Modelo de literariedad usando redes semánticas y n-gramas

Jonatan Ahumada Fernández

Tesis para el título de Ingeniero de Sistemas

Facultad de Matemáticas E Ingeniería
Fundación Universitaria Konrad Lorenz
Bogotá, Colombia
2022

Contents

1 FORMULACIÓN DEL PROBLEMA

1.1 Introducción

¿Qué constituye la esencia de un texto? ¿Qué diferencia un texto considerado 'literario' de aquél que no lo es? Esta pregunta se ha planteado en áreas como los estudios literarios y la lingüística [1]. Particularmente, la escuela denominada 'formalismo ruso' planteó que el objeto de estudio de la literatura, no *podría* ser la belleza, la relevancia histórica o el valor pragmático de un texto. Más bien, su objeto de estudio *debe* recaer en un aspecto más 'objetivo': su *literariedad*. Como su nombre sugiere, los formalistas se abocaron a formular una definición 'objetiva' y 'concreta' del fenómeno literario y adoptaron los –en ese entonces– modernos métodos de la buyente disciplina de la linguística.

Siendo este el caso, ¿no es, por consiguiente, factible que un autómata pueda medir y presentar tales características presuntamente formales con las actuales herramientas informáticas? ¿Cómo se podría traducir la noción de *literariedad* a un algoritmo que pueda ejecutar una máquina?

1.2 Planteamiento del problema

Roman Jakobson propone que la *literariedad* de un texto está dada por dos componentes de lenguaje: la diacronía y la sincronía. Estos elementos fueron expandidos de la teoría linguística de Saussure. Más tarde, puestos en el contexto del análisis de la poesía, Jakobson renombró esos dos ejes como *metáfora* y *metonímia*, en su texto "Linguística y poética".

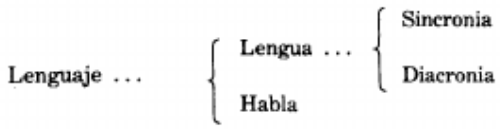

Figure 1: Distinción entre sincronía y diacronía

¿Es posible modelar algorítmicamente tales conceptos? Según Jakobson, en el estudio de la *literariedad* se omite el factor emisor y factor receptor. Tan solo se centra en el mensaje. Representado únicamente a través de un *medio* particular: en este caso, la palabra escrita. Es, por lo tanto, *factible* que un autómata pueda medir y presentar tales características.

Saussure ofrece ya un modelo cualitativo muy bien esbozado en teoría, que es el que luego Jakobson utilizará para definir la literariaded. Sin embargo, aunque existe un planteamiento cualitativo del problema, no se halló en la bib-

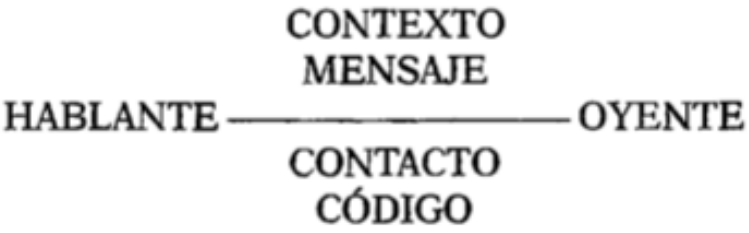

Figure 2: Factores de comunicación de Roman Jakobson [2]

liografía consultada un modelo computacional que modelara el concepto y lo implementara.

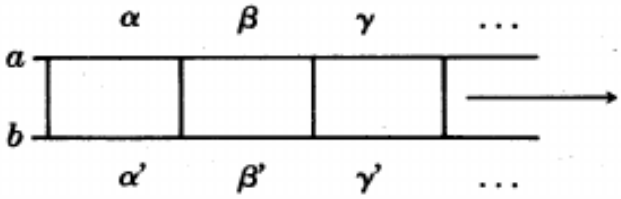

Figure 3: Modelo cualitativo inicial expuesto por Ferdinand De Saussure tomado de [1]

Preliminarmente, se puede observar que el modelo de Saussure se fundamenta en una estructura bastante familiar en la computación: la secuencia [3]. Así, el objetivo de este trabajo es modelar e implementar el modelo de *literariedad* de Roman Jakobson utilizando redes semánticas y n-gramas.

1.3 Justificacón

Si bien existen infinitas operaciones realizables sobre un texto computarizado, hay pocas que tengan un enfoque humanístico, sea este linguístico, literario o estético. Este enfoque busca generar una mayor comprensión del fenómeno literario, en contraposición a los enfoques 'típicos' –y hoy en día indispensables– de procesamiento de lenguaje: extracción de información, clasificación con base a un modelo predictivo, entre muchos otros [4].

Más aún, dentro de este subcunjunto reducido, pocos están guiados por aquello que Gelbuhk llama 'la ciencia fundamental'. A saber, la linguística. En otras palabras, hay un vacío en el campo de la linguística computacional en lo que se refiere a modelos que procuran cuantificar esta perspectiva.

Tal vacío genera que el estudio académico de la literatura no pueda sustentarse en datos 'duros' o ,por lo menos, cuantitativos propias del método científico. Por otro lado, los diversas y posibles formas de calcular la 'creatividad', la 'rima' o la 'belleza' de un texto, propuesto por otros investigadores, pueden considerarse casuísticos, acoplados a las objetivos y circunstancias de cada investigación en particular, desde la perspectiva de la linguística general.

Así, se necesita un modelo de la *literariedad* que exprese concretamente la metáfora y la metonimia. Bien sea para ampliar las aplicaciones de la linguistica computacional o para someter a escrutinio los planteamientos de la teoría.

En esta investigación se formulará y evaluará un modelo para obtener una medida cuantitativa para el concepto de *literariedad* de Roman Jakobson utilizando redes semánticas y n-gramas. De este modo, la presente investigación respondería a la pregunta ¿Cómo medir computarizadamente la *literariedad* de un texto según el marco de la lingüística de Jakobson?

1.3.1 Palabras clave:

NLP, computational linguistics, literariness,literary theory, poetics, theory of formal method

1.3.2 Área de conocimiento:

Lingüística computacional

1.4 Alcances y delimitaciones:

Para computar una métrica de *literariedad* será necesario comparar un *corpus objetivo* con respecto a un *corpus de referencia,* este último representará el 'uso corriente de la lengua'. La primera limitación de este trabajo es que no se compilará un corpus propio, sino que partirá de los de acceso libre. La mayoría de estos se encuentran en inglés. Por este motivo, el corpus de referencia más a la mano es WordNet, que al ser una ontología ya contiene las anotaciones necesarias para mi objetivo. A saber, una lista de sinónimos por palabras. Por otro lado, el corpus objetivo no tiene que estar anotado (utilizaré un PlainTextCorpus), pero de algún modo tiene que ser razonable su comparación con el corpus objetivo. Por ejemplo, los resultados del modelo serían muy difíciles de evaluar si la relación entre corpus objetivo y de referencia sobrepasa los 2 siglos, dada la naturaleza fluida de la lengua.

La segunda limitación concierne a la formulación de los algoritmos en sí mismos. Me limitaré a formular los modelos más naive posibles. Por ejemplo, (retomando el ejemplo previo) dada una palabra se considerará un sinónimo todas las palabras listadas como tal en el corpus de referencia, sin considerar los sub-problemas que esto podría conllevar.

En general, el alcance de este proyecto es formular e implementar un modelo general que muestre cómo sería viable implementar el concepto de *literariedad*, sin ahondar en los detalles que se desprenden de cada fase del flujo de NLP

(por ejemplo, ¿cómo tokenizar?, ¿Qué peso tendrían las diferentes partes de una oración en el computo final, etc).

2 OBJETIVO GENERAL

Diseñar e implementar un modelo que, dado un corpus de texto, produzca indicadores para el concepto de *literariedad* que plantea Roman Jakobson.

3 OBJETIVOS ESPECÍFICOS

1. Construir el corpus necesario para representar el *eje diacrónico*
2. Diseñar e implementar el algoritmo para calcular la *metáfora* sobre un corpus
3. Diseñar e implementar algoritmo para calcular la *metonimia* sobre un corpus
4. Seleccionar y unir los textos que serán procesados (corpus objetivo) por el algoritmo
5. Correr el algoritmo sobre los corpus objetivo
6. Evaluar el algoritmo de manera cuantitativa y cualitativa

4 MARCO TEÓRICO

4.1 Literariedad

La *literariedad* es, según Jakobson, la cualidad de un objeto literario en cuanto tal. Por lo tanto, la *literariedad* no depende de ningún factor extrínseco, como su emisor, su valor histórico, las ventas de tal o cual libro, las citaciones, etc. La *literariedad* se da exclusivamente por atributos propios del fenómeno del lenguaje.

Para analizar la *literariedad*, se deben analizar las dos operaciones más básicas de la conducta verbal: *la selección* y *la combinación.*

4.1.1 Selección (ver linguística sincrónica):

La selección estudia qué palabra selecciona un hablante entre las palabras existentes de la lengua, más o menos similares y hasta cierto punto equivalentes. La selección se basa en la sinonimia o antonimia de una palabra. En otros términos, en su semántica.

4.1.2 Combinación (ver linguística diacrónica):

La combinación estudia el "entramado de la secuencia" de un mensaje. Es decir, el mensaje considerado como una secuencia temporal y/o ordenada de palabras. La combinación se basa en la proximidad o, en otras palabras, en la relación de una palabra con la que la sucede o antecede en un mensaje.

4.2 Poética

La poética procura responder a la pregunta de ¿qué hace que un mensaje (verbal o de otra naturaleza) sea una obra de arte? Lidia principalmente con cuestiones estéticas del lenguaje. Sin embargo, para hacer un analisis exhaustivo, la poética debe hacer uso de la linguística, puesto que esta última estudia el lenguaje en todo su conjunto. La *literariedad* podría, entonces, considerarse un concepto enmarcado en la poética, porque se preguntá qué hace que un texto sea literario y por qué es distinto de otro que no lo es.

4.3 Linguística

La lingüística es la ciencia que estudia el lenguaje. Tradicionalmente, esta ciencia se subdivide en las ramas de fonética, fonología, morfología, sintaxis, semántica y pragmática.

La lingüística es un campo de estudio interdisciplinar e involucra disciplinas heterogéneas como la lógica y la neurolingüistica. Sin embargo, se considera que hay un núcleo común llamado *linguística general*.

4.3.1 Lingüística General:

Se conoce como lingüística general al paradigma lingüístico establecido por Ferdinand De Saussure, también llamado *modelo diferencial del lenguaje*.

El modelo diferencial se caracteriza porque propone dos ejes principales existentes en todo fenómeno lingüístico: el *eje de sincronía* y el *eje de diacronía*.

Estos dos ejes son la base de lo que Jakobson considera *selección* y *combinación*.

4.3.2 Linguística sincrónica

La linguística sincrónica se ocupa de las operaciones que realiza un hablante, sean lógicas o psicológicas, para formar un sistema linguístico. En el marco de esta investigación el *eje sincrónico* se referirá a las posibles palabras que un hablante pudo haber seleccionado para expresar una misma idea. Por ejemplo, para referirse a un niño, un hablante puede utilizar la las palabras "niño", "chico", "jovencito", o "párvulo".

4.3.3 Linguística diacrónica

La linguística diacrónica estudia los cambios sucesivos en el lenguaje, producidos por la actividad constante del *eje sincrónico*. En la perspectiva de Jakobson, un *mensaje* tiene en sí mismo un eje diacrónico. Tal eje mide la similaridad entre cada término del mensaje entindido como secuencia. Un ejemplo se puede apreciar en la oracion "I like Ike". An esta se evidencia una repetición de sonidos similares: [ay layk ayk]. La similaridad, no está dada por el significado, sino que aquí se proyecta a lo largo del tiempo:"(...) para decirlo de un modo más técnico: todo secuencia es un símil."

4.4 Lenguaje

En términos simples, el lenguaje es la facultad de formular y comprender signos o símbolos, ya sean hablados, escritos, imágenes, etc. En otros términos, el lenguaje es una capacidad general. Sin embargo, para Saussure, la lengua tiene una característica doble: que es al mismo tiempo un sistema establecido y la constante evolución de tal sistema. Estos dos componentes son la *lengua* y el *habla*.

4.4.1 Lengua

La lengua (*langue*) es uno de los dos componentes del *lenguaje*. La lengua es fenómeno social y se equipara a una *cristalización* o un producto de la suma de asociaciones entre conceptos e imagenes acústicas en la mente de los hablantes. Por ejemplo, la lengua es lo que permite que dos hablantes bogotanos puedan asociar en su mente el sonido de la palabra "chino" con el concepto de "niño" o "infante", mientras que en otras partes del mundo hispanohablante no existe tal asociación común. En términos simples, la lengua es un entendimiento compartido de lo que significan las palabras. La contraparte de la lengua, es el habla.

4.4.2 Habla

El habla (*parole*) es uno de los dos componentes del *lenguaje*. El habla es el uso individual de la lengua. Evidentemente, cuando un individuo habla puede modificar la lengua a su antojo, porque posee la facultad del lenguaje y jamás meramente repite el consenso de la lengua. Como consecuencia de esto, la lengua está continuamente siendo transformada por el habla. En términos simples, la suma de los actos individuales de comunicacion lentamente terminan por transformar el consenso social sobre cómo hablar. Por este motivo la linguística debe tener una perspectiva doble: *diacrónica* y *sincrónica*.

4.5 Lingüística Computacional

Es la intersección entre la computación y la lingüística. Por lo general, se preocupa acerca de cómo procesar automáticamente el lenguaje material, para lo

cual genera modelos lingüísticos sobre los que luego se pueden definir operaciones comunes [4].

La lingüística computacional es en sí misma un campo amplio y heterogéneo, pero en términos de este trabajo, me limitaré a señalar una herramienta:

4.5.1 NLTK

El Natural Language Toolkit (NLTK) es un módulo de Python que ofrece una interfaz para tareas comunes en la lingüística computacional. La ventaja principal de NLTK es que se considera a sí mismo un *toolkit*. Esto significa que no impone una estructura de procesamiento definida a la vez que ofrece un extenso abanico de herramientas, tales como: tokenizacion, filtros, generación de n-gramas, análisis sintáctico de oraciones, entre otras.

4.5.2 Corpus

Un corpus es una colección de textos auténticos que pueden ser leídos por una máquina. Estos pueden estructurarse de muchas formas, dependiendo de los objetivos de la investigación [5]. Por ejemplo, pueden ser aislados (una colección arbitraria), categorizados (una colección escogida según algún criterio), temporales (una colección organizada cronológicamente) o solapados (un documento puede pertenecer a varias colecciones) [6]. Además, el formato del corpus varía significativamente de acuerdo al objeto de la investigación. Por ejemplo, si se desea hacer un análisis sintáctico (de la estructura de una oración), se debe hacer un corpus anotado con POS (Part Of Speech tag); para hacer un análisis pragmático se utiliza una anotación pragmática, etc.

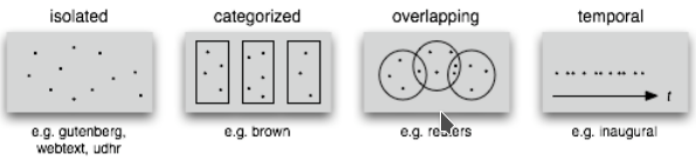

Figure 4: Diferentes estructuras de corpus

5 MARCO REFERENCIAL

El trabajo de Delmonte [7] presenta a SPARSAR, un sistema para calcular automáticamente el estilo de la poesía. SPARSAR funciona sobre sistemas previos del mismo autor, como, por ejemplo, un analizador semántico [8]. Delmonte tiene una larga trayectoria en el modelamiento de conceptos lingüísticos "difíciles", como la prosodia y la rima en términos cuantitativos.

El aporte principal de Delmonte fue su innovación al momento de aplicar herramientas comunes de NLP (tokenizadores, splitters y NER) con el fin de analizar aspectos estilísticos de un texto. Los modelos de Delmonte son muy cercanos a la teoría lingüística y propone soluciones a aspectos complejos del análisis lingüístico. Esta proximidad me llevo a plantearme la pregunta ¿qué otros aspectos del lenguaje valdría la pena modelar que aún no hayan sido abordados desde una perspectiva computacional? Así mismo, Delmonte reporta que hay pocos trabajos en el área con este mismo enfoque. Esta fue una inspiración para explorar más en el tema y ofrecer un enfoque distinto, tal como él lo hizo.

Sin embargo, Delmonte no revela detalles de implementación de sus sistemas en los artículos revisados. Además, sus sistemas tienen una alcance mucho mayor que el dispuesto para este trabajo, por lo que para mayores detalles tuve que referirme a otros trabajos.

El trabajo de [9] establece una métrica para medir el grado de creatividad en la poesía, basándose en qué tanto de la rima se conserva en la traducción de un poema con respecto al original. Tomé de Zuñiga la idea de establecer una métrica para un aspecto tradicionalmente cualitativo (la creatividad). Lo que diferencia este trabajo del de Delmonte, es su aproximación matemática. Particularmente, Zuñiga ofreció una forma naive de calcular similitud en rima, sin necesidad de recurrir a construcciones que requieren de recursos léxicos complejos como una ontología para fonemas, etc.

Por último, el trabajo de [10] es una tesis de pregrado sobre el cálculo del estilo de la poesía desde una perspectiva estadística. Kaplan fue una inspiración para Delmonte, por lo tanto debía formar parte de mi revisión bibliográfica. Kaplan formuló un modelo que media 84 métricas distintas para cada documento, luego transformó el modelo de 84 métricas para visualizarlo en un espacio 3D y poder comparar distintas obras literarias. Esto inspiró mi idea inicial de obtener una métrica más general para analizar un texto, que no tenga que recurrir un trabajo de compilación de métricas existentes, como lo hizo Kaplan. Tal métrica debería estar sustentada en conceptos linguísticos, para lo cual recurre en los conceptos presentados en el marco teórico.

6 DISEÑO METODOLÓGICO

El diseño metodológico seguirá –a grandes rasgos– los pasos de la metodología CRISP-DM, que se considera un estándar *de facto* para proyectos de minería de datos. Esta metodología ayudará organizar el proceso de mi investigación, que vá desde el acceso a los corpus (los datos disponibles) hasta el despliegue (la visualización de los resultados).

6.1 Entendimiento del negocio

El resultado tangible del modelo de literariedad propuesto son dos métricas cuantitativas: *metáfora* y *metonímia*. Estas métricas juntas constituiran una representación 'objetiva' del concepto cualitativo de *literariedad*.

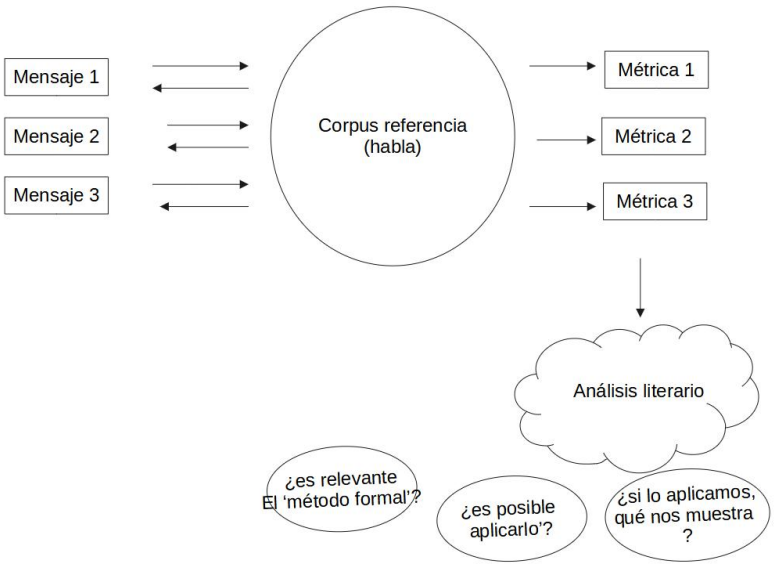

Figure 5: Entradas y salidas del algoritmo

¿Cuál sería el beneficio de obtener este resultado? Se podría comparar las métricas de n mensajes cualesquiera y tener una medida objetiva con las cuales compararlas. Algunos casos de uso posible serían:

- determinar si un mensaje que yo he escrito es más metáforico o metonímico que otro.

- determinar si un mensajes de una misma categoria (por ejemplo, del mismo autor, o del mismo género) tienen medidas de métadora y metonímia similares.

- correr grandes grupos de mensajes, por ejemplo, 'poemas de la escuela simbolistas' y compararlo con 'poemas realistas' y verificar si hay o no una diferencia sustancial desde el punto de vista linguístico .

Como se puede apreciar (ref:fig:posibles$_{\text{usos}}$), las aplicaciones del modelo en principio supondrían un factor adicional para ser considerado para el estudio literario, cuya naturaleza es cualitativa. Sin embargo, si el modelo demuestra ser efectivo, podría llegar a ser una medida de similitud para un texto, lo que implicaría que se podría clasificar un texto con base en su metáfora y metonímia,

6.2 Entendimiento de los datos

En esta sección, se enumeraran las distintas fuentes de datos, que en este caso vendrían a ser los diferentes tipos de corpus.

6.2.1 El corpus de referencia

El corpus de referencia es un compendio de muestras que terminará por representar un consenso sobre el uso de la *lengua*. Su correlación teórica es el eje de diacronía y cumple la función de cristalizar una lengua en un lugar y un tiempo establecido. A nivel de implementación, se trata de un cadena muy larga compuesto de muestras seleccionadas según criterios aptos (ver sección sobre preparación de los datos).

6.2.2 El corpus objetivo

El corpus objetivo serán los mensajes sobre los cuales se computarán las dos medidas de *metáfora* y *metonimia*. Su correlativo teórico es el *habla* y son los textos que el usuario final del final del sistema desea someter a análisis. A nivel de implementación, cada mensaje es una cadena (que corresponde a un documento real), pero en su totalidad el corpus objetivo es mucho más pequeño que el corpus de referencia, del mismo modo en que una persona que profiere una oración utiliza un subconjunto mucho más pequeño de la lengua a la que pertenece.

6.2.3 La red semántica

La red semántica es un tipo de corpus particular que no solamente consta de palabras anotadas, como el de Brown, sino que vincula las palabras por su relación conceptual con otras palabras. La red semántica correspondería a la facultad de asociar conceptos con las "imágenes acústicas" (las palabras) de Saussure. En esta investigación, la red semántica se utilizará para obtener sínonimos de palabras, que representarán conceptos. Tal red no será implementada, sino que será un servicio utilizado por el algoritmo.

6.2.4 Resumen de entendimiento de los datos

6.3 Preparación de los datos

La tarea de preparación de los datos consitirá principalmente en seleccionar los distintos tipos de corpus de manera significativa y coherente. A continuación, describiré cómo se conformaron los corpus y qué criterios se utilizaron.

6.3.1 Corpus de referencia

El corpus de referencia representa la *lengua* (*langue*). Por lo tanto debe estar compuesto de una muestra de textos comparativamente mucho más grande los

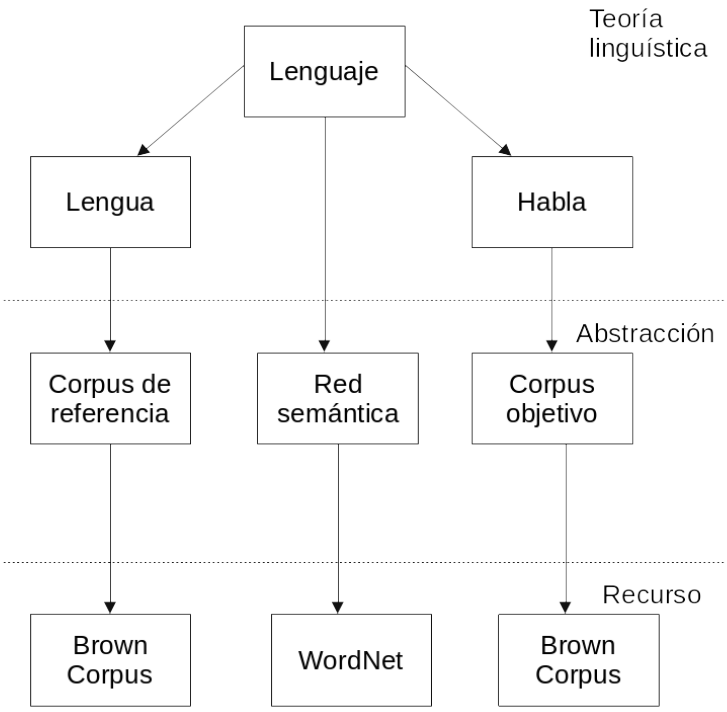

Figure 6: Resumen de las fuentes de datos utilizadas para cada concepto

mensajes individuales que serán contrastados con este. ¿Cómo construir un corpus tal?

En primer lugar, se descartó la idea de modelar la *lengua* en su totalidad, pues como lo indica la teoría linguística, esta tarea es imposible puesto que esta se encuentra en constante cambio. Así, el primer criterio para construcción del corpus fue restringirlo diacrónicamente al espacio de un año y a un idioma específico.

El siguiente criterio fue armar un corpus *balanceado*. Es decir, el corpus de referencia no puede estar compuesto de muestras de un mismo tipo (un estilo, un género, un autor), porque esto sesgaría la comparación de el corpus objetivo con respecto a este. Así, se optó por partir de un corpus *categorizado* y tomar partes iguales de cada una de las categorias. Esto es, cada categoría tiene igual peso en cuanto a número de textos y palabras que lo representan.

El tercer criterio fue utilizar un corpus fácilmente accesible, de origen libre y avalado por la comunidad científica. Por todos los motivos anteriores, se escogió el corpus de Brown, que presenta las siguientes características:

- todas las muestras del corpus pertenecen al año 1961
- todas las muestras del corpus se imprimieron en Estados Unidos durante ese año
- todos los autores son hablantes nativos de inglés
- la categorización de las muestras fue hecha por un comité de expertos de la universidad de Brown
- la intención declarada del corpus es la de ser una muestra representativa del inglés de aquel año
- tiene una lista amplia de categorías que podrían ser útiles para observar diferencias entre las categorías
- los resultados obtenidos del modelo podrían ser replicados porque el corpus es ampliamente conocido

En la tabla 1 se muestra lo que se utilizará como corpus de referencia.

cód.	nombre	categoria
a01	Political Reportage	reportage
a11	Sports Reportage	reportage
a19	Spot News	reportage
a26	Financial Reportage	reportage
a40	People, Art & Education	reportage
b03	Editorials	editorial
b08	Columns	editorial
b15	Letters to the editor	editorial

b19	The Voice of the people	editorial
b24	Reviews	editorial
d15	Zen:A Rational critique	religion
d11	War & the Cristian Conscience	religion
d13	The New Science & The New Faith	religion
d04	The Shape of death	religion
d02	Christ Without Myth	religion
e05	The Younger Generation/Use of Common Sense Makes Dogs Acceptable	skills & hobbies
e06	The American Boating Scene	skills & hobbies
e10	The New Guns of 61	skills & hobbies
e19	How to Own a Pool and Like It	skills & hobbies
e23	The Watercolor Art or Roy Mason	skills & hobbies
f07	How to Have a Successful Honeymoon/Attitudes Toward Nudity	popular lore
f12	New Methods of Parapsychology	popular lore
f13	Part-time Farming	popular lore
f14	The Trial and Eichmann	popular lore
f33	Slurs and Suburbs	popular lore
g15	Themes and Methods: Early Storie of Thomas Mann	belles lettres
g13	Sex in Contemporary Literature	belles lettres
g18	Verner von Heidenstam	belles lettres
g26	Two Modern Incest Heroes	belles lettres
g28	William Faulkner, Southern Novelist	belles lettres
j18	Linear Algebra	learned
j17	Prolegomena to a Theory of Emotions	learned
j28	Perceptual Changes in Psycopathology	learned
j39	Stock, Wheats and Pharaohs	learned
j35	Semantic Contribution of Lexicostatistics	learned
k18	Midcentaury	general fiction
k25	The Prophecy	general fiction
k04	Worlds of Color	general fiction
k23	The Tight of the Sea	general fiction

k17	Mila 8	general fiction
l05	Bloodstain	mistery and detective fiction
l11	The Man Who Looked Death in the Eye	mistery and detective fiction
l04	Encounter with Evil	mistery and detective fiction
l19	Make a Killing	mistery and detective fiction
l20	Death by the Numbers	mistery and detective fiction
m01	Stranger in a Strange Land	science fiction
m03	The Star Dwellers	science fiction
m04	The Planet with no Nightmare	science fiction
m05	The Ship who Sang	science fiction
m06	A Planet Named Shayol	science fiction
n01	The Killer Marshall	adventure and western fiction
n05	Bitter Valley	adventure and western fiction
n15	Sweeny Squadron	adventure and western fiction
n20	The Flooded Deares	adventure and western fiction
n26	Toughest Lawman in the Old West	adventure and western fiction
p29	My Hero	romance and love story
p27	Measure of a Man	romance and love story
p22	A Husband Stealer from Way Back	romance and love story
p16	A Secret Between Friends	romance and love story
p12	A Passion in Rome	romance and love story

Table 1: Corpus de referencia

6.3.2 Corpus objetivo

En contrapartida al corpus de referencia, el corpus objetivo representa el *habla* (*parole*). Así, estos son considerados mensajes que serán interpretados por el receptor con relación al consenso de la lengua compartida entre emisor y recep-

tor.

El primer criterio para construir el corpus de referencia es que este tenga una delimitacion diacrónica igual a la de el corpus objetivo. El segundo criterio, que las categorías fueran comparables a las categorias establecidas del corpus de referencia.

El tercer criterio es que cada muestra del corpus del corpus objetivo tuviera un tamaño similar entre sí, para descartar que diferencias en la longitud del mensaje afectaran sustancialmente los resultados del algoritmo

Por estos motivos, se optó por tomar tomar muestras del mismo corpus de Brown. La diferencia radica en que cada categoría solo tiene una muestra y la muestra seleccionada para la categoría está ausente en el corpus objetivo. Así, el corpus objetivo presenta las siguientes características:

- es una muestra 'miniatura' del corpus de Brown
- la relación de tamaño entre el corpus objetivo y el corpus de Brown es de 1:5
- Cada categoría en el cropus objetivo tiene su correlativo en el de referencia y viceversa
- el tamaño de cada muestra es de cerca de 2000 palabras

A continuación, se presenta un resumen del corpus objetivo en las tablas 2, 3, 4, 5 y 6.

cód	nombre	categoría
a40	People. Art & Education	reportage
b27	Letters to the Editor	editorial
c17	Reviews	reviews
d09	Organizing the Local Church	religion
e36	Renting a Car in Europe	skills & hobbies
f48	Christian Ethics & the Sit-In	popular lore
g75	A Wreath for Garibaldi	belles lettres
h30	Annual Report of Year Ending June 30:1961	miscellaneous
j80	Principles of Inertial Navigation	learned
k29	The Sheep's in the Meadow	general fiction
l24	The Murders	mistery and detective fiction
m02	The Lovers	science fiction
n29	Riding the Dark Train Out	adventure and western fiction
p20	Dirty Dig Inn	romance and love story

Table 2: Corpus objetivo 1

cód	nombre	categoría
a02	The Dallas Morning News	reportage
b01	The Atlanta Constitution	editorial
c01	Chicago Daily Tribune	reviews
d01	William G. Pollard Physicist and Christian	religion
e02	Organic Gardening and Farming	skills & hobbies
f01	How Much Do You Tell When You Talk?	popular lore
g01	Northern Liberals and Southern Bourbons	belles lettres
h01	Handbook of Federal Aids to Communities	miscellaneous
j01	Radio Emission of the Moon and Planet	learned
k01	First Family.	general fiction
l02	Bachelors Get Lonely	mistery and detective fiction
m01	Stranger in a Strange Land	science fiction
n02	The Valley	adventure and western fiction
p01	A Cup of the Sun	romance and love story

Table 3: Corpus objetivo 2

cód	nombre	categoría
a03	Chicago Daily Tribune	reportage
b02	The Christian Science Monitor	editorial
c02	The Christian Science Monitor	reviews
d03	Christian Unity in England	religion
e03	Will Aircraft or Missiles Win Wars?	skills & hobbies
f02	America's Secret Poison Gas Tragedy	popular lore
g02	Toward a Concept of National Responsibility	belles lettres
h02	An Act for International Development	miscellaneous
j02	Proceedings of the 1961 Heat	learned
k02	The Ikon	general fiction
l03	Encounter with Evil	mistery and detective fiction
m03	The Star Dwellers	science fiction
n03	Trail of the Tattered Star	adventure and western fiction
p02	Seize a Nettle	romance and love story

Table 4: Corpus objetivo 3

6.4 Modelamiento

6.4.1 Selección de técnica de modelado

Esta investigación se enmarca dentro de un enfoque mixto, en donde se utilizan métodos tanto cualitativos (el marco teórico) como cuantitativos, por lo tanto, hay varias técnicas implicadas en el modelado.

Desde el aspecto cuantitativo, se utilizan técnicas conocidas dentro del NLP, como tokenizacion, n-gramas y bag-of-words. Estas técnicas se utilizan como

cód	nombre	categoría
a04	The Christian Science Monitor	reportage
b04	The Miami Herald:September	editorial
c03	The New York Times	reviews
d05	Theodore Parker: Apostasy within Liberalism	religion
e04	High Fidelity	skills & hobbies
f03	I've Been Here before!	popular lore
g03	The Chances of Accidental War	belles lettres
h03	87th Congress: 1st Session. House Document No. 247.	miscellaneous
j03	The Normal Forces and Their Thermodynamic Significance	learned
k03	Not to the Swift	general fiction
l06	Hunter at Large	mistery and detective fiction
m04	The Planet with No Nightmare	science fiction
n04	The Shadow Catcher	adventure and western fiction
p03	The Fairbrothers	romance and love story

Table 5: Corpus objetivo 4

cód	nombre	categoría
a05	The Providence Journal	reportage
b05	Newark Evening News	editorial
c04	The Providence Journal	reviews
d06	Tracts published by American Tract Society	religion
e07	How to design your Interlocking Frame	skills & hobbies
f04	North Country School Cares for the Whole Child	popular lore
g04	The Invisible Aborigine	belles lettres
h04	Rhode Island Legislative Council. Research Report Number 1	miscellaneous
j04	Proton magnetic resonance study	learned
k05	The Judges of the Secret Court	general fiction
l07	Deadlier Than the Male.	mistery and detective fiction
m05	The Ship Who Sang	science fiction
n06	Here Comes Pete Now.	adventure and western fiction
p04	The Moon and the Thorn.	romance and love story

Table 6: Corpus objetivo 5

medios de vectorización, mediante lo cual se logra un transformación de un texto (una variable cuantitativa) a una representación númerica, (la matriz de uso).

Desde el aspecto cualitativo, se hizo una revisión de la literatura y de la intuición para acotar los planteamientos de la teoría, los conceptos de *lengua* y *habla*, hasta una formulación cuantificable con los métodos descritos.

6.4.2 Diseño experimental

Una vez formulado el modelo, se conduce un experimento que evaluará si produce resultados satisfactorios. El objetivo del experimento es escudriñar si los valores arrojados para los índices propuestos son coherentes con las intuiciones detrás del marco teórico y/o con el 'juicio experto'.

El experimento se basa en una cualidad del corpus de referencia seleccionado: su categorización. Por lo tanto, como se explica en la sección 6.3, se seleccionaron muestras del Corpus de Brown de tal modo que cada categoría está representada igualmente en cada muestra. Así, luego de procesar las muestras, se compararán los resultados por cada categoría.

El modelo se considerará existoso si los valores del índice metafórico e índice metonímico son consistentes a lo largo de las muestras para cada categoría.

Además, dentro de cada muestra, se espera que se cumplan ciertas hipótesis:

- H1: Se espera que las categorías de ficción tengan un índice metafórico significativamente mayor que los de no-ficción
-
- H2: Se espera que las categorias 'Reportage' y 'Editorial' tengan índices metafóricos similares a través de las muestra
- H3: Se espera que la categoría 'Learned' tenga un indíce metafórico más alta entre las categorías de no-ficción
- H4: Se espera que la categoria 'Learned' tenga un indice metonímico bajo en general

No se formularán más hipotesis acerca del índice metonímico, pues según los planteamientos teóricos este indicador es sensible especialmente al género de poesía, que no está presente en la muestra por las limitaciones del corpus seleccionado.

6.4.3 Presentación del modelo

1. Presentacion de las ecuaciones

$$mensaje = \{w_1, w_2, w_3, \ldots, w_j\} \tag{1}$$

$$vector\ semantico(w) = \{s_1, s_2, s_3, \ldots, s_j\} \tag{2}$$

$$vector\ uso(w) = \{freq(s_1), freq(s_2), freq(s_3), \ldots, freq(s_j)\} \tag{3}$$

$$uso(w) = \frac{freq(w)}{uso} \tag{4}$$

$$indice\ metaforico(mensaje) = \Sigma_i^j \frac{uso(w_i)}{\mu(vector\ semantico(w_i))} \tag{5}$$

$$N = \{n_1, n_2, n_3, \ldots, n_j\} \tag{6}$$

$$met(n) = \frac{letras\ iguales}{set(letras(n_i 1) + letras(n_i 2)} \tag{7}$$

$$indice\ metonimia = \Sigma_i^j met(n_i) \tag{8}$$

2. Procedimientos para indicadores
3. Matrices de semántica y de uso
4. Índice Metonímico

6.5 Despliegue

Presentación de resultados:

Table 7: Muestra 1

categoria	metafora	metonimia	w
reportage	880514.226605173	232.266917233093	2340
editorial	880324.393897166	245.719531857031	2262
reviews	929802.38416219	242.953762332438	2370
religion	850127.6846531	264.683072130827	2314
skills & hobbies	831781.725628903	242.632252469752	2232
popular lore	833825.825225262	265.83988095238	2222
belles lettres	877690.52541314	229.785869685869	2288
miscellaneous	782613.273615479	278.192915417915	2214
learned	863208.047211933	266.998263827676	2254
general fiction	891211.57527208	249.95016095016	2264
mistery and detective fiction	1032943.85669407	244.615023865023	2446
science fiction	1064426.54657215	235.067805233981	2412
adventure and western fiction	1234204.19460692	229.817769158945	2560
romance and love story	993413.094671098	217.506968031968	2428

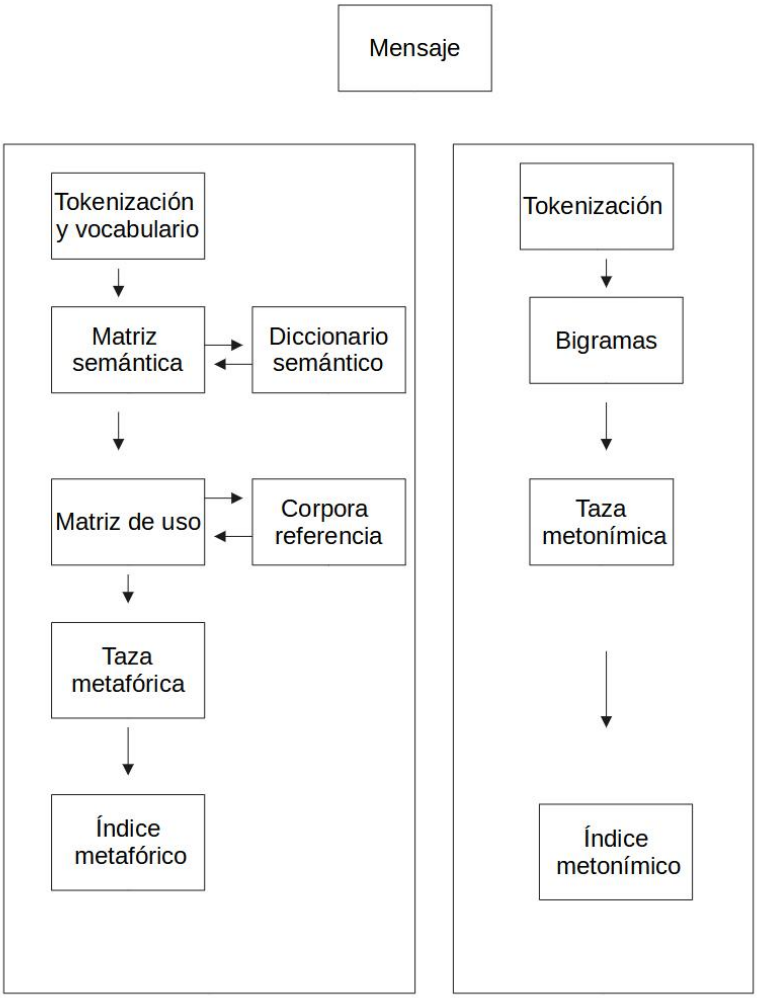

Figure 7: Procesamiento de corpus objetivo

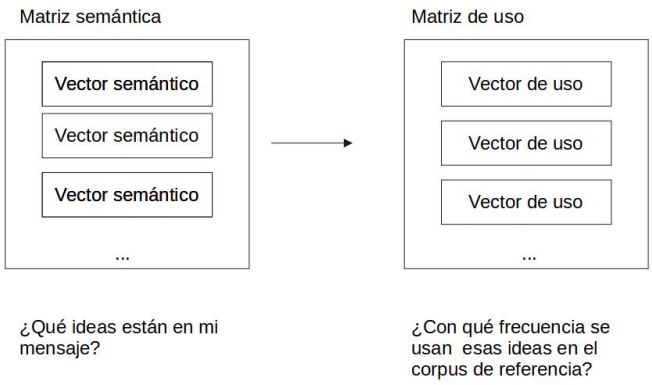

Figure 8: Abstracciones necesarias para el índice metafórico

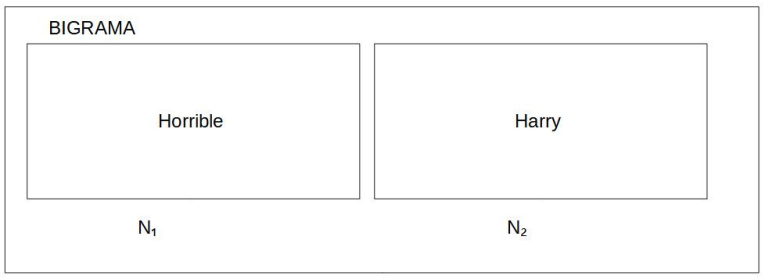

Figure 9: Abstraccciones necesarias para el indice metonímico

Table 8: Muestra 2

categoria	metafora	metonimia	w
reportage	869205.2371696023	233.99592490842463	2277
editorial	777241.5394134748	252.29809496059465	2200
reviews	978095.225396233	242.3226565101564	2415
religion	831466.3628116096	234.21091131091077	2213
skills & hobbies	833209.3790445685	237.43338605838585	2279
popular lore	965391.1906183016	270.5444999444997	2369
belles lettres	863139.7507327744	279.74454989454966	2289
miscellaneous	873426.7117151126	302.2738428238428	2416
learned	912477.0323082526	241.59998334998312	2189
general fiction	1025249.8452137534	243.0625180375174	2440
mistery and detective fiction	959584.2017381956	231.74134476634435	2370
science fiction	1049847.7175834612	260.93059440559404	2486
adventure and western fiction	1079790.9124281127	232.90989288489175	2383
romance and love story	969075.2121776282	261.1946331446324	2332

Table 9: Muestra 3

categoria	metafora	metonimia	w
reportage	832961.122494042	253.461402486402	2275
editorial	798751.012651529	266.66209346209246	2234
reviews	884194.0844699917	249.01867299367268	2320
religion	831865.8440237658	266.0598665223664	2332
skills & hobbies	850383.4965037219	263.1010350760349	2257
popular lore	869221.9181097293	245.8761655011648	2264
belles lettres	871094.3935751553	275.37426046176046	2311
miscellaneous	839155.9869742717	295.0817980222388	2360
learned	781733.2618728676	246.0817654567651	2182
general fiction	924678.68595826	258.49646187146146	2325
mistery and detective fiction	1123420.1486319497	259.7061299811289	2428
science fiction	935994.4646234306	248.55044955044897	2364
adventure and western fiction	1032713.1638679344	250.64708347208267	2380
romance and love story	997559.1771764176	251.74584582084492	2320

Table 10: Muestra 4

categoria	metafora	metonimia	w
reportage	739005.545665808	273.2918525918524	2217
editorial	839392.6586708553	252.962795537795	2230
reviews	897166.8448193009	267.3208680208676	2356
religion	971902.397216239	265.22606282606193	2410
skills & hobbies	913636.3833983988	260.77830780330754	2295
popular lore	827298.639753781	263.91099178599177	2256
belles lettres	948168.5408124946	263.5388195138189	2403
miscellaneous	863483.173212439	246.39977799977743	2207
learned	842569.1577530246	231.37843986079253	2205
general fiction	917557.8900258496	230.44950882450823	2296
mistery and detective fiction	866731.5026959036	245.56009546009463	2288
science fiction	1102841.6209263606	248.0798007548002	2461
adventure and western fiction	976789.2077744814	253.20416527916453	2349
romance and love story	1111028.8409040042	248.49708902208823	2422

Table 11: Muestra 5

categoria	metafora	metonimia	w
reportage	804307.8590497638	254.57564380064355	2244
editorial	797847.982604727	256.40300255300195	2241
reviews	926295.4083615864	234.46358363858295	2342
religion	935931.8321572712	233.24144189144172	2317
skills & hobbies	916884.62774593	232.22511377511276	2370
popular lore	796816.1152101667	263.7263361638353	2258
belles lettres	861343.6692835388	239.3655889861766	2359
miscellaneous	863173.038736266	279.4144463379755	2316
learned	907069.3580927892	255.3453282828281	2334
general fiction	870179.8901159727	224.0298867798861	2345
mistery and detective fiction	914219.7991227966	256.1841630591622	2331
science fiction	1000556.046812526	255.7852647352645	2369
adventure and western fiction	835693.3281863902	228.3971750471748	2279
romance and love story	1113220.902539808	261.2546370296359	2546

7 CONCLUSIONES

References

[1] Boris Eijembaum. La teoría del" método formal". In *Textos de teorías y crítica literarias:(del formalismo a los estudios postcoloniales)*, pages 33–62. Anthropos, 2010.

[2] Roman Jakobson and Ana María Gutiérrez Cabello. *Lingüística y poética*. Cátedra España, 1981.

[3] Ferdinand De Saussure. Curso de lingüística general. *Buenos Aires: Losada. Original de Ferdinand de*, 1945.

[4] Igor A. Bolshakov and Alexander Gelbukh. *Computational Linguistics: Models, Resources, Applications*. Mexico City: Centro de Investigacioǿn en Computacióǹ, Instituto Polit´cnico Nacional, 1981.

[5] Nitin Indurkhya and Fred J Damerau. *Handbook of natural language processing*, volume 2. CRC Press, 2010.

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[7] Rodolfo Delmonte. Computing poetry style. In *ESSEM@ AI* IA*, pages 148–155, 2013.

[8] Rodolfo Delmonte, Sara Tonelli, Marco Aldo Piccolino Boniforti, and Antonella Bristot. Venses–a linguistically-based system for semantic evaluation. In *Machine Learning Challenges Workshop*, pages 344–371. Springer, 2005.

[9] Daniel F Zuñiga, Teresa Amido, and Jorge E Camargo. Automatic computation of poetic creativity in parallel corpora. In *Colombian Conference on Computing*, pages 710–720. Springer, 2017.

[10] D Kaplan. Computational analysis and visualized comparison of style in american poetry. *Unpublished undergraduate thesis*, 2006.