Siguientes iteraciones:** Utilización del algoritmo fonético (apartado 5.3) con la lista de palabras de las iteraciones anteriores y revisión de los candidatos / corrección manual de las palabras.
4. Unificación de las listas de palabras.
5. Volver al paso 1.

La figura 32 ilustra el proceso descrito más arriba.

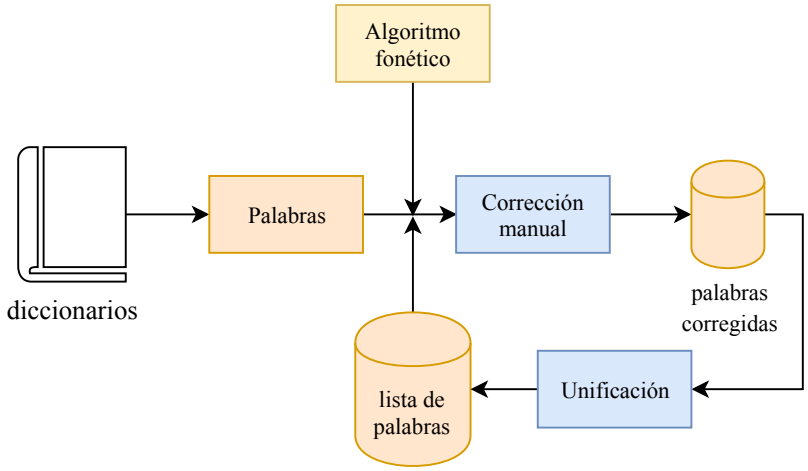


Figura 32: Flujo de actividades de la creación de la lista de palabras.

La lista final de palabras está compuesta 13.168 palabras. Por falta de tiempo, no fue posible incorporar las palabras de los diccionarios:

- Ariston, E. M. (2002) English-Chabacano Dictionary with TAGALOG and SPANISH equivalents together with a SIMPLE GRAMMAR and a CHABACANO-ENGLISH WORDLIST. Manila, Filipinas: Ms.
- Chambers, J. (2003). English-Chabacano Dictionary. Ciudad de Zamboanga, Filipinas: Ateneo de Zamboanga University Press.

5.5 IMPLEMENTACIÓN DE UN CORRECTOR ORTOGRÁFICO CON HUNSPELL

Utilizando la lista de palabras creada a partir de algunos diccionarios y obras disponibles (ver apartado 5.4), creamos un diccionario para el corrector ortográfico *hunspell*, que habría de servirnos de *baseline* en la evaluación de nuestra aproximación.

En el fichero de afijos, incluimos los sufijos y prefijos de mayor frecuencia que tenían comportamiento relativamente regular. Como prefijos se implementaron: aka--[2], ika- -[2], ma-, maka-, pagka-, ka-, pinaka-; y como sufijos: -han, las terminaciones verbales del infinitivo (-r que suele aparecer junto a preposiciones en el habla de los mayores), gerundio (-ando, -iendo) y participio (-ao, -ido) y de la sustantivación (-ada, -ida).

No se implementaron reglas de sustitución ni fonéticas por las razones ya mencionadas en el apartado 5.3.

5.6 DATOS DE ENTRENAMIENTO

Para crear los datos de entrenamiento, se obtuvieron 17 muestras del corpus CWZCC. Cada muestra está compuesta de aproximadamente 6.000 palabras y contienen sendas submuestras del 0,1 % de todas las fuentes, excepto de Twitter. Para garantizar el máximo de diversidad, en lugar de seleccionar textos completos, decidimos trabajar con una unidad mínima coherente. En los poemas y letras de música, esta es la estrofa; para los demás textos, es el párrafo. Al extraer la unidad mínima del texto, aunque se excediera el 0,1 % predeterminado, se include la unidad completa. Una vez obtenido el 0,1 % de todas las fuentes, se completaba la muestra con material de Twitter hasta alcanzar las 6.000 palabras. Esto, en media, dejaba aproximadamente 0,05 % de material total de Twitter dentro del corpus, o un 40 % del tamaño total de cada muestra de 6.000 palabras.

La figura 33 muestra el flujo de las diferentes actividades de la creación de los datos de entrenamiento.

[2] La doble - - indica un guion que obligatoriamente debe ser escrito después del prefijo.

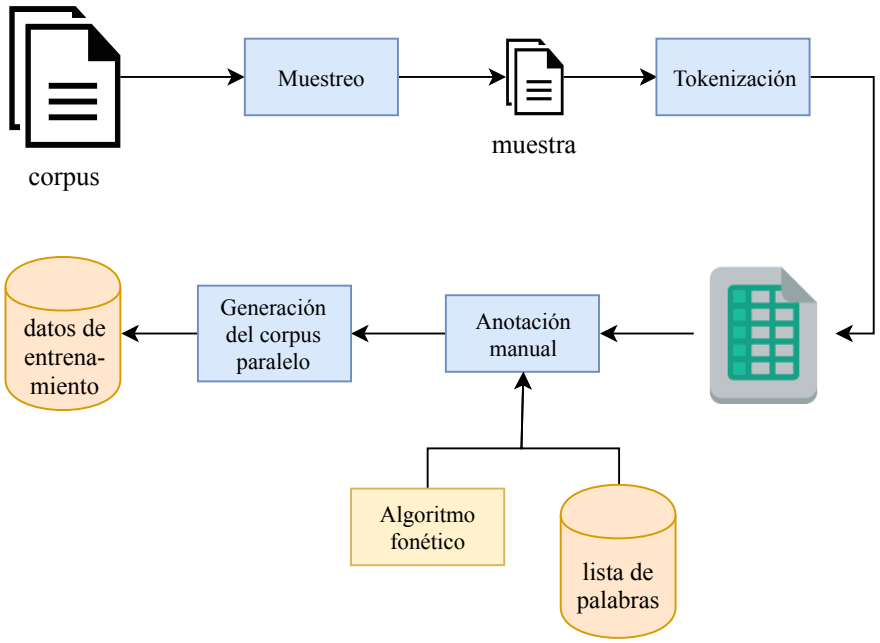


Figura 33: Flujo de actividades de la creación de los datos de entrenamiento.

5.6.1 *Tokenización*

Para tokenizar las muestras, utilizamos una versión modificada del tokenizador *TweetTokenizer* de la biblioteca NLTK. La modificación principal fue la implementación de la función `span_tokenize()`, que permite obtener el índice de los *tokens* en la cadena de caracteres original y no estaba implementada en nltk.tokenize.casual.

Otra mejora atañe a la expresión regular utilizada para segmentar emoticones, que no incluía muchos de los patrones existentes en nuestro corpus, haciendo que no se segmentaran de manera correcta. También agregamos una expresión regular para segmentar individualmente los *emoji*, además de hacer alteraciones en la forma en la que se separan palabras y números (para el caso de las reduplicaciones en CZ), palabras con guion y apóstrofe.

5.6.2 *Anotación*

Para la etapa de anotación, elegimos utilizar ficheros CSV. Esta decisión se tomó debido a la facilidad de manipulación de esos ficheros en cualquier editor de textos, editor de hojas de cálculo y lenguaje de programación *Python*. Cada línea del fichero representa un *token*, con los siguientes campos:

- **token:** la palabra en la forma en la que aparece en el corpus;
- **correction:** la palabra con la grafía corregida (en blanco si la forma es correcta);

- **errType:** tipo de error (en blanco si el campo `correction` está en blanco);
- **context:** concordancia de 30 caracteres como máximo antes y después de la palabra;
- **file:** fichero original del corpus de donde se extrajo el texto;
- **#:** identificador del texto;
- **p:** posición de la palabra dentro del texto;
- **T_correction:** lista de candidatos para su corrección sugerida por el algoritmo fonético y la lista de palabras.

La inclusión de un campo de concordancia permite verificar el contexto en el que se utiliza la palabra, sin que sea necesario consultar el corpus. El identificador del texto es un valor numérico único dentro de la muestra en cuestión. La posición de la palabra dentro del texto permite localizar la palabra dentro de la cadena de caracteres del texto.

En la anotación, decidimos trabajar con unigramas (n-gramas de tamaño 1, o sea, una única palabra) porque, de lo contrario, en la tabla de frases generada por *Moses* se perdería la distinción entre comienzo y fin de palabra. La limitación que esto impone es que las correcciones se hacen *token* a *token*, y por ende, no se pueden corregir errores que afecten a dos *tokens* en el origen.

Tomemos la expresión «quiere de sir» («quiere decir») como ejemplo: son tres *tokens*, el primero de los cuales es correcto, pero los otros dos presentan un problema de segmentación. En este caso, no es posible corregir la palabra «de sir» con «decir», ya que afecta a dos *tokens*. Por otro lado, casos como el de «taabla» («ta habla») en los que la forma incorrecta está compuesta por un único *token*, pero cuya forma correcta está compuesta por dos, sí pueden ser corregidos.

5.6.3 *Generación del corpus paralelo*

La generación de un corpus paralelo es, como ya explicado en el apartado 3.1.4.1.1, un paso esencial para el entrenamiento de un modelo de TA estadística. Para que el modelo tradujera caracteres en lugar de palabras, debimos separar las letras por espacios. Como se menciona en el apartado 5.6.2, en nuestros datos de entrenamiento trabajamos con unigramas en lugar de frases. Agregamos como delimitadores de principio y fin de cada palabra los caracteres # y $ respectivamente. Donde era necesario utilizar espacios, estos se sustituyeron por el carácter de espacios multi-byte, representado en codificación UTF-8 por U+3000. Puesto

que *Moses* utiliza XML internamente en algunos procesos, para evitar problemas, se convertieron algunos caracteres especiales a entidades HTML utilizando las herramientas de tokenización de NLTK. Igualmente, los *emoji* fueron convertidos en texto utilizando la biblioteca *emoji* de *Python*. Finalmente, cambiamos todos los caracteres a minúsculas. Esto se hace para evitar una dispersión en los datos a la hora de entrenar el modelo de TA estadística, ya que este distingue mayúsculas y minúsculas.

El corpus paralelo que sirvió de datos de entrenamiento para *Moses* está formado por dos ficheros: uno, con extensión .cbk, que contiene la forma original de las palabras, y otro, con extensión .cbk-zam, que contiene las mismas palabras corregidas. Se incluyen tanto las palabras corregidas como las que no necesitan de corrección, para que el modelo entrenado sea capaz tanto de generar las palabras correctas como de corregir las que son incorrectas. En la figura 34 mostramos un ejemplo mínimo de cada uno de los ficheros.

```
# a b l a $              # h a b l a $
# y a l a n g $          # y a   l a n g $
# e r a $                # e r a $
# e l $                  # e l $
# d e b e r a s a n $    # d e v e r a s a n $
```

Figura 34: Ejemplo de ficheros .cbk y cbk-zam. (Verso extraído de: Pondong (2009). Promesa.)

5.7 MODELO DE TA ESTADÍSTICA DE CARACTERES

Utilizando el corpus paralelo creado a partir de las muestras anotadas (véase el apartado 5.6), procedimos al entrenamiento de un modelo de TA estadística. Utilizamos SRILM para crear el modelo lingüístico de cuadrigramos (n=4, determinado empíricamente para nuestro corpus) basado en la versión corregida de los datos. Luego, utilizando GIZA++, alineamos lexicalmente el corpus (recordando que, en lugar de palabras, estamos trabajamos con caracteres, y los alineamientos pueden ser tanto simétricos como asimétricos (es decir, 1:1, 1:n, n:1 o m:n, con m, n $\in \mathbb{N}$). La figura 35 ilustra ambos casos.

Figura 35: Ejemplo de alineamiento de caracteres de dos palabras. (Título de una de las canciones del grupo Cheeze de Sal (2010).)

Una vez creado el modelo de traducción, es posible ejecutar *Moses* pasando como parámetro la ruta de acceso hacia el modelo, o ejecutando `mosesserver`, una implementación como servidor que acepta consultas remotas por medio de XMLRPC.

5.7.1 *Extracción de reglas*

Por falta de tiempo, en este trabajo no hemos podido avanzar en el análisis de la tabla de frases generadas por *Moses*. Es posible podar la tabla de frases aplicando criterios estadísticos para eliminar las parejas redundantes y/o espúreas causadas por datos con ruido. A partir de ahí, creemos que sería posible extraer de ella reglas fonéticas más exactas que las que utilizamos en la implementación del algoritmo fonético (véase el apartado 5.3).

Si en realidad se deseara incorporar esas reglas a un corrector ortográfico actual como *hunspell*, puede que fuera más interesante eliminar o corregir los errores aleatorios de los datos de entrenamiento, pero habría que realizar más experimentos para determinar si este paso sería necesario. Para tareas de depuración, llegamos a implementar modificaciones en *mosesserver* que permiten verificar, para cada solicitud de traducción enviada a *Moses*, los alineamientos realizados, y que podrán ser de utilidad en trabajos futuros.

CAPÍTULO 6

EVALUACIÓN DE RESULTADOS

En este apartado presentamos la tipología de errores ortográficos (apartado 6.1), las estadísticas de los datos de entrenamiento (apartado 6.1.1), las dos versiones del corpus CWZCC (apartado 6.2) y los resultados del corrector ortográfico con *hunspell* y de la aproximación propuesta en este trabajo (apartado 6.3).

6.1 LA TIPOLOGÍA DE ERRORES ORTOGRÁFICOS

La tipología de errores ortográficos desarrollada a lo largo de este trabajo de investigación es resultado del proceso descrito en el apartado 5.2. Como se puede ver en el diagrama de la figura 36, se trata de una estructura jerárquica en forma de árbol no balanceado. La complejidad de la clasificación crece a medida que se baja en el árbol, y disminuye subiendo hacia el nodo raíz. Las líneas horizontales punteadas especifican los criterios de partición aplicados en cada rama. Los errores ortográficos, por lo tanto, se clasifican con los tipos de errores en los nodos hoja y van asociados a una etiqueta compuesta de 3 caracteres a lo sumo. Una ventaja de esa estructura es que, en caso de ser necesaria una clasificación menos detallada, es posible subir niveles en el árbol y adoptar los respectivos nodos padre para clasificar algunos tipos de errores.

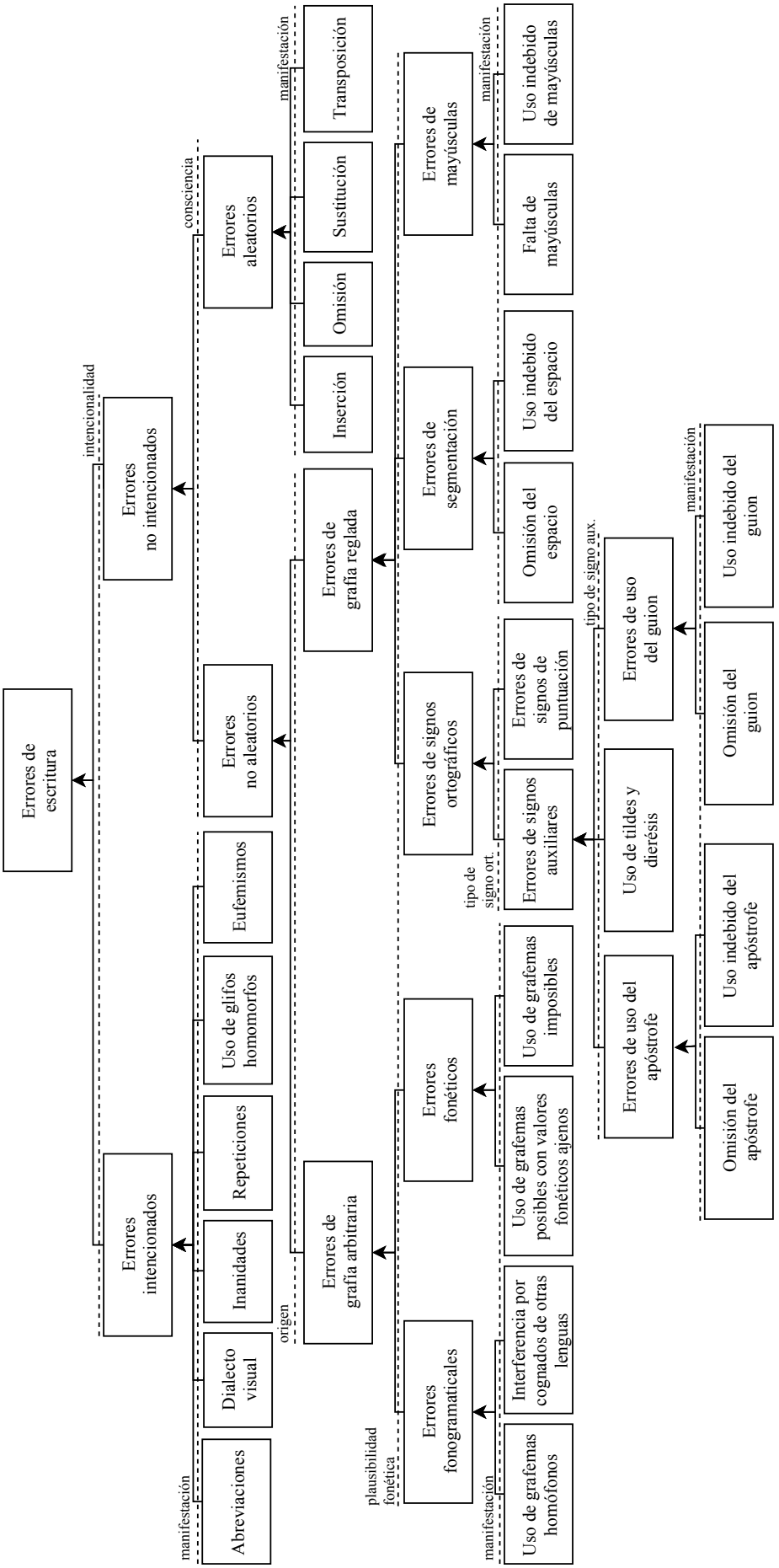


Figura 36: Tipología de errores ortográficos del CZ

A continuación, describimos cada uno de los tipos de errores de acuerdo a la estructura de la figura 36. La tabla 8 incluye ejemplos apropiados para cada uno de ellos.

Errores de escritura

1. **Errores intencionados:** Ocurren de manera consciente por parte del hablante en la forma de recursos expresivos o con el objetivo de acortar mensajes, pero su uso no se admite en el lenguaje formal.
   1.1. **Abreviaciones (ABR):** Omisión de letras o uso de letras y/o números para reemplazar sílabas.
   1.2. **Dialecto visual (ED):** Representación escrita de una realización fonética específica.
   1.3. **Inanidades (INN):** Combinación de varias palabras o transformación intencionada de palabras sin ninguna explicación aparente.
   1.4. **Repeticiones (REP):** Recurso expresivo utilizado para expresar emociones como sorpresa o enfado.
   1.5. **Uso de glifos homomorfos (HMM):** Sustitución de un carácter por otro de forma semejante.
   1.6. **Eufemismos (EPH):** Cambio intencional de una palabra para evitar palabras malsonantes o blasfemia. Se clasifican como tal cuando la sustitución implica de cambios menores y/o se genera una palabra inexistente en la lengua.
2. **Errores no intencionados:** Ocurren por descuido o por falta de conocimiento del hablante.
   2.1. **Errores no aleatorios:** Producidos con gran probabilidad por una causa ajena al azar.
      2.1.1. **Errores de grafía arbitraria:** Errores de aspectos ortográficos sin reglas prácticas que permitan deducir la escritura correcta.
         2.1.1.1. **Errores fonogramaticales:** Uso de grafemas que representan una pronunciación correcta de la palabra, pero en desacuerdo con la escritura correcta establecida.
            2.1.1.1.1. **Interferencia por cognados de otras lenguas (COG):** Aplicación parcial o total de la grafía de un cognado procedente de una de las lenguas que están en contacto con el CZ.
            2.1.1.1.2. **Uso de grafemas homófonos (HOM):** Aplicación de un grafema cuya realización es la que se desea representar, pero en desacuerdo con la grafía correcta establecida.

2.1.1.2. **Errores fonéticos:** Uso de grafemas inválidos o que no admiten el valor fonético que se desea representar.

2.1.1.2.1. **Uso de grafemas posibles con valores fonéticos ajenos (XPG):** Uso de un grafema que no permite la realización que se desea representar.

2.1.1.2.2. **Uso de grafemas imposibles (XIG):** Uso de un grafema no aceptado en CZ.

2.1.2. **Errores de grafía reglada:** Errores de aspectos ortográficos regidos por reglas.

2.1.2.1. **Errores de signos ortográficos:** Errores de uso de signos auxiliares y de puntuación.

2.1.2.1.1. **Errores de signos auxiliares:** Errores de uso del apóstrofe, tildes y diéresis o del guion.

2.1.2.1.1.1. **Errores de uso del apóstrofe:** Violación de las reglas de uso del apóstrofe.

2.1.2.1.1.1.1. **Omisión del apóstrofe (OA):** Falta del apóstrofe en una palabra que debería llevarlo.

2.1.2.1.1.1.2. **Uso indebido del apóstrofe (XA):** Uso del apóstrofe en una palabra o entre dos palabras que no deberían llevarlo.

2.1.2.1.1.2. **Uso de tildes y diéresis (XD):** Uso de tildes o diéresis por influencia del castellano o del tagalo.

2.1.2.1.1.3. **Errores de uso del guion:** Violación de las reglas de uso de guion.

2.1.2.1.1.3.1. **Omisión del guión (OH):** Falta del guion en una palabra que debería llevarlo.

2.1.2.1.1.3.2. **Uso indebido del guión (XH):** Uso del guion en una palabra o entre dos palabras que no deberían llevarlo.

2.1.2.1.2. **Errores de signos de puntuación (XP):** Uso de los signos de puntuación invertidos del castellano. No serán clasificados otros casos de falta o uso indebido de signos de puntuación.

2.1.2.2. **Errores de segmentación:** Violación de las reglas de uso de espacios en blanco.

2.1.2.2.1. **Omisión del espacio (OS):** Falta de un espacio entre dos palabras que no deberían ir unidas.

2.1.2.2.2. **Uso indebido del espacio (XS):** Uso del espacio en una palabra que no debería escribirse segmentada.

2.1.2.3. **Errores de mayúsculas:** Violación de las reglas de uso de mayúsculas.

2.1.2.3.1. **Omisión de mayúsculas (OC):** Falta de mayúsculas en una palabra que debería empezar por mayúscula.

2.1.2.3.2. **Uso indebido de mayúsculas (XC):** Uso de mayúsculas en una palabra que no debería empezar por mayúscula.

2.2. **Errores aleatorios:** Errores que ocurren por descuido a la hora de teclear.

2.2.1. **Inserción (INS):** Adición de una letra en una palabra.

2.2.2. **Omisión (OMS):** Eliminación de una letra en una palabra.

2.2.3. **Sustitución (SUB):** Cambio de una letra en una palabra.

2.2.4. **Transposición (TRS):** Cambio de posición entre dos letras en una palabra.

Cabe comentar algunos detalles al respecto de la tipología:

- En la mayoría de los casos, la clasificación es múltiple, ya que en una palabra escrita incorrectamente en CZ suele ocurrir más de un tipo de error.
- Dentro de «Errores aleatorios», a su vez, se encuentran errores cuya causa aparenta ser un simple descuido a la hora de teclear, como una letra que sobra, que falta, que reemplaza o que cambia de posición con otra.
- Dentro de «Errores intencionados» agrupamos diferentes tipos de textismos (véase el apartado 3.1.2) encontrados en CZ. Aunque no son errores ortográficos propiamente dichos, muchas veces acaban apareciendo en escritos formales de algunas personas, por descuido o por hábito. De la clasificación de textismos de Craig (2003), adoptamos la denominación «Inanidades» para clasificar algunos casos problemáticos.
- La distinción entre «Interferencia por cognados de otras lenguas» y «Uso de grafemas homófonos» es bastante sutil y se hace por eliminación: si el primer tipo de error no es aplicable (no existe cognado en las lenguas locales o en inglés que pueda haber influido directamente en la forma utilizada), se aplica el segundo.
- Los conceptos de «grafía reglada» y de «grafía arbitraria» son prestados de la tipología de Galí (citado en Cristóbal Rojo, 1982).

- La nomenclatura de «signos ortográficos», «signos de puntuación» y «signos auxiliares» está tomada de las definiciones de la Real Academia Española (RAE) en su Diccionario panhispánico de dudas:

  > SIGNOS ORTOGRÁFICOS. Son todas aquellas marcas gráficas que, no siendo números ni letras, aparecen en los textos escritos con el fin de contribuir a su correcta lectura e interpretación. Cada uno de ellos tiene una función propia y unos usos establecidos por convención. Hay signos de puntuación y signos auxiliares.
  >
  > a) Signos de puntuación. Sus funciones son marcar las pausas y la entonación con que deben leerse los enunciados, organizar el discurso y sus diferentes elementos para facilitar su comprensión, evitar posibles ambigüedades en textos que, sin su empleo, podrían tener interpretaciones diferentes, y señalar el carácter especial de determinados fragmentos de texto —citas, incisos, intervenciones de distintos interlocutores en un diálogo, etc.—. La información relativa al uso específico de cada signo se ofrece en su entrada correspondiente (→ coma; comillas; corchete; dos puntos; interrogación y exclamación (signos de); paréntesis; punto; puntos suspensivos; punto y coma; raya).
  >
  > b) Signos auxiliares. Sus funciones son muy variadas y se explican en las entradas correspondientes a cada uno de ellos (→ apóstrofo; asterisco; barra; diéresis; guion o guión; llave; párrafo; tilde).
  >
  > (Real Academia Española y Asociación de Academias de la Lengua Española, 2005, p. 605)

- La tipología de Sánchez-Jiménez (2010) también influenció enormemente a la nuestra, sobre todo en la esencia de los errores de las ramas de «Errores de grafía arbitraria», aunque en muchos casos utilicemos una nomenclatura un poco diferente.

En la tabla 8 se incluyen los tipos utilizados efectivamente para clasificar los errores ortográficos (nodos hoja de la figura 36) acompañados de algunos ejemplos ilustrativos. Dado que, como se ha mencionado, en una palabra suelen ocurrir diferentes tipos de errores, la parte que interesa en cada ejemplo va subrayada.

Tabla 8: Tipos de errores ortográficos con ejemplos.

Cód.	Nombre	Ejemplos
ABR	**Abreviaciones**	*kme → kame *solo2 → solo-solo *d2u → de tuyo *cge → segui
ED	**Dialecto visual**	*nema → no hay mas *pehcaw → pescao
INN	**Inanidades**	*kers → quiere dormz → dormi

Cód.	Nombre	Ejemplos
REP	**Repeticiones**	*porkeee → por que
HMM	**Uso de glifos homomorfos**	*k0sa → cosa
EPH	**Eufemismos**	*pota → puta kunyununani → coño vos nana
COG	**Interferencia por cognados de otras lenguas**	*attende → atende *technico → tecnico
HOM	**Uso de grafemas homófonos**	*sapatos → zapatos *talya → talla
XPG	**Uso de grafemas posibles con valores fonéticos ajenos**	*kunumon → kanamon
XIG	**Uso de grafemas imposibles**	*qiere → quiere *itsura → hechura
OA	**Omisión del apóstrofe**	*tan → ta'n *unoy otro → uno'y otro
XA	**Uso indebido del apóstrofe**	*sesenta'y nueve → sesenta y nueve
OH	**Omisión del guion**	*kosa kosa → cosa-cosa
XH	**Uso indebido del guion**	*alas-8 → a las 8 *man-viaje → man viaje
XD	**Uso de tildes y diéresis**	*mío → mio *olê → ole galè → gale *vergüenza → verguenza
XP	**Errores de signos de puntuación**	*¿cosa? → cosa? *¡gracias! → gracias!
OS	**Omisión del espacio**	*manmirahan → man mirahan *kunambre → con hambre *yalang → ya lang
XS	**Uso indebido del espacio**	*o hala → ojala
OC	**Omisión de mayúsculas**	*filipino → Filipino *pedro → Pedro

Cód.	Nombre	Ejemplos
XC	**Uso indebido de mayúsculas**	Onde *Ustedes ta queda? → Onde ustedes ta queda?
INS	**Inserción**	*karsa → casa
OMS	**Omisión**	*Chaacano → Chavacano
SUB	**Sustitución**	*cpmpra → compra
TRS	**Transposición**	*beuno → bueno

6.1.1 *Estadísticas de los datos de entrenamiento*

En este apartado, analizaremos la distribución de los tipos de errores dentro de los datos de entrenamiento.

En los datos de entrenamiento, encontramos una media de aproximadamente un 20 % de las palabras con algún tipo de error. Como podemos observar en el gráfico de la figura 37, hay una predominancia de errores no intencionados dentro de las muestras extraídas del corpus.

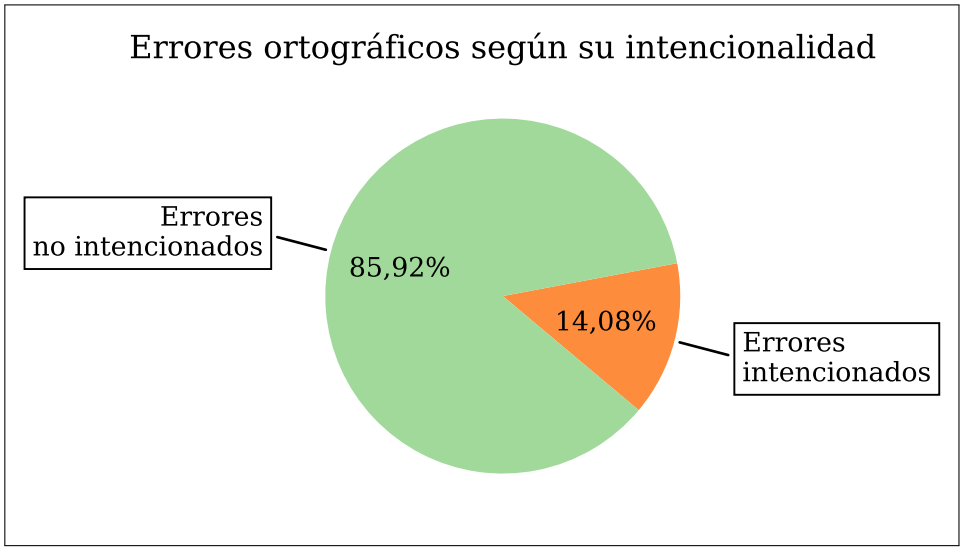


Figura 37: Distribución de errores ortográficos intencionados y no intencionados.

Como puede verse en el gráfico de la figura 38, entre los errores no intencionados, la mayoría de los errores son no aleatorios. Es decir, hay una posible explicación para su ocurrencia que no es la casualidad.

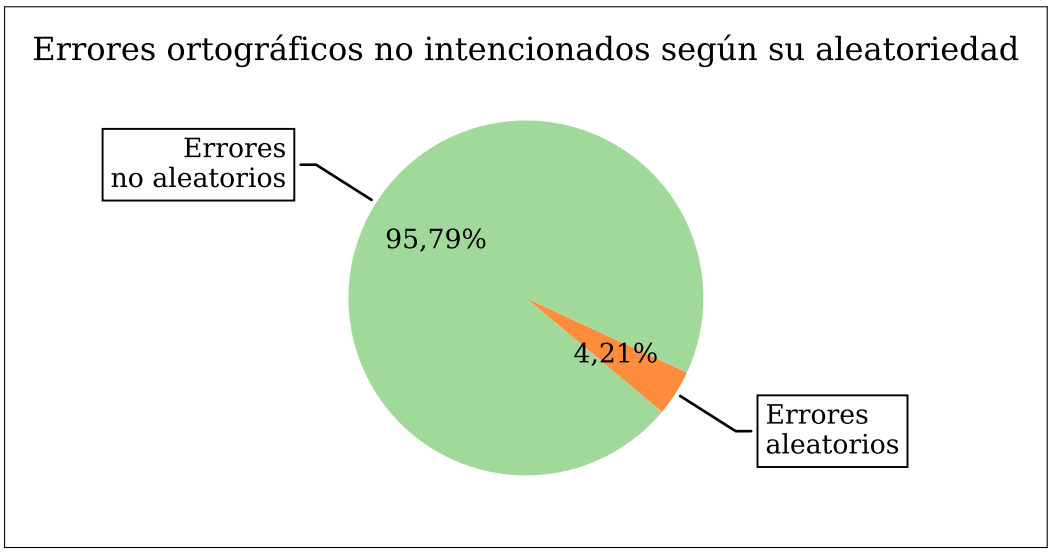


Figura 38: Distribución de errores ortográficos no intencionados, aletorios y no aleatorios.

Finalmente, el gráfico de la figura 39 muestra que, aunque los errores de grafía arbitraria sean mayoritarios, los de grafía reglada representan también una parte importante del total de errores.

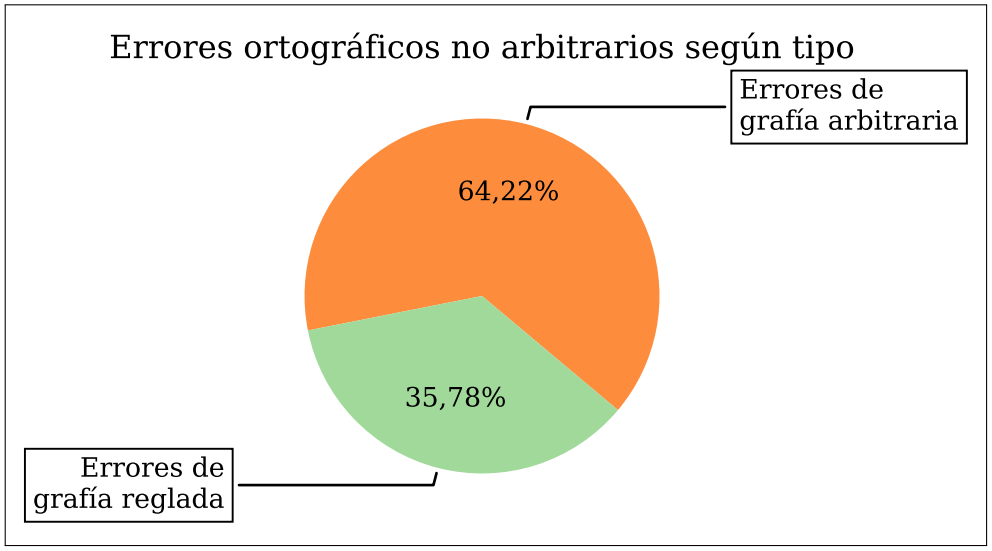


Figura 39: Distribución de errores ortográficos no aleatorios de grafía arbitraria y reglada.

Con esta visión general, podemos interpretar más fácilmente las estadísticas de la tabla 9. Como se observa, la inmensa mayoría de los errores que ocurren en nuestros datos son de tipo «Sustitución por grafemas homófonos», es decir, un tipo de error de grafía arbitraria. Esto es esperado, dada la gran cantidad de grafemas posibles que el hablante de CZ tiene a su disposición para representar los sonidos de la lengua. Tras ellos, de manera más fragmentada, aparecen diversos errores de grafía reglada, lo que nos indica que, una vez interiorizadas las reglas de la nueva grafía, una parte sustancial de los errores ortográficos debería dejar de

producirse. Los errores de «Dialecto visual» también son bastante numerosos, lo cual resalta la distancia entre la forma escrita y la forma oral de algunas palabras y expresiones. Finalmente, aparecen las abreviaciones, error de tipo intencionado y bastante frecuente en contextos informales, que se explica por el hecho de que gran parte de los datos de las muestras provienen de redes sociales.

Tabla 9: Número de ocurrencias de cada tipo de error en los datos de entrenamiento.

Cód.	Ocurrencias
HOM	13.605
OA	2.933
OS	2.623
ED	1.909
ABR	1.610
OC	1.351
COG	1.088
XC	588
OMS	450
REP	389
XD	347
INS	307
XIG	234
XS	223
SUB	161
OH	143
TRS	104
XA	54
EPH	39
XH	38
INN	27
XP	21
XPG	10
HMM	4

6.2 EL CWZCC

En este apartado, presentamos algunos ejemplos del corpus convertido y anotado en los dos formatos: NIF y TEI-XML.

6.2.1 *La versión NIF*

El corpus en formato NIF está dividido en 9 ficheros de extensión .ttl, cada uno correspondiente a uno de los géneros del corpus.

En las primeras líneas, se incluyen las declaraciones de los espacios de nombres, que definen las propiedades que serán utilizadas en el fichero RDF. En el espacio de nombres *cwzcc* se encuentra definida una ontología que contiene los tipos de errores definidos por la tipología.

```
@prefix cwzcc: <http://research.chavacano.org/cwzcc.owl#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>
    .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

El bloque siguiente representa los documentos existentes dentro del fichero. Cada documento va identificado unívocamente por un URI.

```
<http://research.chavacano.org/cwzcc> a nif:ContextCollection ;
   nif:hasContext <http://research.chavacano.org/cwzcc/news/1-1-bilyones-de-
       pesos-budget-na-security-ya-hace-bandera-el-alcalde-climaco>,
      <http://research.chavacano.org/cwzcc/news/1-2-kilo-shabu-confiscao>,
      ...
   dcterms:conformsTo <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/
       nif-core/2.1>
```

He aquí un ejemplo de texto del corpus, extraído de Twitter:

```
<http://research.chavacano.org/cwzcc/social/twitter/14> a nif:Context,
     nif:OffsetBasedString ;
```

```
   nif:beginIndex "0"^^xsd:nonNegativeInteger ;
   nif:endIndex "140"^^xsd:nonNegativeInteger ;
   nif:isString "pirmi man iyo ta sinti dwele.. nusabe yo porke pirmi ansina..
        kwando iyo keda alegre? kwando gaha pasa el diya o mes na hinde iyo
       triste? =(" ;
   dcterms:date "2009-08-27 00:30"^^xsd:string .
```

Luego, siguen las correcciones con la respectiva etiqueta de tipo de error. Para etiquetar los errores, utilizamos la propiedad `classAnotation` de *nif*. La propiedad `correction` y los tipos de errores de la tipología están definidos en la ontología de tipos de errores, en el espacio de nombres *cwzcc*.

```
<http://research.chavacano.org/cwzcc/social/twitter/14#offset_0_5> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "pirmi" ;
   nif:beginIndex "0"^^xsd:nonNegativeInteger ;
   nif:endIndex "5"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:COG ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
       > ;
   cwzcc:correction "firme"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_10_13> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "iyo" ;
   nif:beginIndex "10"^^xsd:nonNegativeInteger ;
   nif:endIndex "13"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
       > ;
   cwzcc:correction "yo"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_17_22> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "sinti" ;
   nif:beginIndex "17"^^xsd:nonNegativeInteger ;
   nif:endIndex "22"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
```

```
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "senti"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_23_28> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "dwele" ;
   nif:beginIndex "23"^^xsd:nonNegativeInteger ;
   nif:endIndex "28"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "duele"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_31_37> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "nusabe" ;
   nif:beginIndex "31"^^xsd:nonNegativeInteger ;
   nif:endIndex "37"^^xsd:nonNegativeInteger ;
   nif:classAnotation [cwzcc:HOM, cwzcc:OS] ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "no sabe"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_41_46> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "porke" ;
   nif:beginIndex "41"^^xsd:nonNegativeInteger ;
   nif:endIndex "46"^^xsd:nonNegativeInteger ;
   nif:classAnotation [cwzcc:HOM, cwzcc:OS] ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "por que"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_47_52> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "pirmi" ;
```

```
   nif:beginIndex "47"^^xsd:nonNegativeInteger ;
   nif:endIndex "52"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:COG ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "firme"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_62_68> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "kwando" ;
   nif:beginIndex "62"^^xsd:nonNegativeInteger ;
   nif:endIndex "68"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "cuando"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_69_72> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "iyo" ;
   nif:beginIndex "69"^^xsd:nonNegativeInteger ;
   nif:endIndex "72"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "yo"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_73_77> a nif:
    OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "keda" ;
   nif:beginIndex "73"^^xsd:nonNegativeInteger ;
   nif:endIndex "77"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "queda"^^xsd:string .
```

```
<http://research.chavacano.org/cwzcc/social/twitter/14#offset_86_92> a nif:
    OffsetBasedString,
       nif:Phrase ;
   nif:anchorOf "kwando" ;
   nif:beginIndex "86"^^xsd:nonNegativeInteger ;
   nif:endIndex "92"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
       > ;
   cwzcc:correction "cuando"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_106_110> a nif:
    OffsetBasedString,
       nif:Phrase ;
   nif:anchorOf "diya" ;
   nif:beginIndex "106"^^xsd:nonNegativeInteger ;
   nif:endIndex "110"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
       > ;
   cwzcc:correction "dia"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_120_125> a nif:
    OffsetBasedString,
       nif:Phrase ;
   nif:anchorOf "hinde" ;
   nif:beginIndex "120"^^xsd:nonNegativeInteger ;
   nif:endIndex "125"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
       > ;
   cwzcc:correction "hende"^^xsd:string .

<http://research.chavacano.org/cwzcc/social/twitter/14#offset_126_129> a nif:
    OffsetBasedString,
       nif:Phrase ;
   nif:anchorOf "iyo" ;
   nif:beginIndex "126"^^xsd:nonNegativeInteger ;
   nif:endIndex "129"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:HOM ;
```

```
   nif:referenceContext <http://research.chavacano.org/cwzcc/social/twitter/14
      > ;
   cwzcc:correction "yo"^^xsd:string .
```

Cuando el modelo de TA estadística corrige un error todavía no anotado, se aplica la etiqueta «UND». A continuación se muestra un ejemplo de corrección que, en realidad, no es adecuada.

```
<http://research.chavacano.org/cwzcc/news/alcalde-ya-manda-chekia-estado-de-
    arroz-na-pueblo#offset_196_202> a nif:OffsetBasedString,
      nif:Phrase ;
   nif:anchorOf "chekia" ;
   nif:beginIndex "196"^^xsd:nonNegativeInteger ;
   nif:endIndex "202"^^xsd:nonNegativeInteger ;
   nif:classAnotation cwzcc:UND ;
   nif:referenceContext <http://research.chavacano.org/cwzcc/news/alcalde-ya-
      manda-chekia-estado-de-arroz-na-pueblo> ;
   cwzcc:correction "checia"^^xsd:string .
```

6.2.2 *La versión TEI-XML*

El corpus en formato TEI-XML está dividido en múltiples ficheros con extensión .xml, cada uno correspondiente a un documento. Por cuestiones de simplicidad, se ha utilizada la estructura mínima de documento TEI-XML.

Las palabras corregidas llevan la etiqueta `error` con el atributo `@type`, que puede ser multivaluado, y el atributo `@correction`, que contiene la corrección.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>twitter14</title>
        <author />
      </titleStmt>
      <editionStmt />
      <publicationStmt>
```

```
        <p>Published as a part of the Contemporary Written Zamboangue&ntilde;
            o Chabacano Corpus (CWZCC) TEI-XML Version.</p>
        <date when="2009-08-27 00:30"/>
      </publicationStmt>
    </fileDesc>
    <revisionDesc>
      <listChange>
        <change>
          <name />
          <date />
        </change>
      </listChange>
    </revisionDesc>
  </teiHeader>
  <text>
    <body>
      <p><error type="COG" correction="firme">pirmi</error> man <error type="
          HOM" correction="yo">iyo</error> ta <error type="HOM" correction="
          senti">sinti</error> <error type="HOM" correction="duele">dwele</
          error>.. <error type="HOM,OS" correction="no sabe">nusabe</error> yo
           <error type="HOM,OS" correction="por que">porke</error> <error type
          ="COG" correction="firme">pirmi</error> ansina.. <error type="HOM"
          correction="cuando">kwando</error> <error type="HOM" correction="yo"
          >iyo</error> <error type="HOM" correction="queda">keda</error>
          alegre? <error type="HOM" correction="cuando">kwando</error> gaha
          pasa el <error type="HOM" correction="dia">diya</error> o mes na <
          error type="HOM" correction="hende">hinde</error> <error type="HOM"
          correction="yo">iyo</error> triste? =(</p>
    </body>
  </text>
</TEI>
```

6.3 EVALUACIÓN DE LOS CORRECTORES ORTOGRÁFICOS

En este apartado se analiza el desempeño del corrector ortográfico con *hunspell* y de nuestra aproximación utilizando TA estadística de caracteres.

6.3.1 *Métricas utilizadas*

Para evaluar el desempeño de los correctores ortográficos, utilizamos 4 métricas: exactitud, precisión, exhaustividad y medida-F.

En el caso de los correctores ortográficos, tenemos 4 tipos de predicciones:

- **Verdadero Positivo (VP):** el corrector detectó la palabra como *correcta* y la palabra está *correcta*.
- **Falso Positivo (FP):** el corrector detectó la palabra como *correcta* y la palabra está *incorrecta*.
- **Verdadero Negativo (VN):** el corrector detectó la palabra como *incorrecta* y la palabra está *incorrecta*.
- **Falso Negativo (FN):** el corrector detectó la palabra como *incorrecta* y la palabra está *correcta*.

La **exactitud** es dada por la razón entre las prediciones correctas (VP + VN) sobre el total de observaciones (VP + FP + VN + FN). Esta métrica indica el índice de aciertos del corrector en relación a la salida esperada.

$$\text{Exactitud} = \frac{\text{VP}+\text{VN}}{\text{VP}+\text{FP}+\text{VN}+\text{FN}}$$

La **precisión** es dada por la razón entre las predicciones positivas correctas (VP) sobre el total de predicciones positivas (VP+FP). Esta métrica indica, de todas las palabras que el corrector ortográfico detectó como correctas, cuántas realmente lo eran.

$$\text{Precisión} = \frac{\text{VP}}{\text{VP}+\text{FP}}$$

La **exhaustividad** es dada por la razón entre las predicciones positivas correctas (VP) sobre el total de observaciones positivas (VP + FN). Esta métrica indica, de todas las palabras realmente correctas, cuántas el corrector ortográfico detectó como correctas.

$$\text{Exhaustividad} = \frac{\text{VP}}{\text{VP}+\text{FN}}$$

La **medida-F** es una media armónica de la **precisión** y de la **exhaustividad**.

$$\text{Medida-F} = 2 \cdot \frac{\text{Precisión} \cdot \text{Exhaustividad}}{\text{Precisión}+\text{Exhaustividad}}$$

6.3.2 *hunspell (baseline)*

Esta implementación se realizó con el fin de comparar sus resultados con los de la aproximación propuesta en este trabajo. Para los datos de prueba, utilizamos las 17 muestras extraídas del CWZCC al completo, lo que sumaba un total de aproximadamente 102.000 palabras, sin contar los signos de puntuación. Para la evaluación, se tuvieron en consideración los tres primeros candidatos de corrección ofrecidos por *hunspell* y se descartaron los demás. En la tabla 10 pueden observarse los valores de las métricas de exactitud, precisión, exhaustividad y medida-F para uno, dos y tres candidatos. El valor de la exhaustividad permanece casi inalterado, variando más allá de los dos lugares decimales.

Tabla 10: Valores de exactitud, precisión, exhaustividad y medida-F para el corrector ortográfico con hunspell (baseline).

	top 1	**top 2**	**top 3**
Exactitud	59,67 %	61,62 %	63,84 %
Precisión	77,94 %	80,16 %	82,84 %
Exhaustividad	68,9 %	68,9 %	68,9 %
medida-F	73,14 %	74,1 %	75,23 %

6.3.3 *TA estadística de caracteres*

Para la evaluación del modelo de TA estadística de caracteres con *Moses*, definimos cuatro conjuntos de datos: el primero, compuesto por 2 muestras (aproximadamente 12.000 palabras); el segundo, con 4 muestras (aproximadamente 24 mil palabras); el tercero compuesto por 8 muestras (aproximadamente 48.000 palabras); y por fin el cuarto, con 17 muestras (aproximadamente 102.000 palabras).

Para la evaluación, utilizamos la **validación cruzada** *k-fold*, con *k* igual a 10. Esta técnica de evaluación consiste en dividir los datos en *k* particiones (en este caso, 10) y utilizar cada una de ellas para validar el modelo: para cada partición, se utilizan las demás *k-1* particiones para entrenar el modelo, y la partición en cuestión para la evaluación. Luego, una vez calculadas la exactitud, la precisión, la exhaustividad y la medida-F para todas las particiones, se calculan sus medias aritméticas y se obtienen así los valores promedios para cada una de las métricas.

En este caso, también consideramos únicamente los tres primeros candidatos de corrección proporcionados por el modelo y calculamos las métricas para uno, dos y tres candidatos. Las tablas 11, 12 y 13 muestran los resultados obtenidos para los cuatro conjuntos de datos, considerando uno, dos y tres candidatos respectivamente.

Tabla 11: Valores de exactitud, precisión, exhaustividad y medida-F para el corrector ortográfico con TA estadística de caracteres para cada uno de los conjuntos de datos, considerando un candidato.

top 1	12k	24k	48k	102k
Exactitud	84,14 %	85,22 %	87,01 %	88,26 %
Precisión	91,55 %	92,27 %	92,69 %	93,01 %
Exhaustividad	87,21 %	89,08 %	90,86 %	92,22 %
medida-F	89,31 %	90,64 %	91,77 %	92,61 %

Tabla 12: Valores de exactitud, precisión, exhaustividad y medida-F para el corrector ortográfico con TA estadística de caracteres para cada uno de los conjuntos de datos, considerando dos candidatos.

top 2	12k	24k	48k	102k
Exactitud	88,52 %	90,48 %	92,24 %	93,6 %
Precisión	94,7 %	95,46 %	95,99 %	96,38 %
Exhaustividad	90,89 %	92,55 %	94,2 %	95,58 %
medida-F	92,75 %	93,98 %	95,08 %	95,98 %

Tabla 13: Valores de exactitud, precisión, exhaustividad y medida-F para el corrector ortográfico con TA estadística de caracteres para cada uno de los conjuntos de datos, considerando tres candidatos.

top 3	12k	24k	48k	102k
Exactitud	90,85 %	92,47 %	93,8 %	94,9 %
Precisión	95,46 %	96,18 %	96,77 %	97,19 %
Exhaustividad	93,10 %	94,37 %	95,4 %	96,41 %
medida-F	94,26 %	95,27 %	96,08 %	96,8 %

Los valores de las métricas muestran un desempeño más que aceptable en todas las métricas. Se nota, como esperado, una mejora del desempeño a medida que se incrementa la cantidad de datos de entrenamiento, aunque esta se haga cada vez más sutil a medida que nos acercamos de la limitación natural de desempeño de esta aproximación. En el apartado siguiente comparamos el desempeño de las dos aproximaciones y se discuten sus puntos fuertes y débiles.

6.3.4 *Discusión*

Para empezar, cabe decir que los valores encontrados para el corrector ortográfico desarrollado con *hunspell* son bastante más bajos que los del modelo de TA estadística de caracteres. El valor de la precisión (tabla 10) indica que el corrector de *hunspell*

tiene un desempeño razonable al corregir errores ortográficos. Sin embargo, la baja exhaustividad muestra que muchos de los errores ortográficos detectados por ese mismo corrector en realidad no son errores. En todas las métricas, el modelo de TA estadística supera al de *hunspell* para el conjunto de datos utilizado.

La causa principal para el bajo desempeño de la solución basada en *hunspell* es, como se ha explicado en el apartado 1.3, la frecuencia con que ocurren la mezcla y el cambio de código en CZ. Al no estar presentes en el diccionario, muchas palabras de otras lenguas, como el inglés y el tagalo o el cebuano, son marcadas como incorrectas. Pese a que la incidencia de esos casos es menor en los registros formales de la lengua, ni siquiera los escritores escapan de la dificultad de encontrar equivalentes en CZ para ciertos términos no establecidos y muchas veces acaban por adoptar el término en inglés. Esto refuerza la necesidad de una normalización terminológica adecuada para el CZ, que defina cuando se deben aceptar términos ingleses o tagalos, tomar préstamos del castellano actual o crear nuevas palabras nativas. En ese sentido, el modelo de TA supera de lejos al basado en *hunspell*, ya que es capaz de generar palabras tanto del CZ como las de otras lenguas, siempre y cuando éstas hayan aparecido en los datos de entrenamiento. También es capaz de ofrecer correcciones plausibles, basadas en patrones que aparecen en diferentes palabras, aunque es cierto que también llega a ofrecer como candidatos palabras que no existen, o bien deforma palabras correctas por una generalización de las reglas o al encontrarlas por primera vez.

Otra razón para el desempeño inferior del corrector basado en *hunspell* es la distancia gráfica entre la forma incorrecta y la correcta. Observamos diversos casos en los cuales la forma correcta no viene listada entre los tres primeros candidatos, además de otros casos en que ésta ni siquiera figura en la lista. Para los primeros, un conjunto de reglas de sustitución ayudaría a subsanar el problema, dado que provocaría el reordenamiento de los candidatos. Para los segundos, la definición de reglas fonéticas podría ser útil para recuperar de los diccionarios los candidatos que las heurísticas de *hunspell* no logren encontrar. No obstante, el éxito de esta aproximación dependería enormemente de la calidad de las reglas y de una jerarquización adecuada de las mismas. Aunque la dependencia de los diccionarios sea una limitación de *hunspell*, ya que las correcciones ofrecidas están limitadas a las palabras que figuran en su diccionario, al mismo tiempo es también su fortaleza, ya que impide que se ofrezcan palabras que no existen o que son incorrectas como candidatos de corrección. En este punto, la inclusión de nuevas palabras en el modelo de TA llega a ser más costosa que simplemente incrementar una lista de palabras, ya que requiere alimentar el modelo con más datos anotados manualmente con anterioridad.

En cambio, *hunspell* parece obtener mejores resultados para errores aleatorios, o sea, aquellos en los que se inserta, se elimina, se transpone o se sustituye una letra en la palabra. En estos casos, el modelo de TA estadística acaba por deformar la palabra y generar candidatos que no existen, a no ser que hayan aparecido casos semejantes en los datos de entrenamiento. *hunspell* también podría superar al modelo de TA en la corrección de palabras poco usadas o limitadas a ciertos registros. Además, *hunspell* es escalable, y su velocidad permite una respuesta en tiempo real a medida que se escribe en texto. Con *Moses*, sin embargo, obtener una respuesta en tiempo real sería inviable, lo cual obliga al usuario a escribir el texto completo y, una vez terminado, revisarlo todo a la vez con la herramienta. Finalmente, la ventaja principal de *hunspell* es su grado de interoperabilidad con otros programas y aplicaciones, lo que elimina la necesidad de desarrollo adicional.

Considerando todo esto, la situación ideal sería poder combinar ambas aproximaciones. Como se ha mencionado en el apartado 5.7.1, parece realista la posibilidad de extraer reglas fonéticas y de sustitución de las tablas de frases generadas por *Moses* e implementarlas dentro del fichero de afijos del diccionario de *hunspell*. Sin embargo, quedaría por saber cuán relevante sería la mejora de desempeño de *hunspell* y cómo esto impactaría en la velocidad del algoritmo.

CAPÍTULO 7

CONCLUSIONES

7.1 APORTACIONES DEL TRABAJO

Este trabajo de investigación se divide en dos partes. La primera, de carácter introductorio, pretende mostrar, aunque de manera rápida y superficial, la situación del chabacano zamboangueño por medio de declaraciones de los propios hablantes en las redes sociales. Después, con la ayuda de los resultados de un cuestionario difundido en las redes sociales, se detalla en qué contextos utilizan los hablantes el zamboangueño y el grado de familiaridad del hablante corriente respecto a la ortografía que ahora se utiliza para enseñar la lengua en las escuelas. Este estudio identificó que el zamboangueño se escribe sobre todo en contextos informales, y que la mayoría de los hablantes, aunque no conocían la ortografía, están dispuestos a aprender más al respecto. La mayoría de los hablantes también cree que un corrector ortográfico del zamboangueño sería de gran utilidad.

La segunda parte del trabajo se concentra en encontrar una aproximación de corrección ortográfica para el zamboangueño que sea capaz de alcanzar un desempeño aceptable, teniendo en cuenta la gran diversidad de grafías en uso hoy día para escribirlo. Construimos para ello un corpus del zamboangueño escrito y, tomando como referencia la ortografía, estudiamos los errores ortográficos cometidos por los hablantes y desarrollamos una tipología de los mismos. Utilizando un modelo de traducción automática (TA) estadística de caracteres entrenado con datos extraídos del corpus y corregidos manualmente, demostramos que su desempeño es superior al que puede proporcionar *hunspell*, considerando el caso del zamboangueño. Sin embargo, todavía hay mucho camino por recorrer en ese aspecto, como se menciona en el apartado 7.3.

Durante toda la ejecución de este trabajo, nos hemos encontrado con vacíos en diferentes campos de estudios, tanto acerca de la lengua como de la sociolingüística

del zamboangueño. Esto se debe, en parte, a que durante mucho tiempo las lenguas criollas fueron ignoradas por los investigadores. Hoy en día, pese al creciente interés por lenguas de ese tipo, la mayoría de los estudios acerca del zamboangueño se concentran en investigar su formación o aspectos específicos de su morfosintaxis y son hechos por extranjeros. Esperamos que este trabajo, además de cumplir con el propósito de estudiar la práctica escrita del zamboangueño y de proponer una aproximación alternativa de corrección ortográfica automática, haya contribuido a arrojar luz sobre otros aspectos de la lengua y de la sociedad zamboangueñas que a menudo son obviados por los investigadores, y que los zamboangueños mismos también se animen a estudiar la lengua de su lugar de origen.

7.2 LIMITACIONES

La limitación principal de la aproximación propuesta en este trabajo es el hecho de trabajar con unigramas. Algunos errores ortográficos pueden afectar a más de un *token* o dependen del contexto, como es el caso de palabras homófonas y, por ende, quedan fuera de su alcance. No se corrigen errores de signos de puntuación, excepto el uso de los puntos de interrogación y exclamación invertidos, que no forman parte del zamboangueño. Tampoco corregimos barbarismos, casos de interferencia o de hipercorrección, excepto cuando se trata de errores de naturaleza ortográfica. Finalmente, tampoco se corrigen errores de mayúsculas, aunque estos vayan corregidos en los datos de entrenamiento, para evitar la dispersión de los datos dentro del modelo.

Respecto a las limitaciones tecnológicas, no es posible ofrecer una respuesta en tiempo real al usuario, lo que le obligaría a introducir todo el texto que desea comprobar antes de obtener las posibles correcciones. Tampoco hay manera de comprobar si el candidato ofrecido por el modelo de TA estadística existe, a menos que, por medio de un posprocesamiento, se consulte una lista de palabras válidas.

Respecto a las limitaciones lingüísticas, como se ha mencionado anteriormente, durante la ejecución de este trabajo no hemos tenido acceso al diccionario publicado por el gobierno de la ciudad de Zamboanga. Algunas de las formas utilizadas pueden diferir de las que constan en esa obra, lo que requeriría algún esfuerzo adicional para adecuar las formas de nuestros datos de entrenamiento si en algún momento se tuviera acceso al diccionario para volver a entrenar el modelo.

7.3 TRABAJOS FUTUROS

Como continuación de este trabajo de investigación, proponemos la extracción de reglas fonéticas y de sustitución a partir de las tablas de frases generadas por *Moses* y su introducción en el fichero de afijos de *hunspell*. También sería interesante encontrar maneras de lidiar con palabras homófonas o errores que afecten múltiples *tokens*. Algunas posibilidades serían entrenar un modelo mixto con unigramas y bigramas, o bien dos modelos por separado y cruzar sus respectivas correcciones. Otra tarea esencial es la de ajustar las formas utilizadas tanto en la lista de palabras como en los datos de entrenamiento de acuerdo a las del diccionario normativo. Como tareas secundarias, proponemos revisar de manera incremental el corpus anotado por el modelo con el fin de obtener mayor cantidad de datos de entrenamiento, bien como expandir las listas de palabras a partir de los datos de entrenamiento. También se podría considerar la expansión del corpus con otras fuentes.

En lo que concierne a los estudios de ámbito general acerca del zamboangueño, sugerimos estudios que intenten cuantificar qué porcentaje de la población de la ciudad de Zamboanga es capaz de hablar la lengua, el impacto de la introducción del CZ en la enseñanza, las actitudes de los habitantes de la ciudad hacia el zamboangueño y los efectos del contacto del zamboangueño con otras lenguas, promoviendo así una mejor comprensión de la situación del zamboangueño en su totalidad, así como una mejora de la política lingüística y del proceso de normativización de la lengua.

BIBLIOGRAFÍA

Abregú Tueros, J. L. G. T. D. L. F. (2011). Fundamentos para la intervención en el aprendizaje de la ortografía. *Audición y Lenguaje*, (96), 4-9.

Atkinson, K. (2019, 29 de julio). (Ver. 0.60.7). Recuperado desde http://aspell.net/

Badilla, L. (2014). Los grammar nazis y su intransigencia textual, ¿vale la pena? Recuperado el 20 de agosto de 2019, desde https://www.eldefinido.cl/actualidad/plazapublica/1917/Los_grammar_nazis_y_su_intransigencia_textual_vale_la_pena/

Balmaseda Neyra, E., Osvaldo; Molina Almeida. (2001). *La importancia del diagnóstico para la enseñanza aprendizaje de la ortografía*. Ciudad de La Habana: Editorial Pueblo y Educación.

Bautista, M. L. S. (2004). Tagalog-english code switching as a mode of discourse. *Asia Pacific Education Review*, *5*(2), 226-233. doi:10.1007/BF03024960

Beinborn, L., Zesch, T. & Gurevych, I. (2013). Cognate Production using Character-based Machine Translation. En *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya.

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Bringhurst, R. (1992). *The Elements of Typographic Style*. Point Roberts, Washington, USA: Hartley & Marks.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., … Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Comput. Linguist. 16*(2), 79-85. Recuperado desde http://dl.acm.org/citation.cfm?id=92858.92860

Canadian Institute for Research in Computing and the Arts. (2010). CIRCA:TEI XML. Recuperado desde http://circa.cs.ualberta.ca/index.php/CIRCA:TEI_XML

Catach, N. (2011). *L'orthographe: « Que sais-je ? » n° 685*. Que sais-je ? Presses Universitaires de France. Recuperado desde https://books.google.com.br/books?id=DrAICwAAQBAJ

Cook, E. (2018). «How the Philippine media's use of code switching stands apart in Asia.» en Splice. Recuperado desde https://www.thesplicenewsroom.com/philippines-code-switching-media/

Craig, D. (2003). Instant messaging: the language of youth literacy. En *The Boothe Prize Essays 2003*, Stanford University.

Cristóbal Rojo, M. (1982). Mi descubrimiento sobre Alexandre Galí. *Maina*, (5), 56-59.

DeptEd, Division of Zamboanga City. (2016). *Revised Zamboanga Chavacano Orthography (Guia para na Enseñanza de Chavacano)*. Zamboanga: Zamboanga City Local Government.

Earnest, L. (2016). The First Cursive Handwriting Recognizer Needed a Spelling Checker and so did the Rest of the World. Recuperado desde https://web.stanford.edu/~learnest/les/spelling.htm

Al-enzi, A. A. A. Ẽ. K. (2011). EFL Teachers' Feedback to Oral Errors in EFL Classroom: Teachers' Perspectives. *Arab World English Journal*, *2*(1), 214-232.

Fernández, M. (2015). La emergencia del chabacano en Filipinas: pruebas, indicios, conjeturas. En J. M. S. Rovira (Ed.), *Armonía y contrastes: estudios sobre variación dialectal, histórica y sociolingüística del español* (pp. 175-196).

Fernández, M. A. & Sippola, E. (2017). A new window into the history of Chabacano: Two unknown mid-19th century texts. *Journal of Pidgin and Creole Languages*, *32*(2), 304-338. doi:https://doi.org/10.1075/jpcl.32.2.04fer

Galí, A. (1971). *L'ensenyament de l'ortografia als infants, Barcelona*. Barcelona: Ed. Barcino.

Gasparin Nobile, S., Gislaine;Domingos BarreraI. (2009). Análise de erros ortográficos em alunos do ensino público fundamental que apresentam dificuldades na escrita. *Psicologia em Revista*, *15*(2), 36-55.

Gómez-Pérez, A. [A.], Fernandez-Lopez, M. & Corcho, O. (2004). *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition*. Advanced Information and Knowledge Processing. Springer London. Recuperado desde https://books.google.com.br/books?id=UjSoN1W7GSEC

Gómez-Pérez, A. [Asuncion] & Benjamins, V. R. (1999). Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods. En *IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends* (Vol. 18, pp. 1-15). CEUR Publications.

Gorin, R. E. (1974). SPELL: Spelling Check and Correction Program. Recuperado desde https://www.saildart.org/allow/SPELL.REG%5BUP,DOC%5D

Goulet, R. M. (1971). *English, Spanish, and Tagalog: A Study of Grammatical, Lexical, and Cultural Interference*. Manila: Linguistic Society of the Philippines.

Government of the Philippines. (2011). DepEd develops learning supplements using mother- tongue. Recuperado el 19 de enero de 2017, desde https://www.officialgazette.gov.ph/2011/11/28/deped-develops-learning-supplements-using-mother-tongue/

Government of the Philippines. (2014). About The Philippines. Recuperado el 26 de noviembre de 2017, desde https://www.gov.ph/about-the-philippines

Grant, A. P. (2007). Some aspects of NPs in Mindanao Chabacano: Structural and historical considerations. En M. Baptista & J. Guéron (Eds.), (Cap. 6, Vol. 31, pp. 173-204). Creole Language Library. doi:https://doi.org/10.1075/cll.31.10gra

Grant, A. P. (2013). On the (dis)unity of the Manila Bay Creoles: some lexical strata in Ternateño. *Revista de Crioulos de Base Lexical Portuguesa e Espanhola*, *4*(2), 26-47.

Gruber, T. R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *Int. J. Hum.-Comput. Stud. 43*(5-6), 907-928. doi:10.1006/ijhc.1995.1081

Hasan, S., Heger, C. & Mansour, S. (2015). Spelling Correction of User Search Queries through Statistical Machine Translation. En *EMNLP*.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. En *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation* (pp. 187-197). Edinburgh, Scotland, United Kingdom. Recuperado desde https://kheafield.com/papers/avenue/kenlm.pdf

Heisel, A. (2015). Stop shaming people on the Internet for grammar mistakes. Its not there fault. Recuperado el 25 de julio de 2019, desde https://www.washingtonpost.com/posteverything/wp/2015/04/17/stop-shaming-people-on-the-internet-for-grammar-mistakes-its-not-there-fault/

Himoro, M. Y. (s.f.). *Linguistic variation in Zamboanga Chabacano: a pilot study*.

Jones, E. & Uribe-Jongbloed, E. (2013). *Social Media and Minority Languages: Convergence and the Creative Industries*. Multilingual Matters. Channel View Publications. Recuperado desde https://books.google.com.br/books?id=8UlOnNdP5WMC

Jurafsky, D. & Martin, J. H. (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Third Edition draft).

Karimi, S. (2008). *Machine transliteration of proper names between English and Persian* (Tesis doctoral).

Kim, E. (2006). Reasons and Motivations for Code-Mixing and Code-Switching. *Issues in EFL*, *4*(1).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. En *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 177-180). Prague, Czech Republic: Association for Computational Linguistics. Recuperado desde https://www.aclweb.org/anthology/P07-2045

Komisyon sa Wikang Filipino (KWF). (2015). Mapa ng mga Wika ng Filipinas. Recuperado el 3 de agosto de 2018, desde http://kwf.gov.ph/mapa-ng-mga-wika-ng-filipinas/

Korchagina, N. (2017). Normalizing Medieval German Texts: from rules to deep learning. En *NoDaLiDa 2017 Workshop on Processing Historical Language*, Gothenburg.

LimeSurvey GmbH. (2018, 3 de mayo). Limesurvey (Ver. 3.7.1). Recuperado desde https://www.limesurvey.org/

Lipski, J. M. (1992). New thoughts on the origins of Zamboangueño (Philippine Creole Spanish). *Language Sciences*, *14*(3), 197-231. doi:https://doi.org/10.1016/0388-0001(92)90005-Y

Lipski, J. M. (2003). *Chabacano/Spanish and the Philippine linguistic identity*.

Lipski, J. M. (2013). Remixing a mixed language: The emergence of a new pronominal system in Chabacano (Philippine Creole Spanish). *International Journal of Bilingualism*, *17*(4), 448-478. doi:10.1177/1367006912438302

Martínez Agudo, J. d. D. (2008). Oral communication in the EFL classroom. (Cap. Linguistic risk-taking and corrective feedback, pp. 165-193). Sevilla: Alfar.

Miede, A. (2016). *A Classic Thesis style*. Recuperado desde http://www.ctan.org/tex-archive/macros/latex/contrib/classicthesis/ClassicThesis.pdf

Nakov, P. & Tiedemann, J. (2012). Combining Word-Level and Character-Level Models for Machine Translation Between Closely-Related Languages. (Vol. 2, pp. 301-305).

National Census and Statistics Office (NCSO). (1974). *1970 Census of Population and Housing, Final Report - Vol. 1 - Zamboanga del Sur*. Manila. Recuperado desde https://psa.gov.ph/content/census-population-and-housing-report

National Census and Statistics Office (NCSO). (1983). *1980 Census of Population and Housing, Volume 1, Final Report - Zamboanga del Sur*. Manila, Manila, Filipinas. Recuperado desde https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (1992). *1990 Census of Population and Housing, Report No. 3 - 86 I - Socio-Economic and Demographic Characteristics*. Manila. Recuperado desde https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (2003a). *2000 Census of Population and Housing, Report No. 2 Vol. 1 - Demographic and Housing Characteristics*. Manila. Recuperado desde https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (2003b). *2000 Census of Population and Housing, Report No. 2 Vol. 1 - Demographic and Housing Characteristics - Zamboanga City*. Manila. Recuperado desde https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (2014a). *2010 Census of Population and Housing, Report No. 2B - Population and Household Characteristics (Sample Variables)*. Manila. Recuperado desde https://psa.gov.ph/content/census-population-and-housing-report

National Statistics Office (NSO). (2014b). *2010 Census of Population and Housing, Report No. 2B - Population and Household Characteristics (Sample Variables) - Zamboanga City*. Manila. Recuperado desde https://psa.gov.ph/content/census-population-and-housing-report

Németh, L. (2019, 29 de julio). (Ver. 0.60.7). Recuperado desde http://aspell.net/

Neves, M. (2015). Devemos apontar publicamente os erros de português dos outros? Recuperado el 20 de agosto de 2019, desde https://www.certaspalavras.net/devemos-apontar-publicamente-os-erros-de-portugues-dos-outros/

Nocon, N., Cuevas, G., Gopez, J. & Suministrado, P. (2014). NORM: A Text Normalization System for Filipino Shortcut Texts Using the Dictionary Substitution Approach. En *Proceedings of the 10th National Natural Language Processing Research Symposium* (pp. 87-92).

Nocon, N., Cuevas, G., Magat, D., Suministrado, P. & Cheng, C. (2014). NormAPI: An API for normalizing Filipino shortcut texts. (pp. 207-210). doi:10.1109/IALP.2014.6973494

Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*(1), 19-51.

Octaviano, M., Go, M. P., Borra, A. & Oco, N. (2016). A corpus-based analysis of Filipino writing errors. En *2016 International Conference on Asian Language Processing (IALP)* (pp. 95-98). doi:10.1109/IALP.2016.7875943

Pantieri, L. (2017a). *LaTeXpedia*. Recuperado desde http://www.lorenzopantieri.net/LaTeX_files/LaTeXpedia.pdf

Pantieri, L. (2017b). *LaTeX per l'impaziente*. Recuperado desde http://www.lorenzopantieri.net/LaTeX_files/LaTeXimpaziente.pdf

Pantieri, L. & Gordini, T. (2017). *L'arte di scrivere con LaTeX*. Recuperado desde http://www.lorenzopantieri.net/LaTeX_files/ArteLaTeX.pdf

Peterson, J. L. (1980). Computer Programs for Detecting and Correcting Spelling Errors. *Communications of the ACM*, *23*(12), 676-687.

Philippine Statistics Authority (PSA). (2014). «Statistical Tables on Sample Variables from the Results of 2010 Census of Population and Housing» en Census of Population and Housing. Recuperado desde https://psa.gov.ph/population-and-housing/statistical-tables/2010

Philips, L. (2000). The Double Metaphone Search Algorithm. *C/C++ Users Journal*, *18*, 38-43.

Priberam Informática. (2019). «Tipos de Erro - Ortografia» en FLiP. Recuperado el 31 de agosto de 2019, desde https://www.flip.pt/Modulos/Corrector-Sintactico/Tipos-de-Erro-Ortografia

Python Software Foundation. (2019). Python Language Reference. Ver. 3.7.4. Recuperado desde https://www.python.org

Real Academia Española & Asociación de Academias de la Lengua Española. (2005). *Diccionario panhispánico de dudas*. Real Academia Española. Recuperado desde https://books.google.com.br/books?id=3FCTQgAACAAJ

Rojas, J. E. N. O. M. (2009). ¿Pueden tener dificultades con la ortografía los niños que leen bien? *Revista Española de Pedagogía*, *242*(67), 45-60.

Russell, R. C. (1918). *Soundex*. US1261167.

Saint-Exupéry, A. (1943). *Le Petit Prince*. Ebooks libres et gratuits. Recuperado desde https://www.cmls.polytechnique.fr/perso/tringali/documents/st_exupery_le_petit_prince.pdf

Saint-Exupéry, A. (2018). *El Diutay Principe (trad. Jerome Herrera)*. Jerome Herrera.

Sánchez-Jiménez, D. (2010). El análisis de errores ortográficos de estudiantes filipinos en el aprendizaje de español como LE y su aplicación didáctica. En *I Congreso de Español como Lengua Extranjera en Asia-Pacífico, El Currículo de E/LE en Asia Pacífico*.

Sanchez, V. (2013). *Bakit-Why? An analysis of the Sociolinguistic Motivations behind Taglish* (Trabajo de Fin de Grado, Swarthmore College. Dept. of Linguistic).

Santarnarina, A. (1995). Estudios de sociolingüística galega: sobre a norma do galega culto. En H. Monteagudo (Ed.), (Cap. Norma e estándar). Vigo: Galaxia.

Santos, F. M. (2014), En *Facebook [Perfil]*. Recuperado el 19 de enero de 2017, desde https://www.facebook.com/felino.santos.1/posts/746614055388086

Scherrer, Y., Samardžić, T. & Glaser, E. (2016). Normalizing orthographic and dialectal variants in the ArchiMob corpus of spoken Swiss German. ID: unige:90850. Recuperado desde https://archive-ouverte.unige.ch/unige:90850

Schneider, G., Pettersson, E. & Percillier, M. (2017). Comparing Rule-based and SMT-based Spelling Normalisation for English Historical Texts. En *NoDaLiDa 2017 Workshop on Processing Historical Language*.

Simon, J. (1952). Étude psychopédagogique de l'orthographe. doi:10.3406/enfan.1952.1223

Singla, K. (2015). *Methods for Leveraging Lexical Information in SMT* (Tesis doctoral). doi:10.13140/RG.2.1.2138.7367

Soroa, A., Rigau, G., Porta, J., Atserias, J., Gómez Guinovart, X. & Saggion, H. (2017). *Plataformas y sistemas de procesamiento lingüístico de alto rendimiento*. Informe Plataformas NLP.

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. En *IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP 2002* (pp. 901-904).

Text Encoding Initiative (TEI). (2019). The TEI Guidelines. Recuperado desde https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

The World Wide Web Consortium (W3C). (2004). OWL Web Ontology Language. Overview. Recuperado desde https://www.w3.org/TR/owl-features/

The World Wide Web Consortium (W3C). (2013). Web Ontology Language (OWL). Recuperado desde https://www.w3.org/OWL/

Tiedemann, J. (2009). Character-based PSMT for Closely Related Languages. (pp. 12-19).

Tiedemann, J. & Nabende, P. (2009). Translating Transliterations. *International Journal of Computing and ICT Research*, *3*, 33-41.

Tobar Delgado, E. & Fernández, M. A. (2019). Hacia una ortografía para el chabacano zamboangueño: Prácticas escritas y propuestas de estandarización. *Language Problems and Language Planning*, *1*(43), 32-54. doi:https://doi.org/10.1075/lplp.00031.tob

Vilar, D., Peter, J.-T. & Ney, H. (2007). Can We Translate Letters? En *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 33-39). StatMT '07. Prague, Czech Republic: Association for Computational Linguistics. Recuperado desde http://dl.acm.org/citation.cfm?id=1626355.1626360

Whinnom, K. (1956). *Spanish Contact Vernaculars in the Philippine Islands*. Hong Kong University Press.

Willisson, P. (2015, 8 de febrero). ispell (Ver. 3.3.03).

APÉNDICE A

SURVEY ON THE USE OF ZAMBOANGA CHAVACANO

(Si quiere tu contesta con este survey na Chavacano, favor man click aqui)

This survey is part of my thesis for **Master's degree in Information Technologies and Communication in Language Education and Processing** at **National University of Distance Education** (Madrid, Spain).

Purpose of this study: Identify how Chavacano de Zamboanga speakers use the language and know more about their attitudes towards it.

Confidentiality: All answers are anonymous. The information collected will be used solely for research purposes.

By answering this survey, you will receive a FREE gift! If you have friends or relatives who can speak Chavacano, please take some time to share this link to help us reach as many people as possible. It will remain open until June 12th only.

If you have any questions, please do not hesitate to contact me:
Marcelo Yuji Himoro
mhimoro1@alumno.uned.es

Muchas gracias!

A. ABOUT YOU

1. **What is your age?***
   _____ [6; 100]

2a. **What is your gender?***

♀ Female	♂ Male	⚥ Other

2b. **(Optional) How do you define your gender?**
_____
(Esta pregunta solo se mostrará a los que hayan elegido «Other» en la pregunta 2a.)

3. **Where do you live?***
   - ☐ In Zamboanga City
   - ☐ In Basilan
   - ☐ Other*: _____

4. **Approximately how many years have you been living away?***
   - ☐ less than 1 year
   - ☐ 1-2 years
   - ☐ 3-5 years
   - ☐ 6-10 years

- ☐ 11-20 years
- ☐ 21-30 years
- ☐ 30+ years

5. What industry do you belong to?*

- ☐ Accommodation and Food Services
- ☐ Agriculture, Forestry and Fishing
- ☐ Arts, Entertainment and Recreation
- ☐ Construction
- ☐ Education (including students)
- ☐ Electricity, Gas, Steam and Air Conditioning Supply
- ☐ Financial and Insurance
- ☐ Health Care and Social Work
- ☐ Information and Communication
- ☐ Manufacturing
- ☐ Mining and Quarrying
- ☐ Professional, Scientific and Technical Services
- ☐ Real Estate
- ☐ Transportation and Storage
- ☐ Water Supply, Sewerage, Waste Management and Remediation Activities
- ☐ Wholesale and Retail Trade
- ☐ Other Services

6. What is the highest level of education you have attained?*

- ☐ None
- ☐ Elementary School
- ☐ High School
- ☐ College
- ☐ Postgraduate

B. ABOUT YOUR USE OF CHAVACANO

1. **What were the mediums of instruction when you were in Elementary School?***

   ❗ Choose ll that apply

   □ Filipino

   □ English

   □ Chavacano

   □ Spanish

   □ I studied at a Muslim or Chinese school

2. **Have you ever studied Spanish?***

   □ Yes □ No

3. What language(s) do you speak...?*	Chavacano	Filipino	English	Visaya	Tausug	Other
a. at home	□	□	□	□	□	_____
b. when talking to your Zamboangueño friends	□	□	□	□	□	_____
c. at your workplace / school	□	□	□	□	□	_____

❗ Check all that apply

4. What language(s) do you use...?*	Chavacano	Filipino	English	Visaya	Tausug	Other
a. to take down notes	□	□	□	□	□	_____
b. to email a Zamboangueño	□	□	□	□	□	_____
c. to text or chat with a Zamboangueño	□	□	□	□	□	_____

❗ Check all that apply

5. **Please choose the reasons why you feel more comfortable writing in other languages:***

   □ Chavacano is not my mother tongue

   □ I did not have a chance to study Chavacano at school

   □ I have a hard time writing in Chavacano because it does not have clear spelling rules

   □ Other*: _____

   (Esta pregunta solo se mostrará a los que no hayan marcado «Chavacano» en ninguna de las opciones de la pregunta 4.)

C. ABOUT THE ZAMBOANGA CHAVACANO ORTHOGRAPHY

Since 2014, an official orthography (**Zamboanga Chavacano Orthography**) has been implemented at the schools in Zamboanga City. It establishes as a general rule that:

> *Chavacano words of Spanish origin are written following the original Spanish form. Chavacano words of local origin are likewise spelled according to its origin.*
>
> (Department of Education: Division of Zamboanga City & Local Government of Zamboanga City (Eds.). (2014). Revised Zamboanga Chavacano Orthography. Zamboanga City, Philippines: Local Government of Zamboanga City, 15.)

6. **Did you know about the official orthography?***
   □ Yes □ No

Now try to read the paragraph below:

TEXT A

> Bien flojo gayod este mio marido. Todo'l dia ya lang ta'n tomahan junto con de suyo maga amigo. Ni no quiere anda busca trabajo. Falta gad juicio. No hay 'le cosa sabe sino maga fanfarronadas. Cuando ya habla yo con ele kay ya coge yo con ele junto na otro mujer, ya re que re lang 'le conmigo. Habla 'le kay loca ya daw yo. Yo pa ahora el loca. Pero no hay yo cosa puede hace. Tres ya el anak de amon. No puede yo con ele deja.

TEXT B

> Bien ploho gayot este miyo marido. Todol dia ya lang tan tomahan hunto con disuyu maga amigo. Ni nukere anda busca trabaho. Palta gat wisyo. Nuay le cosa sabe sino mga pamparonadas. Cuando ya abla yo conele kay ya kuhi yo cunele huntu na otro muher, ya rikiri lang le kumigo. Abla le kay loka ya daw yo. Iyo pa ara el loka. Pero nuay io kosa pwede ase. Tres ya el anak diamun. No puede yo kunele deha.

7. **In your opinion, which one is easier to read?***
   □ Text A □ Text B

8. **Do you feel confident enough about applying the official orthography when writing in Chavacano?***
   Not confident at all □—□—□—□—□ Very confident

9. **Would you be willing to learn more about the official orthography?***
   I am not interested □—□—□—□—□ I would love to

10. **In your opinion, is it important for Chavacano to have an official orthography?***
    □ Yes □ No

11. **Do you think a spell checker for Chavacano would be useful?***
    □ Yes □ No ❓ A spell checker is a software that automatically attempts to identify possible spelling mistakes and suggest corrections to them.

Try to read the following joke and look at the highlighted words.

> Ta porfia si Juan con su maga uban tomador por causa del color del luna.
> JUAN: El color del luna ahora como yellow y tiene un poco colorao.
> UBAN: Hende, el color del luna ahora yellow green.
> JUAN: Mali ustedes, sobra ya **siguro** el vino na cabeza de ustedes. Bien claro gayod kay yellow el color del luna y tiene un poco colorao.
> UBAN: Basta ya 'se porfiahan. Taqui ya si Pedro. Ele el puede habla kanaton cosa el verdadero color del luna kay hende 'le tomao. **Bene** daw anay aqui, Pedro.
> JUAN: Pedro, favor habla kanamon cosa el color del luna ahora?
> Enseguidas ya man tanga para arriba, y ya habla:
> PEDRO: El donde? El na derecha o el na **esquierda**?
>
> (Adapted from the text by Franklin Cañizares Alibasa)

12. **Can you identify which of the words above are incorrectly spelled?***
    ❗ Check all that apply
    - □ siguro
    - □ bene
    - □ esquierda
    - □ I do not know

    ❓ Please refrain from using dictionaries or online translators. This is **NOT** an exam.

D. ABOUT THIS SURVEY

1. **How have you found out about this survey?***
   ❗ Check all that apply
   - ☐ friends or relatives
   - ☐ Facebook groups
   - ☐ Twitter
   - ☐ blogs
   - ☐ forums
   - ☐ Other*: _____

2. **(Optional) If you have any comments, opinions or anything you would like to add, feel free to use the comments box below:**

   _____

   _____

Thank you for taking time to answer our survey.

Please click here to download your free Chavacano ebook!

"**Cosa adentro del caldero?**" is the Zamboanga Chavacano edition of the children's book "**What's in the pot?**", written by Hayley Alonzo, Crystal Warren and Rat Western for Book Dash and licensed under Creative Commons Attribution licence.

The Chavacano edition was produced by Colorin Colorao, a non-profit initiative that aims at spreading literacy in Chavacano by providing its young speakers with reading materials in their own mother tongue. This is just the first of many books we plan to publish. If you want to know more about us, please click here.

APÉNDICE B

SURVEY ACERCA DEL USADA DEL CHAVACANO DE ZAMBOANGA

(If you wish to answer this survey in English, please click here)

Parte este *survey* del mio *thesis* del **Master's degree in Information Technologies and Communication in Language Education and Processing** na **National University of Distance Education** (Madrid, España).

Objetivo del estudio: Identifica que laya el maga sabe conversa Chavacano de Zamboanga ta usa con ese y aumenta el de aton saber acerca na de ila actitud con el lenguaje.

Confidencialidad: Secreto todo el maga contestacion. Para lang na *research* kame usa con el maga colectao informacion.

Contesta con este *survey*, acabar tiene kame cosa ay dale contigo LIBRE! Si tiene tu maga amigo o familia sabe conversa Chavacano, ta roga kame que ay man *share* tu con este link para puede este llega na mas mucho gente. Abierto este hasta na 12 de Junio lang.

Si tiene tu cosa quiere pregunta, no tu tiene huya man email conmigo:
Marcelo Yuji Himoro
mhimoro1@alumno.uned.es

Muchas gracias!

A. ACERCA CONTIGO

1. **Cuanto año ya tu?***
   _____ [6; 100]

2a. **Cosa de tuyo *gender*?***

♀ Mujer	♂ Hombre	⚥ Otro

2b. **(Opcional) Que laya tu ta defini con el de tuyo *gender*?**
   _____
   (Esta pregunta solo se mostrará a los que hayan elegido «Other» en la pregunta 2a.)

3. **Donde tu ta queda?***
   - □ Na Zamboanga City
   - □ Na Basilan
   - □ Otro*: _____

4. **Maga cuanto año ya tu ta queda afuera?***

- □ no hay pa 1 año
- □ 1-2 año
- □ 3-5 año
- □ 6-10 año
- □ 11-20 año
- □ 21-30 año
- □ 30+ año

5. Na cosa *industry* tu ta trabaja?*

- □ Accommodation and Food Services
- □ Agriculture, Forestry and Fishing
- □ Arts, Entertainment and Recreation
- □ Construction
- □ Education (entrao maga estudiante)
- □ Electricity, Gas, Steam and Air Conditioning Supply
- □ Financial and Insurance
- □ Health Care and Social Work
- □ Information and Communication
- □ Manufacturing
- □ Mining and Quarrying
- □ Professional, Scientific and Technical Services
- □ Real Estate
- □ Transportation and Storage
- □ Water Supply, Sewerage, Waste Management and Remediation Activities
- □ Wholesale and Retail Trade
- □ Other Services

6. Cosa de con todo alto *level of education* que tu ya puede alcanza?*

- □ No hay
- □ Elementary School
- □ High School
- □ Colegio
- □ Postgraduate

B. ACERCA NA DE TUYO USADA DEL CHAVACANO

1. **Cosa-cosa el maga *medium of instruction* cuando na *Elementary School* pa tu?***

   ❗ Man *check* el todo deverasan para contigo

   □ Filipino

   □ Ingles

   □ Chavacano

   □ Español

   □ Ya estudia yo na un *Chinese* o *Muslim school*

2. **Ya puede ba tu estudia Español antes?***

   □ Si □ Hende

3. Cosa-cosa lenguaje tu ta conversa...?*	Chavacano	Filipino	Ingles	Visaya	Tausug	Otro
a. na de ustedes casa	□	□	□	□	□	_____
b. si ta'n cuento na de tuyo maga amigo Zamboangueño	□	□	□	□	□	_____
c. na de tuyo trabajo / escuela	□	□	□	□	□	_____

❗ Man *check* el todo deverasan para contigo

4. Na cosa-cosa lenguaje tu ta escribi...?*	Chavacano	Filipino	Ingles	Visaya	Tausug	Otro
a. si ta escribi tu **notes**	□	□	□	□	□	_____
b. ta escribi email na un Zamboangueño	□	□	□	□	□	_____
c. si ta'n text o si ta'n chat na un Zamboangueño	□	□	□	□	□	_____

❗ Man *check* el todo deverasan para contigo

5. **Favor escoge el razon porque mas quiere tu escribi na otro lenguaje:***

   □ Hende mio *mother tongue* el Chavacano

   □ No hay yo puede estudia Chavacano na escuela

   □ Malisud escribi na Chavacano kay no hay este maga claro reglamento na deletreada (spelling)

   □ Otro*: _____

   (Esta pregunta solo se mostrará a los que no hayan marcado «Chavacano» en ninguna de las opciones de la pregunta 4.)

C. ACERCA NA ZAMBOANGA CHAVACANO ORTHOGRAPHY

Del año 2014, tiene ya un oficial ortografia (**Zamboanga Chavacano Orthography**) que ta implementa na maga escuela del ciudad. Este ta establece como reglamento principal kay:

> *Chavacano words of Spanish origin are written following the original Spanish form. Chavacano words of local origin are likewise spelled according to its origin.*
>
> (Department of Education: Division of Zamboanga City & Local Government of Zamboanga City (Eds.). (2014). Revised Zamboanga Chavacano Orthography. Zamboanga City, Philippines: Local Government of Zamboanga City, 15.)

6. Ya puede ya ba tu oi por causa del oficial ortografia?*
□ Si □ Hende

Ahora proba tu lee con el *paragraph* abajo:

TEXTO A

> Bien flojo gayod este mio marido. Todo'l dia ya lang ta'n tomahan junto con de suyo maga amigo. Ni no quiere anda busca trabajo. Falta gad juicio. No hay 'le cosa sabe sino maga fanfarronadas. Cuando ya habla yo con ele kay ya coge yo con ele junto na otro mujer, ya re que re lang 'le conmigo. Habla 'le kay loca ya daw yo. Yo pa ahora el loca. Pero no hay yo cosa puede hace. Tres ya el anak de amon. No puede yo con ele deja.

TEXTO B

> Bien ploho gayot este miyo marido. Todol dia ya lang tan tomahan hunto con disuyu maga amigo. Ni nukere anda busca trabaho. Palta gat wisyo. Nuay le cosa sabe sino mga pamparonadas. Cuando ya abla yo conele kay ya kuhi yo cunele huntu na otro muher, ya rikiri lang le kumigo. Abla le kay loka ya daw yo. Iyo pa ara el loka. Pero nuay io kosa pwede ase. Tres ya el anak diamun. No puede yo kunele deha.

7. Para contigo, cosa el mas facil lee?*
□ Texto A □ Texto B

8. **Ta senti ba tu que ay puede tu usa el oficial ortografia si ay escribi tu na Chavacano?***
   Hende gayod □—□—□—□—□ Puede gayod

9. **Quiere ba tu aprende mas acerca del oficial ortografia?***
   Hende yo interesao □—□—□—□—□ Bien quiere yo aprende

10. **Ta mira tu, importante ba tiene un oficial ortografia del Chavacano?***
    □ Si □ Hende

11. **Bueno ba gaha tiene un *spell checker* de Chavacano?***
    □ Si □ Hende
    ❓ El *spell checker* un software que ta precura identifica pati corregi automaticamente con el maga posible mali na deletreada (spelling).

Lee daw tu con este broma y pone tu atencion na maga *highlighted* palabra.

> Ta porfia si Juan con su maga uban tomador por causa del color del luna.
> JUAN: El color del luna ahora como yellow y tiene un poco colorao.
> UBAN: Hende, el color del luna ahora yellow green.
> JUAN: Mali ustedes, sobra ya **siguro** el vino na cabeza de ustedes. Bien claro gayod kay yellow el color del luna y tiene un poco colorao.
> UBAN: Basta ya 'se porfiahan. Taqui ya si Pedro. Ele el puede habla kanaton cosa el verdadero color del luna kay hende 'le tomao. **Bene** daw anay aqui, Pedro.
> JUAN: Pedro, favor habla kanamon cosa el color del luna ahora?
> Enseguidas ya man tanga para arriba, y ya habla:
> PEDRO: El donde? El na derecha o el na **esquierda**?
>
> (Adaptao estaba na texto de Franklin Cañizares Alibasa)

12. **Puede ba tu habla cosa-cosa palabra arriba el tiene kamali na deletreada (spelling)?***
    ❗ Man *check* el todo deverasan para contigo
    □ siguro
    □ bene
    □ esquierda
    □ No sabe yo

❓ Favor no usa maga diccionario o maga traductor na internet. **HENDE** este un *exam*.

D. ACERCA CON ESTE *survey*

1. **Que laya tu ya puede encontra con este *survey*?***
   ❗ Man *check* el todo deverasan para contigo
   - ☐ maga amigo o familia
   - ☐ maga grupo na Facebook
   - ☐ Twitter
   - ☐ maga blogs
   - ☐ maga forums
   - ☐ Otro*: _____

2. **(Opcional) Si tiene pa tu maga comento, opinion o cosa quiere aumenta, favor escribi abajo:**

   _____

   _____

Gracias na tuyo contestacion y tiempo.

Favor man click aqui para man download tuyo libre Chavacano ebook!

"**Cosa adentro del caldero?**" amo el version na Chavacano de Zamboanga del libro "**What's in the pot?**", escribido de Hayley Alonzo, Crystal Warren and Rat Western para na Book Dash y distribuido bajo de un licencia Creative Commons Attribution.

El edicion na Chavacano ya produci Colorin Colorao, un iniciativa sin ganancia que el objetivo amo produci maga historia na Chavacano para ay tiene el maga bata el costumbre lee na de ila nativo lenguaje. Este el primero lang na mucho pa otro libro ta planea kame publica. Si quiere tu sabe mas acerca kanamon, favor man click aqui.