

# Modelo de literariedad basado en la lingüística de Roman Jakobson

## Fundación Universitaria Konrad Lorenz

Jonatan Ahumada

24 de mayo de 2022

# Outline

- 1 Introducción
- 2 Objetivos
- 3 Marco teórico y referencial
- 4 Diseño Metodológico
- 5 Entendimiento del negocio
- 6 Entendimiento de los datos
- 7 Preparación de los datos
- 8 Modelamiento
- 9 Despliegue
- 10 Evaluación
- 11 Conclusiones

# ¿Qué es *literariedad*?

Es una presunta *característica* que distingue un texto literario de otro no literario.

Por ejemplo:

- Manual de un carro vs. un poema de Jose Asunción Silva (fácil)
- Artículo periodístico vs. cuento corto (normal)
- *50 sombras de Gray* vs. *Ulysses* (difícil)

En los estudios clásicos, se encuentran teorías acerca de qué constituye un texto 'poético' o una buena 'narración'.

- Fedro, Platón
- Póetica, Aristóteles
- Carta a los pisones, Horacio

Sin embargo, el enfoque que toman estos autores no es **metódico o sistemático**

En el siglo 20, Ferdinand de Saussure fundó la lingüística general (también llamada estructural). Se considera al fenómeno del lenguaje como una estructura compuesta de varios componentes interdependientes, pero identificables.

Roman Jakobson fue un lingüista ruso americano. Se considera una figura clave tanto en movimiento del *formalismo ruso*, así como en el *estructuralista*. La lingüística de Jakobson se basa en los postulados de la lingüística de Saussure, **pero** propuso una crítica a las ideas de Saussure.

## Cita

*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.*  
[1, p. 99]

¿Cómo medir computarizadamente la *literariedad* de un texto según el marco de la lingüística de Jakobson?

## General

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

- 1 Construir el corpus necesario para representar el *eje sincrónico*.
- 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.

- 1 Modelos **naive**
- 2 Corpus de acceso libre (Brown Corpus)
- 3 Herramientas 'básicas' de NLP. (no Machine Learning)

# Introducción a metáfora y metonimia

**Figura:** *Cisnes reflejando elefantes*  
de Salvador Dalí



**Figura:** *9 grados* de Denise Green

# Metáfora y metonimia

...	chico	primoroso
...	puer	lindo
el	niño	bello
...	párvulo	bonito
...	infante	excelso

Figura: Ejemplo de metáfora

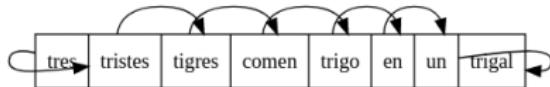


Figura: Ejemplo de metonimia

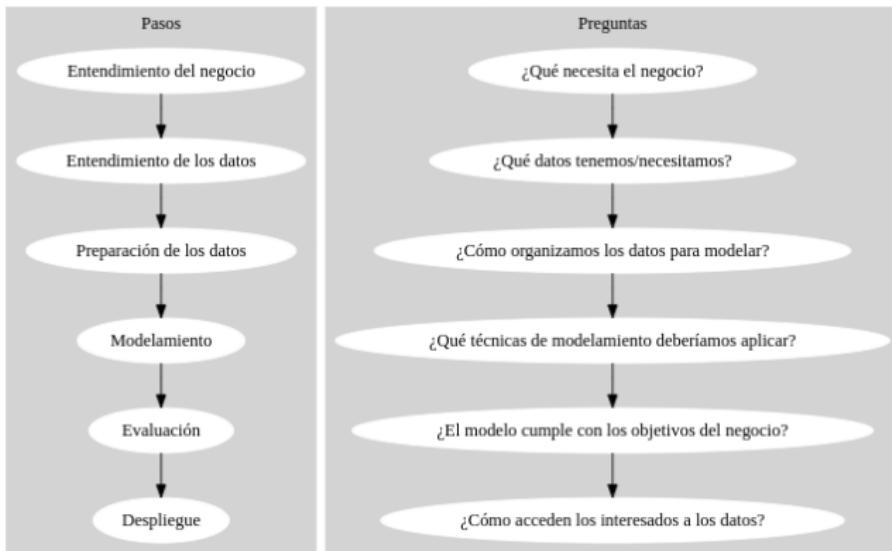
## Selección/Metáfora

A selection between alternatives implies the possibility of substituting one for the other, equivalent in one respect and different in another. Actually, selection and substitution are two faces of the same operation. [1, p.98]

## Combinación/Metonimia

Any linguistic sign involves two modes of arrangement: Any sign is made up of constituent signs and/or occurs only in combination with other signs. This means that any linguistic unit at one and the same time serves as a context for simpler units and/or finds its own context in a more complex linguistic unit. [1, p.99]

Figura: Pasos de CRISP-DM



Cada algoritmo recibe un mensaje  $m$  de entrada con:

## Entrada

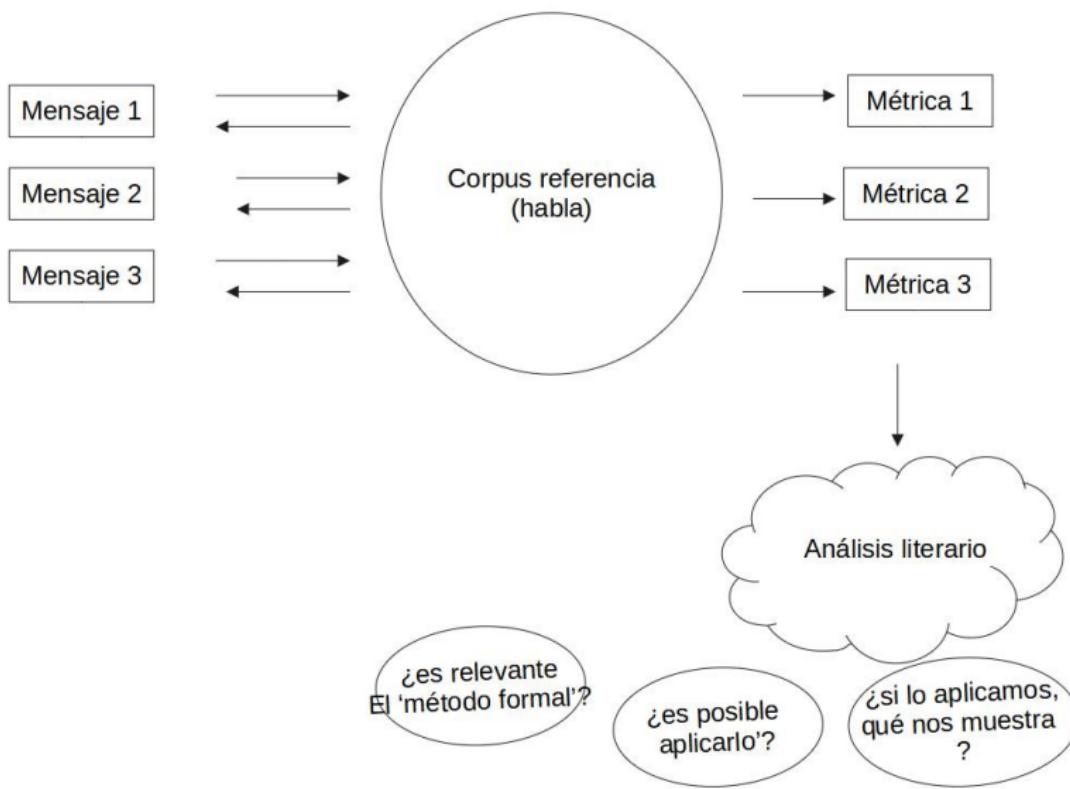
- cadena de cualquier longitud
- sin POS
- sin set de entrenamiento

produce:

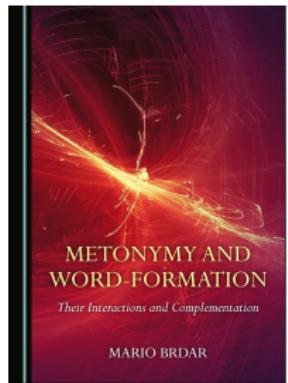
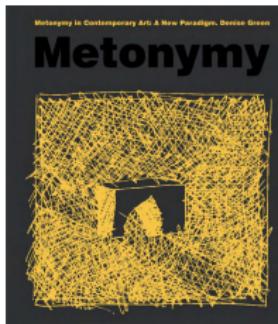
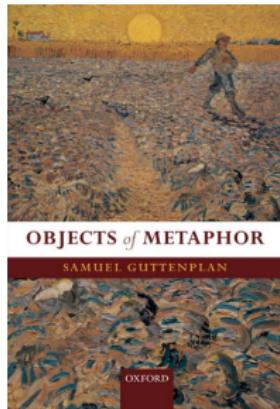
## Salida

- Un valor continuo para dicho mensaje (no es categórico)
- Entre más alto el valor, más fuerte es esa característica (metáfora y/o metonimia)

# Casos de uso



# Usuarios



# Entendimiento de los datos

Son esencialmente 3 componentes:

## Corpus de referencia

Modela el estado actual de la *lengua*. Eje de sincronía en Saussure.

## Red semántica

Modela el lenguaje: la capacidad de asociar ideas con símbolos.

## Corpus objetivo

Modela el *habla*. El mensaje que será sometido a análisis.

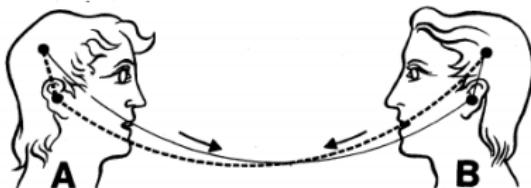


Figura: El circuito lingüístico.  
Tomado de [2]

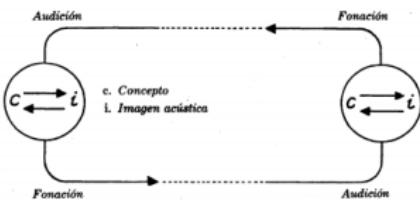
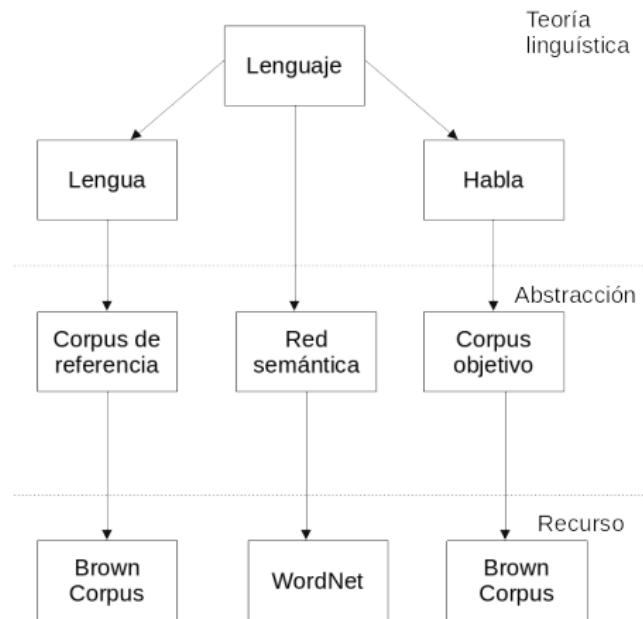


Figura: El circuito lingüístico (visión alterna). Tomado de [2]

# Resumen



Se seleccionó porque:

- 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.
- el número de textos por categoría guarda la relación entre los textos publicados de esa categoría durante ese año y
- los resultados obtenidos del modelo podrían ser replicados porque el corpus es ampliamente conocido.

## Definición

Es una base de datos de *relaciones conceptuales* entre palabras.

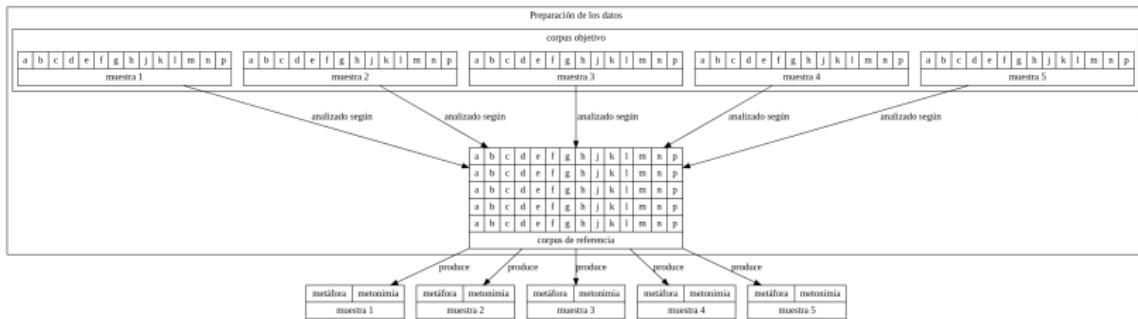
## Cita

Wordnet's design resembles that of a thesaurus in that its building block is a synset consisting of all the words that express a given concepts (...) The synsets are linked by means of a number of relations, including, hyponymy, meronymy and entailment. [3, p.8]



Figura: Ejemplo de relaciones entre conceptos en Wordnet. Tomado de [3, p.30]

# Preparación de los datos



¿En qué consistió la preparación?

- Conformar el corpus de referencia
- Conformar los corpus objetivo
- Controlar la mayor cantidad de variables

Atributo	Cantidad
Textos en corpus de referencia	60
Categorías en corpus de referencia	13
Textos en corpus objetivo	70
Textos en muestra de corpus objetivo	14
Muestras de corpus objetivo	5
Categorías por muestra	14
Total de textos usados	130

Cuadro: Resumen de datos utilizados

# Modelamiento

## Metáfora

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

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

$$vector\ uso(w) = \{freq_{ref}(s_1), freq_{ref}(s_2), freq_{ref}(s_3), \dots, freq_{ref}(s_j)\} \quad (3)$$

$$\mu = \frac{\sum_i^j freq_{referencia}(s_i)}{j} \quad (4)$$

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

$$indice\ metaforico(mensaje) = \sum_i^j uso(w_i) \quad (6)$$

```
In [8]: from jakobson.base import vector_semantico
In [8]: vector_semantico('bucolic')
Out[8]:
['peasant',
 'idyll',
 'provincial',
 'pastoral',
 'bucolic',
 'eclogue',
 'aristocrat',
 'idyl']

In [9]: vector_semantico('astonishing')
Out[9]:
['astounding',
 'staggering',
 'stupefying',
 'amazing',
 'astound',
 'astonish',
 'astonishing']
```

Figura: Ejemplo de implementación

```
In [26]: vector_semantico('then')
Out[26]: ['and_so', 'then', 'and_then', 'so']

In [27]: vector.uso(vector_semantico('then'), f_d)
Out[27]: [0, 177, 0, 228]

In [28]: from jakobson import prom_vector_uso

In [29]: prom_vector_uso(vector.uso(vector_semantico('then'),
f_d), 'then')
Out[29]: 101.25
```

Figura: Ejemplo de implementación

# Modelamiento

## Metonimia

$$N = \{n_1, n_2, n_3, \dots, n_j\} \quad (7)$$

$$met(n_i) = \frac{\text{letras iguales}}{set(\text{letras}(n_i1) + \text{letras}(n_i2))} \quad (8)$$

$$\text{indice metonimia} = \sum_i^j met(n_i) \quad (9)$$

```
In [56]: from jakobson.metonimia import metonimia
In [57]: metonimia(["tu", "pie"])
Out[57]: 0.0
In [58]: metonimia(["Arbol", "Argot"])
Out[58]: 0.42857142857142855
In [59]: metonimia(["arbol", "arbol"])
Out[59]: 1.0
```

Figura: Ejemplo de implementación

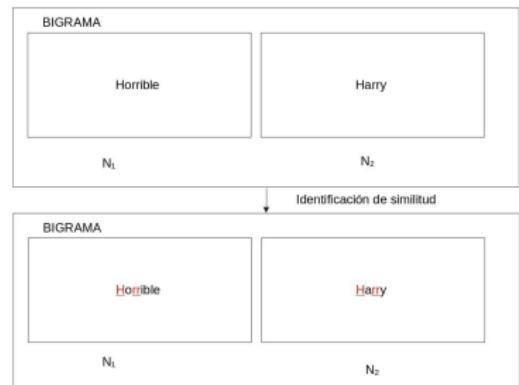


Figura: Concepto de metonimia

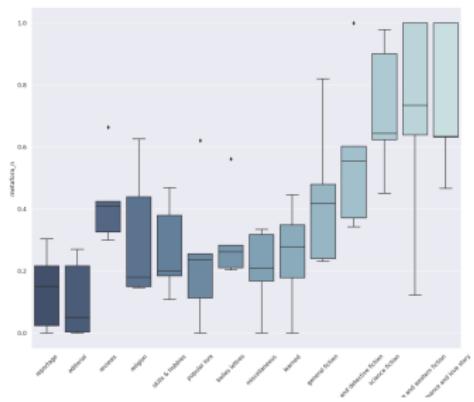
## Criterios cualitativos

- H<sub>1</sub>: Se espera que las categorías de ficción tengan un índice metafórico significativamente mayor que los de no-ficción.
- H<sub>2</sub>: Se espera que las categorías 'Reportage' y 'Editorial' tengan índices metafóricos similares a través de las muestras.
- H<sub>3</sub>: Se espera que la categoría 'Belles Lettres' tenga un índice metafórico más alta entre las categorías de no-ficción.
- H<sub>4</sub>: Se espera que la categoría 'Learned' tenga un índice metonímico bajo en general.

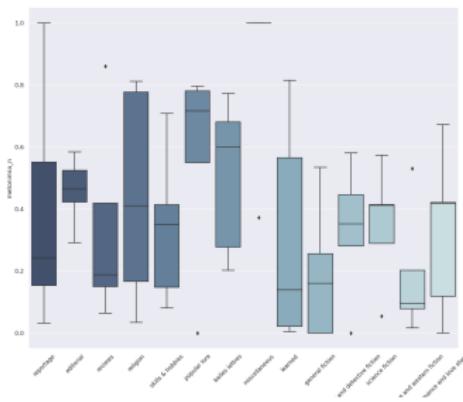
## Criterios cuantitativos

Prueba ANOVA: ¿Los resultados que se obtuvieron son aleatorios?

## Resultados por categorías



## Figura: Metáfora través de las muestras



## Figura: Metonimia través de las muestras

# Resultados por metacategorías

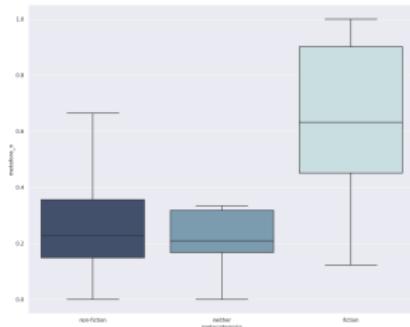


Figura: Índice metafórico por metacategorías a través de muestras

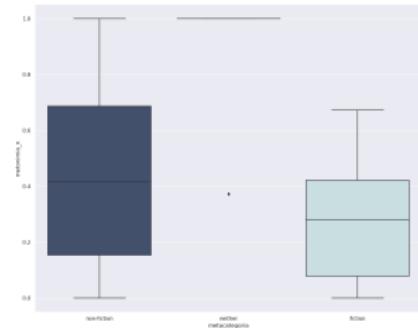


Figura: Índice metonímico por metacategoría a través de muestras

# Evaluación

## Criterios cualitativos

Criterio	Evaluación
H <sub>1</sub>	Cumplió
H <sub>2</sub>	Cumplió
H <sub>3</sub>	No cumplió
H <sub>4</sub>	Cumplió

## Criterios cuantitativos

Indicador	F	p-value
Metafora	51.41	9.81 <sup>-10</sup>
Metonimia	4.32	0.04

El modelo propuesto:

- 1 produce valores cuantitativos capaces de 'distinguir' significativamente entre dos metacategorías: los textos de ficción y los de no-ficción,
- 1 parece avalar las observaciones de Jakobson en torno a la relación de la metonimia con el polo 'Realista' (periódicos, reportes, artículos, etc) y la metáfora con el polo del 'Romanticismo' (historias, fábulas, fantasía, etc),
- 2 tiene algunas ventajas y desventajas con respecto a un enfoque de Machine Learning. Como ventaja, no se requiere un *training set*. Como desventaja, el valor de los índices debe ser comparado entre textos según un contexto dado por el corpus de referencia.

## Mejoramiento del algoritmo

- Utilizar Tf-idf para matizar mejor el índice metafórico.
- Hacer el índice metafórico sensible a los casos donde una palabra se use por debajo del promedio de uso.
- Hacer el cálculo de similitud del índice metonímico con sílabas o fonemas.
- Ofrecer la capacidad de personalizar la red semántica.

## Mejoramiento de evaluación

- Aumentar el número de muestras hasta agotar el Corpus de Brown.
- Repetir el mismo diseño experimental en el Corpus de LOB, que tiene las mismas categorías que el de Brown.
- Utilizar los valores de los índices dentro como una característica en un escenario de Machine Learning.



# Bibliografía

-  R. Jakobson, "Two aspects of language and two types of aphasic disturbances," *Fundamentals of language*, vol. 1, pp. 69–96, 1956.
-  F. De Saussure, "Curso de lingüística general," *Buenos Aires: Losada. Original de Ferdinand de*, 1945.
-  C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.