Centro Nacional de Investigación y Desarrollo Tecnológico

Subdirección Académica

Departamento de Ciencias Computacionales

TESIS DE MAESTRÍA EN CIENCIAS

Método para la Identificación Automática de Albures Cortos en Textos

presentada por
Ing. Roberto Villarejo Martínez

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis
Dr. Noé Alejandro Castro Sánchez

Cuernavaca, Morelos, México. Noviembre de 2016.

Cuernavaca, Morelos a 22 de septiembre del 2016
OFICIO No. DCC/195/2016
Asunto: Aceptación de documento de tesis

C. DR. GERARDO V. GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **Ing. Roberto Villarejo Martínez**, con número de control M14CE071, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado **"Método para la identificación automática de albures cortos en textos"** y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806

REVISOR 1

Dra. Alicia Martínez Rebollar
Doctora en Informática
7399055

REVISOR 2

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la Computación
7820329

REVISOR 3

Dr. Gerardo Eugenio Sierra Martínez
Doctor en Ingeniería del Lenguaje
SAC/292/2016

C.p. M.T.I. María Elena Gómez Torres.- Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 04 de octubre de 2016
OFICIO No. SAC/297/2016

Asunto: Autorización de impresión de tesis

ING. ROBERTO VILLAREJO MARTÍNEZ
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Método para la identificación automática de albures cortos en textos"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

SEP TecNM
CENTRO NACIONAL
DE INVESTIGACIÓN
Y DESARROLLO
TECNOLÓGICO
SUBDIRECCIÓN
ACADÉMICA

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

C.p. M.T.I. María Elena Gómez Torres .- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

Resumen

En el presente trabajo se aborda un tipo de humor muy peculiar y propio de México llamado albur. Éste es comúnmente denominado como un juego de palabras que utiliza el "doble sentido" para someter sexualmente de manera verbal al escuchante. Anteriormente este tipo de humor ya había sido estudiado en otras áreas del conocimiento como la lingüística, en donde se han identificado algunos de los fenómenos más importantes (polisemia, homofonía, calambures, etc.), sin embargo, no se habían implementado a nivel computacional.

Esta tesis describe un método para realizar la identificación automática del albur y la implementación de éste en un programa que analiza un texto de entrada y determina si existe este tipo de humor que podría haber sido escrito involuntariamente. Los tipos de albures que este método aborda son aquellos basados en la ambigüedad léxica (polisemia) y semántica. A pesar de que existen otros tipos de albur, este trabajo representa una contribución significativa al humor computacional y al estudio de este tipo de humor.

Orientar este trabajo de investigación a la identificación computacional de los albures, se debió a varias razones: su uso en México está muy extendido, sin importar estrato social, nivel educativo o zona geográfica; se utilizan en diferentes formatos tales como: conversación, rima, juego de palabras, etc.; su particular naturaleza (frase sexual oculta) impide encontrar un análogo en otras comunidades lingüísticas del mundo; y finalmente porque han sido de gran interés para el estudio en ésta y otras áreas. El método que se presenta en esta investigación está basado en el análisis de varios albures escritos por personas de distintos estados de la república mexicana, lo que permite tener una amplia variedad de éstos.

Una de las aplicaciones que puede tener este método implementado es identificar albures en textos en los que se emplean de forma inadvertida. Esto resulta útil para evitar la aparición de éstos en textos de naturaleza formal, en los que la existencia de albures comprometería la seriedad o importancia del escrito.

La evaluación realizada muestra que este método identifica adecuadamente ciertos tipos de albures con una precisión de 0.91 y cobertura de 0.81. Aunque la variedad de albures es amplia y no se logran cubrir todos, el método puede ser expandido y mejorado para cubrir el resto de los casos.

Abstract

In this work a kind of Mexican humor is addressed, the albur. This is commonly referred to as a wordplay that uses the double entendre for "sexually subdue" a listener. This type of humor has been studied previously in other areas of knowledge such as linguistics, in which some important phenomena were identified (such as polysemy, homophony, calambur, etc.), however, they had not been computationally implemented.

This thesis describes a method for automatic identification of albur and its implementation in a program that analyzes an input text and determines if exists this type of humor (written intended or unintended). The types of albures that addresses this method are those based on lexical ambiguity (polysemy) and semantic ambiguity. Although not all types of albures are tackled, this work represents a significant contribution to computational humor and the study of albur.

This work is oriented to computational identification of albures for several reasons: 1) its use is widespread in Mexico, regardless social status, education level or region, 2) they are used in different formats: conversation, rhyme, wordplay, etc., 3) its particular nature (hidden sexual meaning) prevents find an analogue in other regions of the world and, 4) they have been of great interest to the study in this and other areas. The method presented in this research is based on the analysis of several albures written by people from different states of the Mexican republic, which allows have a wide variety of these.

One of the applications that this implemented method can have is to identify unintended albures on texts, this is especially useful in formal texts, in which the existence of albures would compromise the seriousness or importance of the document.

The evaluation performed shows that this method adequately identifies certain types of albures (0.91 precision, 0.81 recall). The variety of albures is wide and method fails to identify all of them, though, this can be expanded and enhanced to cover the remaining cases.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por su Programa Nacional de Posgrados de Calidad (PNPC) por medio del cual me apoyó económicamente como estudiante de tiempo completo.

A la red Temática en Tecnologías del Lenguaje por medio de la cual conocí a otros investigadores y estudiantes del área quienes directa o indirectamente ayudaron a la realización de este trabajo con críticas, ideas y sugerencias.

A mi comité tutorial conformado por el Dr. Gerardo Eugenio Sierra Martínez, la Dra. Alicia Martínez Rebollar y el Dr. Juan Gabriel González Serna, quienes dedicaron parte de su tiempo a las revisiones de este trabajo y ayudaron a la mejora del mismo.

A mi director de tesis, el Dr. Noé Alejandro Castro Sánchez, quien me enseñó sobre la emocionante área del Procesamiento del Lenguaje Natural y quien siempre apoyó mis iniciativas sobre el desarrollo de esta tesis.

Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), que creyó en mis capacidades y habilidades al permitirme ser parte de esta gran comunidad.

Dedicatoria

A ustedes, padres, que siempre me han apoyado moral, espiritual y económicamente en mis estudios, que me han enseñado a ser responsable y ordenado; que me alientan a superarme y ser mejor persona. Ustedes me escuchan y me dan consejos acertados.

A ti, Lluvia, que me apoyas en todo momento, me escuchas y aconsejas, tú me brindas la seguridad para cumplir mis metas y enfrentarme a situaciones difíciles. Te amo.

A mis hermanos Mauricio, Raquel y Marco... y Jacky, gracias por su apoyo. Los quiero mucho.

Contenido

Índice de ilustraciones

Índice de tablas

Capítulo 1 Introducción

En los últimos años, se ha puesto especial atención al desarrollo de sistemas capaces de procesar el lenguaje natural. Se ha trabajado en comunicación humano-computadora, en la comprensión de narrativas escritas, búsqueda de información en la Web y conversaciones humanas (Taylor, 2008). Esta tendencia ha tomado tanta fuerza que incluso existe un paradigma experimental en el cual se afirma que la relación humano-computadora es fundamentalmente social: Las Computadoras Son Actores Sociales (*Computers Are Social Actors, CASA*) (Nass, Steuer, & Tauber, 1994). En este paradigma se muestra que "principios existentes provenientes de la literatura de psicología social, comunicación y sociología son relevantes para el estudio de la interacción humano-computadora y tienen claras implicaciones en el diseño de interfaces de usuario".

Una de las características humanas que tiene gran importancia en la manera en que socializamos es el humor, la cual se caracteriza por ser sumamente compleja. Ha sido estudiado en diversas áreas del conocimiento como la psicología, filosofía, lingüística, sociología y literatura (Mulder & Nijholt, 2002). Diversos autores han hecho aseveraciones sobre la importancia de éste: Reyes, Rosso & Buscaldi (2012) afirman que "la principal función del humor es liberar emociones, sentimientos o sensaciones que impactan positivamente en la salud humana" (p. 2); Barsoux (1993) afirma que se usa para entretener, liberar tensión, incrementar la unión, disfrazar la ignorancia, velar las críticas y lograr la cooperación; Binsted (1996) dice que una computadora podría usar el humor para los mismos fines que un humano.

A veces se piensa que el estudio del humor en el área computacional es simplemente curioso o divertido pero no importante. Binsted (1996) dice que la mayoría de los investigadores dedicados a la Inteligencia Artificial (IA) estaría de acuerdo con lo que dice Minsky (1961) que un objetivo adecuado para la investigación en IA es tener una computadora que realice "…una tarea la cual, si se hace por humanos, requiere inteligencia para desempeñarla". Siendo así, entonces la mayoría de la experiencia humana está abierta a la investigación. La fluidez lingüística requiere la habilidad de usar y entender lenguaje no literal, como metáforas, humor, exageración, etc. Si queremos ser capaces de hablar fácilmente con las computadoras (y que ellas nos hablen también), éstas deben ser capaces de usar y entender el humor (Binsted, 1996).

En el área computacional hay pocos estudios sobre el humor. Sin embargo, recientemente los lingüistas computacionales han hecho progresos considerables en el modelado de formas

lingüísticas relacionadas al humor como la metáfora y la analogía. Con esto se han logrado progresos concretos hacia el entendimiento de cómo funcionan estos fenómenos y cuál es el rol que juegan en el lenguaje. Por tal razón es tiempo de hacer lo mismo con el humor (Binsted, 1996).

Con el presente trabajo se aborda un tipo de humor muy peculiar y propio de México: el albur, el cual comúnmente es definido como un juego de palabras que utiliza el "doble sentido" para someter sexualmente de manera verbal al escuchante. En los últimos años se han hecho algunas investigaciones sobre este fenómeno en distintos campos del conocimiento como la lingüística y la sociología, sin embargo, no se había provisto de un método formal que describiera al albur y que por tanto sea posible implementarlo computacionalmente.

En esta tesis se aborda el problema de identificación automática de albur desde una perspectiva lingüística computacional. Se retoma el conocimiento generado en otras áreas en las que ya se ha estudiado anteriormente el albur y se lleva al campo computacional. Se identifican las características clave que conforman al albur, así como las estructuras más utilizadas en este tipo de humor.

1.1 Planteamiento del problema

El humor se ha estudiado en distintas áreas del conocimiento como la psicología, la filosofía, la lingüística, sociología y literatura (Mulder & Nijholt, 2002), sin embargo, en el campo computacional hay pocos estudios sobre éste. Algunos de ellos se han hecho sobre tipos específicos de humor: Binsted estudió los *punning riddles,* Taylor & Mazlack lo hicieron con los chistes *knock, knock*, Mihalcea & Strapparava con los *one-liners*. En este trabajo se hace lo propio con un tipo de humor en específico del que se presume no tiene equivalente en ningún idioma: el albur, el cual nace y se practica en México desde hace mucho tiempo.

El albur ha sido objeto de investigaciones en diversas áreas como Ciencias Sociales y Humanidades (Olguín Martínez, n.d.), Comunicación (Díaz, 2001), Lingüística (Lavertue, 1998), Ciencias Políticas y Sociales (Durán González, 2012), entre otras y, recientemente en el área computacional, siendo muy escasos los trabajos y apenas se tratan de aproximaciones generales (Guzmán, Beltrán, Tovar, Vázquez, & Martínez, 2014; Ocampo Pólito, 2010). Aunque en los trabajos mencionados ya se habían identificado la mayoría de las características clave del albur, éstas no habían sido formalizadas y por tanto no se habían implementado en un programa computacional (a excepción de Ocampo Pólito que realizó una primera aproximación), meta que se logra con esta investigación.

A menudo, el albur se define como un juego de palabras en "doble sentido", el cual, para comprenderlo y generarlo, es necesario: conocer todas aquellas palabras de argot sexual; además, de manejar otros fenómenos lingüísticos como los metaplasmos, usados como técnicas para formar nuevas palabras, por adición, supresión y transposición; e incluso la reestructuración de oraciones (calambures) y habilidad para identificar la contracción, sinéresis y sinalefa. Todo esto con el fin de que el interlocutor pase por inadvertida la connotación sexual de la frase.

En esta investigación se desarrolló un método para la detección automática de albures escritos en textos. Para resolver este problema se abordó un enfoque lingüístico que abarca los niveles fonético y semántico del lenguaje principalmente. Con la implementación de este método se realiza la identificación automática de albures ya sean escritos deliberadamente o no en un texto.

1.2 Justificación

Investigadores de los tiempos de Aristóteles y Platón hasta nuestros días se han esforzado por descubrir y definir el origen del humor. Hay casi tantas definiciones del humor como teorías (Latta, 1999). El humor ha sido estudiado en distintas áreas del conocimiento como la lingüística, psicología, sociología, etc. Sin embargo, en el área computacional se han realizado apenas algunos trabajos y pocos han sido para el idioma español.

La utilidad del humor computacional y sus motivaciones han sido discutidas varias veces. Se dice que "el humor computacional puede hacer que las computadoras sean más amigables con el usuario, más persuasivas, simpáticas y competentes mejorando la interacción humano-computadora en general, desarrollando mejores agentes inteligentes, mejorando sistemas de aprendizaje de un segundo idioma, publicidad electrónica y comercio electrónico"(Taylor, 2009, p. 429). Es importante que los sistemas que reconocen el lenguaje natural sean capaces de identificar cuándo un usuario está usando lenguaje figurativo, ya sea ironía o humor, para manejar adecuadamente su petición (Barbieri & Saggion, 2014).

Raskin sugiere que "una aplicación buscará humor no intencional, perversión del texto, por ejemplo, en una dirección presidencial, una nota diplomática o cualquier otro asunto sumamente importante. Por otro lado, las mismas aplicaciones de detección de humor pueden ser usadas para determinar los puntos vulnerables en un texto que puedan ser denigrados, por ejemplo en una campaña política y entonces trabajar en conjunto con la generación del humor para crear humor apropiado y efectivo" (Raskin, 2002). Taylor (2010) menciona que la detección de humor puede ser usada para descartar información relevante en una búsqueda en la Web.

En el área computacional se han realizado investigaciones sobre el humor en idiomas como el inglés, chino y catalán. Sin embargo, en el idioma español no existen muchos trabajos relacionados a este campo, a pesar de ser uno de los idiomas más hablados en el mundo. Según el Instituto Cervantes (2015), el español es una lengua hablada por más de 540 millones de personas como lengua nativa, segunda o extranjera. Por lo tanto es importante realizar más trabajos sobre este idioma en el área computacional. En el caso del albur, Beristáin menciona que cada vez hay más interés en él, además de otros tipos de lenguaje coloquial como las groserías y el dicho, entre otros, sobre todo por traductores literarios.

En esta tesis se realiza una investigación para la detección automática del albur abordando un enfoque lingüístico computacional. Los científicos de este campo se interesan en proveer modelos de distintos tipos de fenómenos lingüísticos. Estos modelos pueden ser *knowledge-based* (hecho a mano) o *data-driven* (estadístico o empírico).

El trabajo en la lingüística computacional es en algunos casos motivado por una perspectiva científica que trata de proveer una explicación computacional de un fenómeno en particular ya sea lingüístico o psicolingüístico; y en otros casos la motivación puede ser puramente tecnológica que desea proveer un componente funcional de un sistema de voz o de lenguaje natural (ACL, 2016). En este trabajo, se identifican las características clave que componen al albur: argot sexual, metaplasmos, estructuras sintácticas, polisemia, ambigüedad fonética y léxica, y se implementan en un programa computacional a fin de realizar identificación automática de éste en textos.

1.3 Objetivos

1.3.1 Objetivo general

- Diseñar un método para la identificación automática de albures en textos escritos en español, basado en un enfoque fonético y semántico.

1.3.2 Objetivos específicos

- Conformar un diccionario de argot sexual a fin de consultarlo durante el proceso de identificación automática de albures.
- Desarrollar un módulo de procesamiento fonético para manejar los fenómenos que se efectúan en ese nivel del lenguaje en el fenómeno del albur.
- Desarrollar un módulo de procesamiento semántico a fin de encontrar posibles significados sexuales en las oraciones y palabras de un texto.

Capítulo 2 Marco teórico

En este capítulo se presentan algunos conceptos que el lector debe conocer para comprender el resto del documento.

2.1 Conceptos lingüísticos

2.1.1 Pronombres átonos

Los pronombres personales átonos son aquellos que funcionan como complemento verbal no preposicional (Ya TE lo he dicho) o como formante de los verbos pronominales (Ahora ME arrepiento). Precisamente por su carácter átono, se pronuncian necesariamente ligados al verbo, con el que forman una unidad acentual.

Estos pronombres carentes de independencia fónica se denominan, en general, *clíticos*: cuando anteceden al verbo (ME encanta; LO dijo; SE fue) se llaman *proclíticos*; cuando siguen al verbo (ayúdame, díselo, vete) se llaman *enclíticos* (Diccionario panhispánico de dudas: Real Academia Española, 2005).

2.1.2 Español convencional

A lo largo de este documento se utiliza el término "español convencional" para referirse a todas aquellas palabras que se consideran "válidas" en este idioma y que son oficialmente reconocidas por las autoridades de la lengua, como la Real Academia Española y el Colegio de México.

2.1.3 Argot sexual mexicano

Este término se utilizará para referirse a todas aquellas palabras que utilizan los hablantes mexicanos como parte del argot sexual y que normalmente no se encuentran en los diccionarios del español.

2.1.4 Homografía

Coincidencia en la escritura y la pronunciación de dos palabras que tienen distinto significado y distinta etimología.

2.2 Conceptos de Lingüística Computacional

2.2.1 Token

Es la aparición concreta de cada palabra en un texto. Por ejemplo, la separación en tokens de la frase "Ciudad de México" sería: "Ciudad", "de", "México". Sin embargo con procesos especializados como el reconocimiento de entidades nombradas estas palabras serían "tokenizadas" como una sola: "Ciudad_de_México".

2.2.2 Atributos de la palabra

Según la Real Academia Española (2016), la palabra es una unidad lingüística dotada generalmente de significado, que se separa de las demás mediante pausas potenciales en la pronunciación y blancos en la escritura.

En este trabajo se adopta la nomenclatura que comúnmente se utiliza en el área de PLN. Así, además de la definición antes mencionada se emplean otros atributos que son útiles computacionalmente. Debido a que en la implementación se utiliza la librería de *Freeling* se adopta el objeto *Word* de ésta y sus atributos se describen a continuación.

Tabla 2.1 Atributos de la palabra

Atributo	Descripción
Forma de palabra	Es la forma "original" en que una palabra aparece en un texto y necesaria para crear un nuevo objeto *Word*.
Lema	Es la forma única común a todas las posibles variaciones de una misma palabra. En sustantivos se utiliza el masculino en singular de la palabra y para los verbos su forma en infinitivo. Por ejemplo, el lema de "tomar", "tomemos" o "tomen" es "tomar"; el lema de "perro", "perros" o "perrito" es "perro".
Etiqueta gramatical	Es una serie de símbolos que proveen información gramatical de la forma de palabra. Indica por ejemplo si es un verbo, sustantivo, adjetivo, etc.
Forma fonética	Es la representación del sonido de la palabra por medio de símbolos de un alfabeto fonético. Se utiliza una serie de reglas fonológicas para generar la forma fonética, a este proceso se le llama "transcripción fonética".
Forma semántica	Es la representación del significado de la palabra. Esta forma semántica no es descriptiva, es decir, no es un texto en el que se describe el significado de la palabra. La forma más común de representar la forma semántica es a través de un synset, el cual es un código único para cada concepto del mundo.

2.2.3 Part-of-Speech Tagging

Es la asignación de etiquetas gramaticales a cada una de las palabras de un texto. Este proceso se realiza dependiendo del contexto en que se encuentra cada una.

2.2.4 Synset

Según Saussure (1945), un signo lingüístico se compone de dos elementos: el significante y el significado. El primero corresponde a la representación escrita o hablada de un concepto, es decir, la palabra. El segundo corresponde al concepto en sí, a la imagen mental que se tiene del concepto.

A nivel computacional, para representar los significados en lugar de imágenes mentales se utilizan códigos llamados *synsets*. Un *synset* es un código único que representa un concepto en específico, por ejemplo, el *synset* de "perro" es 02084071-n. A este *synset* se le asocia un conjunto de lemas que adquieren este significado. En el caso de este *synset*, "perro" y "can" serían los lemas asociados.

2.2.5 Freeling

Freeling es una librería que provee funcionalidades de análisis del lenguaje (análisis morfológico, detección de entidades nombradas, etiquetado *PoS (Part of Speech)*, *parsing*, desambiguación semántica, etc.) para una variedad de lenguajes (inglés, español, portugués, italiano, francés, alemán, ruso, catalán, galés, croata, esloveno, entre otros). Aunque *Freeling* está escrita en C++ también se proveen *API*'s para utilizarla desde otros lenguajes como PHP, Python y Java.

2.3 Nivel semántico

2.3.1 Sinonimia

Relación de igualdad que hay entre el significado de dos o más palabras o enunciados.

2.3.2 Polisemia

Es un concepto que se refiere a que una palabra tiene más de un significado. Debido a esto una palabra puede ser ambigua si no se considera el contexto en el que aparece.

2.3.3 Metáfora

Figura retórica de pensamiento por medio de la cual una realidad o concepto se expresan por medio de una realidad o concepto diferentes con los que lo representado guarda cierta relación de semejanza. Las metáforas tratan siempre de explicar de mejor manera la realidad, por ejemplo: "tus dientes son como perlas".

2.3.4 Eufemismo

Un eufemismo es una palabra o una expresión utilizada para sustituir una palabra que socialmente se considera ofensiva o de mal gusto. Pueden sustituir términos de diversos tipos, por ejemplo en palabras que pueden resultar groseras, escatológicas u obscenas.

2.3.5 WordNet

WordNet es una extensa base de datos en la que verbos, sustantivos, adjetivos y adverbios se encuentran agrupados en conjuntos de sinónimos cognitivos llamados *synsets*. Cada synset expresa un concepto distinto, tal como se explica en 2.2.4. La estructura de WordNet la convierte en una herramienta útil para la lingüística computacional y el procesamiento del lenguaje natural.

2.3.6 Multilingual Central Repository

Multilingual Central Repository (MCR) integra en el mismo *framework EuroWordNet,* wordnets de seis lenguajes diferentes: inglés, español, catalán, vasco galés y portugués. El índice interlengua (*Inter-Lingual-Index, ILI)* permite la conexión entre palabras de un lenguaje a traducciones equivalentes de cualquiera de los otros lenguajes gracias a las correspondencias generadas automáticamente entre las versiones de WordNet.

2.4 Nivel fonético

2.4.1 Calambur

El calambur consiste en realizar un agrupamiento de sílabas diferente al original para formar una nueva frase. Por ejemplo, la oración “yo lo quito” puede reestructurarse para formar la oración “yo loquito” la cual adquiere un significado totalmente diferente. La comicidad de este tipo de juego de palabras consiste en exponer la frase que inicialmente se encontraba oculta debido a la homofonía y al significado de dicha frase inicial.

2.4.2 Sinalefa

La sinalefa se produce al pronunciar en una sola sílaba las vocales final e inicial de dos palabras contiguas. La combinación y coarticulación de los sonidos da a lugar a una nueva palabra. Ejemplo: al pronunciar la frase “Memo Herdez” los sonidos /mo/ y /herdez/ dan lugar a “muerdes” (Lavertue, 1998, p. 33).

2.4.3 Sinéresis

”Reducción a una sola sílaba, en una misma palabra, de dos vocales contiguas que normalmente se pronuncian en sílabas distintas; p. ej. “aho-ra” por “a-ho-ra” (Real Academia Española, 2014).

2.4.4 Contracción

Ésta sucede al contraer y suprimir una de las "dos" vocales idénticas que entran en contacto por sinalefa. Ejemplo: al pronunciar la frase "te echo" se da lugar a la palabra "techo" (Lavertue, 1998).

2.4.5 Homofonía

"Dicho de una palabra: Que suena igual que otra, pero que tiene distinto significado y puede tener distinta grafía. Aunque se distinguen ortográficamente, tubo y tuvo son homófonos" (Real Academia Española, 2014).

2.5 Albur

En México, existe un fenómeno lingüístico bastante peculiar denominado "albur". A pesar de ser muy popular, a menudo se le confunde con otras formas de expresión coloquial como el doble sentido, los chistes léperos, los chistes blancos, los retruécanos, la grosería, el insulto y algunos elementos del caló (Lavertue, 1998).

A pesar de que no es del agrado de todas las personas, se trata sin duda de un tipo de humor. Beristáin (2000) aclara que "no se trata de un intercambio de insultos u ofensas, pues no provoca enojo y, por el contrario, funge como una especie de fiesta improvisada que adopta la forma de una competencia a base de filosa pericia imaginativa, pero esgrimida con espíritu deportivo" (p. 410).

2.5.1 Definición del albur

La mayoría de las personas lo definen como un juego de palabras en doble sentido, pero el albur va mucho más allá de la conjunción de estos dos elementos. Algunos exponentes y estudiosos han propuesto diferentes definiciones, sin embargo, la más acertada es la de Beristáin (2000, p. 410) que se cita a continuación:

> "El verdadero albur es el juego de esgrima intelectual, verbal, regido por normas situacionales, que funciona en grupos masculinos configurados por antagonistas y jueces (es decir, dotado de jugadores), que se realiza a base de expresiones de doble sentido que aparentan manifestar una idea anodina, inocua y al alcance de todos, cuando en realidad operan como detonadores al desatar el inicio de la construcción interactiva de un mensaje secreto, cifrado, que alude a las funciones del cuerpo y al acto sexual, que está dirigido a quienes sean capaces de descifrarlo y que tiene la estructura del diálogo".

Esta definición es bastante completa puesto que abarca distintos elementos del albur que otras no lo hacen. A continuación se hace un breve análisis de los puntos más importantes de ésta:

- "El albur es un juego de esgrima intelectual, verbal": indica que éste tiene un formato de conversación.
- "Funciona en grupos masculinos": aunque actualmente se dice que el albur también es utilizado por grupos femeninos, la definición hace referencia a que se trata de que el vencedor humilla al vencido porque simbólicamente lo somete y lo sitúa en un rol sumiso sexual. Según Beristáin es un juego "machista" porque el vencedor constata y reafirma su calidad masculina al "rebajar" a la feminidad la calidad viril del vencido.
- "Se realiza a base de expresiones de doble sentido que aparentan manifestar una idea anodina, inocua y al alcance de todos": esto corresponde al conjunto de palabras de argot sexual que se utiliza en el albur. Lo que sucede es que algunas palabras del español convencional adquieren un significado sexual que no todos conocen, esto provoca que las personas que no conocen este significado alterno no se percaten del mensaje real que se transmite.
- "Construcción interactiva de un mensaje secreto": en el albur, además del argot sexual se utilizan otros recursos para lograr que el mensaje real pase desapercibido por los escuchantes. Comúnmente de manera general se les refiere como juegos de palabras, sin embargo, éstos a su vez utilizan otros fenómenos lingüísticos como la sinéresis, contracción, sinalefa, calambures, etc.
- "Alude a las funciones del cuerpo y al acto sexual": existen muchos tipos de juegos de palabras en el mundo, sin embargo, el albur es exclusivamente sexual. Es decir, los mensajes que se transmiten de manera oculta en el albur son siempre de carácter sexual y son en su mayoría sobre actos sexuales.

2.5.2 Connotación sexual

A menudo se confunde al albur con el "doble sentido", sin embargo, son fenómenos distintos. El albur utiliza el "doble sentido" para lograr el ocultamiento y pasar desapercibido o para ser más cómico. Pero el "doble sentido" no es necesariamente albur, de hecho, el primero puede existir sin ningún trasfondo sexual como lo aclara Lavertue (1998, p.46) con los siguientes ejemplos:

> *El **doble sentido** se asemeja al albur por la carga de connotación o de doble significado de las palabras. Esta connotación puede ser sexual como por ejemplo:*
>
> *Kimono Oyito > ¡Qué mono hoyito! > ¡Qué bonito es tu ano!*

Pero también puede haber doble sentido sin ningún trasfondo sexual como por ejemplo el que se hacer con el nombre del general japonés:

Sasse Komokenoye > Se hace como que no oye

2.5.3 Juegos de palabras

Existen varios tipos de humor, algunos más complejos que otros, pero en la mayoría se observa que los juegos de palabras se utilizan a menudo para provocar el efecto deseado. Estos juegos de palabras son un recurso tan utilizado que incluso se pueden encontrar en anuncios publicitarios, marcas reconocidas, nombres de negocios, etc., a fin de atraer la atención del espectador. Esto aplica no sólo para el idioma español, sino también para otros idiomas como el inglés (Binsted, 1996; Mihalcea & Strapparava, 2005; Partington, 2009) y el chino (Ren, Kaji, Yoshinaga, & Kitsuregawa, 2013).

Los juegos de palabras son un tipo de humor verbal ampliamente utilizado en todo el mundo. Comúnmente se valen de la homofonía o de frases muy similares en sonido, además de la ambigüedad en los significados de las palabras. En el área de Procesamiento de Lenguaje Natural normalmente esto representa una deficiencia. Sin embargo, en el humor computacional puede verse como una ventaja que puede ser aprovechada en la generación e identificación de estos textos.

Los chistes, normalmente, tienen una preparación y un remate. La preparación crea ciertas expectativas y el remate las rompe, conduciendo así a diferentes interpretaciones de la preparación (Ritchie, 1997). Cuando se trata de chistes que involucran juegos de palabras sucede lo mismo. Los chistes de juegos de palabras, o chistes que involucran juegos verbales, son una clase de chistes que dependen de palabras que son similares en sonido, pero se usan con diferentes significados. La diferencia entre los significados crea un conflicto que rompe la expectativa y es humorística. El juego de palabras puede ser creado entre dos palabras con la misma pronunciación y ortografía, entre dos palabras con diferente ortografía pero misma pronunciación y, con dos palabras con diferente ortografía y pronunciación similar (Taylor & Mazlack, 2004).

Como ejemplo tenemos el chiste “le pedí un café y me dijo ‘sólo queda té’, fue hermoso”. La primera frase “le pedí un café” establece cierta expectativa; la segunda frase “me dijo ‘sólo queda té’” es una respuesta aceptable y encaja con la primera; sin embargo, la última frase “fue hermoso” fuerza a interpretar la segunda de manera diferente, se sitúa ahora en un contexto romántico. La segunda frase, entonces, debe leerse como “sólo quédate” que es homófona a la original. Se ha logrado pues romper la expectativa y se produce así un efecto cómico o divertido.

Algunos juegos de palabras están basados en la homofonía entre dos palabras diferentes en significado y/o diferente ortografía, o bien, entre dos frases que cumplan estas condiciones. De esta manera, los juegos de palabras se prestan a diferentes interpretaciones e incluso a construcciones léxicas diferentes dependiendo, de cómo suena la frase. En la mayoría de los casos esta homofonía es producida por los fenómenos lingüísticos que se describen enseguida.

Los fenómenos más frecuentes en los juegos de palabras son la sinalefa, la sinéresis y la contracción. Éstos consisten en pronunciar como uno solo dos sonidos contiguos dentro de una palabra (sinéresis) o, del final de una y el principio de otra (sinalefa). La contracción se presenta cuando dos sonidos contiguos iguales se pronuncian como uno solo. Por ejemplo /t//e//ch//o/ en lugar de "te echo" que, en el plano acústico, son difíciles de desambiguar. Incluso algunos sistemas de reconocimiento de voz tienen problemas para hacerlo.

En los juegos de palabras también deben considerarse los metaplasmos, como aféresis y apócope. Y es que no siempre la frase latente coincide completamente con la frase original, ya sea en su forma escrita o a nivel fonético. La aféresis consiste en la desaparición de uno o más fonemas al principio de una palabra, mientras que la apócope se realiza al final de una palabra. Como ejemplo tenemos un chiste en el que un panqué de chocolate hincado frente a una fresa le dice "mi corazón chocolate por ti". El juego de palabras está en usar "chocolate" por "late", siendo la primera el sabor del panqué. Este juego de palabras es lo que hace divertida a la imagen.

Por tanto, para comprender los juegos de palabras hay que realizar una construcción léxica diferente. La nueva frase debe ser homófona a la inicial o muy similar. También puede ser que la representación fonética de esta nueva frase coincida solo en parte a la de la original. Esta última condición correspondería a apócope y aféresis.

2.5.3.1 Metaplasmos

Un metaplasmo es una figura de dicción que consiste en alterar la escritura o pronunciación de las palabras sin alterar su significado. Éstos pueden ser de tres tipos: por adición, por supresión o por transposición.

En el albur, los metaplasmos se usan frecuentemente para formar eufemismos que se refieren a palabras sexuales. En la tabla siguiente se muestran los diferentes tipos de metaplasmos.

Tabla 2.2 Tipos de metaplasmos (Janer, 1919)

Metaplasmos		
Por adición	Prótesis	Se agrega una letra o sílaba al comienzo de la palabra.
	Epéntesis	Se agrega una letra o sílaba al medio de la palabra.
	Paragoge	Se agrega una letra o sílaba al final de la palabra.
Por supresión	Aféresis	Consiste en quitar una letra o sílaba al principio de la palabra.
	Síncopa	Consiste en quitar del medio de una palabra alguna de sus letras o sílabas.
	Apócope	Consiste en quitar del final una palabra una letra o sílaba.
	Haplología	Consiste en la eliminación, en una palabra, de una sílaba semejante o parecida a la que sigue.
Por transposición	Metátesis	Se aplica al intercambio de la posición de los fonemas vocálicos o consonánticos.
	Disimilación	Cuando dos sonidos de una misma palabra tienden a diferenciarse.
	Asimilación	La asimilación consiste en que un segmento se articula al sonido de otro segmento adyacente o cercano.

2.5.4 Argot Sexual Mexicano

El argot (también llamado jerga o caló) es una modalidad lingüística usada en contextos específicos. El argot es hablado por personas con algo en común, por ejemplo ocupación, profesión, región geográfica, status social, etc. Hay una amplia variedad de argots para cada lenguaje alrededor del mundo. Algunas veces el argot es usado para ocultar el significado real de una palabra o frase pero también se usa con fines humorísticos en chistes, conversaciones y otros escenarios de la vida diaria.

En México, uno de los más populares es el argot sexual, y es tan rico que incluso puede tener algunas variaciones dependiendo de la región. Esto se debe principalmente a la cultura local o la mezcla del español convencional con idiomas locales o ambas. El argot sexual es definitivamente una parte fundamental del idioma y la cultura popular. En el albur, el argot sexual es un elemento fundamental para su comprensión y, el conocimiento de éste es determinante para su identificación. Por lo tanto, para desarrollar un sistema automático de identificación de albures es necesario conformar un diccionario de argot sexual mexicano que pueda utilizarse computacionalmente. A continuación se hace un breve repaso de los fenómenos lingüísticos más utilizados en el albur a nivel fonético, morfológico y semántico.

En el argot sexual existen algunas palabras que son producto de la creatividad y el ingenio mexicano y no se encuentran en los diccionarios oficiales del idioma español por ser

consideradas tabú. Sin embargo, la mayoría de las palabras del argot sexual provienen del español convencional, es decir, existen en los diccionarios pero su significado no es sexual. De hecho, es por esto que muchas veces pasan desapercibidas por los hablantes que desconocen este argot.

Las palabras del argot sexual han sufrido modificaciones, adaptaciones y/o corrupciones al transmitirse entre los hablantes que las usan. Algunas veces para mantener oculto el significado sexual de éstas y otras veces para hacerlas más cómicas. A continuación se describen brevemente las técnicas y fenómenos lingüísticos más utilizados en el argot sexual mexicano.

2.5.4.1 Eufemismos

Un eufemismo es una palabra que se usa en lugar de otra por considerarse ofensiva, vulgar o inapropiada. Por ejemplo, para decir que alguien ha muerto se utilizan frecuentemente los siguientes eufemismos: "colgó los guantes/tenis", "pasó a mejor vida", "se nos adelantó", "se petateó", por nombrar algunos.

La forma más común de formar un eufemismo es hacer analogías o metáforas. Para esto debe existir una relación entre ambos conceptos, ya sea similitud en forma, en tamaño, en uso, color, etc. Esto es algo que depende mucho de la cultura y creatividad de los hablantes.

En el caso del argot sexual se encuentran ejemplos como "lavar a mano", que es una metáfora de la masturbación, o "huevos", que es una metáfora para referirse a los testículos. A veces una palabra eufemística se vuelve tan popular que su significado sexual predomina sobre el original, entonces se dice que se ha vuelto un disfemismo. A continuación se muestran algunos ejemplos de eufemismos que muy probablemente el lector identifique como disfemismos según su conocimiento de argot sexual.

Tabla 2.3 Ejemplos de eufemismos

Eufemismo	**Significado original**	**Significado sexual**
Papaya	Fruta	Órgano sexual femenino
Pajarito	Ave	Falo
Verga	Vara (palo largo y delgado)	Falo
Panocha	Pan dulce/piloncillo	Órgano sexual femenino
Huevo	Cuerpo redondeado que contiene el germen de un embrión	Testículo

2.5.4.2 Corrupciones

Una corrupción se realiza cuando en una palabra se cambian una o más letras (y por tanto fonemas) para producir otra diferente. Esta nueva palabra conserva el significado de la palabra original. En el argot sexual, "ano" ha sido corrompida para formar una nueva palabra: "anís". De esta manera, "anís" adquiere, además del propio, un significado sexual.

Estas técnicas se aplican tanto en palabras del español convencional como en las que ya forman parte del argot sexual. De esta manera se realizan combinaciones entre éstos, por ejemplo, eufemismos y metaplasmos: "Aniceto" (nombre propio) es el resultado de aplicar paragoge a "anís", el cual a su vez es una corrupción de "ano".

Capítulo 3 Estado del arte

A continuación, se presentan los trabajos relacionados más representativos en la revisión de la literatura. Los dos primeros (Guzmán et al., 2014; Ocampo Pólito, 2010) están directamente relacionados con el albur. Los cinco siguientes abordan el problema del reconocimiento automático de algún tipo de humor en específico, por lo cual se muestran ejemplos de textos humorísticos en cada uno. Enseguida, un interesante trabajo sobre recolección de textos con profanidades e insultos provenientes de proyectos FLOSS (*Free/Libre Open Source Software*). Luego se encuentra DEviaNT, que es un enfoque con el que se trata de resolver el problema del doble sentido aplicado a los chistes TWSS (That's What She Said). Por último se encuentra JAPE (Binsted, 1996), el cual a pesar de ser un trabajo de generación de humor y no de reconocimiento, resulta interesante el enfoque lingüístico que se aborda para resolver el problema.

3.1 Clasificación de frases obscenas o vulgares dentro de tweets

En este trabajo Guzmán et. al (2014) propusieron modelos para clasificar frases obscenas y vulgares en textos cortos en español. También generaron estadísticas sobre cuáles son las palabras de este tipo más usadas y además se muestran cuáles son los estados de la República que más utilizan este tipo de lenguaje.

Con la ayuda de un programa desarrollado en *Python* y la *API* de *Twitter*[1] conformaron un *corpus* de 548,243 tuits. Estos textos fueron escritos por un total de 173,339 usuarios, publicados durante un cierto día y geo-localizados dentro de un radio no mayor a 10 kilómetros de la capital de cada estado de la República Mexicana.

Con la ayuda de un diccionario de mexicanismos y la herramienta WEKA construyeron dos modelos de clasificación: *obscenidad y ninguna* y, *vulgaridad y ninguna*. El diccionario de

[1] Red social https://twitter.com/

mexicanismos está dividido en dos partes: la primera, formada de palabras vulgares y etiquetadas con la leyenda *VULG*; la segunda, compuesta de palabras obscenas etiquetadas con la leyenda *OBSC*.

Como resultados de la clasificación obtuvieron un 91.07% en el caso de *obscenidad y ninguna* y un 98.90% en el caso de *vulgaridad y ninguna*, utilizando el algoritmo SMO[2] con la herramienta WEKA.

En la construcción de los albures muchas veces se utilizan palabras vulgares y obscenas. A pesar de que este trabajo no resuelve ningún problema de detección de humor, resulta interesante la utilización de un diccionario de mexicanismos etiquetado con las leyendas mencionadas.

3.2 DAHTCE

Uno de los pocos trabajos realizados sobre el humor en el idioma español es el de Ocampo (2010), en el cual se presentó la herramienta DAHTCE (Detector Automático de Humor en Textos Cortos en Español).

Este software procesa por separado dos archivos: uno humorístico y otro no humorístico y le asigna una ponderación a cada palabra de entrada. El peso que se le asigna a cada una está dado por los módulos con los que cuenta la herramienta: albur, contenido adulto, rima y aliteración. A continuación se muestra en la tabla 3.1 la ponderación que utiliza este software para clasificar los tipos de humor.

[2] *Sequential Minimal Optimization* es un algoritmo para entrenar máquinas de soporte vectorial con el que se resuelve el problema de programación cuadrática (*Quadratic Programming, QP*).

Tabla 3.1 Ponderación que asignan los módulos de DAHTCE a las palabras

Valor de la palabra	Atributos de la palabra
-1	Caracter especial
0	Sin atributos
1	Albur
2	AdultSlang
3	Aliteración
4	Albur, Aliteración
5	AdultSlang, Aliteración
6	Rima
7	Albur, Rima
8	Adulto, Rima
9	Aliteración, Rima
10	Albur, Aliteración, Rima
11	AdultSlang, Aliteración, Rima

El módulo de albur realiza la identificación con base en un diccionario de términos frecuentemente utilizados en dicho tipo de humor. El número de palabras del diccionario es muy reducido (alrededor de 200 palabras). Este módulo no realiza más que una búsqueda de palabras mayores a tres caracteres en el texto y luego compara cada palabra con el diccionario antes mencionado para hacer la ponderación.

La detección de albur en este trabajo es muy básica. *Para el caso del albur, no implementa el cálculo para cuando existen juegos de palabras ni semejanza homófona* (p.64) menciona el autor. Es decir, simplemente se cuentan las palabras que se encuentran en el diccionario de términos de albures pero no se hace un análisis de pronunciación, ortografía ni mucho menos semántica.

3.3 Reconocimiento de chistes knock, knock

En este trabajo Taylor & Mazlack (2004) propusieron una metodología para el reconocimiento de chistes *knock, knock*. Realizan la detección de estos chistes utilizando únicamente técnicas estadísticas de reconocimiento del lenguaje. Es decir, no utilizan ningún tipo de estructura abstracta para representar el conocimiento del mundo exterior.

Al igual que los albures, los chistes *knock, knock* son un tipo de humor verbal que utiliza juegos de palabras para producir el efecto hilarante. Dicho juego verbal se basa principalmente en la paronimia, homonimia y la homofonía.

Los chistes *knock, knock* son textos cortos que tienen una estructura bien definida y por tanto son más fáciles de estudiar. Los autores identificaron durante su investigación tres tipos de chistes *knock, knock*. En la tabla 3.2 se muestra un ejemplo del tipo de chiste que es objeto de estudio de su investigación.

Tabla 3.2 Ejemplo de chiste knock, knock

Chiste *knock, knock*	**Explicación**
Persona 1: *knock, knock* Persona 2: *Who's there?* Persona 1: *Water* Persona 2: *Water who?* Persona 1: *Water you doing tonight?*	(Debido a la naturaleza del chiste no se muestra traducción del mismo) El efecto cómico se produce por la semejanza en la pronunciación de la palabra *water* con *what are* en inglés, por lo que en la última línea se complementa la línea 3 con la frase *water you doing tonight?* que suena muy similar a *what are you doing tonight?*

La herramienta[3] desarrollada cuenta con un generador de secuencias de juegos de palabras, la cual, con la ayuda de una tabla de similitud, produce una frase B similar en pronunciación a una frase A pero con diferente significado. La tabla de similitud es producto de una modificación heurística de la *tabla de Frisch* (Frisch, 1997).

Para probar la eficiencia del programa reunieron dos conjuntos de textos: uno conformado por 130 chistes *knock, knock* y otro conformado por 66 textos no humorísticos sintéticos, es decir, textos que originalmente eran chistes de este tipo pero a los que les modificaron la última línea para que ésta tuviera sentido con la línea 3, esto con el fin de evitar que el texto fuera gracioso y así mantener una estructura similar a los demás textos.

El programa fue capaz de encontrar el juego de palabras (contenido en la última línea del chiste) en 85 de 122 chistes. La prueba con el segundo conjunto de textos (los no humorísticos) es mucho más alentadora, puesto que el programa reconoció exitosamente 62 textos no humorísticos de los 66 disponibles.

[3] http://www.azkidsnet.com/JSknockjoke.htm

3.4 Reconocimiento automático de one-liners humorísticos

En este trabajo Mihalcea & Strapparava (2005) realizaron un trabajo de reconocimiento automático de *one-liners* humorísticos utilizando un enfoque de *aprendizaje automático*. Es decir, utilizaron técnicas de clasificación de textos para distinguir entre aquellos que son humorísticos y los que no.

Los *one-liners* humorísticos son oraciones con una longitud de entre 10 y 15 palabras. Tienen una interesante estructura lingüística: sintaxis simple, uso deliberado de recursos retóricos y creativas construcciones lingüísticas. *La estructura simple de este tipo de humor garantiza que las características que lo producen se encuentren en la primera (y única) oración (p.1)*, mencionan los investigadores. En la tabla 3.3 se presenta un ejemplo de *one-liner* humorístico y la explicación de su comicidad.

Tabla 3.3 Ejemplos de one-liners humorísticos

One-liner	**Significado**
Take my advice; I don´t use it anyway.	*Toma mi consejo; de todas maneras no lo uso.* Es común usar la frase *toma mi consejo* cuando se sugiere una solución sobre algún problema a una persona. Sin embargo, la frase *de todas maneras no lo uso* fuerza el significado de la oración anterior a tratar el sustantivo *consejo* como si éste se tratase de un objeto. La incoherencia de la frase produce el efecto cómico.

Para el conjunto humorístico recolectaron de la Web 20,000 *one-liners* diferentes por medio de un algoritmo de *bootstrapping*. A dicho algoritmo se le introdujo como *semilla* un total de diez *one-liners* identificados manualmente. Como contraejemplos, recolectaron, a través de distintas fuentes, tres conjuntos de oraciones similares en estructura a los *one-liners*: títulos del sitio de noticias Reuters[4]; proverbios; y oraciones del *BNC (British National Corpus)*; con un tamaño de 20,000 sentencias cada uno.

Los autores realizaron varios experimentos para caracterizar dos aspectos: calidad de los datos, un conjunto de 200 *one-liners* diferentes recolectados manualmente; cantidad de datos,

[4] http://www.reuters.com/news

un conjunto de 20,000 *one-liners* recolectados por medio del algoritmo de *bootstrapping* con un ruido estimado del 9%.

Los autores llevaron a cabo algunos experimentos para obtener información sobre varios aspectos de la identificación automática de humor: precisión de clasificación, tasas de conocimiento, impacto de los tipos de datos negativos usados en el proceso de aprendizaje y el impacto de la metodología de clasificación.

Como sentencias no humorísticas seleccionaron aleatoriamente 200 títulos de Reuters, 200 oraciones del BNC y 200 proverbios del conjunto contraejemplo antes mencionado. La primera serie de experimentos la realizaron sobre el conjunto de los 200 *one-liners* recolectados manualmente contra cada uno de los tres conjuntos. Llevaron a cabo seis experimentos de clasificación: *one-liners vs Reuters*, *one-liners vs BNC* y *one-liners vs proverbios*. Es importante resaltar que, por ser en cada caso un problema de decisión binario, la base de precisión del experimento es del 50%.

Los mejores resultados de precisión obtenidos en la primera fase usando pequeños conjuntos de datos no fueron tan altos: 89.75% en el caso de *one-liners vs Reuters* usando *Naive Bayes*[5], 63.75% en *one-liners vs BNC* y 70% en *one-liners vs proverbios*, ambos con el algoritmo SVM[6].

En la segunda fase de experimentos se usaron los conjuntos de datos con tamaño de 20,000 antes mencionado. En esta ocasión los resultados fueron más alentadores: 96.89% en el caso de *one-liners vs Reuters* usando *Naive Bayes* y 77.84% con el algoritmo *SVM*, los proverbios fueron omitidos por no conformar un conjunto con tal número de contraejemplos. Lo cual sugiere que, en este caso, a mayor cantidad de datos se obtiene mayor precisión en la clasificación.

[5] Un clasificador *Naive Bayes* es un clasificador probabilístico fundamentado en el teorema de Bayes.

[6] *Support Vector Machine* o máquina de soporte vectorial es un conjunto de algoritmos de aprendizaje supervisado.

3.5 Reconocimiento automático del humor en textos escolares en catalán

En este trabajo Reyes et. al (2009) realizaron un estudio acerca del Reconocimiento Automático del Humor. En éste no se desarrolló ninguna herramienta, lo que se pretendía era descubrir cuáles son las características más relevantes de los textos humorísticos, esto con el fin de utilizar dichas características para diferenciarlos de los no humorísticos.

Utilizaron un *corpus* de textos escritos en el idioma catalán por niños y adolescentes de entre 6 y 16 años de edad. Este *corpus* lleva por nombre *CesCa*[7] *(Català escolar escrit a Catalunya)* del cual se seleccionaron dos particiones: los chistes, con un tamaño de 1,867 ejemplos como datos positivos (humorísticos) y las narraciones, con un tamaño de 2,172 ejemplos como datos negativos (no humorísticos). En la tabla 3.4 se muestran ejemplos de textos (traducidos del catalán al español) extraídos del *corpus*.

Tabla 3.4 Ejemplos de textos del corpus CesCa

Chistes	Llega el niño con su padre, y le pregunta: Papá, papá, ¿cómo se escribe campana? El padre le responde: así como suena. Y el niño en su tarea escribe tan, tan, tan.
	¿Acaso piensas que me caso contigo por tus ocho millones de dote? Cómo te equivocas. Igual me casaría contigo si tuvieses nueve.

Los autores realizaron 6 tipos de experimentos: *i) perplejidad*: para saber cuán diferente es la estructura de los chistes y las narraciones; *ii) palabras clave*, se extrajeron las 100 unidades cuyo valor de *keyness* fuera lo suficientemente elevado como para ser considerada como palabra clave; *iii) información mutua*, para evaluar la probabilidad de que dos unidades formasen un patrón recurrente y no fueran un producto de la casualidad o el estilo; *iv) etiquetado taxonómico*, con el fin de obtener elementos para construir una taxonomía de esta clase de humor para trabajos futuros; *v) patrones semánticos*, para comprobar si existía algún patrón conceptual subyacente se seleccionaron las 100 palabras más frecuentes sin *stopwords* y se agruparon en las siguientes categorías: agente, tema, acción, lugar, partes del cuerpo, entes animados y otros; *vi) orientación*, buscaron evaluar si con la presencia de verbos o adjetivos con carácter negativo es posible establecer una característica útil para discriminar el conjunto positivo del negativo.

[7] http://clic.ub.edu/corpus/cesca

Para saber cuáles eran las características más relevantes de los chistes que ayudaran al reconocimiento automático de éstos, realizaron un proceso de evaluación por medio de dos clasificadores: *Bayes* y el modelo de regresión logística multinomial[8] aplicando, en ambos casos, el método de validación cruzada. Esta evaluación arrojó resultados favorables en el caso de las características de etiquetado taxonómico con casi 80% de precisión y orientación con un 85.2% de precisión.

Identificación de humor en microblog chino

En este trabajo Ren et al. (2013) realizaron un estudio de la complejidad del problema de clasificación de textos humorísticos contra no humorísticos; analizaron cuáles son las mejores técnicas y algoritmos para resolver el problema que se plantea. El objetivo era categorizar textos humorísticos de las publicaciones en *Sina Weibo* (un popular servicio de *microblogging* en China, muy similar a *Twitter*). En la tabla 3.5 se muestra un ejemplo de texto humorístico el cual muestra la complejidad del problema.

Tabla 3.5 Ejemplo de publicación humorística en el microblog Sina Weibo

Publicación humorística	**Explicación**
One cobras with high myopia dates with one elephant, after a brief greeting, the cobras says to the nose of that elephant 'you are so polite to bring me such a big pig' *Una cobra con alto grado de miopía tiene una cita con un elefante, después de un breve saludo, la cobra le dice a la nariz del elefante 'eres tan amable de traerme tan enorme cerdo'*	La cobra confunde la nariz del elefante con su contraparte. Además en el idioma chino la palabra *cobra* se escribe muy similar a *lentes*, así que el simple hecho de leer la frase *cobra con alto grado de miopía* produce una sensación divertida. Finalmente, el texto conecta con los lectores puesto que, algunos niños chinos conocen la historia de un elefante que perdió su nariz y fue confundido con un cerdo.

Los autores exploraron algoritmos de clasificación supervisada *(SVM)* y semi-supervisada *(LP),* además de paradigmas como *traducción automática, expansión de sinónimos y selección de características*, con el fin de evaluar qué combinación de método y algoritmo es la mejor para resolver el problema.

[8] La regresión logística multinomial es útil en aquellas situaciones en las que se desea poder clasificar a los sujetos según los valores de un conjunto de variables predictoras.

Para los datos de entrenamiento conformaron dos conjuntos: uno positivo (humorístico) y otro negativo (no humorístico). Para el primero, utilizaron dos fuentes diferentes: *Sina Weibo*, de donde fueron seleccionadas nueve cuentas famosas con miles de millones de seguidores y aunque la mayoría de sus publicaciones son humorísticas no se hizo ninguna depuración para eliminar las que no lo son; sitios web, portales en línea que publican historias humorísticas. Para el segundo, no humorístico, se seleccionaron aleatoriamente mensajes de *Weibo Pameng*[9].

Uno de los valores más altos de exactitud (0.733) lo lograron utilizando clasificación con un algoritmo supervisado y el método de traducción automática al idioma inglés en conjunto con la expansión de sinónimos usando la herramienta *WordNet*.

El valor más alto de exactitud (0.752) lo obtienen utilizando la combinación del método de selección de características en conjunto con un algoritmo supervisado *(SVM).*

En vista de los resultados obtenidos, los investigadores concluyen que el reconocimiento de humor en *Sina Weibo* es una tarea bastante difícil porque involucra inferir sobre temas culturales y la experiencia del lector más allá de las palabras contenidas en el texto.

Reconocimiento de ironía y humor en textos

En este trabajo, Reyes et al. (2012) realizaron múltiples experimentos a fin de determinar cuáles son las características más representativas de los textos irónicos y humorísticos. Además de realizar una discriminación entre estos dos tipos de textos. Las características que se evaluaron son: ambigüedad, polaridad, *unexpectedness* y escenarios emocionales.

Conformaron un *corpus* de 50,000 textos obtenidos de la red social *Twitter*, dividido en 5 conjuntos diferentes de 10,000 textos cada uno, cuatro de ellos etiquetados con las *hashtags #humor, #irony, #politics y #technology.* Para evaluar las características de los textos se realizaron dos fases de experimentos: *features representativeness* y relevancia de características.

Los autores realizaron cuatro experimentos de clasificación binaria para el conjunto humorístico contra cada uno de los otros y para el conjunto de ironía contra cada uno de los restantes. En estos experimentos se obtuvieron resultados alentadores, sobre todo en los que se tomaron en cuenta las cuatro características: hasta 93.13% de exactitud y 0.93 de precisión en el caso del humor; 91.97% de exactitud y 0.90 de precisión en el caso de la ironía.

Además, se realizó un experimento de clasificación multinomial el cual arrojó como evidencia la presencia de patrones subyacentes en el humor y la ironía. Como conclusión,

[9] http://cnpameng.com

demostraron que es posible clasificar el humor y la ironía de acuerdo con las características que contienen. Por ejemplo, cuando se clasifican textos de los conjuntos de humor y política, las características más informativas son perplejidad, empatía, complejidad de la oración y dispersión semántica; mientras que cuando se clasifican textos de ironía y política, las más relevantes son empatía, activación, perplejidad y desbalance contextual. También comprobaron que la estructura subyacente en el lenguaje figurado es menos predecible y probabilísticamente más ambigua que en el lenguaje literal.

3.8 FLOSS COMO FUENTE DE PROFANIDAD E INSULTOS: RECOPILACIÓN DE LOS DATOS

En este trabajo Squire & Gazda (2015) describieron un método para elaborar conjuntos de datos de prueba y de entrenamiento a fin de utilizarlos en la tarea de reconocimiento de diferentes tipos de lenguaje humano tales como el humor, sarcasmo, insultos y profanidad. También se describe el proceso de construcción de conjuntos auxiliares de datos relevantes, tales como listas de blasfemias, listas de insultos y listas de proyectos con sus códigos de conducta.

Para crear conjuntos de datos específicamente enfocados en blasfemias e insultos, los autores analizaron proyectos *FLOSS* (*Free/Libre and Open Source Software*, *Software* libre y de código abierto) puesto que la mayoría de éstos se desarrollan usando medios de comunicación que son archivados y transparentes tales como listas de distribución de correo electrónico y chat *IRC* (*Internet Relay Chat*). Algunos de los proyectos *FLOSS* involucrados en este análisis son: Apache, Debian, Ubuntu, Wordpress, Joomla!, Django, Drupal, entre otros. Se buscaron los códigos de conducta de estos proyectos *FLOSS* para saber cuáles comportamiento son considerados inadecuados e inaceptables.

En cuanto a la detección de blasfemias, analizaron algunas estadísticas básicas en sus listas de distribución de correo y canales *IRC*. Utilizaron un método para calcular la profanidad de un medio basado en la presencia de las categorías de las palabras; iniciaron con las "siete malas palabras" ("seven dirty words") que eran el fundamento de la demanda FCC v. Pacifica Foundation (438 U.S. 726, 1978). Estas últimas fueron utilizadas para explorar la profanidad en la *LKML* (*Linux Kernel Mailing List,* Lista de distribución de correo del kernel de Linux*)* y los chats *IRC*.

La parte de la detección de insultos en la que se enfoca este trabajo es en distinguir entre insultos hacia el código de *software* (de alguien más) y los insultos personales. Para tal fin, crearon una lista de oraciones de insultos extraídos de publicaciones de la *LKML* por Linus

Torvalds entre 1995-2014. Para crear esta lista, leyeron la totalidad de estas publicaciones, después crearon un conjunto de datos de la oración o frase insultante. Le entregan la hora y fecha del mensaje y su enlace a la base de datos *MarkMail*[10], así como la oración entera o el clúster de la oración en la cual ocurre el insulto. También etiquetaron si el insulto se refería al código, la personalidad, o ambas. Todo lo anterior con la intención de documentar los incidentes en los que se encuentran insultos personales en la *LKML* para aprender la diferencia entre insultos al código e insultos personales; y para crear una lista de insultos que podrían usarse para entrenar un clasificador de insultos "estilo Linus".

Finalmente, crearon conjuntos de datos para tres tipos de insultos de género: los chistes TWSS (That's What She Said) que contienen doble sentido sexual; los insultos maternales, que son aquellos en los que se utiliza a las madres o abuelas como insulto personal; y otros en los que se utiliza a las mujeres mayores (abuelas) para representar a una persona no inteligente o no sofisticada ("Even Grandma can use the software!").

DEviaNT

En este trabajo Kiddon & Brun (2011) propusieron un novedoso enfoque llamado Double Entendre via Noun Transfer (DEviaNT) con el que se aplican técnicas de identificación de metáfora para resolver el problema del doble sentido y evaluarlo en el problema TWSS (That's What She Said). DEviaNT analiza individualmente las oraciones y clasifica cada una como positiva dependiendo si ésta es chistosa cuando se le añade la frase "that's what she said".

Un chiste "that's what she said" (TWSS) es un tipo de doble sentido que se ha vuelto nuevamente popular gracias al show de televisión "The Office" (Daniels, Gervais, & Merchant, 2005). Este tipo de chistes consisten en decir "that's what she said" luego de que alguien más profiere una afirmación en un contexto no sexual que podría haber sido usado también en un contexto sexual. En la tabla 3.8 se muestra un ejemplo de chiste TWSS y la razón por la que se considera gracioso.

[10] http://markmail.org/

Tabla 3.6 Ejemplo de chiste TWSS (That's What She Said)

Chiste TWSS	Explicación
Un hombre hablando sobre su práctica de basquetbol nocturna: - "I was trying all night, but I just could not get it in" Alguien completa el chiste diciendo: - "Tha's what she said"	La expresión "get it in" en un contexto sexual se podría interpretar como tener un encuentro sexual con alguien.

Los autores identificaron dos importantes características en los chistes TWSS: 1) este tipo de chistes son más propensos a contener sustantivos que son eufemismos para sustantivos explícitamente sexuales; 2) estos chistes tienen una estructura en común con oraciones del dominio erótico. El enfoque para la resolución de este problema está centrado en un modelo SVM (*Support Vector Machine*, Máquina de Soporte Vectorial) que usa características diseñadas para modelar 1 y 2.

DEviaNT usa dos corpus: el corpus de erotismo que está formado de 1.5 millones de oraciones provenientes de *textfiles.com/sex/EROTICA*; y el corpus Brown (Francis & Kucera, 1979) que contiene 57 mil oraciones que representan literatura estándar (no erótica).

DEviaNT explora un particular enfoque para resolver el problema TWSS: identificar relaciones eufemísticas y estructurales entre el dominio origen y el dominio erótico. No utiliza ningún modelo léxico de n-gramas para los datos de entrenamiento TWSS. La precisión se encuentra por encima del 71.4% aunque los autores afirman que si DEviaNT clasificara un subconjunto balanceado (aleatoriamente seleccionado) de los datos de prueba, la precisión sería de 0.995.

JAPE

JAPE (*Joke Analysis and Production Engine*) es un software capaz de generar *punning riddles* a partir de un léxico no humorístico. En este trabajo Binsted (1996) desarrolló un modelo formal de este tipo de humor, el cual se compone de cuatro partes: los esquemas, que especifican las relaciones entre las unidades léxicas utilizadas para construir un chiste; el generador *SAD (Small Adequate Description),* el cual genera descripciones cortas acerca del mundo o de unidades léxicas construidas; las plantillas, que convierten unidades léxicas y sus descripciones en un formato de *pregunta-respuesta;* y los recursos léxicos, los cuales proveen toda la información léxica requerida para que las otras partes funcionen.

Los *punning riddles* son un subtipo de humor basado en juegos de palabras, pero con un formato de acertijo (pregunta-respuesta) y cuentan con ciertos mecanismos de construcción y estructuras regulares. Este humor es bastante común y además ya ha sido estudiado por los lingüistas anteriormente. En la tabla 3.6 se muestra un ejemplo de un *punning riddle* generado por *JAPE*.

Tabla 3.7 Ejemplo de punning riddle *generado por JAPE*

Punning riddle	**Significado**
Pregunta:	Pregunta:
What kind of tree is nauseated?	*¿Qué tipo de árbol tiene náuseas?*
Respuesta:	Respuesta:
A sick-amore	*El sicómoro*
	Existe homofonía de la palabra *sick* con las tres primeras letras de la palabra *sycamore,* que es el nombre de un árbol (sicómoro en español).
	El efecto cómico se produce debido a que *sick* es un sustantivo que se usa para referirse a alguien que sufre de nauseas, concepto que coincide con la pregunta planteada en la primera línea del chiste.

Con el fin de comprobar que todos los acertijos que genera *JAPE* son, en efecto, *punning riddles*, les pidieron a 122 niños de entre 8 y 11 años de edad que realizaran una evaluación. Se evaluaron tres características distintas del conjunto de textos: *jokiness*, el texto parece o no un chiste (0,1), *funiness*, el grado de comicidad (nada gracioso, no tan gracioso, no estoy seguro, gracioso, muy gracioso); *heard before*, si ya habían escuchado el chiste antes o no (0,1).

En la tabla 3.7 se observa que los textos generados por *JAPE* fueron calificados por los infantes con un puntaje de 0.6 en promedio, frente al 0.81 que recibieron los textos generados por humanos, en cuanto a *jokiness*. La característica de comicidad (*funiness)* obtuvo 3.14 unidades en el caso de *JAPE*, frente a los 3.57 de los generados por humanos.

Tabla 3.8 Resultados de la medición de la calidad de los chistes de JAPE

	Jokiness (Máximo 1)	*Funiness* (Máximo 5)
Textos generados por *JAPE*	0.6	3.14
Textos generados por humanos	0.81	3.57
Textos no absurdos no humorísticos	0.2	2.9
Textos absurdos no humorísticos	0.23	2.95

Como segunda fase de este proyecto se conformó el equipo *STANDUP*, quienes desarrollaron un *software* que lleva el mismo nombre, con la finalidad de usar esta herramienta para proporcionar una experiencia de entretenimiento pero también educacional a niños con discapacidades como parálisis cerebral. Actualmente, se encuentran involucrados varios investigadores en este proyecto que ha evolucionado hasta convertirse en *The Joking Computer*[11].

[11] http://www.abdn.ac.uk/jokingcomputer/home.shtml

3.11 Tabla comparativa de los trabajos relacionados

Trabajo	Idioma	Textos humorísticos	Técnicas de solución del problema
Clasificación de frases obscenas y vulgares dentro de *tweets* (México)	Español	Ninguno, sin embargo se realiza un primer acercamiento al albur.	• Uso de diccionario con palabras obscenas y vulgares (diccionario de mexicanismos) • Uso del clasificador *WEKA*
DAHTCE (México)	Español	Albur, *adult slang*, aliteración, rima	• Uso de diccionario para identificación de cada tipo de humor • Ponderación de las palabras de las oraciones
Reconocimiento de chistes *knock, knock* (EE.UU.AA.)	Inglés	Chistes *knock, knock*	• N-gramas • Procesamiento a nivel fonético de las oraciones • Procesamiento a nivel semántico
Reconocimiento automático de *one-liners* (EE.UU.AA. e Italia)	Inglés	*One-liners*	• Enfoque de aprendizaje automático
Reconocimiento automático del humor en textos escolares en Catalán	Catalán	Chistes de niños de entre 6 y 16 años de edad	• Categorización de los textos humorísticos • Uso de los clasificadores *Bayes* y regresión lógica multinomial.
Identificación del humor en *microblog* chino	Chino	Publicaciones de los usuarios del *microblog Sina Weibo*	• Categorización de los textos humorísticos • Uso de algoritmos de clasificación supervisada y semi-supervisada. • Paradigmas: traducción automática, expansión de sinónimos y selección de características
Reconocimiento de ironía y humor	Inglés	Publicaciones de los usuarios del *microblog Twitter*	• Categorización de los textos humorísticos e irónicos • Clasificación multinomial
FLOSS como fuente de profanidad e insultos: recopilación de los datos	Inglés	Listas de distribución de correo electrónico y chats IRC.	• Es un trabajo de recolección de datos, a fin de utilizarlos en trabajos futuros, para entrenar un clasificador de "insultos personales" e "insultos al código" entre desarrolladores de *software*.
DEviaNT	Inglés	*Doble sentido*	• Uso de sustantivos eufemísticos para sustantivos explícitamente sexuales • Máquina de soporte vectorial (SVM) • Técnicas de identificación de metáforas • Estructuras en común con oraciones del dominio erótico
JAPE (Escocia, Reino Unido)	Inglés	*Punning riddles* (Generación)	Modelado del tipo de humor: • Pequeñas descripciones del mundo exterior • Esquemas • Recursos léxicos • Plantillas
Método para la identificación automática de albures cortos en textos (Esta tesis)	Español	Albures escritos de manera intencional o no intencional	Enfoque lingüístico Procesamiento a nivel • Fonético (comparación de palabras y frases a nivel fonético contra las almacenadas previamente en diccionario) • y semántico (uso de diccionario de palabras usadas en doble sentido)

Capítulo 4 Metodología de solución

Para esta investigación se definió una metodología para el desarrollo de la solución que consta de tres fases: 1) búsqueda y generación de recursos léxicos; 2) desarrollo de los módulos de procesamiento; 3) integración de los módulos de procesamiento. En la ilustración 4.1 se muestra el diagrama de la metodología definida.

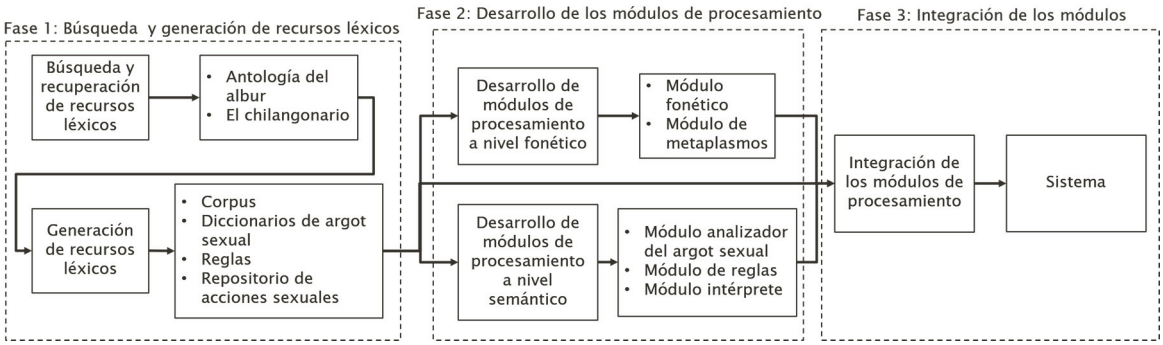

Ilustración 4.1 Metodología de solución

La primera fase consistió en realizar búsquedas en la web a fin de encontrar recursos útiles para esta investigación que ya hubieran sido generadas por la comunidad lingüística. También se generaron recursos propios que fueran útiles para la tarea de identificación automática del albur. Algunos recursos se generaron a partir de los que se identificaron en la búsqueda.

La segunda fase consistió en desarrollar los módulos de procesamiento a partir de los conocimientos adquiridos en las fases anteriores. Estos módulos emulan los procedimientos que realizan los humanos para manejar los fenómenos que más se utilizan en el albur. Para esto fue necesario estudiarlos desde un enfoque lingüístico y adquirir cierto conocimiento que se encuentra de manera resumida en el Capítulo 2 y Capítulo 3 .

La tercera y última fase consistió en integrar los módulos desarrollados en la segunda fase de tal manera que se consolidó el sistema final. Este programa realiza el análisis de un texto de entrada y determina si existen albures o no. Más adelante, en 4.2 se describe por

completo el sistema desarrollado y los módulos que lo integran, además del funcionamiento de cada uno de ellos.

4.1 Búsqueda y generación de recursos léxicos

Como primera fase se realizó una búsqueda de recursos léxicos que ya hubieran sido generados por la comunidad lingüística y que fueran útiles para esta investigación. Se identificaron dos recursos importantes: "El chilangonario" (Peralta de Legarreta, 2012), un diccionario descriptivo del argot que se utiliza en México; y "Antología del albur" (Hernández, 2006), una recopilación de albures escritos por mexicanos.

También se generaron algunos recursos propios para este trabajo tales como: corpus sobre el albur y diccionarios del argot sexual, además de las reglas por medio de las que se realiza la identificación automática. A continuación se hace una breve descripción de estos recursos y se menciona por qué fueron considerados importantes para la investigación.

4.1.1 Recursos léxicos existentes

4.1.1.1 Antología del albur

Durante la búsqueda de recursos existentes no se encontró ningún corpus de albures para usarlo computacionalmente. Sin embargo, sí se encontró una "Antología del albur" (Hernández, 2006) un libro en el que se recopilan albures, entre otros textos humorísticos, que fueron escritos por mexicanos de distintos estados de la república.

En 1996 Hernández inició un sitio web llamado "Albures.net" en el que publicaba sus propias creaciones albureras para sus lectores. Más tarde incluyó una sección de recados a fin de que los lectores le enviaran sus contribuciones. De esta manera el autor recibió tanto material que decidió organizarlo y publicarlo en 2006 con el título de "Antología del albur".

Este libro se considera una excelente muestra representativa del albur porque los textos contenidos en tal fueron escritos por distintas personas del estado de la república mexicana además de que el autor asegura que trató de no repetir ningún albur al realizar la selección del material enviado por sus lectores.

4.1.1.2 El Chilangonario

El segundo recurso léxico recuperado es "El chilangonario" (Peralta de Legarreta, 2012), un diccionario descriptivo del argot mexicano en general, es decir, no sólo argot sexual. En este recurso se provee, para cada palabra, ejemplos de su uso en frases y etiquetas que clasifican a la palabra. Para esta investigación, se seleccionaron solamente las palabras marcadas con las etiquetas "albur", "adaptación", "apócope", "corrupción", "eufemismo" y "sinónimo". En el anexo 1 se muestran algunas de las entradas obtenidas de este recurso.

4.1.2 Recursos léxicos generados

En esta fase de la investigación se generaron algunos recursos importantes para esta investigación. Algunos fueron usados para analizar el fenómeno del albur y otros más para implementarlos computacionalmente en el programa que realiza la detección automática de este tipo de humor.

- Un corpus de albures: a partir del cual se generaron otros dos recursos paralelos a éste: el corpus de frases en doble sentido y el corpus de frases eróticas. La generación de éstos surge de la necesidad de exponer las frases eróticas que los practicantes del albur ocultan de manera deliberada mediante fenómenos fonéticos como la sinalefa, sinéresis, contracción y calambur además del argot sexual. Con estos recursos (disponibles en línea[12]) se logra emular el proceso inverso que lleva a cabo una persona para generar un albur, obteniendo así un conjunto de frases totalmente explícitas y por tanto comprensibles por cualquier hablante del español.
- Un diccionario morfosintáctico del argot sexual: en el cual se recopilan palabras utilizadas en el argot sexual mexicano además de que se provee el lema e información gramatical para cada una.
- Un diccionario semántico del argot sexual: el cual se utiliza para proveer significados representativos (synsets) para las entradas del diccionario morfosintáctico tanto del argot sexual como del español convencional (el que utiliza *Freeling*). Ambos diccionarios (morfosintáctico y semántico) se encuentran disponibles en línea[13].
- Repositorio de acciones sexuales: es un conjunto de pares de synsets (significados) que corresponden a las acciones sexuales que más se utilizan en los albures. Este repositorio es usado por el módulo de reglas para determinar si una frase es albur o no.

4.1.2.1 Creación del corpus de albures

A pesar del título, la "Antología del albur" contiene textos humorísticos que no son albur. Luego de realizar una selección manual de albures se conformó un corpus de 820 albures. La longitud promedio de éstos es de cuatro palabras por oración.

Aunque la colección de albures es pequeña, el corpus conformado se considera una excelente muestra de este tipo de humor debido a que estos textos fueron escritos por

[12] https://github.com/robertovillarejo/corpus-de-albures

[13] https://github.com/robertovillarejo/mexican_sexual_slang

personas de distintos estados de la república mexicana. En el Apéndice B: corpus se muestran algunos ejemplos de albures del corpus conformado.

Los albures que conforman el corpus de albures son en su mayoría juegos de palabras, es decir, las frases sexuales de éstos se encuentran ocultas por medio de fenómenos fonéticos como la sinéresis, sinalefa, contracción y calambur, mismos que no son abordados en este trabajo.

El objetivo de utilizar este tipo de albures como objeto de estudio es identificar y extraer las características clave que los conforman para luego utilizarlas en la construcción del método de identificación automática del albur.

4.1.2.2 Creación del corpus de frases en doble sentido

El albur tiene siempre una connotación sexual y un mensaje oculto que quien no tiene conocimiento del argot sexual ni las habilidades para descifrarlo no es capaz de comprenderlo.

A fin de eliminar dicho obstáculo y realizar un mejor análisis del corpus de albures, se generó un nuevo recurso llamado "Corpus de frases en doble sentido" (véase Apéndice B: corpus). En éste se anotaron las frases sexuales que originalmente se encontraban ocultas en el corpus de albures.

La generación de este nuevo corpus se llevó a cabo de manera manual y consistió principalmente en eliminar aquellos fenómenos que ocultan la frase sexual a nivel fonético. Al exponer la frase originalmente oculta es posible observar más fácilmente las características que componen al albur y, por tanto, la extracción de las características clave se facilita. A continuación se muestra un ejemplo de frase oculta y el albur al que corresponde:

A. Garráis el chico > Agarráis el chico

Para ver más ejemplos de este tipo consulte el anexo B.

4.1.2.3 Creación del corpus de frases eróticas

En el corpus de frases en doble sentido se pueden analizar los albures originales sin todos aquellos fenómenos a nivel fonético que son aprovechados como forma de ocultamiento. Sin embargo, estas frases aún pueden resultar incomprensibles para aquellas personas que no conocen las palabras de argot sexual y por tanto el significado real sigue oculto.

Por esa razón hubo la necesidad de generar, a partir del corpus de frases ocultas, un tercer recurso llamado "Corpus de frases eróticas". Estas frases, además de que no contienen juegos de palabras u otros fenómenos a nivel fonético que oculten el significado real, tampoco contienen argot sexual que pueda usarse para el mismo fin, sino que son frases totalmente

explícitas. A continuación se muestra un ejemplo de frase de este nuevo corpus y su correspondiente frase de origen:

Agarráis el chico > Agarráis el recto

Para ver más ejemplos como éste consulte el Apéndice B: corpus.

Los recursos anteriores, el corpus de albures, el corpus de frases en doble sentido y el corpus de frases eróticas se encuentran disponibles en línea[14].

4.1.2.4 Lemario del argot sexual y sus significados

Con el propósito de conformar un diccionario morfosintáctico y un diccionario semántico del argot sexual mexicano se diseñó un proceso automático que generó una lista de palabras de argot sexual junto con sus significados.

Además se incluyeron las palabras obtenidas del recurso "El chilangonario", que al igual que las anteriores se les anotó su correspondiente palabra en español convencional. En la ilustración 4.2 se muestra un diagrama del proceso realizado y los recursos utilizados para generar el lemario.

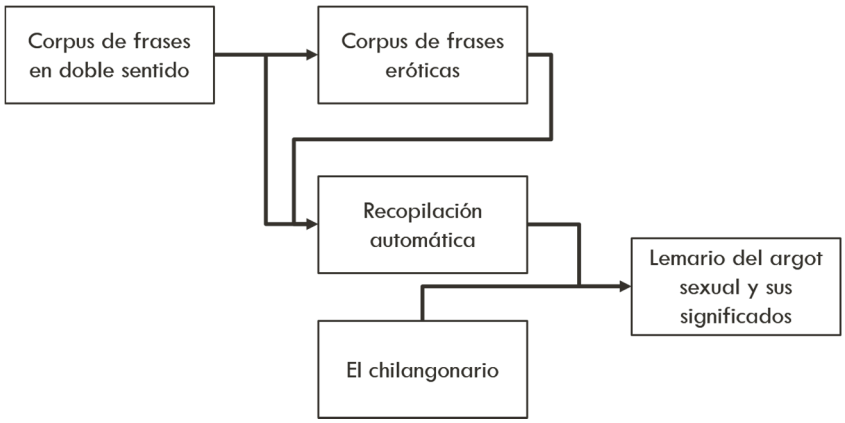

Ilustración 4.2 Recopilación de palabras de argot sexual

En la siguiente tabla se muestran algunas palabras de la lista generada. Del lado izquierdo, las palabras de argot sexual y del lado derecho sus términos correctos que corresponden al significado real de las primeras.

[14] https://github.com/robertovillarejo/corpus-de-albures

Tabla 4.1 Lemario del argot sexual y sus significados (extracto)

Argot sexual	**Español convencional**
Chico	Recto
Pizarro	Falo
Tanates	Testículos
Vergara	Falo
Hoyón	Recto

Se observó que algunas de las palabras recopiladas son el resultado de aplicar metaplasmos a otras. Se considera que estos casos corresponden a otro nivel del lenguaje (fonético/morfológico) y por tanto deben ser procesados de manera separada. Al mismo tiempo se evita que este diccionario contenga un número exagerado de entradas.

4.1.2.5 Creación del diccionario morfosintáctico del argot sexual

El lemario de argot sexual (junto con sus significados) se utilizó para formar el diccionario morfosintáctico. A fin de que este recurso pueda ser utilizado por otros investigadores, se adoptó el formato de los diccionarios de Freeling. Éste, requiere que se anoten todos los posibles pares lema-etiqueta de la palabra. El lema es aquella forma única que es común a todas las posibles variaciones de una palabra, p. ej. para los sustantivos se usa la forma en singular y masculino ("perro" es el lema de "perras", "perritos"); para los verbos se utiliza la forma en infinitivo ("comer" es el lema de "comiendo", "comió"). La etiqueta es una serie de símbolos que provee información gramatical de la palabra; en este caso se utilizaron las etiquetas propuestas por el grupo EAGLES[15]. El formato requerido para los diccionarios es el siguiente:

```
forma_de_palabra lema1 etiqueta1 lema2 etiqueta2
```

En el diccionario morfosintático del argot sexual no se incluyeron diminutivos ni aumentativos. Tales formas son analizadas por el analizador morfosintáctico de *Freeling* el cual, mediante serie de reglas de afijación, lematiza y etiqueta apropiadamente las palabras. A continuación se muestra un ejemplo de entrada de este diccionario:

```
chico chico NCMS000
```

[15] http://www.ilc.cnr.it/EAGLES96/home.html

Para ver más ejemplos de entradas consulte el Apéndice C: diccionarios del argot sexual mexicano.

En total, el diccionario morfosintáctico contiene 410 formas de palabra las cuales corresponden a 350 lemas. Sin embargo, el número de palabras que pueden ser analizadas incrementa si se utilizan las reglas de afijación (de *Freeling*). Este diccionario contiene todas aquellas palabras que son el resultado de la creatividad de los hablantes quienes utilizan las técnicas explicadas en 2.5.4 para producirlas.

4.1.2.6 Creación del diccionario semántico de argot sexual

A fin de dotar de significados a cada palabra del argot sexual, se creó un segundo recurso: el diccionario semántico del argot sexual, el cual provee dicha información a través de *synsets*. Un *synset* es un código único que se utiliza para representar un concepto específico del mundo. Un synset se encuentra asociado a lemas, no a palabras, p.ej. 02084071-n es el *synset* para el concepto de "perro". Los lemas asociados a esta palabra son "perro" y "can". Los *synsets* son independientes del lenguaje, por tanto, en el idioma inglés, los lemas asociados a este *synset* son "dog", "canis_familiaris" y "domestic_dog".

En la Tabla 4.1 se presentaron las palabras de argot sexual y sus significados (lemario). Para la creación del diccionario semántico se buscaron los *synsets* de tales significados (columna "español convencional").

Los *synsets* originalmente provienen de la base de datos *WordNet* (Miller, 1995), sin embargo, ésta solo contiene información para el idioma inglés. Por tanto, fue necesario realizar la búsqueda de los *synsets* en la base de datos *Multilingual Central Repository (MCR)* (Gonzalez-Agirre A., 2012). Este repositorio reutiliza los *synsets* de *WordNet* para otros lenguajes como español, portugués, catalán, vasco y galés. La Tabla 4.2 muestra la asignación de los *synsets* a los significados del argot sexual.

Tabla 4.2 Asignación de códigos synset a las palabras correctas

Palabra correcta (español convencional)	**Synset**
Recto	05538016-n
Falo	05526713-n
Testículos	05524615-n
Vello púbico	05263587-n
Glúteos	05559256-n

Puesto que, en un contexto sexual, todas las palabras del argot sexual se consideran sinónimos de sus significados (palabras del español convencional), entonces éstas pueden agruparse mediante el mismo *synset*. A continuación se muestra un ejemplo de entrada de este diccionario:

05538016-n recto chico hoyo anacleto

Para ver más ejemplos de entradas, consulte el anexo C.

En este diccionario, a un synset se le asocia una lista de lemas. Un lema se puede asociar a más de un *synset*, es decir, una palabra puede tener más de un significado (polisemia).

Se observó que algunas palabras, además de sufrir cambios de significado, en el argot sexual también sufren cambios gramaticales. P.ej. "largo" que es un adjetivo, se nominaliza anteponiéndole un artículo apropiado ("el largo"), de esta manera adquiere el significado de falo. El *synset* 05526713-n (concepto de falo) tiene el número más alto de lemas (112), seguido por 05538016-n (concepto de recto) con 53 lemas.

Ambos recursos (diccionario morfosintáctico y diccionario semántico) son complementarios a los del español, es decir, en la práctica se utilizan para enriquecer y extender el análisis del idioma español. Por lo tanto, es posible incluir lemas del español convencional (del diccionario de *Freeling*) en estos recursos.

4.1.2.7 Estructuras de albures

El motivo de realizar una exploración y análisis de las estructuras del albur fue para identificar cuáles de éstos eran los más comunes y los más fáciles de atacar. Los albures pueden clasificarse en varios tipos dependiendo del enfoque que se aborde: juego de palabras, paronimia, sinonimia, calambur, argot sexual, etc. En este trabajo la clasificación en tipos de albures se realizó según su significado oculto, es decir, según su frase en doble sentido. A continuación se presentan las estructuras del albur identificadas:

- Tipo "verbo + sustantivo": la frase oculta indica que se realiza una acción sobre un sustantivo (normalmente refiere a los genitales).
- Tipo "verbo + pronombre": la frase oculta indica que se realiza un acción sobre algún sustantivo que no se indica explícitamente. La ambigüedad del pronombre hace suponer al escuchante un sustantivo sexual.
- "Otros": Normalmente se encuentra un sustantivo pero no se indica una acción como tal. La mayoría de las veces provee una imagen mental, se plantea una situación sexual sin utilizar un verbo. Este tipo de albures no es abordado en esta tesis porque se considera que no es tan común como los otros dos.

En la Tabla 4.3 se muestran algunos ejemplos de los tipos de albur propuestos en este trabajo.

Tabla 4.3 Ejemplos de estructuras identificadas en las frases en doble sentido[16]

Sustantivo + verbo	**Verbo + pronombre**	**Otros**
Agarráis el chico (A. Garráis el chico)	Se la estaco (Isela Estaco)	Chile pa' su silla (Chile pasusilla)
Agárrame el pizarro (A. Garramel Pizarro)	Te las poncho (Telas "Poncho"	Chile pa' tu ano (Chile patuano)

En el tipo "verbo + pronombre" se encuentra presente una acción (verbo) pero no se indica explícitamente el sustantivo (normalmente genitales) al que se realiza la acción. En este tipo de albures, aunque no aparece un sustantivo como tal, el pronombre lo sustituye y adquiere un significado sexual. A continuación se presenta un conjunto de patrones pronominales que, en conjunto con un verbo alburero, provocan que la frase luzca como albur:

- nos las
- nos la
- nos los
- nos lo
- se las
- se la
- se los
- se lo
- te las
- te la
- te los
- te lo
- me las
- me la
- me los
- me lo

4.1.2.8 Repositorio de acciones sexuales

El significado oculto de un albur es siempre de índole sexual. En la mayoría de los casos este significado oculto (frase en doble sentido) es de tipo "verbo + sustantivo".

A fin de abstraer lo anterior, se conformó un repositorio de acciones sexuales. Una acción sexual consiste en un par de *synsets*, donde uno representa el concepto de un verbo y otro representa el concepto de un sustantivo. Para crear este recurso se diseñó un proceso automático el cual generó estas acciones sexuales (pares de *synsets*) a partir del corpus de

[16] Entre paréntesis se muestra el albur de origen

frases eróticas. A continuación se describen los subprocesos que componen la generación del repositorio.

Para cada frase del corpus de frases eróticas:

1. Se selecciona el verbo y el sustantivo que juntos dan un significado sexual. P.ej. “Agarrar” y “recto”.
2. Cada palabra (verbo y sustantivo) se procesa con tres analizadores:
    a. Analizador morfosintáctico: anota todos los posibles pares lema-etiqueta.
    b. Analizador semántico: anota todos los posibles significados (*synsets*)
    c. Desambiguador semántico: calcula la probabilidad de cada *synset*.
3. Se selecciona el mejor *synset* del verbo y del sustantivo. Es decir, se toman los *synsets* con mayor calificación y se descartan los demás.

En la Ilustración 4.3 se muestra un ejemplo de generación de una acción sexual en *synsets* a partir de una frase erótica.

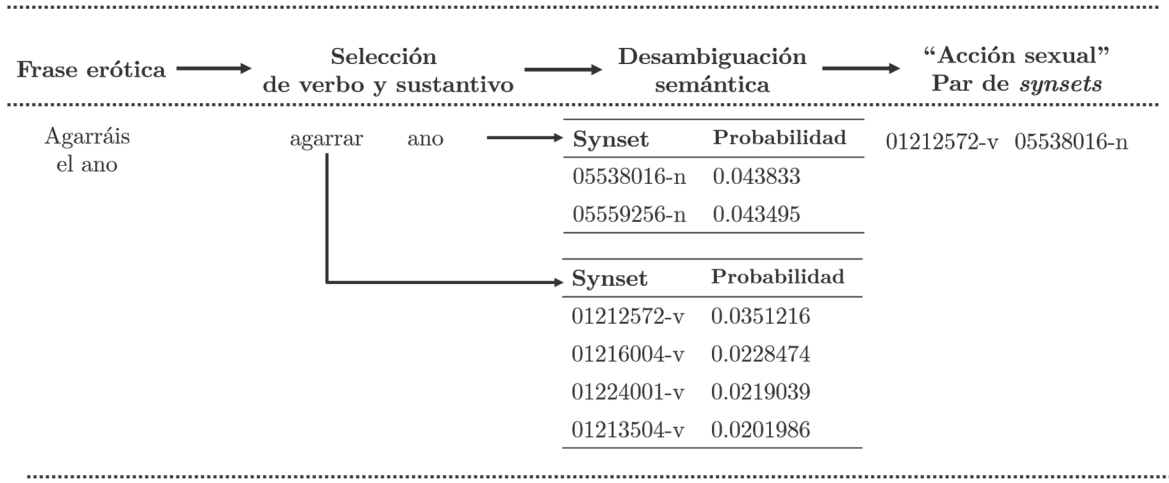

Ilustración 4.3 Ejemplo de generación de acción sexual

El repositorio se compone de pares de synsets y no de palabras porque se desea almacenar los significados y no los significantes (véase 2.2.4). De esta manera se pueden seguir incrementando el diccionario morfosintáctico y el diccionario semántico sin tener que modificar el repositorio de acciones sexuales. Es decir, se continúa asociando palabras a los *synsets* y el método sigue funcionando sin que haya necesidad de modificar este recurso. En la Tabla 4.4 se muestran algunos ejemplos de acciones sexuales del repositorio.

Tabla 4.4 Repositorio de acciones sexuales

Verbo	**Sustantivo**
01212572-v	05538016-n
01308160-v	14173484-n
01212572-v	05526713-n
01346003-v	05538016-n

Estas acciones sexuales (pares de *synsets*) representan el significado oculto de un albur. De acuerdo a las estructuras identificadas anteriormente, corresponden al tipo "verbo + sustantivo". Con un par de *synsets* es posible identificar distintos albures ya que el significado oculto es el mismo tal como se muestra en la Tabla 4.5.

Tabla 4.5 Una misma acción sexual en albures distintos

Albur	**Acción Sexual**
Rózame el fierro	01250908-v 05526713-n
Rózame el tronco	01250908-v 05526713-n
Pito vas que rozas	01250908-v 05526713-n

01250908-v es el *synset* para el concepto "rozar", 05526713-n es el synset para el concepto "falo".

4.1.2.9 Definición de reglas

Por medio de la observación y el análisis del corpus de frases eróticas se identificaron las estructuras más utilizadas en el albur (véase 4.1.2.7). A partir de estas se formularon algunas reglas sintácticas las cuales se utilizan en el "analizador del albur" para determinar si existe albur o no. A continuación se expresan, de manera sencilla y legible, las reglas definidas.

1. ***(synset de sustantivo) + (0 o más palabras) + (synset de verbo)***
2. ***(synset de verbo) + (0 o más palabras) + (synset de sustantivo)***
3. ***("nos", "se", "te" o "me") + ("las", "la", "los" o "lo") + (synset de verbo)***
4. ***(synset de verbo) + ("nos", "se", "te" o "me") + ("las", "la", "los" o "lo")***

Estas reglas se encuentran definidas en un recurso independiente con formato de expresiones regulares. De esta manera es posible aumentarlas o mejorarlas en trabajos futuros sin afectar al programa computacional. En estas reglas se puede hacer uso de distintos atributos de las palabras como la forma de palabra, el lema, etiqueta gramatical y significado.

```
1. \w.*SENSES=SUSTANTIVO.* \w.*SENSES=VERBO.*
2. \w.*SENSES=VERBO.* \w.*SENSES=SUSTANTIVO.*
3. (nos|se|te|me).* (las|la|los|lo).* \w*SENSES=VERBO.*
4. \w*SENSES=VERBO.* (nos|se|te|me).* (las|la|los|lo).*
```

Con estas reglas sintácticas es posible acceder a diferentes atributos de las palabras, tales como la forma de palabra, lema, etiqueta gramatical, forma fonética y significados (*synsets*). Esto permite que la aplicación y la modificación de las reglas sea flexible y mantenible. A continuación se enlistan las palabras reservadas para estas expresiones regulares y se describe su uso.

Las siguientes palabras reservadas sirven para acceder a los distintos atributos de una palabra.

- **TAG:** se refiere a la etiqueta gramatical. P.ej. ***TAG=PP.****, significa que la etiqueta gramatical debe iniciar con el valor "PP", es decir, que la palabra sea un pronombre personal.
- **LEMMA:** se refiere al lema. P.ej. LEMMA=el.*, significa que el lema de la palabra debe ser "el". Éste es el lema de las palabras "la", "el", "los", "las".
- **SENSES:** se refiere al significado (*synset* o forma semántica). P.ej. SENSES=05538016-n.*, significa que la palabra debe tener el significado 05538016-n. Éste es el *synset* del concepto "recto" o "ano".

Cuando se aplica una regla a una oración, se utilizan todas las acciones sexuales del repositorio, es decir, se prueba cada par de *synset*. Las siguientes palabras reservadas se pueden combinar con las anteriores; se refieren al *synset* de una misma acción sexual.

- **SUSTANTIVO:** se refiere al sustantivo *synset* de una acción sexual.
- **VERBO:** se refiere al verbo *synset* de una acción sexual.

P.ej., supongamos que se quiere aplicar la siguiente regla:

*\w.*SENSES=SUSTANTIVO.* \w.*SENSES=VERBO.**

Junto con la siguiente acción sexual:

02593912-v 05538016-n

Entonces se sustituye "SUSTANTIVO" por 05538016-n; y "VERBO" se sustituye por 02593912-v. De esta manera se construye la siguiente expresión regular:

*\w.*SENSES=05538016-n.* \w.*SENSES=02593912-v.**

Con esta expresión regular se busca que exista una palabra con el significado 05538016-n (concepto de “ano”) y otra con el significado 02230247-v (concepto de “transferir”, “transmitir”, “pasar”). Si esta condición se cumple entonces la oración que está siendo analizada se clasifica como albur.

4.2 Desarrollo de los módulos de procesamiento

Correspondiendo a los fenómenos involucrados en el albur se desarrollaron cinco módulos de procesamiento los cuales integran el sistema de cómputo desarrollado: *Analizador de Argot, Módulo Fonético, Módulo de Metaplasmos, Módulo de Reglas* y *Módulo Intérprete.* Adicionalmente, estos cinco módulos fueron integrados en uno sólo nombrado “Analizador del albur”, que es un módulo de conveniencia que sirve para instanciar y llamar a los primeros.

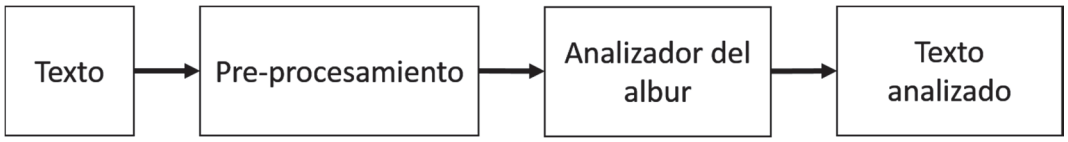

Ilustración 4.4 Vista general del sistema identificador automático de albures

Como se muestra en la Ilustración 4.4 Vista general del sistema identificador automático de albures, el sistema recibe como entrada un texto, ya sea que esté contenido en un archivo plano o escrito por medio de la consola. Antes de procesar el texto para la identificación automática de albures se realiza un pre-procesamiento el cual se describe en 4.2.1.

Como salida, el sistema entrega el análisis de cada oración del texto introducido. En éste se indica si la oración es considerada albur y cuáles son las palabras que lo producen. El formato de dicho análisis se describe en 4.2.2.

4.2.1 Pre-procesamiento

Antes de procesar el texto de entrada con el sistema identificador de albures es necesario realizar un procesamiento previo, el cual consiste en analizar cada oración con los siguientes servicios de *Freeling*:

- Tokenización (*Tokenization*)
- Separación de oraciones (*Sentence splitting*)
- Análisis morfolósintáctico del español convencional

- Etiquetado PoS

Al final del pre-procesamiento se tiene una lista de oraciones cuyas palabras que las conforman poseen información morfosintáctica como todos sus posibles lemas y etiquetas gramaticales.

4.2.2 Texto analizado

Como salida del sistema de identificación automática de albures se entrega una lista de análisis que corresponde a la lista de oraciones del texto introducido. De cada oración se indica si ésta fue clasificada como albur o no; de ser afirmativo cuáles son las palabras que conforman el albur y además una "traducción" de la frase de entrada. En la Tabla 4.6 se muestra en resumen la información que el sistema provee como respuesta al texto introducido.

Tabla 4.6 Formato de salida del sistema

Atributo	**Tipo de variable**	**Descripción**
Oración	Cadena de caracteres	La oración de entrada que fue analizada por el sistema.
Albur	Booleano	Verdadero si el sistema determinó que la oración de entrada es un albur.
Frase traducida	Cadena de caracteres	Si *Albur* es verdadero, entonces se sustituyen las palabras de argot sexual de la oración de entrada.
Acción sexual	Lista de palabras (también puede obtenerse como cadena de caracteres)	Si la oración de entrada es un albur, esta lista contiene las palabras que forman el albur.

La traducción que realiza el sistema se trata de una sustitución de las palabras de argot sexual por las del español estándar. Es decir, es una explicación simple del albur que podría entender cualquier hablante del español.

4.2.3 Módulos de procesamiento

Como última fase de la investigación se desarrollaron cinco módulos de procesamiento necesarios para integrar el sistema final. Estos módulos no son dependientes entre sí, por lo que pueden ser utilizados en otras aplicaciones si así se desea.

En los Objetivos de esta tesis se mencionó que en esta investigación se abordan los niveles fonético y semántico para la resolución del problema. Sin embargo, cabe destacar que durante el pre-procesamiento y con el módulo de reglas también se está abordando los niveles morfológico y sintáctico.

A continuación se describen los módulos desarrollados en el orden en que son utilizados por el programa de cómputo en el que se integraron (véase Ilustración 4.1).

4.2.3.1 Módulo Fonético

Este módulo provee un conjunto de funciones útiles para el procesamiento a nivel fonético. La principal función es la de realizar el análisis fonético de las oraciones, es decir, genera la forma fonética de cada palabra de una oración utilizando una serie de reglas fonológicas.

En este caso se tiene una modificación propia de *SAMPA* (Wells, 1997) cuyo objetivo es generalizar algunos sonidos que son similares entre sí. Por ejemplo, el sonido de “n” antes “b” es muy similar al de “m” antes de “v”. Con estas reglas ambos sonidos se consideran el mismo. De esta manera se produce ambigüedad fonética la cual es aprovechada para atacar aquellos fenómenos que suceden a dicho nivel del lenguaje, como es el caso de los metaplasmos. En la Tabla 4.7 se muestran las reglas fonológicas utilizadas para hacer la transcripción fonética de las palabras, es decir, generar la forma fonética.

Tabla 4.7 Reglas fonológicas de SAMPA modificadas

Categorías		
W=ei (vocales débiles)		
F=aou (vocales fuertes)		
B=pb (bilabiales)		
Reglas fonológicas		
rr/R/_	q/k/_	z/s/_
r/R/^_[18]	c/s/_W	h//_
ch/X/_	c/k/_	v/b/_
x/X/_	g/j/_W	y/i/_
qu/k/_W	gu/g/_W	ñ/J/_
qu/ku/_F	gü/hu/_	n/m/_B
qü/ku/_	ll/i/_	n/m/_f

En la Ilustración 4.5 se muestra que el módulo fonético recibe una oración de entrada y entrega esa misma oración pero enriquecida a nivel fonético, es decir, se genera la forma fonética de cada palabra.

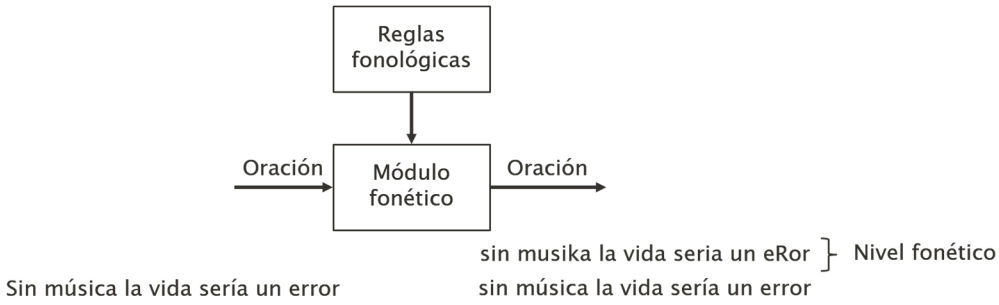

Ilustración 4.5 Ejemplo de procesamiento con el módulo fonético

4.2.3.2 Módulo de Metaplasmos

Este módulo añade pares lema-etiqueta gramatical a una palabra si es que ésta puede ser considerada como un metaplasmo de otra. Como se explica en 2.5.3.1, existen varios tipos de metaplasmos; los que se solucionan con este módulo son prótesis/paragoge y aféresis/apócope.

[18] El símbolo "^" indica que el primer elemento (la letra r) debe encontrarse al inicio de una palabra.

En la Ilustración 4.6 se observa que el módulo de metaplasmos recibe como entrada una oración, misma a la que se realiza un análisis fonético. Es decir, se genera la forma fonética de cada palabra a fin de compararla con las palabras del argot sexual.

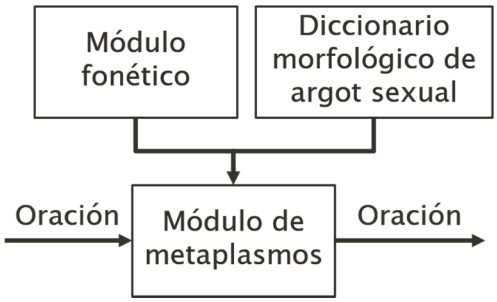

Ilustración 4.6 Módulo de metaplasmos

Freeling, que es la librería en la que se basan los módulos desarrollados, maneja una estructura propia para almacenar información morfosintáctica y semántica en cada palabra. En cuanto a la información morfosintáctica (lema, etiqueta gramatical) ésta se almacena en un objeto llamado "Análisis" a partir del cual se asignan los posibles significados (*synsets*) en un proceso posterior. En la Ilustración 4.7 se muestra un ejemplo gráfico de la información que se almacena en una palabra (objeto *Word*).

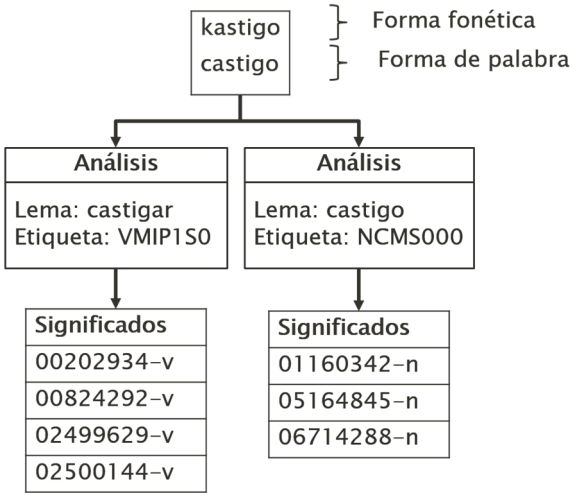

Ilustración 4.7 Ejemplo de información morfosintáctica y semántica de una palabra

El procesamiento realizado por el módulo de metaplasmos consiste en comparar la forma fonética de una palabra de la oración de entrada contra las formas fonéticas del diccionario morfosintáctico. Si alguno de los extremos de la primera coincide con la segunda entonces ésta adquiere, además del propio, el análisis que corresponde a la primera.

Por ejemplo, a la palabra “anónimo” se le asigna el análisis de la palabra “ano” porque es considerada paragoge de esta última. En la Ilustración 4.8 se muestra de manera gráfica lo que se explica.

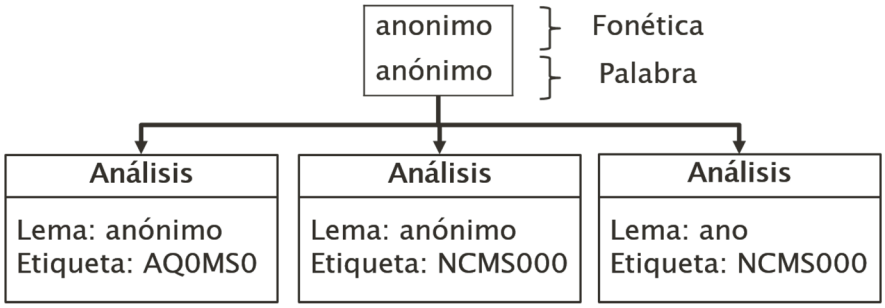

Ilustración 4.8 Ejemplo de adición de información por el módulo de metaplasmos

4.2.3.3 Analizador de argot sexual

Durante el pre-procesamiento se realiza un análisis morfosintáctico correspondiente al español convencional. El módulo analizador de argot sexual añade a cada palabra los posibles análisis (pares lema-etiqueta gramatical) correspondientes al argot sexual. Posteriormente anota a cada palabra los posibles significados tanto del español convencional como del argot sexual.

En la Ilustración 4.9 se muestra la arquitectura de este módulo que utiliza, además del diccionario semántico del español convencional de *Freeling*, los diccionarios del argot sexual (véase 4.1.2.5 y 4.1.2.6) .

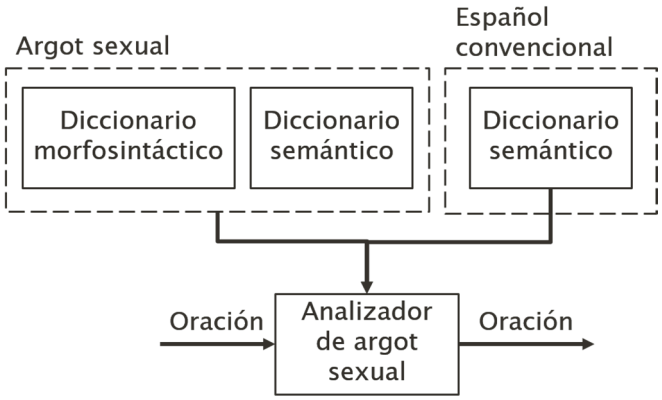

Ilustración 4.9 Módulo analizador del argot sexual

El procesamiento que realiza este módulo para una oración de entrada se realiza en el siguiente orden:

- Análisis morfológico del argot sexual
- Análisis semántico del español convencional
- Análisis semántico del argot sexual

4.2.3.4 Módulo de Reglas

El módulo de reglas aplica las reglas definidas en 4.1.2.9 para realizar la clasificación entre albur o no albur. Éste trabaja a distintos niveles del lenguaje, es decir, puede realizar verificaciones en distintos atributos de una palabra, como la forma de palabra, el lema, la etiqueta gramatical y los significados a través de las palabras reservadas que se utilizan en la declaración de las reglas.

En la siguiente ilustración se muestra la arquitectura del módulo de reglas. Éste utiliza el repositorio de acciones sexuales y las reglas para generar expresiones regulares que se utilizan para la clasificación binaria (albur o no albur).

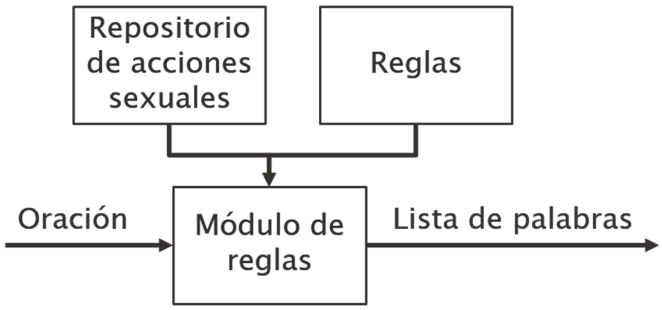

Ilustración 4.10 Módulo de reglas

En caso de que una oración cumpla con al menos una regla entonces ésta se clasifica como albur. Como salida se obtiene una lista de palabras que conforman al albur detectado. En caso contrario (la oración no fue clasificada como albur), entonces la lista se retorna vacía.

En la Ilustración 4.11 se muestra un ejemplo en el que se cumple la regla "verbo + sustantivo" porque se encontró la acción sexual 01168468-v 05526713-n.

Sección de Investigación de Accidentes en el Tránsito**; S.I.A.T.**; Unidad de Carabineros de Chile, que, como su nombre lo indica, su misión es investigar los accidentes del tránsito para determinar sus causas y la de participación de los involucrados para, posteriormente, informar a los Tribunales de Justicia o proceder como se indique.

Chile	como
Significados	**Significados**
05526713-n	01166351-v
07721456-n	01168468-v
07822687-n	01185304-v
12900987-n	

Ilustración 4.11 Ejemplo de aplicación de reglas[19]

4.2.3.5 Módulo Intérprete del albur

Si una oración ha sido clasificada como albur por el módulo de reglas, entonces es posible "traducirla" para que sea comprensible por las personas que no conozcan el argot sexual mexicano. Esta "traducción" consiste en sustituir las palabras de argot sexual por sus correspondientes en el español convencional.

Como se muestra en la Ilustración 4.12, este módulo utiliza el diccionario semántico del argot sexual para sustituir las palabras de este tipo por las del español convencional. Es decir, las palabras que pudieran resultar escatológicas se sustituyen por unas más adecuadas y no ofensivas.

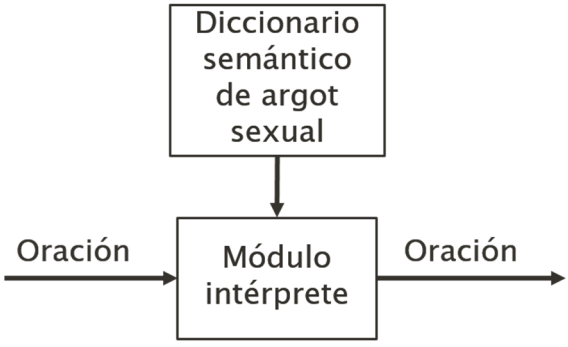

Ilustración 4.12 Módulo intérprete del albur

[19] Por practicidad sólo se muestra la información semántica (*synsets*) de las palabras.

Para hacer la sustitución de palabras se consultan los significados de cada una de las palabras de la oración. Si uno de estos *synsets* corresponde al argot sexual entonces, por medio de éste, se elige una palabra más adecuada y se sustituye por la original. Lo anterior se logra sustituyendo la forma de palabra por el primer lema de la lista del synset en el diccionario semántico.

```
05538016-n recto ano anastasio aniceto anís asterisco chico …
```

4.3 Integración de los módulos (Analizador del albur)

A fin de integrar los procesos que se muestran en la Ilustración 4.1 se creó el "Analizador del albur", un meta-módulo que no realiza ningún procesamiento por sí solo, sino que es un módulo de conveniencia para simplificar las tareas de instanciar y llamar a los sub-módulos descritos en 4.2.3. En el siguiente diagrama se aprecia la composición del "Analizador del albur".

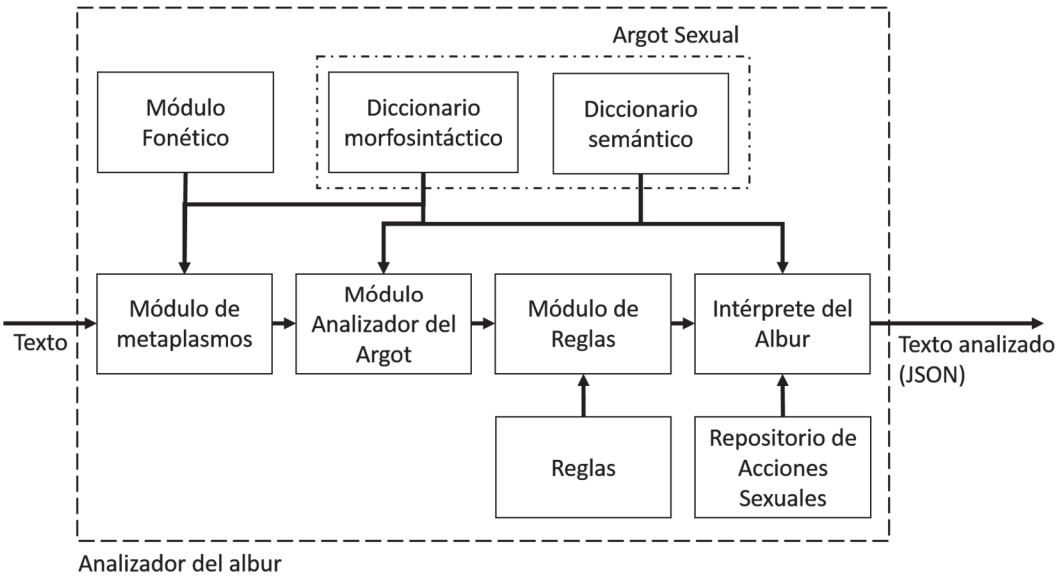

Ilustración 4.13 Meta-módulo "analizador del albur"

En 4.2.2 se muestra la información que contiene la oración o lista de oraciones analizadas en la salida de este módulo.

4.3.1 Desarrollo del servicio web

Finalmente, el sistema completo fue montado como un servicio web. Esto con el fin de que pueda ser utilizado por otras aplicaciones en trabajos futuros. Este servicio entrega como

salida un objeto JSON que contiene la información del análisis realizado a cada una de las oraciones del texto de entrada.

JSON es un formato de intercambio de datos muy ligero. Su escritura y lectura resulta fácil para las personas y la generación automática de éste es fácil de implementar en programas. JSON no depende de ningún lenguaje de programación, lo que lo convierte en el formato de intercambio ideal para usarse en este desarrollo. A continuación se muestra un ejemplo de la respuesta del servicio web ante un texto de entrada.

Entrada: "…unidad de Carabineros de Chile, como su nombre lo indica, su misión es…"

Salida:


```
[{
     "albur": "true",
     "oracion": "Sección de Investigación de Accidentes en el Tránsito; S.I.A.T.; Unidad de Carabineros de Chile, que, como su nombre lo indica, su misión es investigar los accidentes del tránsito para determinar sus causas y la de participación de los involucrados para, posteriormente, informar a los Tribunales de Justicia o proceder como se indique.",
     "accion_sexual": ["Carabineros_de_Chile", "como"]
}]
```

Capítulo 5 Evaluación

En este capítulo se muestran las pruebas del sistema realizadas en muestras positivas y muestras negativas, es decir, albures y no albures. Como primer conjunto se utilizó el corpus de frases en doble sentido que, como se explica en 4.1.2.2, son albures basados en ambigüedad léxica y semántica. Para el segundo conjunto (negativas) se conformó un corpus de frases célebres extraídas de búsquedas en la web. Los temas de este último son diversos: música, matemáticas, política, cine, científicos y escritores.

En resumen, se tienen 553 muestras positivas (albures) y 635 muestras negativas (frases célebres). Como se dijo anteriormente, el sistema desarrollado identifica albures que han sido escritos de modo intencional o no intencional. De hecho, una de las posibles aplicaciones es que el sistema advierta a un usuario cuando se ha escrito un albur sin que el mismo se haya percatado.

5.1 Pruebas con muestras positivas

A continuación se muestran algunas muestras positivas que fueron analizadas en las pruebas del sistema:

- …unidad de Carabineros de Chile, que, como su nombre lo indica, su misión es...
- Cuando un gato hace esta acción de rozar su cabeza con alguna parte de tu cuerpo…
- J. Zagal, C. Fierro, R. Rozas.
- Puede que quieras extraer la leche justo después de tu última sesión de lactancia o de extracción…
- …dos fenómenos totalmente dispares: un hombre metido en luminal, tan tieso que sus pies se extienden más allá de la ventana…
- …un señor estaba agarrando a un chiquilín, que parece que le había robado el celular
- Existe una suerte, un pial especial, que es llamado "de la muerte" o "de la flecha" que consiste no en atorar la reata en la cabeza de la silla

En la Tabla 5.1 se muestran los resultados de los análisis de los albures. Si el sistema determina que una oración es un albur, entonces indica cuál es la acción sexual que encontró y muestra la traducción del albur al español estándar.

Tabla 5.1 Ejemplos de albures analizados

Albur	**¿Es albur?**	**Acción sexual**	**Frase traducida**
unidad de Carabineros de Chile, que, como su nombre lo indica, su misión es	Sí	Chile como	unidad de Carabineros de falo, que, como su nombre lo indica, su misión es
Cuando un gato hace esta acción de rozar su cabeza con alguna parte de tu cuerpo	Sí	Rozar cabeza	Cuando un gato hace esta acción de rozar su falo con alguna parte de tu cuerpo
J. Zagal, C. Fierro, R. Rozas.	Sí	Fierro Rozas	J. Zagal, C. falo, R. rozas.
Puede que quieras extraer la leche justo después de tu última sesión de lactancia o de extracción	Sí	Extraer leche	Puede que quieras extraer el semen justo después de tu última sesión de lactancia o de extracción
dos fenómenos totalmente dispares: un hombre metido en luminal, tan tieso que sus pies se extienden más allá de la ventana	Sí	Metido tieso	dos fenómenos totalmente dispares: un hombre metido en luminal, tan tieso que sus pies se extienden más allá de la ventana
un señor estaba agarrando a un chiquilín, que parece que le había robado el celular	Sí	Agarrando chiquilín	un señor estaba agarrando a un recto, que parece que le había robado el celular
Existe una suerte, un pial especial, que es llamado "de la muerte" o "de la flecha" que consiste no en atorar la reata en la cabeza de la silla	Sí	Atorar reata	Existe una suerte, un pial especial, que es llamado "de la muerte" o "de la flecha" que consiste no en atorar el falo en la cabeza de la silla

5.2 Pruebas con muestras negativas

A continuación se muestran algunas de las frases célebres que fueron analizadas en las pruebas del sistema:

- `Sin música la vida sería un error.`
- `La música es para el alma lo que la gimnasia para el cuerpo.`
- `La música empieza donde se acaba el lenguaje.`
- `Estoy seguro de que la buena música la vida alarga.`
- `No basta con oír la música; además hay que verla.`
- `La música es la armonía del cielo y de la tierra.`
- `La música es un eco del mundo invisible.`

- `Cuando lo hayas encontrado, anótalo.`
- `No me etiquetes, léeme. Soy un escritor, no un género.`
- `El cine es un espejo pintado.`

En la Tabla 5.2 se presentan algunos ejemplos de análisis realizados por el sistema en las frases célebres como muestras negativas.

Tabla 5.2 Ejemplos de frases célebres analizadas[20]

Frase célebre	**¿Es albur?**	**Acción sexual**	**Frase traducida**
Sin música la vida sería un error	No	N/A	N/A
La música es para el alma lo que la gimnasia para el cuerpo	No	N/A	N/A
La música empieza donde se acaba el lenguaje	No	N/A	N/A
Estoy seguro de que la buena música la vida alarga	No	N/A	N/A
Cuando lo hayas encontrado, anótalo.	Sí	Anótalo encontrado	Cuando lo hayas encontrado, ano.
La música es la armonía del cielo y de la tierra	No	N/A	N/A
La música es un eco del mundo invisible	No	N/A	N/A
Cuando lo hayas encontrado, anótalo	No	N/A	N/A
No me etiquetes, léeme. Soy un escritor, no un género	No	N/A	N/A
El cine es un espejo pintado	No	N/A	N/A

Como se observa, la oración “Cuando lo hayas encontrado, anótalo” fue clasificada como albur debido a que cumple con la regla “verbo + sustantivo” explicada en 4.1.2.9. A la palabra “anótalo” se le añadió el significado de recto porque el módulo de metaplasmos lo consideró paragoge de “ano”.

[20] N/A: No aplica

5.3 Resultados

Como resultados de las pruebas realizadas a ambos conjuntos se obtuvo que el sistema identificó correctamente 448 de 553 albures. En cuanto a las muestras negativas el sistema identificó correctamente 529 de 568 frases célebres.

En la Tabla 5.3 se resumen en la matriz de confusión los resultados de las pruebas realizadas a las muestras positivas (albures) y a las muestras negativas (frases célebres).

Tabla 5.3 Matriz de confusión de la clasificación de albures

	Clasificados	
	Positivo	**Negativo**
Positivos reales	448	105
Negativos reales	39	529

A partir de la matriz de confusión se calculó la medida de precisión y cobertura con las siguientes fórmulas:

$$precisión = tp/(tp + fp)$$

$$cobertura = tp/(tp + fn)$$

Siendo, en este caso:

True positive (verdaderos positivos), $tp = 448$

False positive (falsos positivos), $fp = 39$

False negative (falsos negativos), $fn = 105$

Por lo que se obtuvieron las siguientes cifras:

$$precisión = 0.91$$

$$cobertura = 0.81$$

Al hacer una revisión manual de los albures que el sistema no logró identificar se percató de que algunos verbos utilizados en ellos carecían de forma semántica. Es decir, aunque sí se les fueron anotados sus correspondientes lemas y etiquetas gramaticales, ningún synset fue

asignado a estas palabras, p. ej. ablandar, enroscar, enfrascar, zampar, arrimar, jeringar, masturbar, mochar, pepenar, sobar, atorar, rolar, lamber, zambullir, entre otros.

Para hacer el análisis semántico del español convencional se utiliza *Freeling,* el cual a su vez utiliza como fuente la base de datos Multilingual Central Repository (MCR). Al tratarse este recurso de una adaptación de *WordNet* puede carecer de la asignación de *synsets* para algunas palabras como sucede en este caso.

Capítulo 6 Conclusiones

Este trabajo no se limitó solamente a clasificar un texto como albur o no albur, sino que además se identificaron cuáles son los elementos que componen a este tipo de humor. Se estudiaron distintas muestras de albures y se identificaron las características clave que los conforman. Estas características fueron utilizadas para desarrollar un método que busca albures ya sea intencionales o no, en un texto. En un inicio el alcance se fijó en abordar este tipo de humor a nivel fonético y semántico. Sin embargo, debido a la complejidad del fenómeno, hubo la necesidad de contemplar también los niveles morfológico y sintáctico al utilizar diccionarios y enunciar las reglas que realizan la clasificación.

Las características clave fueron identificadas y extraídas de textos que se sabía contenían albures, es decir, albures escritos de modo intencional. Sin embargo, el método propuesto a partir de estas características es útil no solo para identificar albures de esta naturaleza sino también para hacerlo con textos que contienen albures escritos de modo no intencional.

Se identificaron las estructuras más básicas que se utilizan para construir un albur, mismas que fueron aprovechadas para construir reglas que permiten identificar este tipo de humor en textos. Con el método propuesto y los recursos generados se logran identificar albures que se valen de la ambigüedad léxica y semántica para pasar desapercibidos, atacando así la polisemia, además de algunos fenómenos a nivel fonético, como los metaplasmos: apócope, aféresis, prótesis y paragoge.

El fenómeno del albur es bastante complejo. En este trabajo nos enfocamos a resolver el problema de clasificación de cierto tipo de albures: “verbo + sustantivo” y “verbo + pronombre átono”. Sin embargo, es posible extender fácilmente las reglas utilizadas, ya que éstas se proveen de modo independiente al sistema.

El albur tiene algunas variantes en las que se utilizan recursos como la rima y aliteración. A pesar de que en este trabajo no se abordaron estos fenómenos, es posible extender el método propuesto utilizando el repositorio de acciones sexuales, puesto que son la base del sentido oculto de los albures. Así mismo, con las características clave identificadas (repositorio de acciones sexuales y reglas) es posible desarrollar un sistema de generación automática de albures.

El método desarrollado fue implementado en un programa computacional y luego montado como un servicio web del que se pueden desarrollar diversas aplicaciones. Una posible utilidad de este servicio sería para depurar aquellas frases que pueden ser malinterpretadas en un texto de naturaleza formal en el que se desea absoluta seriedad.

Se generaron algunos recursos que pueden ser utilizados en otras aplicaciones de Procesamiento de Lenguaje Natural y, en específico para el humor computacional. Entre ellos destaca un diccionario morfosintáctico que puede ser fácilmente implementado utilizando la librería de *Freeling,* además de un diccionario semántico con el que se proveen significados (*synset*) para cada palabra del diccionario morfosintáctico.

Los recursos generados durante la investigación son independientes de este trabajo y pueden ser reutilizados en otras aplicaciones de PLN si así se desea. A continuación se enlistan dichos recursos los cuales se encuentran disponibles en línea[21]:

- Corpus de albures
- Corpus de frases ocultas
- Corpus de frases ocultas sin argot sexual
- Diccionario morfosintáctico del argot sexual mexicano
- Diccionario semántico del argot sexual mexicano
- Repositorio de acciones sexuales usadas en el albur
- Reglas para identificar el albur

El sistema que se desarrolló en esta investigación resulta efectivo en la identificación automática del albur. Sin embargo, los módulos que integran este sistema, junto con los recursos generados, pueden ser utilizados en la generación automática del albur.

[21] https://github.com/robertovillarejo/mexican_sexual_slang

https://github.com/robertovillarejo/corpus-de-albures

6.1 Trabajos futuros

Es posible utilizar el modelo presentado en este trabajo para identificar otro tipo de albures e incluso para generar chistes de este tipo. Como mejoras o extensiones a la investigación realizada, se proponen los siguientes puntos:

- **Aumentar el número de reglas:** construir una gramática con la que se definan patrones sintácticos que correspondan a las estructuras usadas en el albur. De esta manera se considerarían otros elementos presentes en el albur y la precisión del sistema mejoraría. Además, la gramática puede usarse para la generación automática de albures si se desea.
- **Extender el repositorio de acciones sexuales:** a fin de aumentar la cobertura del sistema, es decir, identificar un mayor número de albures se generarían nuevos pares de synset a partir de las existentes en el repositorio. Esto a través del análisis de relaciones semánticas entre éstos como la hiperonimia y la meronimia.
- **Considerar el uso de locuciones de índole sexual:** en este trabajo se consideraron algunas locuciones (o entradas multipalabra) en los diccionarios del argot sexual. Sin embargo al hacer una búsqueda más detallada de locuciones de este tipo se podría aumentar la cobertura del sistema.
- **Módulo de metaplasmos para enriquecer el diccionario morfosintáctico y reducir el tiempo de procesamiento:** como se detalla en 4.2.3.2, el módulo de metaplasmos añade significados a una palabra si existe cierta coincidencia en la forma fonética de ésta con la forma fonética de una palabra del argot sexual. Esta verificación se realiza en tiempo de ejecución del programa. Una mejora sería enriquecer el diccionario morfosintáctico para evitar la intervención del módulo de metaplasmos en tiempo de ejecución.

  P.ej., en el diccionario de *Freeling* la palabra "máscara" contiene la siguiente información:

  máscara máscara NCFS000

  Siendo que su forma fonética coincide con la palabra "masca", con el módulo de metaplasmos se podría añadir la siguiente información al diccionario morfosintáctico del argot sexual:

  máscara máscara NCFS000 mascar VMIP3S0 mascar VMM02S0

  De esta forma a la palabra "máscara" se le añadiría el significado de "mascar" (01201574-v). Con un diccionario morfosintáctico así no sería necesario realizar un análisis a nivel fonético y por lo tanto el sistema sería más rápido al analizar un texto.

No obstante, en un texto con albures escritos de manera intencional sería necesario que interviniera el módulo de metaplasmos puesto que, en este tipo de textos, es más probable encontrar palabras "inventadas" que son metaplasmos de otras palabras de argot sexual.

Referencias

ACL. (2016). *What is Computational Linguistics: Association for Computational Linguistics.* Obtenido de Association for Computational Linguistics: http://www.aclweb.org/website/what-is-cl

Barbieri, F., & Saggion, H. (2014). Automatic Detection of Irony and Humour in Twitter. *Proceedings of the Fifth International Conference on Computational Creativity*, (1975).

Barsoux, J.-L. (1993). *Funny business: humour, management and business culture.* Londres; Nueva York: Cassell.

Beristáin, H. (2000). El albur. *Acta Poética*, 399–422.

Binsted, K. (1996). *Machine humour: An implemented model of puns.* The University of Edinburgh. Retrieved from http://hdl.handle.net/1842/586

Cervantes, I. (Ed.). (2015). *Quiénes somos: Instituto Cervantes.* Recuperado el 04 de Mayo de 2015, de Instituto Cervantes: http://www.cervantes.es/sobre_instituto_cervantes/publicaciones_espanol/espanol_mundo/anuario_2014.htm

Daniels, G., Gervais, R., & Merchant, S. (2005). The Office. Television series, the National Broadcasting Company (NBC).

Díaz, H. B. (2001). La densidad figurada del lenguaje alburero. *Revista de Retórica Y Teoría de La Comunicación*, *I*, 53–60.

Diccionario panhispánico de dudas: Real Academia Española. (2005). Recuperado el 29 de 10 de 2015, de Real Academia Española: http://lema.rae.es/dpd/srv/search?id=elLl31yYnD65MTS9uF

Durán González, D. G. (2012). *El albur en la televisión: comunicación y entretenimiento para adolescentes.* Universidad Nacional Autónoma de México.

Francis, W. N., & Kucera, H. (1979). *A Standard Corpus of Present-Day Edited American English.* Department of Linguistics, Brown University.

Frisch, S. (1997). Similarity and Frequency in Phonology. *Dissertation*, (December 1996), 1–190. Recuperado de http://www.imamu.edu.sa/Scientific_selections/abstracts/Math/SIMILARITY AND

FREQUENCY IN PHONOLOGY.pdf

Gonzalez-Agirre A., L. E. and R. G. (2012). Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. *In Proceedings of the Sixth International Global WordNet Conference (GWC'12).*

Guzmán, E., Beltrán, B., Tovar, M., Vázquez, A., & Martínez, R. (2014). Clasificación de frases obscenas o vulgares dentro de tweets. *Research in Computing Science*, *85*(2014), 65–74.

Hernández, V. (2006). *Antología del Albur.* Toliro Multimedia and Incógnita|Caja Negra.

Janer, F. (1919). *Gramática castellana, para la enseñanza normal, secundaria y superior.* (B. a. Silver, Ed.) Universidad de Harvard.

Kiddon, C., & Brun, Y. (2011). That's What She Said : Double Entendre Identification. *Computational Linguistics*, 89–94. Recuperado de http://www.cs.washington.edu/homes/brun/pubs/pubs/Kiddon11.pdf

Latta, R. L. (1999). *The Basic Humor Process: A Cognitive-Shift Theory and the Case against Incongruity.* Berlín-Nueva York: Mouton de Gruyter. Recuperado el 25 de 05 de 2016

Lavertue, J. (1998). *El albur en México: descripción y percepción.* Université Laval, Ottawa, Canadá.

Mihalcea, R. (University of N. T., & Strapparava, C. (Istituto per la R. S. e T. (2005). Computational Laughing : Automatic Recognition of Humorous One-liners. In *Proceedings of the Cognitive Science Conference (CogSci)* (pp. 1513–1518). Stresa, Italy.

Miller, G. a. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. http://doi.org/10.1145/219717.219748

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1), 8–30. http://doi.org/10.1109/JRPROC.1961.287775

Mulder, M. P., & Nijholt, A. (2002). Humour research: State of the art. *CTIT Technical Reports Series.* Recuperado de http://doc.utwente.nl/38233/1/0000009e.pdf

Nass, C. I., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Computer-Human Interaction (CHI) Conference: Celebrating Interdependence 1994*, 72–78. http://doi.org/10.1145/259963.260288

Ocampo Pólito, R. (2010). *Detección automática de humor en textos cortos en español.* Instituto Politécnico Nacional.

Olguín Martínez, E. (s.f.). *Una definición lingüística del "albur."* Universidad Autónoma Metropolitana. Recuperado de http://tesiuami.izt.uam.mx/uam/aspuam/presentatesis.php?recno=1192&docs=UAM1192.PDF

Partington, A. S. (2009). A linguistic account of wordplay: The lexical grammar of punning. *Journal of Pragmatics*, *41*(9), 1794–1809. http://doi.org/10.1016/j.pragma.2008.09.025

Peralta de Legarreta, A. (2012). *El chilangonario: Vocabulario de supervivencia para el visitante de la Ciudad de México* (1 ed.). México, D.F.: D.R. Editorial Lectorum, S.A. de C.V.M.; D.R. Editorial Otras Inquisiciones, S.A. de C.V. Recuperado el 2015

Real Academia Española. (2014). *Diccionario de la lengua española (23. ed)*. Recuperado el 2016 de agosto de 31, de http://dle.rae.es/?id=KbZUpzR

Real Academia Española. (2014). *Diccionario de la lengua española (23. ed.)*. Recuperado el 31 de agosto de 2016, de http://dle.rae.es/?id=Xy8jXls

Real Academia Española. (2016). *Real Academia Española*. Obtenido de http://dle.rae.es/?id=RUl938s

Ren, Y., Kaji, N., Yoshinaga, N., & Kitsuregawa, M. (2013). Humor Identification in Microblog. Recuperado de http://www.tkl.iis.u-tokyo.ac.jp/top/modules/newdb/detail.php?id=1376

Reyes, A., Polit, U., Valencia, D., Taul, M., & Via, G. (2009). Características y rasgos afectivos del humor: Un estudio de reconocimiento automático del humor en textos escolares en catalán. *Procesamiento Del Lenguaje Natural*, *43*(Lenguajes y Sistemas Informáticos), 235–243. Recuperado de http://hdl.handle.net/10045/11718

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, *74*, 1–12. http://doi.org/10.1016/j.datak.2012.02.005

Ritchie, G. (1997). Developing the Incongruity-Resolution Theory, (1976).

Saussure, F. d. (1945). *Curso de lingüística general.* Buenos Aires: Editorial Losada.

Squire, M., & Gazda, R. (2015). FLOSS as a source for profanity and insults: Collecting the data. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2015-March*, 5290–5298. http://doi.org/10.1109/HICSS.2015.623

Taylor, J. M. (2008). *Toward Informal Computer Human Communication: Detecting Humor in a Restricted Domain. Academic medicine : journal of the Association of American Medical Colleges.* Recuperado de

http://www.ece.uc.edu/~mazlack/academic.UC/Julia.dissertation.pdf

Taylor, J. M. (2009). Computational detection of humor: A dream or a nightmare? the ontological semantics approach. *Proceedings - 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT Workshops 2009*, *3*, 429–432. http://doi.org/10.1109/WI-IAT.2009.318

Taylor, J. M. (2010). Ontology-based view of natural language meaning: The case of humor detection. *Journal of Ambient Intelligence and Humanized Computing*, *1*, 221–234. http://doi.org/10.1007/s12652-010-0014-2

Taylor, J. M., & Mazlack, L. J. (2004). Humorous Wordplay Recognition. In *International Conference on Systems, Man and Cybernetics Proceedings* (Vol. 4, pp. 3305–3311). The Hague, The Netherlands: IEEE. http://doi.org/10.1109/ICSMC.2004.1400851

Villarejo Martínez, R. (2016). *Github*. Obtenido de https://github.com/robertovillarejo/corpus-de-albures

Wells, J. (1997). *Handbook of standards and resources for spoken language systems* (Vol. 4). Recuperado el 25 de 05 de 2016

Apéndice A: recursos léxicos existentes

Entradas del diccionario "El chilangonario"

Abajeño: vello púbico.
Abierto: en el argot gay, es quien se expone, quien deja ver sus preferencias sexuales.
Activo: en el argot gay, se refiere al hombre que tiene la preferencia de penetrar en lugar de ser penetrado.
Aguacates: en sentido figurado, testículos.
Anacleto: corrupción y adaptación de "ano".
Anastasio: corrupción y adaptación de "ano".
Aníbal: corrupción y adaptación de "ano".
Anillo: corrupción y adaptación de "ano".
Anís: corrupción y adaptación de "ano".
Anófeles: corrupción y adaptación de "ano".
Asterisco: Ano. Corrupción y adaptación del anglicismo *ass*.
Bolas: Testículos.
Botapedos: Nalgas.
Bote: Corrupción de los anglicismos *butt* y *bottom*. Nalgas, trasero, parte baja.
Botiquín: adaptación de "bote". Glúteos.
Bubis: Corrupción del anglicismo *bubbies*. Senos.
Bújero: corrupción de *agujero*. Ano, vagina.
Cabezón: Miembro viril, pene.
Café: excremento
Gayo: homosexual
Huevo/Güevo: testículo
Joyo: corrupción de "hoyo"
Livais: lesbiana
Micrófono: metáfora para el falo

Apéndice B: corpus

Corpus de albures

Atentramente
A. Garrais el Chico
A. Garramel Pizarro
A. Soto Lama Kana
Abrám Eloyo
Agapito A. Prieto
Aida Melano
Ailejo De Mivara
Ajalandro el grande
Alberto Caraz Vergara
Alberto Carlos del Toro
Alfonso Jr. (o sea Poncho Chico)
Alicia la que me acaricia
Alisa Milano
Alma Cano Rosas
Alma Madero Benítez
Alma Madero de Palencia
Alma Madero de Vergara
Alma Marcela Ladez Mayo
Alma Marcela Rico
Alma Marcela Rico Cachorro
Alma Marcela Salta de lo Lindo
Alma Marcela Silva de Alegría
Alma Marcela Silva de Gusto
Alma María FierroBebé San el Grande
Carmela Garro
Damesio John
Elber Gon Sovas
Felipe Dosaco
Gisela Pico Oteo
Herculano Medellín
Isaac Amelo
Jaime Costecho
Yo tengo una casa en Agua Blanca y otra en Zacatepec
¿No vas a Querétaro que Tepic el Culiacán y luego Zacatecas los Pedro el Chico?
Para llegar a la avenida Gerenal Gasponte pasas las de atrás, y por los asentamientos
Despues de ir a Dallas te queda Colorado por Detroit
Hola vivo en Ecatepongo a gatas. El postre de aquí son los cacahuates saconeses y el chorizo en papas y el dulce favorito es el camote en barras de calabaza.
Agarraron al coyote cojo allá por el Cerro Tembroco rumbo a Loma Marías el Palo cerca del Edén pa dentro.
Palomas Ticas
Real Zamesta
Mascadas de caña
Sacos de cacahuate
Gorros de mamey
Alarmas Ticas
Alcóhol Adame
Café Zacoaztla
Cama Maste
Cama Nuela
Clavel negro
Cremería Mezacastle
Lana Algota
Banca Mote ¡La más moderna!
Cerveza la negra
Cerveza Colona y XX lambes

Cerveza la Boa
Cerveza La Negra
Cerveza Melwinni (la melwinni es la mejor)
Cerveza Tecabe
Cerveza Mesta
Cerveza Tetate
Charte Blanca
Dos Equis Larguer
Negra Mordelo
Pitoria
Rubia Chuperior
El Amigo Rito
El Cacarizo de Zacotitlán
El Coyote cojo de Laredo Texas
El Chico de Medellín
El Chico más común
El Chico medalla
El Chico prestamista
El Chico que ha andado mucho
El Chico Santiago
El Chico temido del vecindario
El Chico tímido
El Chico típico
El Gordo Peláez
El Grande Sumiso
El hijo de Javier Soliz: Solicito el chico
El Largo Melambes
El Maistro Nando
El Peladito Encajoso
El Pelón Casas
El Quemón Tolomeo
La Flaca Lagos Aras
La Morena que afloja las nalgas
La Muchacha Queta
La Prieta Cabezona de pechos caidos
La Tía Justa
El Beato Carlos del Toro
El Cardenal Gasponte
El Cura Melano
El Cura Melañonga
El Cura Melchor Izo
El Cura Melo
El Cura Meltrozo
El Niño de Agosto
El Obispo Nemelas
La Beata Carmen
La Hermana Rolanda
La Madre Sota
Pedro el Sacro

Corpus de frases en doble sentido

A continuación se muestran algunas instancias del corpus de frases en doble sentido junto con sus respectivos albures a partir de las cuales fueron generadas.

Instancias del corpus de frases en doble sentido

Albur	**Frase en doble sentido**
A. Garráis el chico	Agarráis el chico
A. Garramel Pizarro	Agárrame el pizarro
Berta Nates	Ver tanates
Damesio John	Dame ese hoyón
Carmela Garro	Me la agarro

Corpus de frases eróticas

A continuación se muestran algunas instancias del corpus de frases e eróticas junto con sus respectivas frases en doble sentido a partir de las cuales fueron generadas.

Instancias del corpus de frases eróticas

Frase en doble sentido	**Frase erótica**
Agarráis el chico	Agarráis el recto
Agárrame el pizarro	Agárrame el falo
Dame ese hoyón	Dame ese recto
Me la agarro	Me la agarro

Apéndice C: diccionarios del argot sexual mexicano

Diccionario morfosintáctico del argot sexual mexicano

abajeño abajeño NCMS000
abajeños abajeño NCMP000
abierto abierto NCMS000
abiertos abierto NCMP000
activo activo NCMS000
activos activo NCMP000
aguacate aguacate NCMS000
aguacates aguacate NCMP000
aguirre aguirre VMIP2S0
aire aire NCMS000
aires aire NCMP000
amargo amargo NCMS000
amargos amargo NCMP000
amigo amigo NCMS000
amigos amigo NCMP000
anacleto anacleto NCMS000
anacletos anacleto NCMP000
anastasio anastasio NCMS000
anca anca NCMS000
ancas anca NCFP000
ancho ancho NCMS000
aníbal aníbal NCMS000
aniceto aniceto NCMS000
anillo anillo NCMS000
anillos anillo NCMP000
anís anís NCMS000
ano ano NCMS000
anselmo anselmo NCMS000
aro aro NCMS000
aros aro NCMP000
arroz arroz NCMS000
asterisco asterisco NCMS000
asteriscos asterisco NCMP000
birote birote NCMS000
blanca blanca NCFS000
blanco blanco NCMS000
blanquillo blanquillo NCMS000
blanquillos blanquillo NCMP000
boa boa NCFS000
bola bola NCFS000
bolas bola NCFP000
botapedos botapedos NCMS000
bote bote NCMS000
botiquín botiquín NCMS000
bubi bubi NCFS000
bubis bubi NCFP000
bújero bújero NCMS000
bújeros bújero NCMP000
busto busto NCMS000
bustos busto NCMP000
cabeza cabeza NCFS000
cabezas cabeza NCFP000
cabezón cabezón NCMS000
cabo cabo NCMS000

cabos cabo NCMP000
cabús cabús NCMS000
caca caca NCFS000
cachagranizo cachagranizo NCMS000
cachagranizos cachagranizo NCMP000
cachete cachete NCMS000
cachetes cachetes NCMP000
cachirul cachirul NCMS000
cachirules cachirul NCMP000
cacho cacho NCMS000
café café NCMS000
cagado cagado NCMS000
cahuil cahuil NCMS000
cajeta cajeta NCFS000
calabaza calabaza NCFS000
camarón camarón NCMS000
camote camote NCMS000
caña caña NCFS000
carne carne NCFS000
carnita carne NCFS00D
cebo cebo NCMS000
cerote cerote NCMS000
chaira chaira NCFS000
chamba chamba NCFS000
chambrita chambrita NCFS00D
chango chango NCMS000
chaqueta chaqueta NCFS000
cheto cheto NCMS000
chicaspiano chicaspiano NCMS000
chicharra chicharra NCFS000
chicharrón chicharrón NCMS000
chichi chichi NCFS000
chichis chichi NCFP000
chichiterio chichiterio NCMS000
chico chico NCMS000
chile chile NCMS000
chimuelo chimuelo NCMS000
chiquilín chiquilín NCMS000
chiquistriquis chiquistriquis NCMS000
chiquistriquis NCMP000
chiquito chiquito NCMS000
chirris chirris NCMS000
chocho chocho NCMS000
chorizo chorizo NCMS000
chóstomo chóstomo NCMS000
chóstomos chóstomo NCMP000
cicirisco cicirisco NCMS000
cíclope cíclope NCMS000
cola cola NCFS000
coliflor coliflor NCFS000
colina colina NCFS000
colona colona NCFS000
concha concha NCFS000
gorro gorro NCMS000
gorros gorro NCMP000
coño coño NCMS000
corneta corneta NCFS000
corona corona NCFS000
cosita cosa NCFS00D
cráter cráter NCMS000
crema crema NCFS000
cremas crema NCFP000
cremita crema NCFS00D
cuarentaiuno cuarentaiuno NCMS000
cucaracha cucaracha NCFS000
cucu cucu NCMS000
cuero cuero NCMS000
culiacán culiacán NCMS000
culito culo NCMS00D
culo culo NCMS000
cutis cutis NCMS000
dona dona NCFS000
donas dona NCFP000
donilla dona NCFS00D
donita dona NCFS00D
donota dona NCFS00A
esperma esperma NCMS000
espinazo espinazo NCMS000
estuche estuche NCMS000
felación felación NCFS000

fierro fierro NCMS000
flaca flaca NCFS000
flacas flaca NCFP000
flauta flauta NCFS000
florecita flor NCFS00D
floripondio floripondio NCMS000
frijoles frijoles NCMP000
funda funda NCFS000
fundillo fundillo NCMS000
galleta galleta NCMS000
galletita galleta NCFS00D
garnachas garnachas NCFP000
garrote garrote NCMS000
gas gas NCMS000
gaver gaver NCFS000
gayeta gayeta NCMS000
gayo gayo NCMS000
geisha geisha NCFS000
gemelas gemelas NCFP000
gemelitas gemelas NCFP00D
gemidor gemidor NCMS000
genitales genitales NCMP000
gomas gomas NCFP000
gordo gordo NCMS000
gordos gordo NCMP000
gorro gorro NCMS000
gorros gorro NCMP000
grande grande NCMS000
grandes grande NCMP000
gruesa gruesa NCFS000
grueso grueso NCMS000
güero güero NCMS000
güeros güero NCMP000
güevo güevo NCMS000
güevos güevo NCPS000
gumaro gumaro NCMS000
gumaros gumaro NCMP000
herculano herculano NCMS000
hoyo hoyo NCMS000
hoyote hoyo NCMS00A
huevo huevo NCMS000
huevos huevo NCMP000
ignacias ignacia NCFP000
jalada jalada NCFS000
jamón jamón NCMS000
jediondo jediondo NCMS000
jeringo jeringar VMIP1S0
jeringueo jeringar VMIP1S0
jocoque jocoque NCMS000
jota jota NCFS000
jotita jota NCFS00D
joto joto NCMS000
jotoreto joto NCMS000
joyo joyo NCMS000
jugo jugo NCMS000
julián julián NCMS000
junior junior NCMS000
keiko keiko NCFS000
kundalini kundalini NCMS000
kundalinis kundalini NCMP000
largo largo NCMS000
largos largo NCMP000
larguesa larguesa NCMS000
leandro leandro NCMS000
leche leche NCFS000
lechita leche NCFS00D
leñero leñero NCMS000
leñeros leñero NCMP000
leño leño NCMS000
leños leño NCMP000
lilo lilo NCMS000
livais livais NCFS000
lodo lodo NCMS000
lola lola NCMS000
lolas lola NCFP000
lomo lomo NCMS000
longaniza longaniza NCFS000
macana macana NCFS000
macano macano NCMS000
majada majada NCFS000

mamada mamada NCFS000
mamey mamey NCMS000
mandarria mandarria NCFS000
manguera manguera NCFS000
manuela manuela NCFS000
marica marica NCMS000
maricón maricón NCMS000
marimacha marimacha NCFS000
mariposón mariposón NCMS000
marisco marisco NCMS000
mastique mastique NCMS000
mayate mayate NCMS000
mayatón mayate NCMS00A
mayonesa mayonesa NCFS000
mechona mechona NCFS000
meco meco NCMS000
mecos meco NCMP000
melón melón NCMS000
melones melón NCMP000
micrófono micrófono NCMS000
miembro miembro NCMS000
miércoles miércoles NCMS000
mierda mierda NCFS000
miringo miringo NCMS000
moco moco NCMS000
mocos moco NCMP000
mofle mofle NCMS000
mojón mojón NCMS000
mono mono NCMS000
nabo nabo NCMS000
nacha nacha NCFS000
nachas nacha NCFP000
nailon nailon NCMS000
nalga nalga NCFS000
nalgas nalga NCFP000
nalguillo nalga NCMS000
negra negra NCFS000
negras negra NCFP000
negro negro NCMS000
negros negro NCMP000
nena nena NCFS000
nenas nena NCFP000
niñas niña NCFP000
nola nola NCFS000
ñonga ñonga NCFS000
ñongas ñonga NCFP000
ojal ojal NCMS000
ojete ojete NCMS000
ojos ojo NCMP000
ojo ojo NCMS000
orto orto NCMS000
ostión ostión NCMS000
oyuki oyuki NCMS000
oyukis oyuki NCMP000
paja paja NCFS000
pajarito pájaro NCMS00D
pájaro pájaro NCMS000
pajarraca pajarraca NCFS000
palanca palanca NCFS000
palenque palenque NCMS000
paliacate paliacate NCMS000
palo palo NCMS000
panocha panocha NCFS000
panucho panucho NCMS000
papaya papaya NCFS000
papayón papaya NCMS00A
paquete paquete NCMS000
parada parada NCFS000
parado parado NCMS000
pedernal pedernal NCMS000
pedo pedo NCMS000
pedorro pedorro NCMS000
pedro pedro NCMS000
pelaco pelaco NCMS000
peladito pelado NCMS00D
peláez peláez VMIP2S0
pelícanos pelícano NCMP000
pelón pelón NCMS000
pelona pelona NCFS000
pelotas pelota NCFP000

pene pene NCMS000
penes pene NCMS000
penenoso penenoso NCMS000
pepa pepa NCFS000
pepperoni pepperoni NCMS000
pepino pepino NCMS000
pepita pepa NCFS00D
pescado pescado NCMS000
pescuezo pescuezo NCMS000
pescuezona pescuezón NCFS000
pescuezudo pescuezudo NCMS000
pescuezudos pescuezudo NCMP000
pestilente pestilente NCMS000
pestolete pestolete NCMS000
petacas petaca NCFP000
pewter pewter NCMS000
pija pija NCFS000
pinga pinga NCFS000
piola piola NCFS000
piolas piola NCFP000
pipo pipo NCMS000
pirinola pirinola NCFS000
pistola pistola NCFS000
pitín pito NCMS00D
pito pito NCMS000
pitomate pitomate NCMS000
pizarrín pizarro NCMS00D
pizarro pizarro NCMS000
pizpiote pizpiote NCMS000
plátano plátano NCMS000
pluma pluma NCFS000
pluto pluto NCMS000
polla polla NCFS000
pompas pompa NCFP000
pompis pompa NCFP000
popis popis NCFS000
popó popó NCFS000
pozo pozo NCMS000
pozos pozo NCMP000
prieta prieta NCFS000
prietas prieta NCFP000
pucha pucha NCFS000
púJaro púJaro NCMS000
puma puma NCMS000
puñal puñal NCMS000
puñeta puñeta NCFS000
puqueque puqueque NCMS000
purrún purrún NCMS000
putín puto NCMS00D
puto puto NCMS000
querétaro querétaro VMN0000
rabo rabo NCMS000
raja raja NCFS000
rajada rajada NCFS000
reata reata NCFS000
rifle rifle NCMS000
rinconera rinconera NCFS000
río río NCMS000
rondana rondana NCFS000
sable sable NCMS000
salami salami NCMS000
salchicha salchicha NCFS000
salchichón salchichón NCMS000
sambuto sambuto VMIP1S0
sebo sebo NCMS000
solecito solecito NCMS000
soplapollas soplapollas NCMS000
talento talento NCMS000
tamal tamal NCMS000
tamalón tamal NCMS00A
tanates tanate NCMP000
teclado teclado NCMS000
teclas tecla NCFP000
telera telera NCFS000
tepalcuana tepalcuana NCFS000
tepalcuanas tepalcuana NCFP000
tesorito tesoro NCMS00D
teta teta NCFS000
tetas teta NCFP000
tieso tieso NCMS000

tiesos tieso NCMP000
tola tola NCFS000
tompiates tompiate NCMP000
tornillo tornillo NCMS000
tornillos tornillo NCMP000
tortas torta NCFP000
tortilla tortilla NCFS000
tortillera tortillera NCFS000
tranca tranca NCFS000
trancas tranca NCFP000
trasero trasero NCMS000
tripa tripa NCFS000
tronco tronco NCMS000
troncoso troncoso NCMS000
trozo trozo NCMS000
tuerca tuerca NCFS000
uyuyuy uyuyuy NCMS000
vagina vagina NCFS000
vaginón vagina NCFS00A
vara vara NCFS000
vela vela NCFS000
velas vela NCFP000
venida venida NCFS000
verdolaga verdolaga NCFS000
verdura verdura NCFS000
verga verga NCFS000
vergudiño vergudiño NCMS000
vergudo vergudo NCMS000
vergudón vergudo NCMS00A
verija verija NCFS000
verruga verruga NCFS000
vestida vestida NCFS000
wawis wawis NCMS000
winni winni NCMS000
yoyo yoyo NCMS000
yoyos yoyo NCMP000
zanahoria zanahoria NCFS000
zanahorias zanahoria NCFP000
zanja zanja NCFS000
zanjas zanja NCFP000
zanjón zanja NCFS00A
zorra zorra NCFS000
zorrita zorra NCFS00D
zorro zorro NCMS000
zorros zorro NCMP000

10182913-n homosexual activo cuarentaiuno choto flor floripondio galleta gayeta gayo geisha jota joto leandro lilo marica mariposón marisco mayate nena pescado pluto púJaro puma puñal puqueque puto soplapollas vestida abierto cachagranizo gemidor maricón mesero_sin_charola

05538016-n ano anacleto anastasio aníbal aniceto anillo anís ano anselmo aro asterisco bújero cagado centro chicaspiano chico chimuelo chiquilín chiquistriquis chiquito chirris cicirisco corona cráter cueva dona eño fundillo herculano hoyo jediondo joyo julián junior kundalini miringo mofle negro ojal ojete ojo orto oyuki pedorro pestilente pestolete pozo rondana solecito tuerca uyuyuy yoyo tercer_ojo

05558717-n espalda espinazo

05457469-n esperma

14854262-n excremento caca café cahuil cajeta calabaza cerote frijoles keiko lodo majada mastique miércoles mierda mojón popis popó

00855169-n felación chamba mamada mamey wawis

05514081-n genitales

05368278-n glande cabeza

05570129-n glúteo cachete nacha nalga nalguillo

05559256-n glúteos bizcocho botapedos bote botiquín cachetes cabús cola coliflor colina colona cucu culiacán culo cutis garnachas ignacia jamón lola nacha nailon nalga petaca pompa rabo telera tepalcuana torta trasero

10254965-n lesbiana livais marimacha tortilla tortillera

00855674-n masturbación chaira chambrita chaqueta jalada manuela paja pajarraca puñeta lavar_a_mano

00839597-n flatulencia pedo aire gas pedernal pedro pewter pluma purrún

05526384-n pene sable salchichón amigo ancho birote blanca boa cabezón cabo cachirul cacho camarón camote caña chaparro carne cheto chile chorizo chóstomo cíclope corneta cosa cuero elote fierro flaca flauta garrote gaver gordo grande gruesa grueso güero largo larguesa leñero leño longaniza macana macano mandarria manguera micrófono miembro mil_arrugas mión moreno nabo negra nola ñonga pájaro palanca palenque paliacate palo paquete parada parado patas_de_bolillo pelado pelón pelona pene penenoso pepperoni pepino pescuezo pescuezón pescuezudo pija piola pipo pirinola pirinolo pistola pito pitomate pizarro pizpiote plátano polla prieta reata rifle rinconera rubia salami salchicha sin_hueso sin_uña tieso tola tornillo tripa tranca tronco troncoso trozo vara vela verdolaga verdura verga vergudiño vergudo verruga winni zanahoria

05560787-n pierna anca

05278714-n pubis verija

05404336-n semen agua amargo arroz blanco cebo crema jocoque jugo leche mayonesa meco moco ostión sebo venida
05554189-n seno busto chichi teta bubi chicharra chicharrón chichiterio gemelas gomas melón nena niña talento teclado tecla teta
05524615-n testículo aguacate bola huevo blanquillo bola güevo gumaro pelota tanate tompiate
05521514-n vagina concha coño estuche funda río tesoro
05263587-n vello_púbico abajeño chino pelaco pelícano
05521636-n vulva araña chango chimuela chino chocho coneja coño cucaracha higo hoyo mechona mono moñoñongo osa paloma panocha panucho papaya peluche pepa pliego pucha punk quesadilla raja rajada raya sapo tamal trompuda verija yoyo zanja
01212572-v agarrar aguirre
01382083-v pelar peláez
01825237-v querer querétaro
01577093-v zambullir sambuto
03088164-n condón gorro
01170983-v lamber
01432176-v lamber
02044278-v rolar
01525177-v atorar
02119874-v sobar
01381357-v pepenar
01442779-v ponchar
00463778-v alisar
01227088-v mochar
01430952-v masturbar
07436100-n jeringar
00990392-v arrimar
01507407-v zampar
01502279-v enfrascar
01523986-v enroscar
00840264-v ablandar
14855724-n orina amarillo
00846021-n cochar
05526957-n prepucio
04508489-n calzoncillo tanga

cenidet®

Centro Nacional de Investigación
y Desarrollo Tecnológico