

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	3
1.1	Introducción	3
1.2	Planteamiento del problema	3
1.3	Justificación	4
1.3.1	Palabras clave:	5
1.3.2	Área de conocimiento:	5
1.4	Alcances y delimitaciones:	5
2	OBJETIVO GENERAL	6
3	OBJETIVOS ESPECÍFICOS	6
4	MARCO TEÓRICO	6
4.1	Literariedad	6
4.1.1	Selección (ver lingüística sincrónica):	7
4.1.2	Combinación (ver lingüística diacrónica):	7
4.2	Poética	7
4.3	Lingüística	7
4.3.1	Lingüística General:	7
4.3.2	Lingüística sincrónica	7
4.3.3	Lingüística diacrónica	8
4.4	Lenguaje	8
4.4.1	Lengua	8
4.4.2	Habla	8
4.5	Lingüística Computacional	9
4.5.1	NLTK	9
4.5.2	Corpus	9
5	MARCO REFERENCIAL	9
6	DISEÑO METODOLÓGICO	10
6.1	Entendimiento del negocio	11
6.2	Entendimiento de los datos	12
6.2.1	El corpus de referencia	12
6.2.2	El corpus objetivo	12
6.2.3	La red semántica	12
6.2.4	Resumen de entendimiento de los datos	12
6.3	Preparación de los datos	12
6.3.1	Corpus de referencia	14
6.3.2	Corpus objetivo	17
6.4	Modelamiento	18
6.4.1	Selecting a modeling technique (no tengo, estoy traduciendo un modelo cualitativo –investigacion mixta–)	18
6.4.2	Generating a test desing	18
6.4.3	Building the models	18

6.5	Despliegue (los notebooks?? creo que no hay despliegue)	21
-----	---	----

7	CONCLUSIONES (Creo que esto se solapa con lo que crisp-dm llama despliegue)	21
---	---	----

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 lingüística.

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 lingüística 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 "Lingüística y poética".

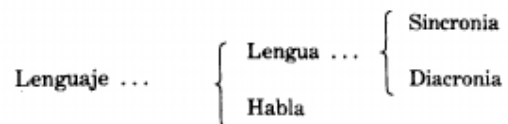


Figure 1: Distinción entre sincronía y diacronía

¿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.

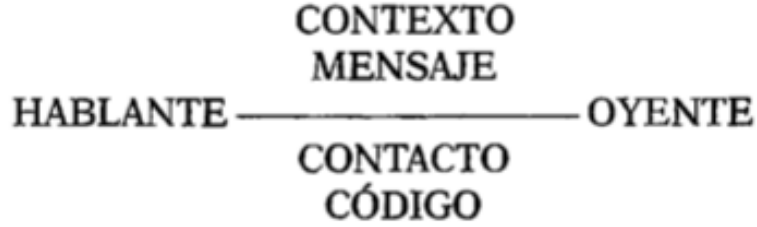


Figure 2: Factores de comunicación de Roman Jakobson [2]

Saussure ofrece ya un modelo cualitativo muy bien esbozado en teoría, que es el que luego Jakobson utilizará para definir la literariedad. Sin embargo, aunque existe un planteamiento cualitativo del problema, no se halló en la bibliografía consultada un modelo computacional que modelara el concepto y lo implementara.

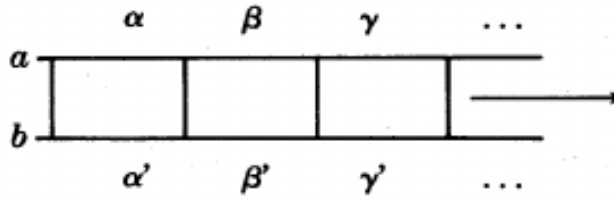


Figure 3: Modelo cualitativo inicial expuesto por Ferdinand De Saussure tomado de [1]

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 Justificación

Si bien existen infinitas operaciones realizables sobre un texto computarizado, hay pocas que tengan un enfoque humanístico, sea este lingüístico, 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 subconjunto reducido, pocos están guiados por aquello que Gelbukh llama 'la ciencia fundamental'. A saber, la lingüística. En otras palabras, hay un vacío en el campo de la lingüística 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 lingüística 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 lingüística 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 lingüística 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 lingüística 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 análisis exhaustivo, la poética debe hacer uso de la lingüística, 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 Lingüística

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 neurolingüística. Sin embargo, se considera que hay un núcleo común llamado *lingüística 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 Lingüística sincrónica

La lingüística sincrónica se ocupa de las operaciones que realiza un hablante, sean lógicas o psicológicas, para formar un sistema lingüístico. 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 las palabras "niño", "chico", "jovencito", o "párvulo".

4.3.3 Lingüística diacrónica

La lingüística 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 entendido como secuencia. Un ejemplo se puede apreciar en la oración "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: toda 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 *crystalización* o un producto de la suma de asociaciones entre conceptos e imágenes 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 comunicación lentamente terminan por transformar el consenso social sobre cómo hablar. Por este motivo la lingüística 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: tokenización, 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.

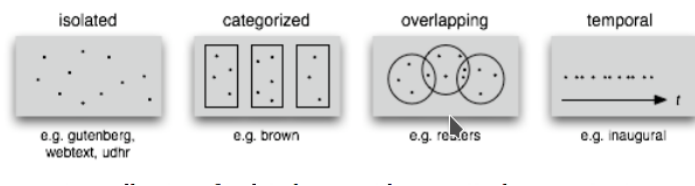


Figure 4: Diferentes estructuras de corpus

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 lingüísticos, 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 *metonimia*. Estas métricas juntas constituirán una representación 'objetiva' del concepto cualitativo de *literariedad*.

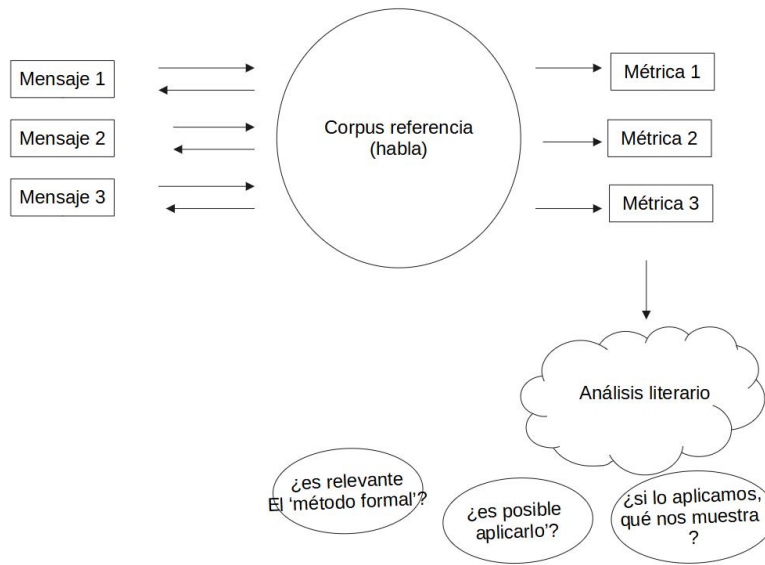


Figure 5: Entradas y salidas del algoritmo

¿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 metafórico o metonímico que otro.
- determinar si un mensajes de una misma categoria (por ejemplo, del mismo autor, o del mismo género) tienen medidas de metadora y metonimia 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 lingüístico .

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 metonimia,

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 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 sinónimos 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 consistirá 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.

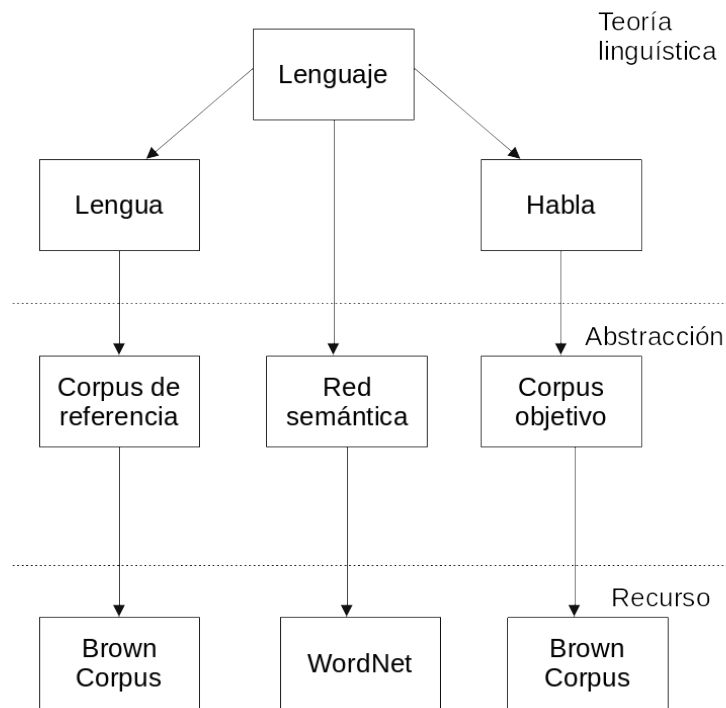


Figure 6: Resumen de las fuentes de datos utilizadas para cada concepto

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 lingüística, 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 categorías. 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

A continuación, se presenta la tabla que constituye lo que se utilizará como corpus de referencia.

cód.	nombre	categoría
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 Psychopathology	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	mistry and de- tective fiction
l11	The Man Who Looked Death in the Eye	mistry and de- tective fiction
l04	Encounter with Evil	mistry and de- tective fiction
l19	Make a Killing	mistry and de- tective fiction
l20	Death by the Numbers	mistry and de- tective 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 receptor.

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

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	mystery 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

6.4 Modelamiento

6.4.1 Selecting a modeling technique (no tengo, estoy traduciendo un modelo cualitativo –investigacion mixta–)

6.4.2 Generating a test desing

- Describing the criteria for "goodness" of a model
- Defining the data on which these criteria will be tested

1. Sampleo de la muestra

- qué textos voy a someter a procesamiento
- por qué escogí estos textos en particular

6.4.3 Building the models

1. Presentacion de las ecuaciones

$$mensaje = \{w_1, w_2, w_3, \dots, w_j\} \quad (1)$$

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

$$uso(w) = \frac{freq(w)}{freqMedia} \quad (3)$$

$$freqMedia = \mu(freq(corpora\ ref)) \quad (4)$$

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

$$N = \{n_1, n_2, n_3, \dots, n_j\} \quad (6)$$

$$met(n) = \frac{letras\ iguales}{set(letras(n_i1) + letras(n_i2))} \quad (7)$$

$$indice\ metonimia = \sum_i^j met(n_i) \quad (8)$$

2. Procedimientos para indicadores

3. Índice Metafórico

4. Matriz semántica

5. Matriz de uso

6. Índice Metonímico

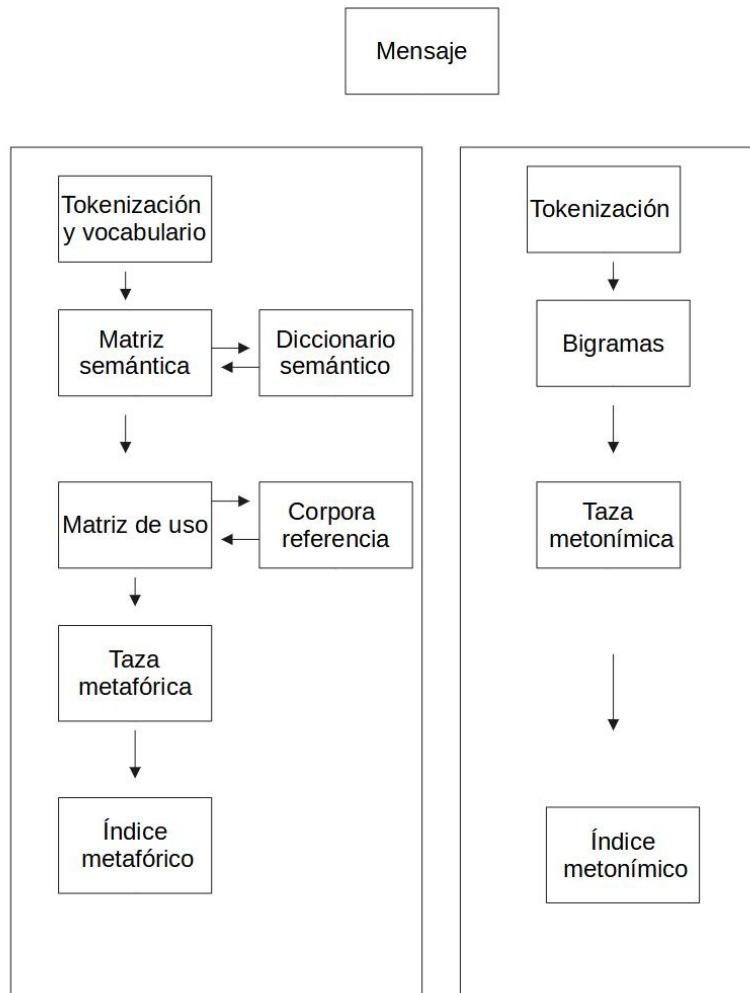


Figure 7: Procesamiento de corpus objetivo

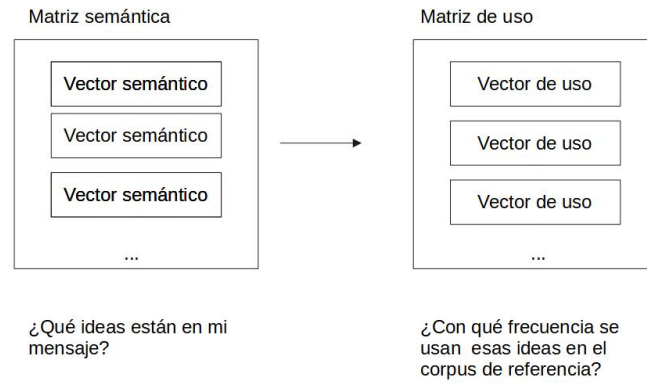


Figure 8: Abstracciones necesarias para el índice metafórico

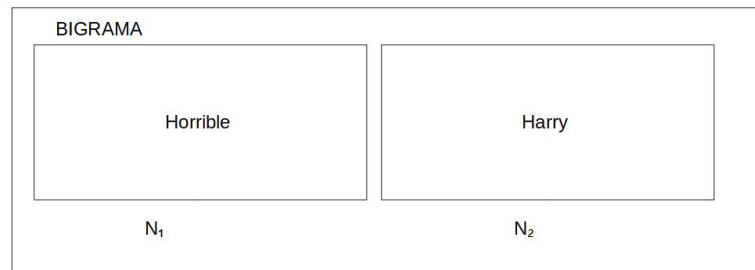


Figure 9: Abstracciones necesarias para el índice metonímico

6.5 Despliegue (los notebooks?? creo que no hay despliege)

Deployment is the process of using your new insights to make improvements within your organization. This can mean a formal integration such as the implementation of a IBM® SPSS® Modeler model producing churn scores that are then read into a data warehouse. Alternatively, deployment can mean that you use the insights gained from data mining to elicit change in your organization. For example, perhaps you discovered alarming patterns in your data indicating a shift in behavior for customers over the age of 30. These results may not be formally integrated into your information systems, but they will undoubtedly be useful for planning and making marketing decisions.

categoria	índice metafórico	w
reportage	7524.541176398631	2340
editorial	7273.001005175399	2262
reviews	7253.977846579716	2370
religion	6454.547716937541	2314
skills & hobbies	6177.00807941614	2232
popular lore	6369.248078015795	2222
belles lettres	6958.927634792041	2288
miscellaneous	7255.178115968617	2214
learned	4137.726478820393	2254
general fiction	6898.164190465642	2264
mistery and detective fiction	5882.878135583104	2446
science fiction	6763.667294465371	2412
adventure and western fiction	5208.702749914055	2560
romance and love story	7068.521208726773	2428

Table 3: Resultado índice metafórico

7 CONCLUSIONES (Creo que esto se solapa con lo que crisp-dm llama despliege)

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 Investigación en Computación, Instituto Politécnico Nacional, 1981.
- [5] Nitin Indurkha 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.