FUNDACIÓN UNIVERSITARIA KONRAD LORENZ

MODELO DE LITERARIEDAD USANDO REDES SEMÁNTICAS Y N-GRAMAS

JONATAN AHUMADA FERNÁNDEZ
Trabajo de grado para optar el título de:
Ingeniero de Sistemas

ANDRÉS RAMÍREZ GAITA
Director de trabajo de grado

Programa de Ingeniería de Sistemas
Facultad de Matemáticas E Ingeniería
Bogotá D.C., Colombia Mayo de 2022

Contents

Resumen

En el presente trabajo se formula un modelo para calcular el concepto de *literariedad*, a través de dos medidas cuantitativas: el *indice metafórico* y el *índice metonímico*. Tanto la *literariedad* como la *metáfora* y la

metonimia no son conceptos *ad hoc*, sino que son modelados a partir del área de la lingüistica estructural y, en particular, del lingüista Roman Jakobson. Luego de formular el modelo, se evalúa a través de un diseño experimental que se basa en el uso del Corpus de Brown. En el experimento, se corren 5 muestras conformadas de 12 textos de categorias diferentes. Los resultados experimentales muestran que el *índice metafórico* reporta consistentemente valores significativamente más altos para las categorías de ficción, que era el resultado esperado. Por otro lado, los resultados del *índice metonímico* muestran consistentemente valores más altos para las categorías de no-ficción y en particular para los comunicados gubernamentales, que era un resultado inesperado, pero consistente con las teorías. El estadístico F y el valor-p de los índices apuntan a que los resultados no son aleatorios, sino consistentes a lo largo de las muestras.

Abstract

The present work formulates a model for the concept of *literariness*, by finding two quantitative measures: the *metaphorical index* and the *metonymical index*. The concepts of *literariness*, *metaphor* and *metonymy* are not *ad hoc* constructs, but are modelled after the tenets of structural linguistics and, in particular, from the works of linguist Roman Jakobson. After formulating the model, it is evaluated through an experimental design based on the use of the Brown Corpus. In the experiment, 5 samples formed by 12 texts of different categories are run. The experimental results show that the *metaphorical index* reports values significatively higher than

non-fiction consistently, which was the expected result. On the other hand, the results of the *metonymical index* show giger values for the non-fiction categories consistently and, in particular, for governmental communications, wich wasn't an expected result but is justifiable from a theoretical perspective. The F-statistic and the p-value for the indexes show that these results are not random, but consistent along the samples.

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, en su conferencia *Lingüística y poética* [2] y en su texto *Dos aspectos del lenguaje y dos tipos de afasia* [3] propone que la *literariedad* de un texto está dada por dos componentes del acto linguístico: la selección y la combinación. Estos dos conceptos se conciben como oposiciones binarias y fueron expandidos de la teoría linguística de Saussure, que en un principio fueron planteados como los componentes de diacronía y sincronía de la lengua (ver figura 1). Puestos en el contexto del análisis de la poesía, Jakobson renombró esos dos ejes como *metáfora* y *metonímia*.

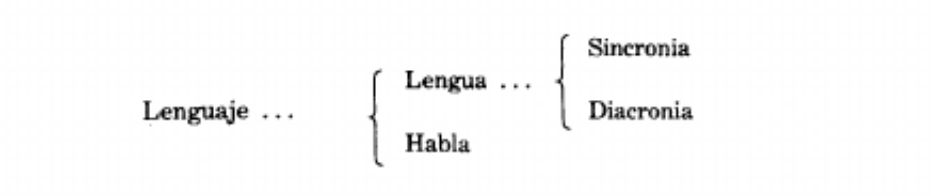

Figura 1: Distinción inicial entre sincronía y diacronía según Saussure, tomado de [4]

¿Es posible modelar algorítmicamente tales conceptos? Según Jakobson, en el acto de lenguaje intervienen 6 factores (ver figura 2). Sin embargo, cuando se estudia la *literariedad*, o la poética en general, Jakobson nos indica que nos debemos centrar nuestra atención en el factor Mensaje: .Esta función no es la única que posee el arte verbal, pero sí es la más sobresaliente y determinante, mientras que en el esto de las actividades verbales actúa como constitutivo subsidiario y accesorio"[2]. Así, al estudiar la *literariedad* deberiamos poner un segundo plano los aspectos que lidian con la intención (Hablante), la

interpretación (el receptor), los problemas de la lógica o referentes (Contexto) o problemas dentro del mismo lenguaje (Código).

La *literariedad* se encuentra, entonces, en el Mensaje. En otros términos, puede considerarse lo que está 'dentro del texto'. Aquello que es patente. En términos más simples: en la secuencia de sonidos que percibimos como un mensaje, expresado a través de la palabra escrita. Siguiendo el análisis de Jakobson, se puede teorizar que si se considera una cadena de textos como un mensaje, podríamos entonces hallar la *metáfora* y la *metonimia* y, por ende, una medida para *literariedad*.

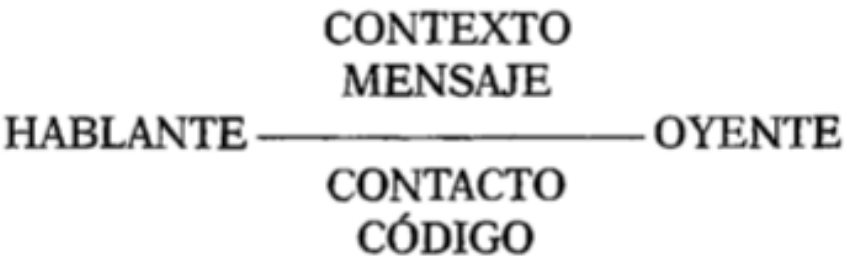

Figura 2: Factores de comunicación de Roman Jakobson [2]

Las similitudes de los conceptos tanto de Saussure como de Jakobson con las estructuras lineales y secuenciales de la computación son evidentes a manera conceptual. Por ejemplo, en la figura 3, el eje AB se considera el *eje de simultaneidades*, mientras que el eje CD se considera el *eje de sucesiones* [4, pg. 106].

Aunque Saussure y Jakobson ofrecen un modelo cualitativo, no se halló en la bibliografía consultada un modelo computacional que modelara el concepto

y lo implementara. Así, el objetivo de este trabajo es modelar e implementar el modelo de *literariedad* de Roman Jakobson utilizando herramientas básicas del procesamiento del lenguaje natural.

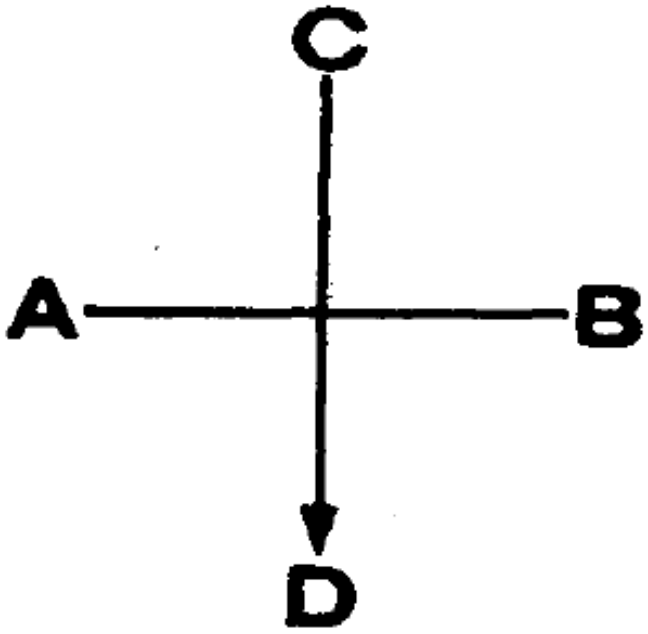

Figura 3: Ejemplo de estructuras secuenciales en el pensamiento de Saussure. Aquí se describe prototípicamente la selección y combinación de Roman Jakobson. Tomado de [4]

1.3 Justificacón

En términos generales, el foco principal de la linguística computacional han sido las aplicaciones que giran en torno a la extracción de información, y su 'comprehensión' por parte de la máquina. Por ejemplo, *text preparation*, *information retrieval*, *automatic translation*, *text classification*, entre otros [5].

Sin embargo, las aplicaciones con un enfoque humanístico, sea este linguístico, literario o estético son relativamente escazos, tal como lo reportan la mayoría de autores consultados (ver sección 5). Más aún, dentro de este subcunjunto reducido, pocos están guiados por aquello que Gelbuhk llama 'la ciencia fundamental', la lingüistica o, desde una perspectiva de analítica de datos, la comprensión del dominio. Más particularmente, no se encuentran modelos que aborden los conceptos de *literariedad*, *selección* y *combinación* de forma explícita, a pesar de que son ideas seminales de la lingüística de Roman Jakobson y, por ende, del llamado enfoque estructuralista.

El vacío de aplicaciones de estos conceptos es una oportunidad para brindarle al estudio académico de la literatura herramientas basadas en datos 'duros' o ,por lo menos, cuantitativos propias del método científico. Por lo tanto, un modelo de la *literariedad*, sustentado en los planteamientos da lingüística diferencial, ampliaría las aplicaciones de la lingüistica computacional y permitiría someter a escrutinio los planteamientos de dicha teoría desde un enfoque experimental.

Por otro lado, las escazas pero variopintas investigaciones en el área muestran un creciente interés en calcular la 'creatividad', la 'rima' o el 'estilo' de un texto. Sin embargo, esta misma diversidad de enfoques evidencia al mismo tiempo una falta de cohesión entre las disciplinas humanísticas y las ciencias (ver sección 5). El autor de esta investigación cree que los conceptos de la lingüistica estructural pueden aportar –si bien modestamente– a formar un mejor diálogo entre estas disciplinas y ofrecer perspectivas que otros investigadores podrían

valorar en un futuro.

Por los motivos expuestos, en esta investigación se formulará y evaluará un modelo para obtener una medida cuantitativa para el concepto de *literariedad* de Roman Jakobson utilizando las herramientas elementales del procesamiento de lenguaje natural y aplicando los postulados de la linguística estructural. 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' (ver sección 4.5.1). La primera delimitación 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, los corpus utilizados son el Corpus de Brown y Wordnet, para que haya una congruencia de

idiomas. Los criterios utilizados para hacer los corpus comparables se detallan en la sección

La segunda limitación concierne a la formulación de los algoritmos en sí mismos. Este trabajo se limitará 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. Por ejemplo, algunos problemas podrían ser que los sinónimos no sean suficientemente cercanos en su significado o que no se encuentren sinónimos suficientes.

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:

> El objeto de la ciencia de la literatura no es la literatura, sino la literariedad (*literaturnost'*), es decir, aquello que hace de una obra determinada una obra literaria. [1, pg. 37]

Por lo tanto, la *literariedad* no depende de ningún factor extrínseco, como su emisor, su valor histórico, sus ventas, número de 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.*

4.2 Roman Jakobson

La lingüistica de Jakobson se basa en los postulados de la lingüistica de Saussure. Sin embargo, es clave resaltar que Jakobson propuso una crítica a las ideas de Saussure y, en particular, postuló que los ejes de diacronía y sincronía corresponden a ’operaciones’ más profundas, que están presentes en todo acto de habla. En el siguiente fragmento, se puede apreciar su diferencia con respecto a Saussure:

> The fundamental role which these two operations play in language was clearly realized by Ferdinand de Saussure. Yet of the two varieties of combination-concurrence and concatenation-it was only the latter, the temporal sequence, which was recognized by the Geneva linguist. Despite his own insight into the phoneme as a set of concurrent distinctive features (*éléments différentiels des phonèmes*), the scholar succumbed to the traditional belief in the linear character of language "which excludes the possibility of pronouncing two elements at the same time ". [3, 99]

4.2.1 Selección:

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. [3]

4.2.2 Combinación:

La combinación estudia el .[en]tramado 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. [3]

4.3 Poética

La poética procura responder a la pregunta de ¿qué hace que un mensaje sea una obra de arte? Lidia principalmente con cuestiones estéticas del lenguaje. Sin embargo, para hacer un análisis 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.

> El objeto principal de la poética es la diferencia específica del arte verbal con respecto a otras artes y a otros tipos de conducta verbal;

> por eso está destinada a ocupar un puesto preeminente dentro de los estudios literarios.[2, pg. 121]

4.4 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. [5]

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*.

4.4.1 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 la lenguo: el *eje de sincronía* y el *eje de diacronía*.[4]

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

4.4.2 Lingüística 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", çhico", "jovencito", o "párvulo". En la perspectiva de Jakobson, el eje de sicronía pasa a ser la selección (ver seleccion en).

4.4.3 Lingüística diacrónica

La linguística diacrónica estudia los cambios sucesivos en el lenguaje, producidos por la actividad constante del *eje sincrónico*. Saussure plantea en un principio a la lingüistica diacrónica como el estudio de los cambios históricos en de la lengua [4].

Sin embargo, en la perspectiva de Jakobson, un *mensaje* tiene en sí mismo un eje diacrónico (ver combinación en). Tal eje mide la similaridad entre cada término del mensaje entindido como secuencia. En *Linguística y Poética*, [2], Jakobson propone como ejemplo la oracion Ï 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.

4.5 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, el lenguaje 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.

4.5.1 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 çhinoçon el concepto de "niño.º ïnfante", 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.

4.5.2 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*.

4.6 Lingüística Computacional

Es la intersección entre la computación y la lingüística. Por lo general, se preocupa acerca de cómo procesar automáticamente el lenguaje natural, para lo cual genera modelos lingüísticos sobre los que luego se pueden definir operaciones comunes [5].

La lingüística computacional es en sí misma un campo amplio y heterogéneo(ver 4). Este trabajo se incribe concretamente dentro del procesamiento del lenguaje natural 4.7, y tiene un fuerte componente de lingüística general .

4.7 NLP

El procesamiento del lenguaje natural (NLP, por sus siglas en inglés), es a menudo considerado sinónimo con la lingüística computacional [5]. Sin embargo, el NLP se refiere concretamente a la aplicación práctica de la linguística computacional para procesar automáticamente (a menudo en enormes cantidades) mensajes de lenguaje natural y obtener de estos alguna información o un acción sin intermedio de un humanano.

En este trabajo, se utilizan algunas herramientas típicas del NLP, como corpus, N-gramas, tokenización y vectorización, explicadas en a continuación. Sin embargo, es necesario hacer explítico de que se parte una herramienta computacional en particular: NLTK.

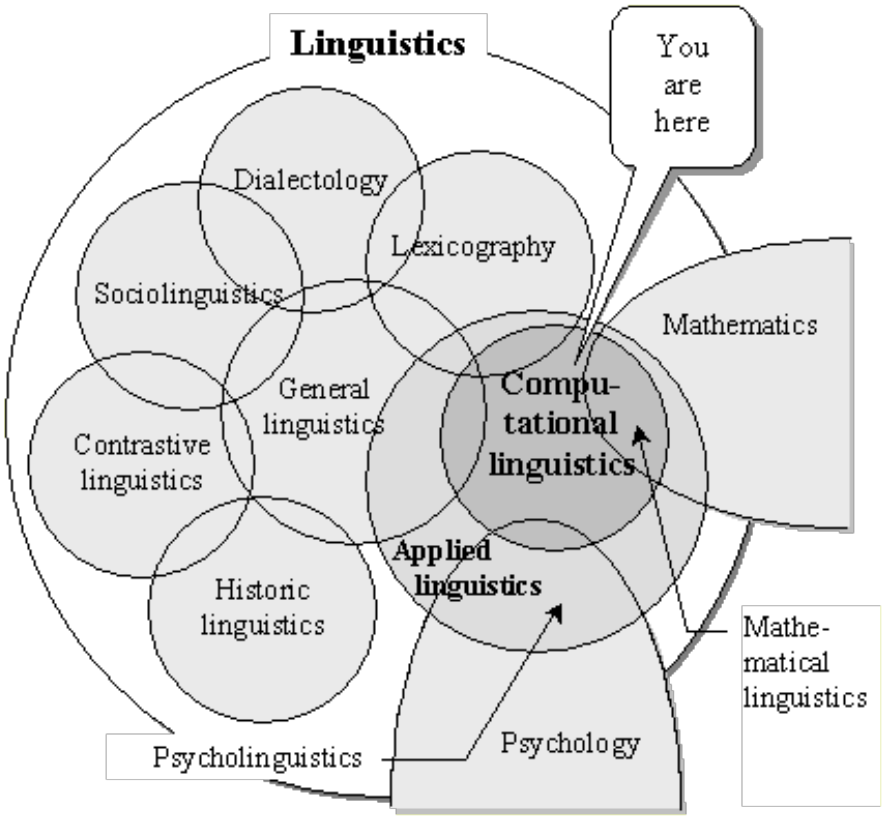

Figura 4: Relación de linguística computacional con otras areas tomado de [5]

4.7.1 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.

Se seleccionó esta herramienta porque no impone una estructura rígida en cuantó que cómo procesar el texto, lo que la hizo idónea para perseguir los objetivos interdisciplinares de esta investigación. Es

4.7.2 N-gramas

Los N-gramas son una herramienta común en el procesamiento de lenguaje natural y tienen diversas aplicaciones. Desde sus inicios [6], los n-gramas se han utilizado para capturar la noción de 'contexto' o 'historia' dentro de una secuencia de tokens. Así los n-gramas, forman una tupla o secuencia de palabras dentro de una secuencia o texto más grande y, delimitado el tamaño o nivel del n-grama, los términos circunscritos dentro del n-grama se entienden como variables aleatorias dependientes entre sí.

Así, los n-gramas se utilizán para tratar de predecir alguna característica con base en algún otro componente del n-grama, utilizando las teorías de cadenas de Markov.

En este trabajo, los n-gramas se utilizan meramente como una herramienta que captura la 'memoria' o 'relación' de dos palabras adyacentes dentro de un mensaje. No se utilizarán funciones de probabilidad, sino que se hará un cálculo de similitud utilizando el algoritmo descrito en la sección 6.4.3, utilizando n-gramas de nivel 2 o **bigramas**.

4.7.3 Tokenización

La tokenización es el proceso mediante el cual se separa la entrada de un programa NLP en unidades de análisis más pequeñas llamadas **tokens**. Un token puede ser una palabra, aunque no necesariamente lo es. Por ejemplo, puede ser un lexema, un signo de puntuación o una unidad sintáctica (un constructo sujeto - verbo, por ejemplo) [6]. El resultado de la tokenización dependerá, por lo tanto, de los objetivos de la investigación.

En esta investigación se tokenizará siguiendo la noción de palabra gráfica (*graphic word*). Esto simplemente se refiere a que cada token corresponde a una palabra separada por un espacio, incluyendo signos de puntuación y otros caracteres alphanumericos.

4.7.4 Vectorización

La vectorización es el proceso de tomar una característica o medida y representarla como una secuencia números reales, como un vector. A menudo, tal representación permite visualizar las características en un espacio vectorial, aunque la visualización no es la ventaja crucial.

La vectorización es una técnica utilizada a lo largo de muchos dominios y tiene una larga historia en el proceso de transformar un concepto a una entrada que sea interpretable por una máquina [7]. Continuamente, catalizadas por el auge del Machine Learning, se desarrollan técnicas de vectorización que ayudan a hacer los cálculos de similitud entre vectores más eficientes, dependiento del objetivo. Un buen ejemplo es el desarrollo del modelo de Google, que codifica las palabras de tal forma que agiliza el cálculo de similitud entre conceptos, conservando la noción de múltiples grados de similaridad [8].

En este trabajo, la técnica de vectorización utilizada es la *bag of words*, que es una técnica basada en la **frequencia**.

4.7.5 Bag of words

Es una técnica de vectorización frecuentemente utilizada en NLP. Se considera de complejidad sencilla, pero funciona exitosamente en muchos casos de uso. Involucra 3 fases: tokenización, creación de vocabulario y, finalmente, creación del vector.

Su funcionamiento es el siguiente: una vez se tiene el la entrada tokenizada se construye un *vocabulario*. Este es set de cada palabra utilizada en la entrada. Luego, se procede a asociar a cada palabra del vocabulario a su frecuencia en el texto, con lo cual se obtiene un histograma de palabras. En la última etápa, usualmente se utiliza una matriz llana en la que cada fila corresponde con una oración y cada columna representa una entrada en el vocabulario [7].

No obstante, para en este trabajo no se utilizará este enfoque tradicional. Sino que el proceso de vectorizacións seguirá los pasos descritos en la sección 6.4.3. Sin embargo, es necesario mencionar que la técnica de bag of words conlleva a los siguientes supuestos: 1) se asume que el orden de las palabras en la entrada no importa, tan solo la frecuencia de cada entrada y 2) la existencia de las palabras en el vocabulario es indenpiente una de la otra.

4.7.6 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 [9]. 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) [10] (ver figura 5). 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.

4.8 Analítica de datos

La analítica de datos es una disciplina heterogenea que auna diversas áreas de de estudio, como la teoría de la computación, la estadística, los negocios y cualquier

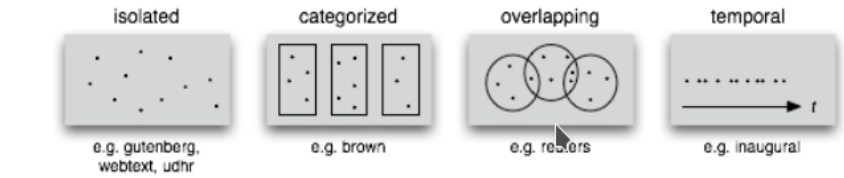

Figura 5: Diferentes estructuras de corpus

otro dominio sobre el cual es aplicada (por ejemplo, química, biología, etc). Una forma sucinta de entender la analítica de datos es el proceso mediante el cual se extrae **información** de los **datos** [11]. Así, se entiende por dato un registro que representa una medida de algún fenómeno observable. Por otro lado, la información se entiende como el conjunto de conclusiones aplicables que se obtienen de los datos luego de ser procesados. Tal proceso es el que se conoce como **análisis de datos**. El análisis de datos es variado y utiliza distintos recursos estadísticos y matemáticos, pero por lo general la análitica de datos tiene por objetivo generar un *modelo* de los datos que tenga capacidad *predictiva*.

Como se ve, la análitica de datos provee, más que un resultado concreto, una metodología para obtener modelos. Este trabajo, por lo tanto, se enmarca dentro de la analítica de datos en la medida en que se propone un modelo y lo evalua haciendo uso de rasgos comunes como: el uso de repositorios de datos (corpus), el uso de la estadística descriptiva para evaluar el modelo, la formalización de un modelo en términos matemáticos y el uso del stack de analítica de datos de Python (Pandas, Numpy, Seaborn, ScikitLearn).

Ahora bien, si bien en este trabajo se enmarca dentro de la analítica de

datos, se debe aclarar que el modelo presentado **no** es producido a partir de ninguna técnica de Machine Learning.

En cuanto a la información específica de la metodología, este proyecto se guió por la metodología CRIPS-DM

4.9 CRISP-DM

El Cross Industry Standard Process for Data Mining (CRISP-DM) es un modelo que sirve de base para cualquier proceso de analítica de datos. Este consta de 6 fases: 1) Entendimiento del negocio (¿Qué necesita el negocio?), 2) Entendimiento de los datos (¿Qué datos tenemos/necesitamos?¿Se necesitan limpiar?), 3) Preparación de los datos (¿Cómo organizamos los datos para modelar?), 4) Modelamiento (¿Qué técnicas de modelamiento deberíamos aplicar?), 5) Evaluación (¿El modelo cumple con los objetivos de negocio?) y 6) Despliegue (¿Cómo acceden a los resultados los interesados?).

CRISP-DM se utiliza, por lo tanto, como una guía para asegurar que cada fase del proceso de análitica de datos tenga las consideraciones adecuadas. Así el Diseño Metodológico de este trabajo está organizado según las fases mencionados. Sin embargo, cabe aclarar que algunas modificaciones debieron ser hechas a las fases, sobre todo a lo concerniente con las fases de Evaluación y Despliegue, pues el objetivo de este trabajo no es producir un modelo utilizado en un entorno empresarial.

5 MARCO REFERENCIAL

En la revisión de la literatura hecha se encontraron, a groso modo, dos tipos de trabajos que se consideran antecedentes cercanos. Esta distinción es importante porque cada categoría tiene un enfoque distinto sobre el problema de la *literariedad*. A continuación, se presentarán estos dos tipos (Tipo I y II) de trabajo y se mencionarán los aspectos relevantes para el presente trabajo.

5.1 Tipo I

El primer tipo de trabajo tiene un enfoque basado en *Machine Learning*, tienen un componente explorativo, y los autores por lo general se muestran escépticos al concepto de *literariedad*. Dentro de estos, los más relevantes son los de Cranenburgh [12] [13] y Louwerse [14]. En ambos trabajos los autores hacen una alusión explícita al concepto (*literariness*). No obstante, estos dos trabajos pasan por alto las bases linguísticas del concepto y se presenta la *literariedad* como una medida percibida por el lector, y poco articulada. Por ejemplo, Cranenbourgh afirma:

> However much debated the topic of literary quality is, one thing we do know: we cannot readily pinpoint what 'literary' means. Literary theory has insisted for a number of years that it lies mostly outside of the text itself (cf. Bourdieu, 1996), but this claim is at odds with the intuitions of readers, of which the [13, pg. 58]

De igual forma, Lowerse coincide y menciona:

> (. . .) whether literary texts overall are linguistically different from non-literary texts is a question that has not been satisfactorily answered.[14, pg. 176]

Como es evidente, esta investigación toma el enfoque opuesto a estos trabajos previos. En concreto: en esta se parte del supuesto de que la *literariedad* está suficientemente descrita por Roman Jakobson y que es algo 'dentro del texto', no dependiente de apreciaciones subjetivas. Por consiguiente, se suspende el juicio con respecto a los contra-argumentos usuales en contra de la *literariedad* [15], el *formalismo* o, de manera más generalizada, el *estructuralismo* y se aboca a proponer y validar un modelo.

La otra divergencia del presente trabajo con respecto a este primer tipo, se da a nivel del uso de tecnologías. En el presente trabajo no se hace uso de Machine Learning, entendiendo este término como el uso de modelos bayesianos (Latent Dirchlet Allocation), modelos de regresión lineal (Suppport Vector Machines), redes neuronales (Paragraph Vectors).

Consiguientemente, en los trabajos citados (exceptuando a [14]), se entiende la literariedad como aquellos patrones que producen la classificación más apta. En el presente trabajo, en contraparte, se formula un modelo basado en la teoría y luego se evalua experimentalmente. Se podría decir, a manera de síntesis, que en los trabajos de tipo I la *literariedad* se encuentra. En este trabajo, en cambio, la *literariedad* se modela.

5.2 Tipo II

Ahora bien, el segundo tipo de trabajo tiene un enfoque basado en estadística y vectorización, pero no emplea de forma explícita el concepto *literariedad* u otra fundamentación de la linguística saussureana. Sus inicios, segun Blei, inician en [16]. En los trabajos de este tipo, los autores, partiendo de un interés muy delimitado buscan medir una característica concreta: determinar el orígen de un texto [16], obtener una herramienta de visualización gráfica [17], determinar el grado de creatividad de una traducción [18] o, en términos más generales, capturar el *estilo* [19] [20].

Su característica principal es que proponen una extensa lista de medidas posibles sobre un texto, forman un espacio vectorial y luego hacen uso de alguna técnica de reducción de dimensionalidad (Principal Component Analysis, Support Vector Machines). Dentro de este tipo de trabajo, el más relevante es Kaplan Blei, en cuyo primer trabajo [17] de tesis de pregrado visualiza 84 métricas distintas en un espacio vectorial y luego formaliza en un artículo científico. [21].

Los trabajos de Kaplan, son luego citados por el trabajo de Delmonte [20] [19]. La trayectoria de Delmonte es bastante amplia en su alcance. Iniciando con módulos que calculan similaridades semánticas en un texto [**?**], luego aprovecha las ideas de Kaplan para desarrollar un sistema multi-modular que abarca prácticamente todas las áreas de estudio de la linguística: semántica, fonética, gramática e incluso aspectos que tienen que ver con la rima (prosodia).

A pesar de el trabajo de Delmonte es el más rico y complejo no solo

dentro de este tipo, sino de toda la bibliografía consultada, realmente nunca hace alusión al concepto de *literariedad*. Lo más relevante del trabajo de Delmonte es el uso constante de las mismas herramientas (tokenizadores, splitters, n-gramas y NER) para construir módulos de creciente complejidad.

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 a lo largo de las distintas áreas de la lingüistica. Por lo tanto, sus modelos son mucho más informados y propone soluciones a aspectos complejos del análisis lingüístico que los autores anteriores no abordan.

Por último, dentro de este segundo tipo de trabajo, tiene mención especial el trabajo de [18]. Aquí se 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. De aquí se tomó la idea de establecer una métrica para un aspecto tradicionalmente cualitativo (la creatividad), desde una perspectiva *hand-crafted*. Lo que diferenció este trabajo del de Delmonte, es su aproximación matemática. Aquí se proveen fórmulas para cada una de las 7 medidas propuestas. El grado de complejidad para cada medida es sencillo, pero se obtienen buenos resultados. Lo que fue un ejemplo tremendo para este trabajo, pues muestra las ventajas del *hand-crafted features*, en contraposición al de *learned-features*. Esto destacó el valor de formular medidas propias por sobre las de un algoritmo no supervisado.

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*.

¿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

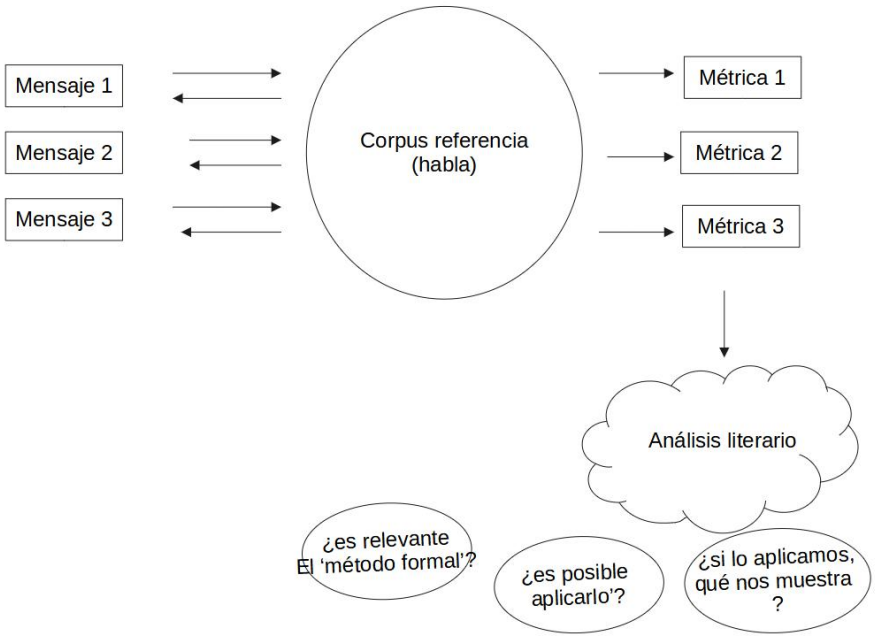

Figura 6: Entradas y salidas del algoritmo

una diferencia sustancial desde el punto de vista linguístico .

Como se puede apreciar (ref:fig:posibles$_{usos}$), 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 ïmágenes 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.

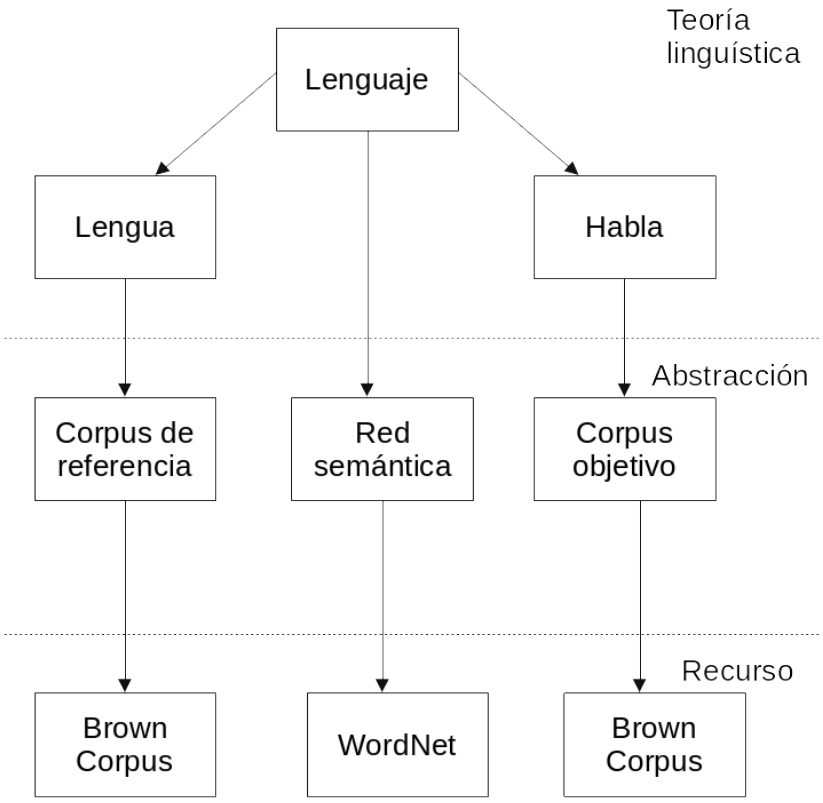

Figura 7: Resumen de las fuentes de datos utilizadas para cada concepto

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 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

Cuadro 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 receptor.

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

Cuadro 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

Cuadro 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

Cuadro 4: Corpus objetivo 3

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

Cuadro 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	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

Cuadro 6: Corpus objetivo 5

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 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 muestras

- H3: Se espera que la categoría 'Belles Lettres' 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

El modelo diseñado se basa en las siguientes ecuaciones. Para una visión a más alto nivel del procedimiento se puede ver la figura 9.

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

(1)

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

(2)

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

(3)

$$\mu = \frac{\Sigma_i^j freq_{referencia}(s_i)}{j} \tag{4}$$

$$uso(w) = \frac{freq_{objetivo}(w)}{\mu} \tag{5}$$

$$indice\ metaforico(mensaje) = \Sigma_i^j uso(w_i) \tag{6}$$

$$N = \{n_1, n_2, n_3, \ldots, n_j\} \tag{7}$$

$$met(n_i) = \frac{letras\ iguales}{set(letras(n_i 1) + letras(n_i 2))} \tag{8}$$

$$indice\ metonimia = \Sigma_i^j met(n_i) \tag{9}$$

En primer lugar, un mensaje es cualquier cadena de texto. Una vez tokenizado, se obtienen las palabras *w* mostradas en la ecuación 6.4.3. Luego, para cada una de las palabras, se hace primero el cálculo del vector semántico. Un vector semantico está compuesto de sinónimos *s* del la palabra inicial (ecuación 6.4.3). Cuando se termina de obtener los campos semánticos de cada palabra del mensaje, se obtiene una matriz semántica. Luego, por cada vector semantico, se calcula un vector de uso que cuenta la frecuencia de cada componente del vector semantico *s* en el corpus de referencia 6.4.3. La suma de todos los vectores de uso de un mensaje se conoce como la matriz de uso. Se puede apreciar las relación de las matrices semántica y de uso con entre sí en la figura 8.

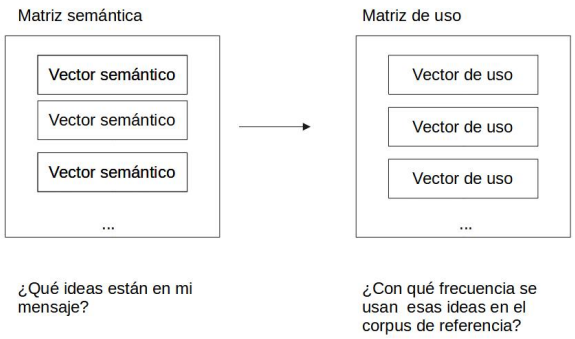

Figura 8: Transformación de matriz semántica a matríz de uso

Seguidamente, para cada vector de uso se calcula el uso, que es la relación entre la media del vector de uso y la frecuencia de la palabra en el corpus objetivo (ecuación 5). Así, si la palabra se utiliza más veces que la media del vector de uso, se considera que la palabra está siendo utilizada de manera más rara (más frecuente que lo indicaría que se debe usar por su vector de uso), por ende el resultado del cociente es mas alto y su aporte al índice metaforico mayor. El índice metafórico es la suma de todos los usos, por lo que el índice en principio solo captura si un mensaje es mas ’metafórico’ que otro si tiene un número más alto que otro mensaje y manteniendo la longitud del mensaje.

Ahora, con respecto al Índice metonímico, se parte de la idea de que un mensaje está compuesto de ngramas *n* (ver ecuación 7). Los ngramas son de nivel 2, es decir, que se toman pares de palabras constiguas (ver figura 10) .

Luego para cada *n* se calcula la metonimia. La metonimia está dado por el numero de letras similares entre los terminos *n* del bigrama 8. Por último, el Índice metonímico está dado por la suma de la metonimima para cada n-grama.

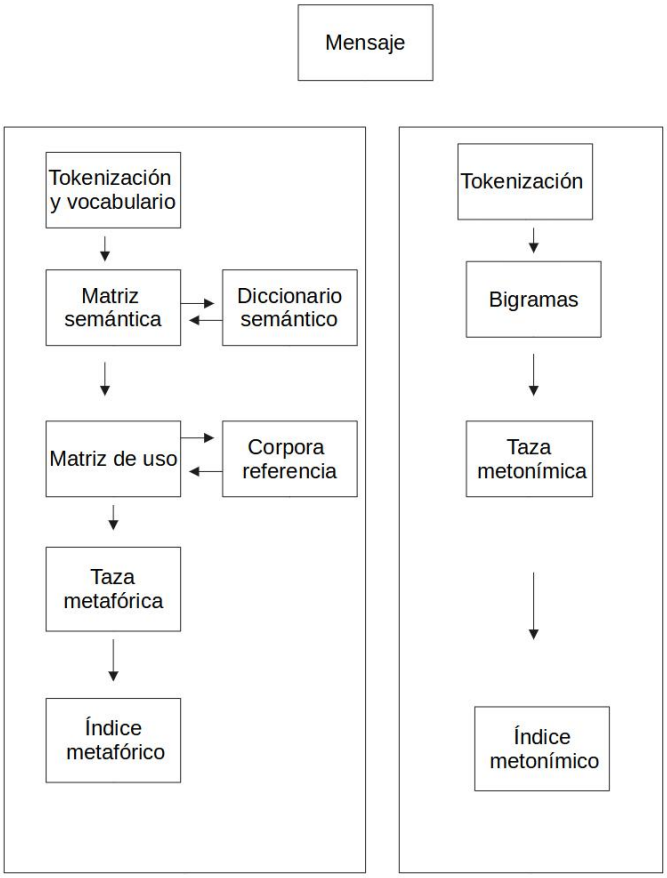

Figura 9: Etapas de procesamiento para cada índice

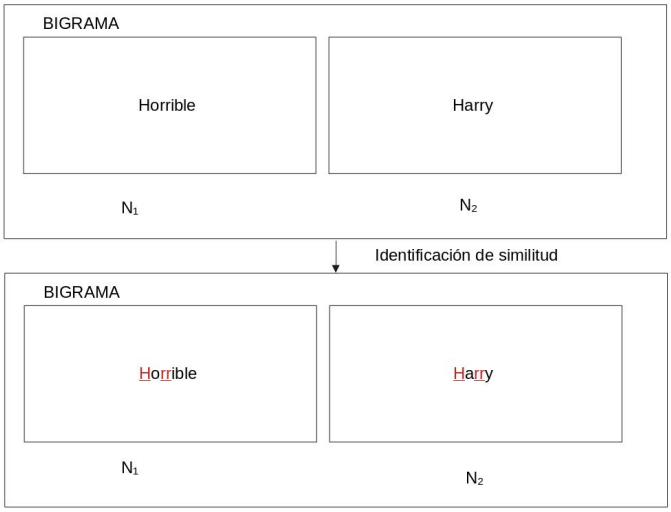

Figura 10: Concepto de metonimia

6.5 Despliegue

En las secciones 6.5.1, 6.5.2 y 6.5.3 se presentarán los resultados del experimento según los parámetros descritos en las secciones anteriores. La presentación va en creciente órden de abstracción, partiendo de los datos brutos, pasando por su visualización, hasta llegar a las Conclusiones

6.5.1 Índices por muestra

En esta sección, se muestran los resultados producidos por el modelo para cada uno de los corpus objetivos definidos en la sección 6.3. En cada tabla se presentan el índice metafórico y el índice metonímico para el representante de cada categoría en las columnas ’metafora’ y ’metonimia’, respectivamente. La

columna 'w' simplemente representa el número de palabras totales en el texto procesado, en caso de que en un futuro se desee hacer comparaciones entre textos de diferentes tamaños.

Estos valores no tienen ningún tipo de procesamiento y para apreciarlos, es mejor consultar las secciones 6.5.2 y 6.5.3.

Cuadro 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

Cuadro 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

Cuadro 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

Cuadro 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

Cuadro 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

6.5.2 Gráficos por muestra

En esta sección se presentan los gráficos para cada uno de los corpus objetivos definidos en 6.3. Cada cúmulo de gráficos consta de 2 filas. La primera fila muestra el puntaje para el **índice metafórico** (izquierda) y el **índice metonímico** (derecha) a través de las categorías, como están definidas en el corpus de Brown. Por otro lado, en la segunda fila se presentan los mismos puntajes para las metacategorías de de **ficción** y **no ficción**. Las metacategorias son agrupaciones de categorías del corpus de Brown y tienen el objetivo de evidenciar más claramente el comportamiento de los dos índices de manera más general.

Para la producción de estos gráficos, se tomaron los resultados

presentados en 6.5.1, y se normalizaron con la técnica Min Max. En cada corpus objetivo, por lo tanto, se evidencia que hay una categoría con el valor mínimo de 0 y otra con el valor máximo de 1. Esto evidencia mejor la relación entre las distintas categorias en cuanto a las dos medidas postulados: la metáfora y la metonimia.

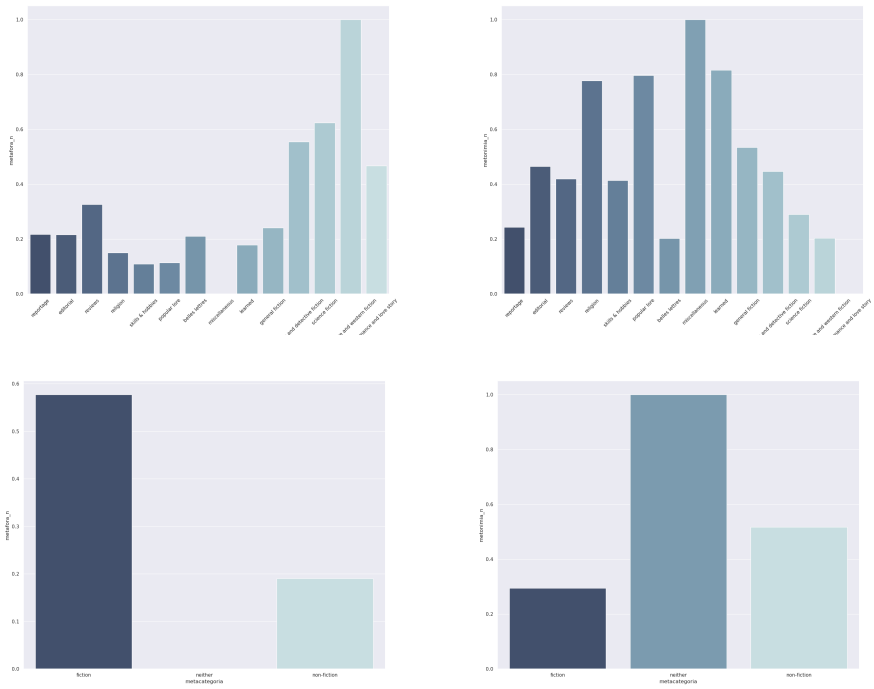

Figura 11: Resultados muestra 1

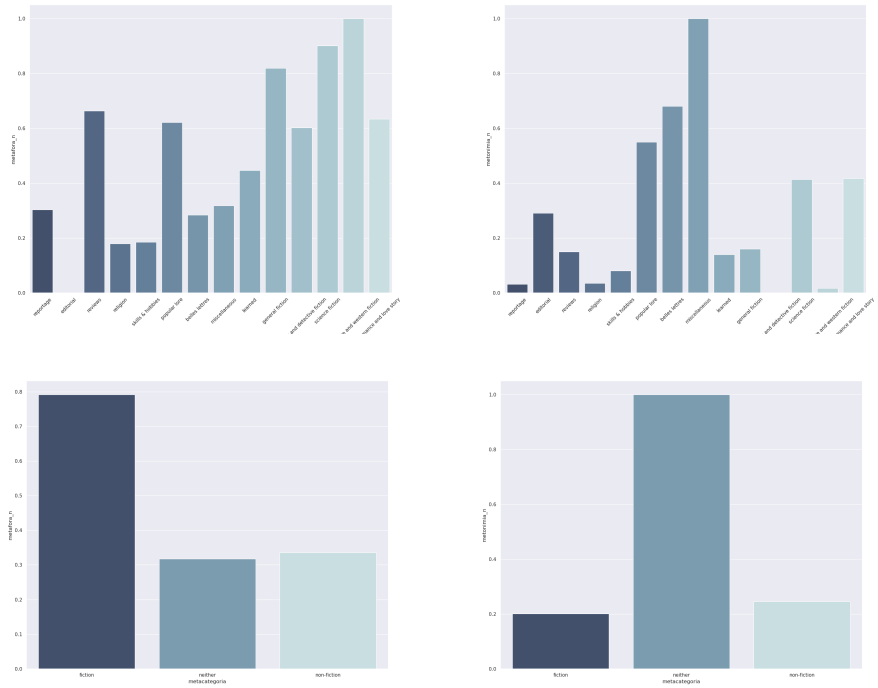

Figura 12: Resultados muestra 2

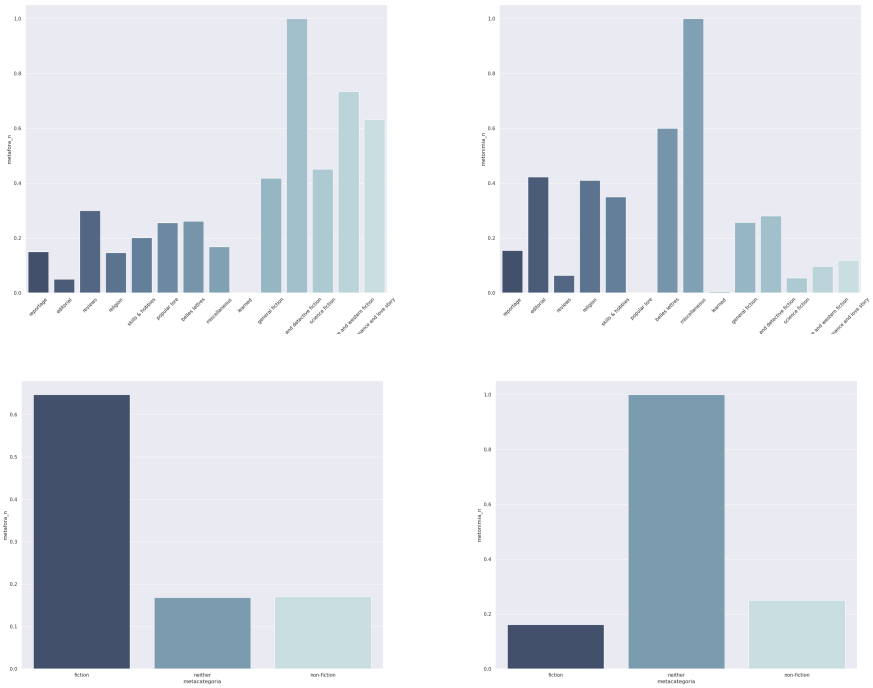

Figura 13: Resultados muestra 3

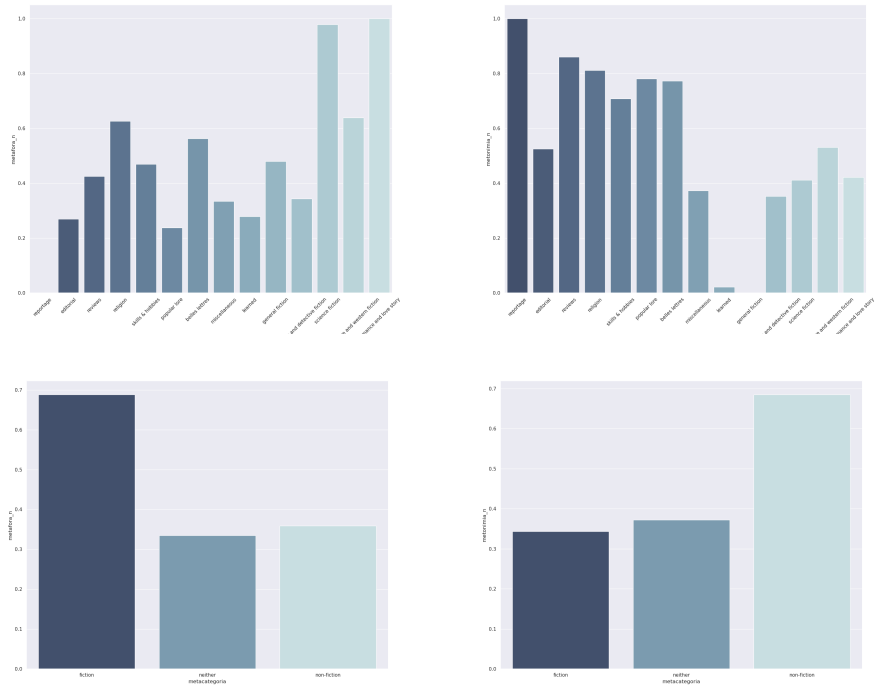

Figura 14: Resultados muestra 4

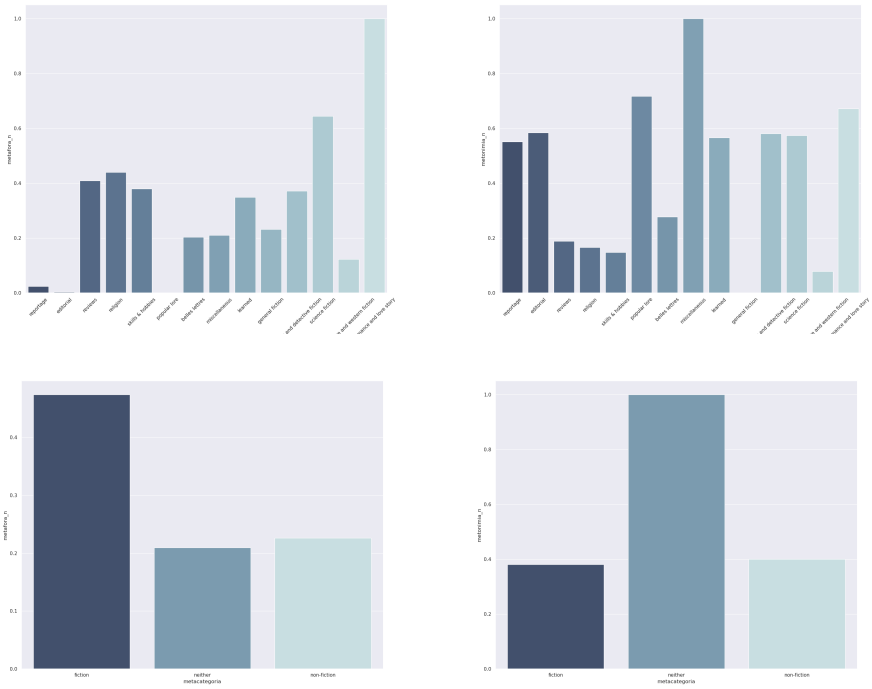

Figura 15: Resultados muestra 5

6.5.3 Gráficos totales

En esta sección se presentan los gráficos para el **índice metafórico** y el **índice metonímico** teniendo en cuenta su comportamiento a lo largo de todos los corpus objetivos. Por lo tanto, cada boxplot está constuido de 5 muestras (correspondientes a 1 representante de cada corpus objetivo) para cada categoría, por lo que se puede obtener una noción más clara del IQR, la mediana , los datos atípicos en cada categoría y metacategoría.

La visualización del comportamiento de los indicadores será necesaria

para las Conclusiones (7).

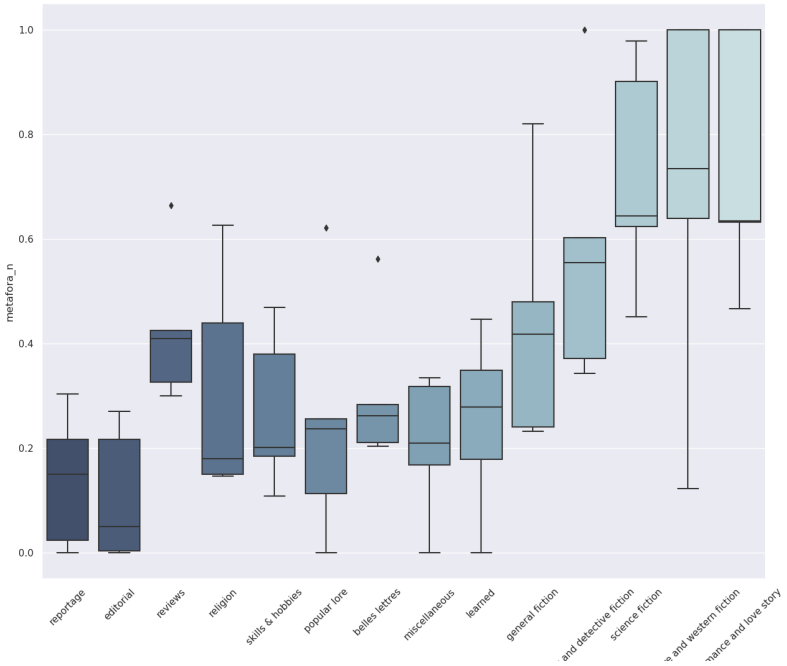

Figura 16: Índice metafórico por categorías a través de las muestras

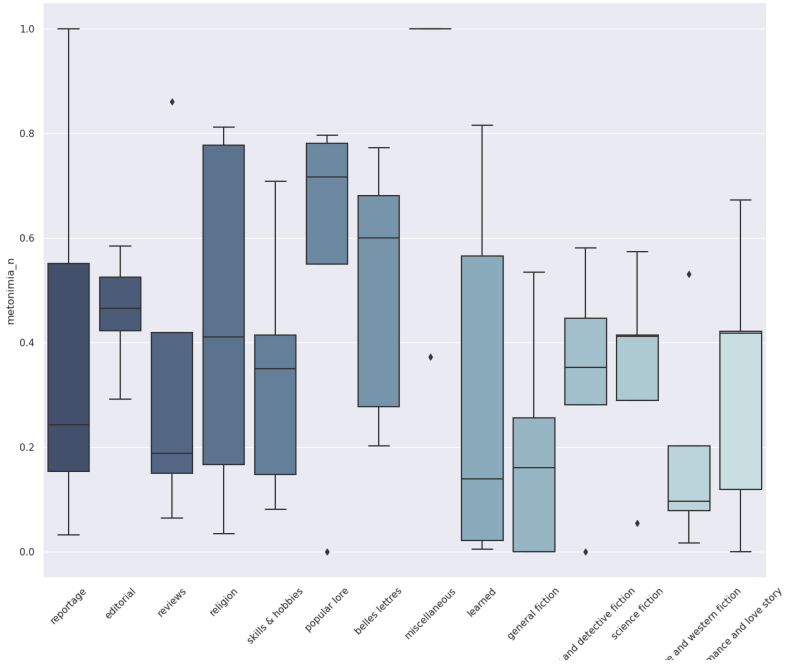

Figura 17: Índice metonímico por categorías a través de las muestras

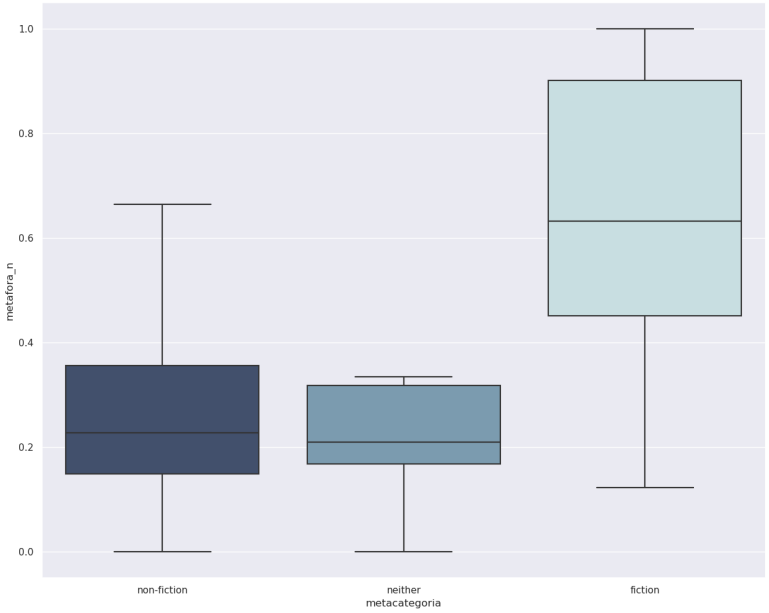

Figura 18: Índice metafórico por metacategorías a través de muestras

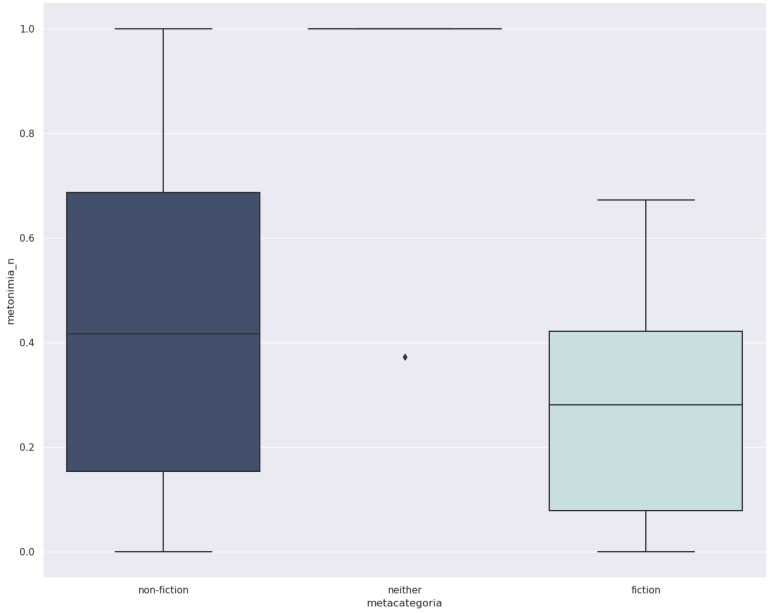

Figura 19: Índice metonimica por metacategoria a través de muestras

6.6 Evaluación

Según lo contempla el proceso de analítica de datos (ver sección 4.8), es necesario someter a prueba los modelos postulados. Sin embargo, como el modelo propuesto no se enmarca dentro de Machine Learning, no se dispone de un algoritmo de clasificación per se, que luego de pueda evaluar con un set de validación.

Sin embargo, teniendo en cuenta los hipótesis planteadas en la sección 6.4.2, se pueden realizar pruebas estadísticas que pueden aportar una

fundamentación cuantitativa, para las hipótesis que lo permitan.

Así, para la H_1, que plantea que las metacategorías de ficción y no ficción tengan un índice metaforico significativamente distinto se formula una prueba ANOVA, a lo largo de todas las muestras, entre los textos de ficción y no ficción, con el siguiente resultado:

```
>>> anova_metafora
F_onewayResult(statistic=51.510567153609514, pvalue=9.812579375438188e-10)
```

Por lo tanto, se como el valor-p para la prueba ANOVA es inferior a 0.01 y el estadístico F es muy alto, se puede afirmar que el algoritmo genera valores significativamente distintos entre las metacategorías de ficción y no ficción, con una confianza mayor al 99 %.

Así mismo, si se hace una prueba ANOVA para el índice metonímico para las metacategorías de ficción y no ficción se obtiene que:

```
  >>> anova_metonimia
F_onewayResult(statistic=4.327636012671773, pvalue=0.04157136345702674)
```

Por lo tanto, como el valor-p para la prueba es inferior a 0.05 y el estadístico F es más alto que 1, se puede aformar que el algoritmo genera valores significativamente distintos entre las metacategorias de ficción y no ficción con una confianza de 95 %.

7 CONCLUSIONES

Para concluir el presente trabajo. Primero se señalarán los resultados del experimento frente a las hipótesis planteadas. Posteriormente, se expondrán las críticas posibles al modelo planteado. Por último, se señalaran trabajos futuros para profundizar más en la pregunta de investigación.

7.1 Las hipótesis planteadas

Para la hipótesis H_1 se observa en 18 que el índice metafórico es, en promedio, más alto para las categorias de no ficción a lo largo la muestras que para las categorias de no ficción. De hecho, en promedio, las obras de ficción reportan un índice metafórico un poco más de 3 veces más alto. Esto es consistente con la intuición, que nos dicta que en las obras de ficción se hace uso de un vocabulario más amplio y distinguido, lo que aporta más al índice metafórico.

En cuanto a la hipótesis H_2, las medias para las categorias *reportage* y *editorial* son cercabis (0.14 y 0.11, respectivamente). En el gráfico 16 se puede apreciar que el rango interquartil (IRQ) es muy similar. Esto es consistente con el resultado esperado, puesto que estas dos categorías son similares entre sí: ambas están conformadas por textos que aparecieron en publicaciones periódicas. Por lo tanto, comparten muchos parámetros linguísticos similares en cuanto al vocabulario. Por lo tanto, sú indice metafórico debe ser similar a lo largo de las muestras.

Luego, para la hipótesis H_3 se puede observar que la categoria *Belles*

Lettres es la tercera más categoría con el índice metafórico más alto (con un 0.30). Queda por debajo de *Religion* (0.31) por un punto y de *Reviews* (0.42). Este resultado no es el esperado, pero es comprensible si se tiene en cuenta que la categoría *Reviews* está compuesta de críticas a obras de arte como música clásica, libros y obras de teatro, cuyo vocabulario puede terminar aportando más al índice metafórico que las biografías y caras de la categoría *Belles Lettres*.

Por último, para la hipótesis H_4, se observa que la categoría *Learned* tiene el segundo índice de metonimia más bajo (0.31), luego de (sorprendentemente) las categorias *General Fiction* y *Adventure & Western Fiction* (0.19 ambas). Si bien este resultado no es estrictamente el esperado a lo largo de todas las categorias, la hipótesis H_4 sí se cumple dentro de la metacategoría de no-ficción. La hipótesis inicial se hizo sobre la base de que los textos técnicos y científicos no deberían tener un enfasis en la metonimia entre cada una de sus palabras. Es decir, no debería haber un énfasis en repetir sonidos a lo largo de una oración, puesto que los factores de comunicación de Jakobson se centran en las funciones conativa o fática.

Ahora bien, en la hipótesis inicial no se contemplo que, según lo encontrado en este experimento, las obras de ficción por lo general tienen un índice metonímico más bajo que las de no ficción (ver 19. Esto parece apuntar a una relación inversa entre el índice metaforico y el índice metonómico. Sin embargo, esa discusión está por fuera de los alcances de la presente investigación.

Con base en el análisis de las hipótesis, se puede señalar que el algoritmo propuesto es capaz de:

- 'distinguir' entre dos metacategorias: los textos de ficción y los de no ficción
- arrojar un índice metaforico consistentemente más alto que los de no-ficción para los textos de ficción
- arrojar, para los textos de no ficción, un índice metonímico consistentemente mas alto que los de ficción

7.2 Crítica del modelo

En términos generales, se considera que el modelo es razonablemente exitoso. En general no hay resultados inconsistentes con la intuición, salvo el comportamiento del índice metonimico a lo largo de una categoria. Sin embargo, si se considera las metacategorias de ficción y no ficción, en la muestras se evidencia claramente que hay una consistencia en los resultados.

Una debilidad significativa del modelo es su dependencia de la con la red semántica. Esto ocasiona que dada una palabra, su vector semantico quede asociado con palabras muy dificiles de encontrar,lo que incide en su puntuación total. Esto particularmente se evidencia en algunas palabras comunes que no se encuentran tan solo porque son pronombres o adverbios que empiezan con mayúscula cuando la red los espera en minúscula. Es posible que esto esté afectando el modelo, pero no es claro a priori de qué manera lo hace porque se haría necesario un análisis multivariado para cada una de las variables. Ahora bien, esta falencia no parece ser tan pronunciado, ya que la diferencia entra las

metacategorias es evidente y resulta inverosímil atribuir la concordancia con la intuición y el juicio experto a un mero error.

Sin embargo, para tener una fundamentación más rigurosa de la idoneidad del algoritmo se debe diseñar un experimento que mire los resulados de cada vector de uso por palabra y verificar como el número de sinónimos, la media de esos sinónimos y la frecuencia de la palabra original en los corpus se compartan. Una tarea que no es fácil por la dependencia de las variables de otras. Este aspecto es muy importante hacerlo, pero se sale de los alcances de esta investigación.

7.3 Trabajo futuro

7.3.1 El índice metafórico

Para trabajos futuros es necesario repetir más veces este mismo experimento, aumentando el número de muestras. Luego, se podría plantear el mismo documento con un corpus distinto, tal vez con categorias individuales distintas, pero conservando las mismas metacategorías:ficción y no ficción. Si el resultado sigue avalando la hipótesis H_1 fortalecería la validez del modelo para capturar la 'metafora' en documentos.

Por otro lado, como se expuso en el marco teórico y en la presentación del modelo, el índice metafórico es dependiende del corpus de referencia, que es una representación del concepto de *lengua*, y la red semántica, que corresponde con el concepto de *lenguaje*. Así, para obtener resultados más intuitivos se

deberá disponder de la capacidad de configurar tanto el corpus de referencia como la red semántica. Por ejemplo, para asociar palabras entre sí en la red semántica o quitar relaciones espurias.

7.3.2 El índice metonímico

El algoritmo para metonimia fue realizado de la manera más *naive* posible, lo que puede ser una causa de la variabilidad de este índice entre una misma categoria a lo largo de las muestras. Un siguiente paso sería calcular la metonimia no por el número de letras iguales, sino tokenizar por sílabas y contar las sílabas con una misma vocal como un aporte al índice.

Otra posible modificación es parametrizar el la aridad del n-grama, puesto que la repetición de sonidos solo se está teniendo en cuenta para palabras consecutivas, cuando en realidad la metonimia suele darse por elementos sintacticos distintos. Por ejemplo, se puede dar entre oraciones, ente párrafos, entre estrofas etc. Sin embargo, el cálculo de esto necesitaria incorporar POS al modelo, lo que complejizaría significativamente la implementación del índice.

Referencias

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

[2] R. Jakobson and A. M. G. Cabello, *Lingüística y poética*. Cátedra España, 1981.

[3] R. Jakobson, "Two aspects of language and two types of aphasic disturbances," *Fundamentals of language*, vol. 1, pp. 69–96, 1956.

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

[5] I. A. Bolshakov and A. Gelbukh, *Computational Linguistics: Models, Resources, Applications*. Mexico City: Centro de Investigacioǿn en Computaciǿn, Instituto Polit´cnico Nacional, 1981.

[6] C. Manning and H. Schutze, *Foundations of statistical natural language processing*. MIT press, 1999.

[7] A. Jha, "Vectorization techniques in nlp [guide]," Dec 2021.

[8] "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[9] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing*, vol. 2. CRC Press, 2010.

[10] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. .ºReilly Media, Inc.", 2009.

[11] F. Nelli, "Python data analytics," *Apress Media, California*, 2018.

[12] A. van Cranenburgh, K. van Dalen-Oskam, and J. van Zundert, "Vector space explorations of literary language," *Language Resources and Evaluation*, vol. 53, no. 4, pp. 625–650, 2019.

[13] A. van Cranenburgh and C. Koolen, "Identifying literary texts with bigrams," in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pp. 58–67, 2015.

[14] M. Louwerse, N. Benesh, and B. Zhang, "Computationally discriminating literary from non-literary texts," *Directions in empirical literary studies: In honor of Willie Van Peer*, vol. 5, pp. 175–191, 2008.

[15] R. Chuit-Roganovich, "Epistemología de la teoría literaria: objeto y método en el formalismo ruso," *Aisthesis*, no. 66, pp. 13–35, 2019.

[16] E. Klarreich, "Bookish math," Aug 2019.

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

[18] D. F. Zuñiga, T. Amido, and J. E. Camargo, "Automatic computation of poetic creativity in parallel corpora," in *Colombian Conference on Computing*, pp. 710–720, Springer, 2017.

[19] R. Delmonte, S. Tonelli, M. A. P. Boniforti, and A. Bristot, "Venses–a linguistically-based system for semantic evaluation," in *Machine Learning Challenges Workshop*, pp. 344–371, Springer, 2005.

[20] R. Delmonte, “Computing poetry style.,” in *ESSEM@ AI* IA*, pp. 148–155, 2013.

[21] D. M. Kaplan and D. M. Blei, “A computational approach to style in american poetry,” in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 553–558, IEEE, 2007.