Uso de modelos de NLP para el estudio del lenguaje en el cerebro

Clase 4: misc

Dr. Bruno Bianchi
Laboratorio de Inteligencia Artificial Aplicada
Dpto Computación - FCEN - UBA
Instituto Cs Computación - CONICET - UBA
bbianchi@dc.uba.ar

[ELBA]

Julio 2024

Abnar, 2019

Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains

Samira Abnar **Lisa Beinborn** **Rochelle Choenni** **Willem Zuidema**

Institute for Logic, Language and Computation
University of Amsterdam

{abnar,l.beinborn}@uva.nl, rochelle.choenni@student.uva.nl, zuidema@uva.nl

Abstract

In this paper, we define and apply *representational stability analysis* (ReStA), an intuitive way of analyzing neural language models. ReStA is a variant of the popular *representational similarity analysis* (RSA) in cognitive neuroscience. While RSA can be used to compare representations in models, model components, and human brains, ReStA compares instances of the *same* model, while systematically varying single model parameter. Using ReStA, we study four recent and successful neural language models, and evaluate how sensitive their internal representations are to the amount of prior context. Using RSA, we perform a systematic study of how similar the representational spaces in the first and second (or higher) layers of these models are to each other and to patterns of activation in the human brain. Our results reveal surprisingly strong differences between language models, and give insights into where the *deep* linguistic processing, that integrates information over multiple sentences,

is simple: instead of directly trying to map models to brains, we first construct two similarity matrices that record how similar brain responses are to each other for different stimuli, and how similar the computational model's representations for each stimulus are to each other. The representational similarity score is then defined as the similarity (typically: Pearson's correlation) of the two similarity matrices (or equivalently: the similarity of two distance matrices).

RSA can also be applied to deep learning models (Laakso and Cottrell, 2000; Dharmaretnam and Fyshe, 2018; Alvarez-Melis and Jaakkola, 2018; Wang et al., 2018; Chrupała and Alishahi, 2019). In this paper, we present a large-scale study and comparison of both neural language models and fMRI data from brain imaging experiments with human subjects, using RSA. However, we extend standard RSA using an approach we call *Representational Stability Analysis* (ReStA). The idea is

Abnar, 2019

Presentan ReStA (basado en RSA): otra forma de comprar actividad cerebral y procesamiento de los modelos

Modelos analizados:

- ➔ GloVe (tipo Word2Vec)
- ➔ ELMO y GoogleLM (RNN)
- ➔ BERT y UniSentEnc (transformers)

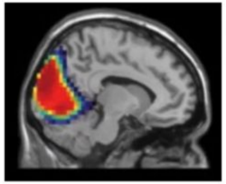
Variables a analizar:

- ➔ Cantidad de contexto
- ➔ Capas

Dataset de fMRI:

- ➔ Whebe et al., 2014
- ➔ Sujetos leyendo Harry Potter

Abnar, 2019

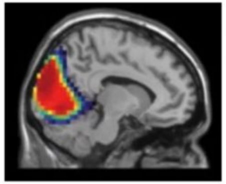


Representational Similarity Matrix from Visual Cortex (fMRI)

	Scene 1	Scene 2	Scene 3	...	Scene 20
Scene 1	1	.84	.09	...	.26
Scene 2	.84	1	.32	...	.17
Scene 3	.09	.32	1	...	.54
⋮	⋮	⋮	⋮		⋮
Scene 20	.26	.17	.54	...	1

https://www.youtube.com/watch?v=oywoPu5ivgs

Abnar, 2019

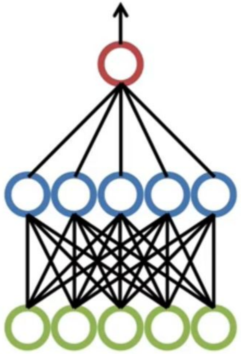


Representational Similarity Matrix from Visual Cortex (fMRI)

	Scene 1	Scene 2	Scene 3	···	Scene 20
Scene 1	1	.84	.09	···	.26
Scene 2	.84	1	.32	···	.17
Scene 3	.09	.32	1	···	.54
⋮	⋮	⋮	⋮		⋮
Scene 20	.26	.17	.54	···	1

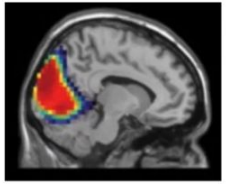
Representational Similarity Matrix from Computational Model

	Scene 1	Scene 2	Scene 3	···	Scene 20
Scene 1	1	.72	.12	···	.31
Scene 2	.72	1	.28	···	.14
Scene 3	.12	.28	1	···	.61
⋮	⋮	⋮	⋮		⋮
Scene 20	.31	.14	.61	···	1



https://www.youtube.com/watch?v=oywoPu5ivgs

Abnar, 2019



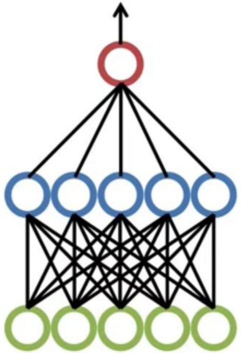
Representational Similarity Matrix
from Visual Cortex (fMRI)

	Scene 1	Scene 2	Scene 3	...	Scene 20
Scene 1	1	.84	.09	...	.26
Scene 2	.84	1	.32	...	.17
Scene 3	.09	.32	1	...	.54
⋮	⋮	⋮	⋮		⋮
Scene 20	.26	.17	.54	...	1

Representational Similarity Matrix
from Computational Model

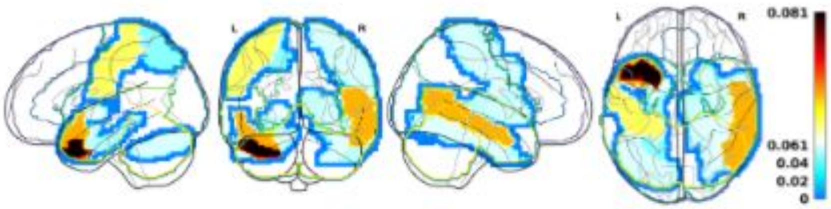
	Scene 1	Scene 2	Scene 3	...	Scene 20
Scene 1	1	.72	.12	...	.31
Scene 2	.72	1	.28	...	.14
Scene 3	.12	.28	1	...	.61
⋮	⋮	⋮	⋮		⋮
Scene 20	.31	.14	.61	...	1

Rank order correlation

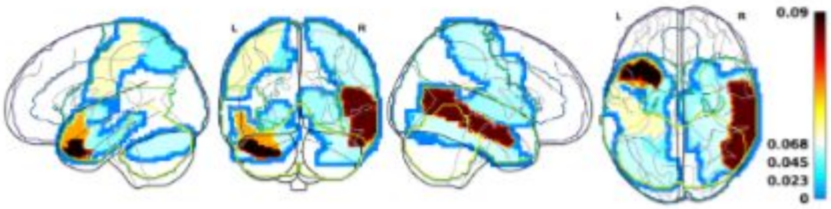


https://www.youtube.com/watch?v=oywoPu5ivgs

Abnar, 2019



(f) BERT (L0)



(g) BERT (L11)

Zhou et al., 2023

Divergences between Language Models and Human Brains

Yuchen Zhou [1] Emmy Liu [1] Graham Neubig [1] Michael J. Tarr [1] Leila Wehbe [1]

Abstract

Do machines and humans process language in similar ways? Recent research has hinted in the affirmative, finding that brain signals can be effectively predicted using the internal representations of language models (LMs). Although such results are thought to reflect shared computational principles between LMs and human brains, there are also clear differences in how LMs and humans represent and use language. In this work, we systematically explore the divergences between human and machine language processing by examining the differences between LM representations and human brain responses to language as measured by Magnetoencephalography (MEG) across two datasets in which subjects read and listened to narrative stories. Using a data-driven approach, we identify two domains that are not captured well by LMs: social/emotional intelligence and physical commonsense. We then validate

Huth, 2018; Toneva & Wehbe, 2019), EEG (Hale et al., 2018), MEG (Wehbe et al., 2014b), and ECoG (Goldstein et al., 2022), can effectively be predicted using representations from language models such as BERT (Devlin et al., 2018) or GPT-2 (Radford et al., 2019). Robust neural prediction is hypothesized to stem from the shared computational objective of both LMs and the human brain: predicting subsequent words based on prior context (Yamins & DiCarlo, 2016; Schrimpf et al., 2021).

Despite the evident behavioral similarities, the extent to which LMs and human brains align functionally for language processing remains an open question. Essentially, the methods that LMs and humans use to acquire language are very different. LMs learn statistical regularities across massive sets of linguistic symbols, whereas humans rely on applying structured linguistic principles across relatively little input. Additionally, LMs that are confined to linguistic data are likely to fail to ground linguistic symbols in real-world contexts (Harnad, 1990; B

Zhou et al., 2023

¿Qué información NO capturan bien los modelos de lenguaje?

Modelos analizados:

- ➔ GPT-2 XL
- ➔ Llama-2

Dataset de MEG:

- ➔ Whebe et al., 2014
- ➔ Sujetos leyendo Harry Potter

Dataset de fMRI:

- ➔ LeBel et al., 2023
- ➔ Sujetos escuchando podcast

Zhou et al., 2023

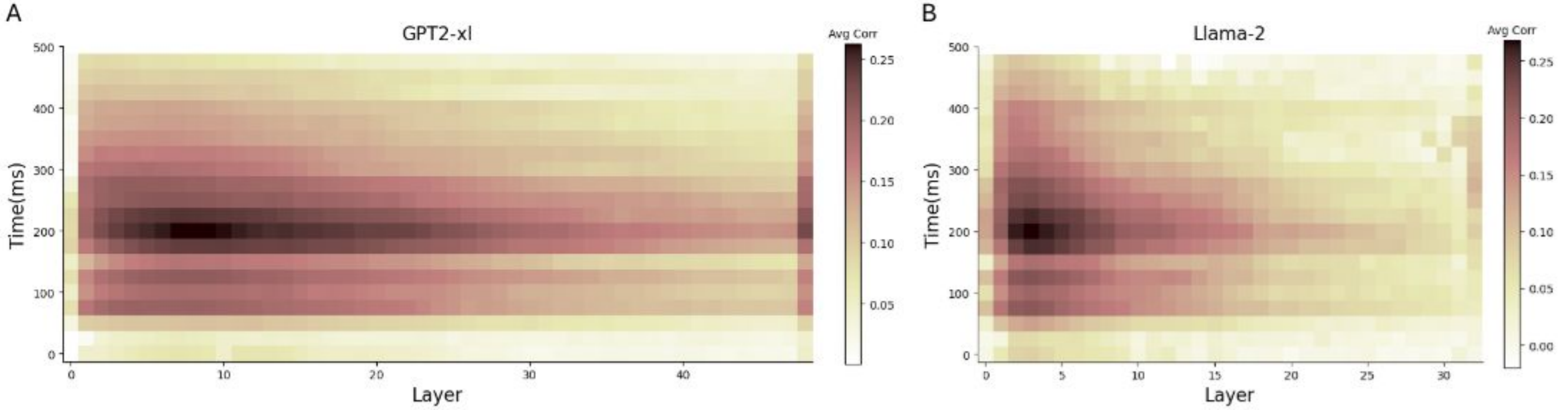


Figure 2. Pearson Correlation between actual MEG responses and predictions from (A) GPT-2 XL and (B) Llama-2 across LM layers and time after word onset. Both models exhibit high correlations in early and intermediate layers at around 200ms. Correlation is computed across words and averaged across MEG channels.

Zhou et al., 2023

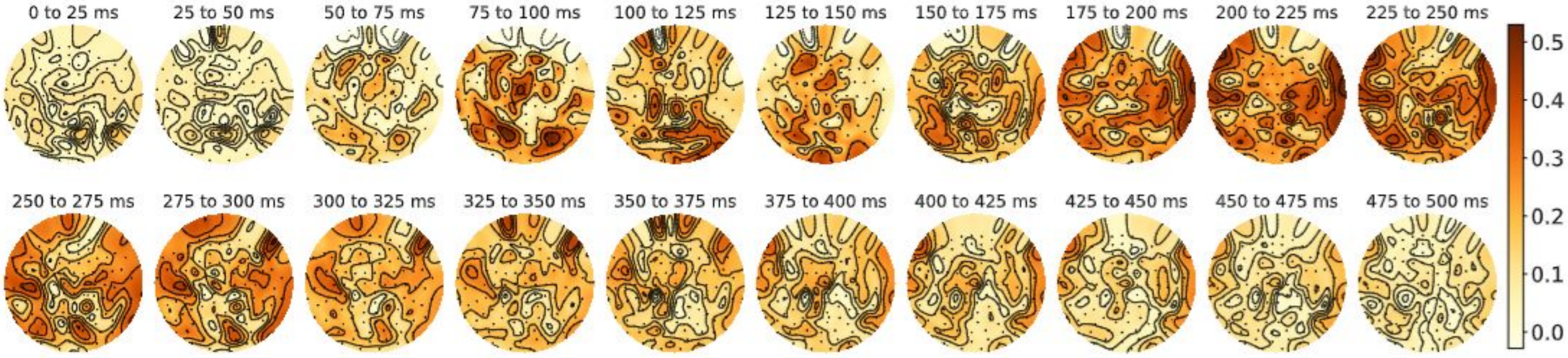


Figure 3. Pearson correlation of actual MEG responses with those predicted by LM embedding from the best layer (layer 7) of GPT-2 XL (evaluated on the test set). The displayed layout is a flattened representation of the helmet-shaped sensor array. Deeper reds indicate more accurate LM predictions. Language regions are well predicted in language processing time windows (refer to §2.4 for more details).

Zhou et al., 2023

Análisis de diferencias:

- → Calculan error de predicción (MSE) de cada palabra
- → Para cada oración calculan el MSE promedio
- → Comparan las peores 100 con las mejores 100
  - ◆ Le piden a otro modelo de lenguaje que las compare y gener hipótesis
  - ◆ Después hacen una validación a mano

Zhou et al., 2023

Table 2. Top 10 hypotheses generated by the best layer of GPT-2 XL for the Harry Potter dataset

Hypothesis	**Validity**	***p*-value**
have a high level of emotional intensity	0.250	0.010
involve complex sentence structures or grammar	0.250	0.015
include emotional language or descriptions	0.238	0.008
have a high level of tension or conflict	0.237	0.023
have characters using body language or non-verbal cues	0.225	0.032
are emotionally charged, making it challenging for language models to accurately interpret the intended tone or sentiment	0.213	0.020
include conflicts between characters	0.200	0.035
have characters interacting with their environment	0.188	0.059
have complex sentence structures	0.175	0.081
have dialogue between characters with varying emotions	0.175	0.022

Table 3. Top 10 hypotheses generated by the best layer of GPT-2 XL for the Moth dataset

Hypothesis	**Validity**	***p*-value**
contain elements of fiction or exaggeration	0.212	0.012
feature emotional or dramatic language	0.150	0.090
refer to cultural or societal norms	0.138	0.107
include sensory details or imagery	0.137	0.107
have strong emotional or dramatic content	0.100	0.173
show a lack of coherence or logical flow	0.100	0.111
contain elements of surprise and unpredictability	0.094	0.201
contain emotional, personal narratives	0.088	0.201
use idiomatic expressions or figurative language	0.088	0.178
refer to specific events or incidents	0.087	0.237

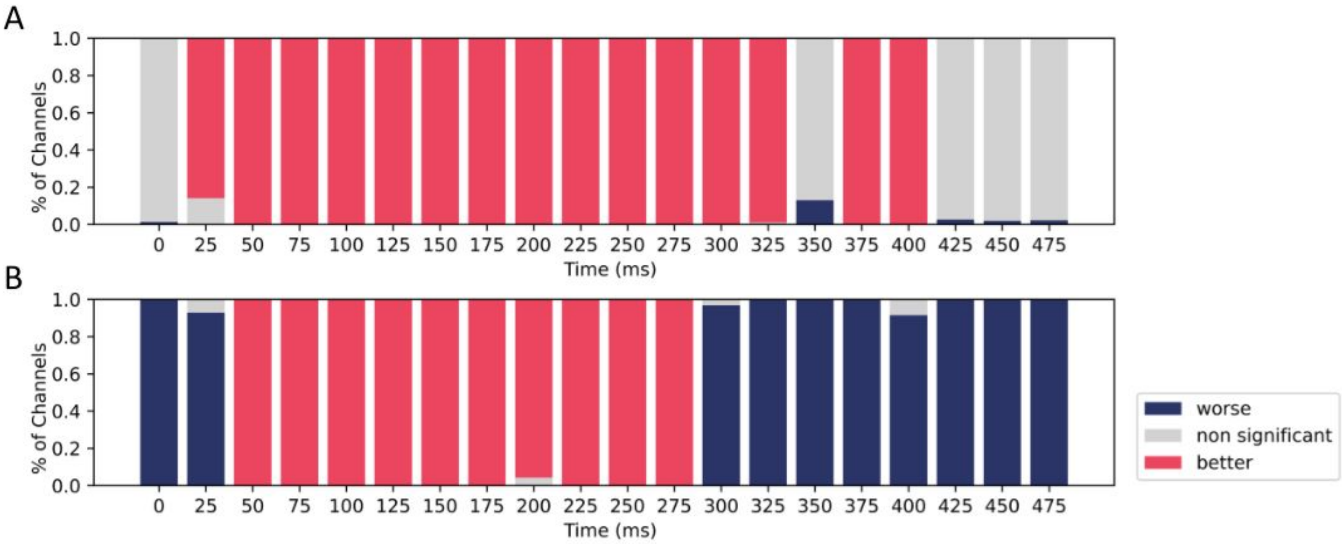
“We identified two primary differences between the language model and the human brain: firstly, the processing of **social and emotional information**, and secondly, the capacity for **interaction with the surrounding environment.**”

Zhou et al., 2023

Reentrenamiento de GPT2

Table 4. Datasets for Fine-Tuning with Sample Questions and Answers (Correct Answer in Bold)

Dataset	Type	Num train	Options	Sample question	Sample answers
Social IQa	Social/Emotion	33.4k	3	Sydney had so much pent up emotion, they burst into tears at work. How would Sydney feel afterwards?	1. affected 2. **like they released their tension** 3. worse
PiQA	Physical	16.1k	2	When boiling butter, when it's ready, you can	1. Pour it onto a plate 2. **Pour it into a jar**



LeBel et al., 2021

Behavioral/Cognitive

Voxelwise Encoding Models Show That Cerebellar Language Representations Are Highly Conceptual

Amanda LeBel,[1] Shailee Jain,[3] and Alexander G. Huth[2,3]

[1]Helen Wills Neuroscience Institute, University of California–Berkeley, Berkeley, California 94720, [2]Department of Neuroscience, University of Texas–Austin, Austin, Texas 78712, and [3]Department of Computer Science, University of Texas–Austin, Austin, Texas 78712

There is a growing body of research demonstrating that the cerebellum is involved in language understanding. Early theories assumed that the cerebellum is involved in low-level language processing. However, those theories are at odds with recent work demonstrating cerebellar activation during cognitive tasks. Using natural language stimuli and an encoding model framework, we performed an fMRI experiment on 3 men and 2 women, where subjects passively listened to 5 h of natural language stimuli, which allowed us to analyze language processing in the cerebellum with higher precision than previous work. We used these data to fit voxelwise encoding models with five different feature spaces that span the hierarchy of language processing from acoustic input to high-level conceptual processing. Examining the prediction performance of these models on separate BOLD data shows that cerebellar responses to language are almost entirely explained by high-level conceptual language features rather than low-level acoustic or phonemic features. Additionally, we found that the cerebellum has a higher proportion of voxels that represent social semantic categories, which include "social" and "people" words, and lower representations of all other semantic categories, including "mental," "concrete," and "place" words, than cortex. This suggests that the cerebellum is representing language at a conceptual level with a preference for social information.

Key words: cerebellum; computational; encoding; fMRI; language; semantic

Significance Statement

Rodd, 2020

Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access

Perspectives on Psychological Science
2020, Vol. 15(2) 411–427

DOI: 10.1177/1745691619885860
www.psychologicalscience.org/PPS
SAGE

Jennifer M. Rodd
Department of Experimental Psychology, University College London

Abstract
Most words are ambiguous: Individual word forms (e.g., *run*) can map onto multiple different interpretations depending on their sentence context (e.g., the athlete/politician/river *runs*). Models of word-meaning access must therefore explain how listeners and readers can rapidly settle on a single, contextually appropriate meaning for each word that they encounter. I present a new account of word-meaning access that places semantic disambiguation at its core and integrates evidence from a wide variety of experimental approaches to explain this key aspect of language comprehension. The model has three key characteristics. (a) Lexical-semantic knowledge is viewed as a high-dimensional space; familiar word meanings correspond to stable states within this lexical-semantic space. (b) Multiple linguistic and paralinguistic cues can influence the settling process by which the system resolves on one of these familiar meanings. (c) Learning mechanisms play a vital role in facilitating rapid word-meaning access by shaping and maintaining high-quality lexical-semantic knowledge throughout the life span. In contrast to earlier models of word-meaning access, I highlight individual differences in lexical-semantic knowledge: Each person's lexicon is uniquely structured by specific, idiosyncratic linguistic experiences.

Keywords
cognition, comprehension, language/communication, lexical ambiguity

Rodd, 2020

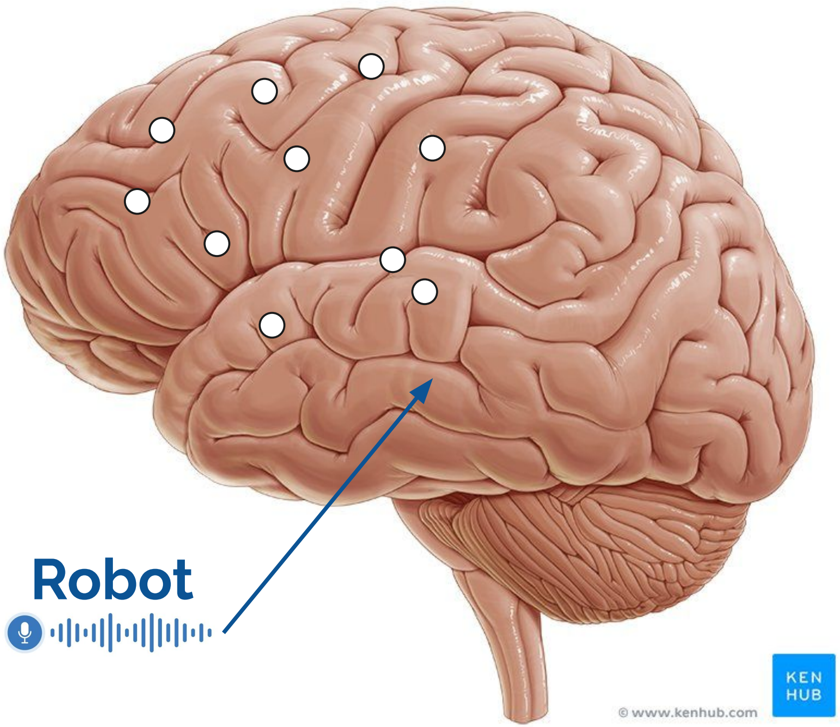
¿Cómo se guarda (y accede) la información léxico-semántica de las palabras en el cerebro?

Rodd, 2020

¿Cómo se guarda (y accede) la información léxico-semántica de las palabras en el cerebro?

Una hipótesis (Rodd 2020):

> **Representaciones distribuidas:** el acceso al significado de una palabra se da al activarse zonas de la corteza relacionadas al significado de la palabra



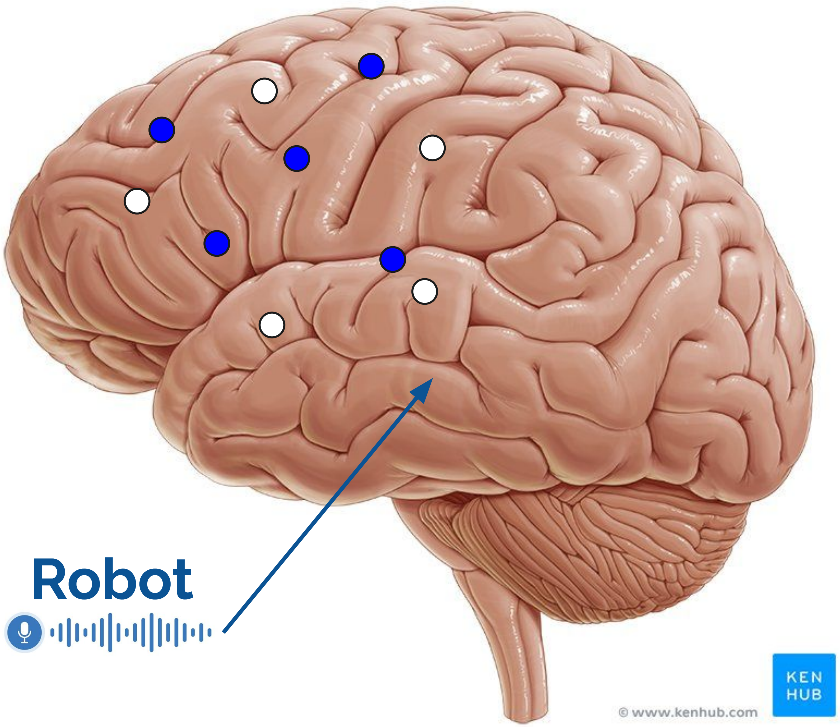
Rodd, 2020

¿Cómo se guarda (y accede) la información léxico-semántica de las palabras en el cerebro?

Una hipótesis (Rodd 2020):

Representaciones distribuidas: el acceso al significado de una palabra se da al activarse zonas de la corteza relacionadas al significado de la palabra

¿



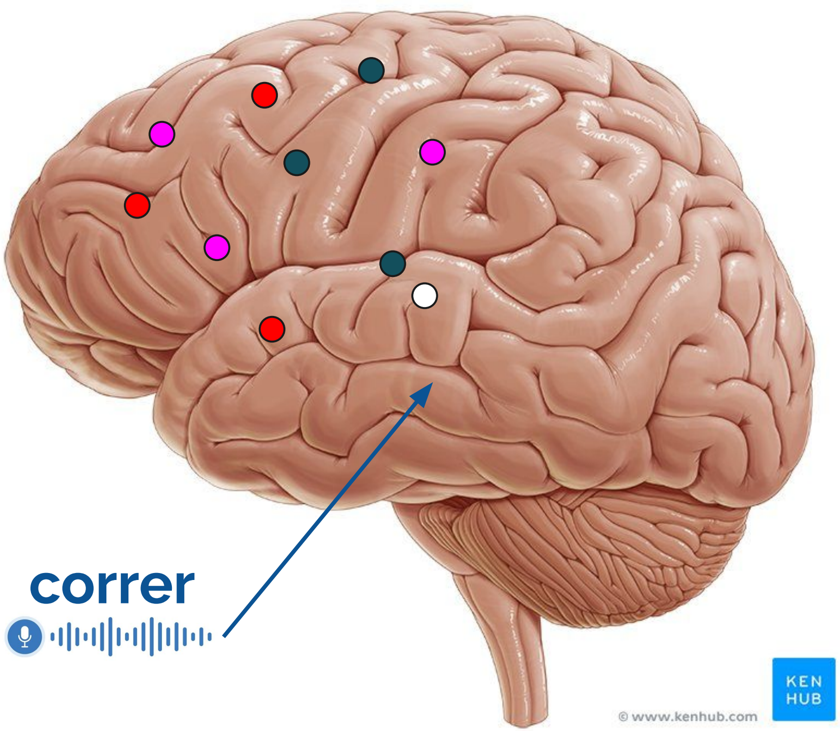
Rodd, 2020

¿Cómo se guarda (y accede) la información léxico-semántica de las palabras en el cerebro?

Una hipótesis (Rodd 2020):

> **Representaciones distribuidas:** el acceso al significado de una palabra se da al activarse zonas de la corteza relacionadas al significado de la palabra

¿Y si es una palabra con más de un significado?



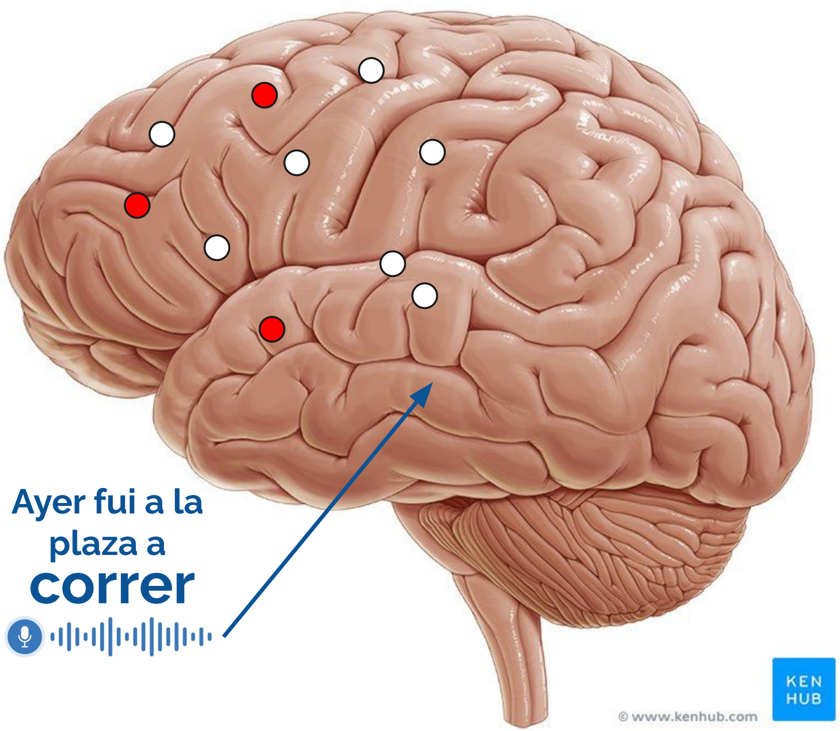
Rodd, 2020

¿Cómo se guarda (y accede) la información léxico-semántica de las palabras en el cerebro?

Una hipótesis (Rodd 2020):

Representaciones distribuidas: el acceso al significado de una palabra se da al activarse zonas de la corteza relacionadas al significado de la palabra

¿Y si es una palabra con más de un significado?



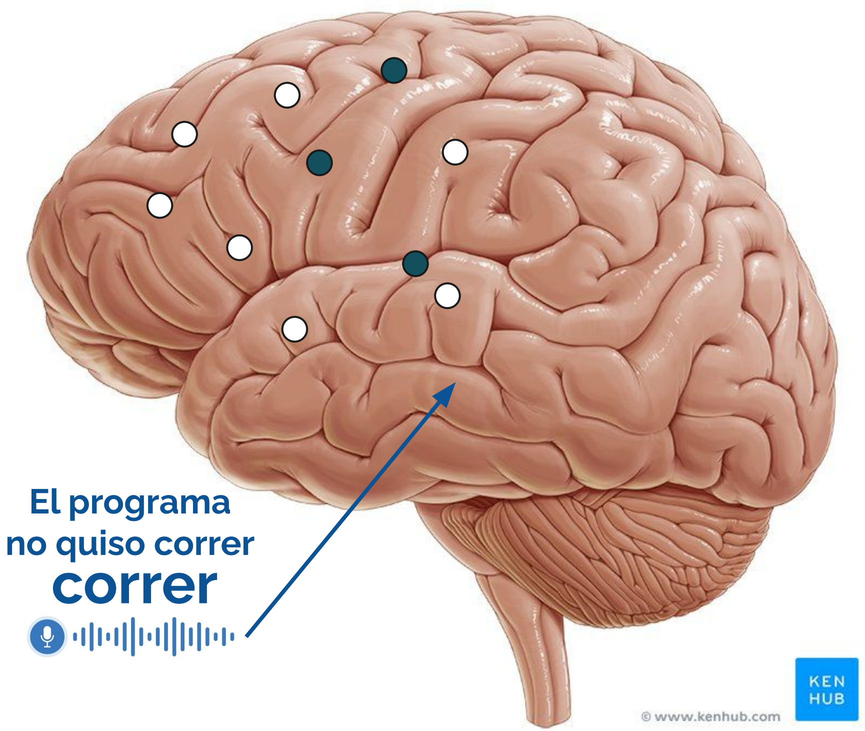
Rodd, 2020

¿Cómo se guarda (y accede) la información léxico-semántica de las palabras en el cerebro?

Una hipótesis (Rodd 2020):

Representaciones distribuidas: el acceso al significado de una palabra se da al activarse zonas de la corteza relacionadas al significado de la palabra

¿Y si es una palabra con más de un significado?



Rodd, 2020

¿Cómo se guarda (y accede) la información léxico-semántica de las palabras en el cerebro?

Una hipótesis (Rodd 2020):

> **Representaciones distribuidas:** el acceso al significado de una palabra se da al activarse zonas de la corteza relacionadas al significado de la palabra

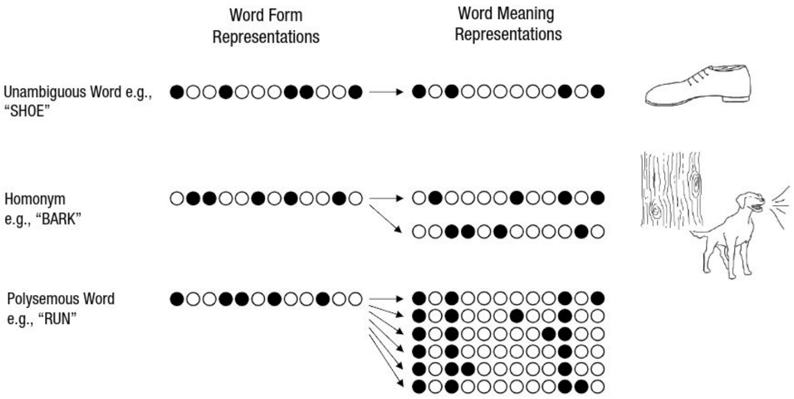


Fig. 1. Representations of different types of words within a distributed framework. Drawings adapted from Betts (2017) under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

¿Y si es una palabra con más de un significado?

Nuestro experimento

El dataset

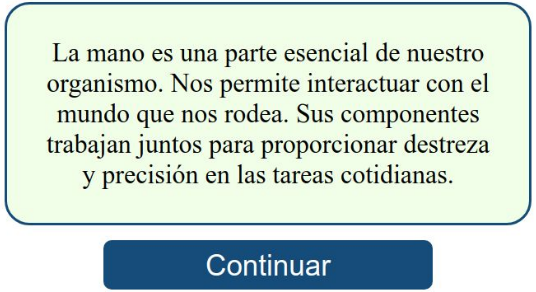
	target	oracion	significado1	Contexto1	significado2	Contexto2	Significado3	Contexto3
0	raya	Daniel observó la raya fijamente durante vario...	Animales	El mar, vasto e inexplorado, es hogar de una a...	Geometría	Las líneas son elementos básicos en el diseño ...	Lingüística	La tilde diacrítica es la que permite distingu...
1	llama	A lo lejos pudo ver la llama moviéndose suavem...	Animales	La puna argentina, con su altiplano y clima in...	Física	La combustión es una reacción química exotérmi...	Educación	La inflación es el aumento generalizado y sost...
2	salamandra	Paula le tenía mucho aprecio a la salamandra q...	Animales	Los pequeños anfibios son criaturas fascinante...	Mobiliario	La cabaña se erguía solitaria en la montaña, c...	Juguete	LEGO, es una empresa danesa, cuyo producto más...
3	muñeca	Carla estaba agarrando su muñeca fuertemente c...	Cuerpo	La anatomía es una de las ciencias fundamental...	Juguete	El niño abrió su regalo con emoción y encontró...	Música	Los Beatles, fue una banda de rock británica f...
4	palma	Enzo movió la palma muy rápido para evitar que...	Cuerpo	La mano es una parte esencial de nuestro organ...	Vegetales	La playa caribeña se presenta como un paraíso ...	Política	El 11 de noviembre de 1951 las mujeres argenti...
5	peso	Al llegar a destino notaron que el peso había ...	Economía	A pesar de los esfuerzos para estabilizar a nu...	Física	La microgravedad crea un ambiente de ingravide...	Armamento	El AK-47 es un fusil de asalto soviético. Fue ...
6	banco	A Lara le gustaba ese banco porque le traía bu...	Economía	El sistema financiero es un conjunto de instit...	Mobiliario	Una plaza es un espacio público al aire libre ...	Frutas	Mientras observaba detenidamente, pude ver cóm...

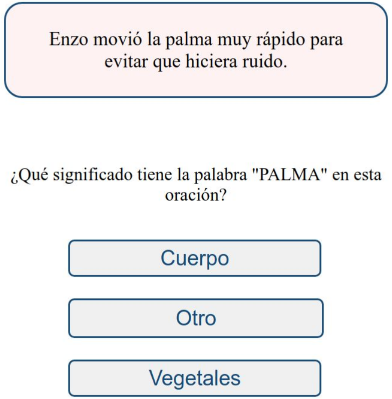
Experimento comportamental (humano)

Se realizó un experimento donde sujetos experimentales leyeron oraciones que contenían palabras ambiguas.

Estas oraciones podían estar precedidas de un contexto sesgador, o un contexto distractor.

Luego se pedía a los participantes que manifestaran el significado de la palabra en la oración.

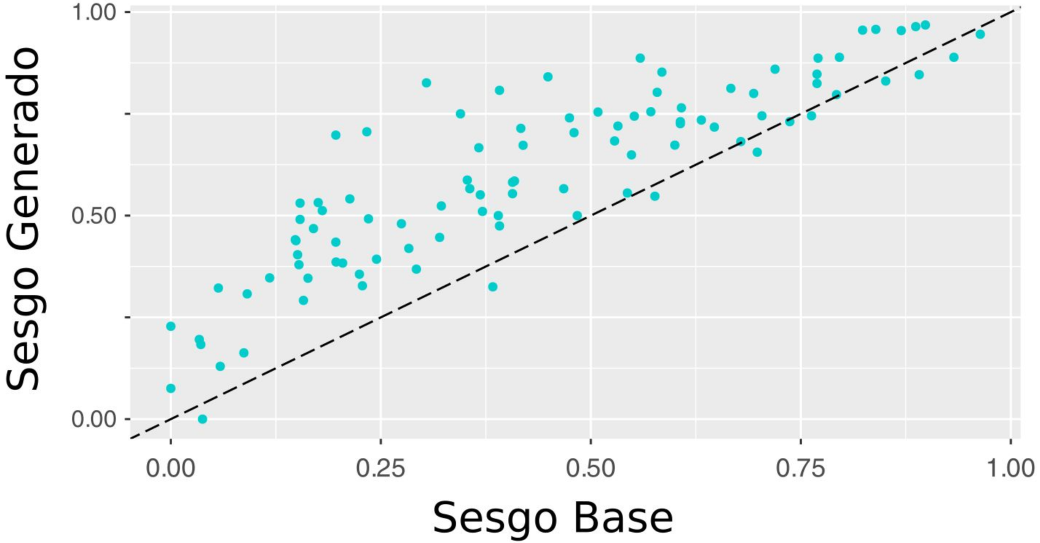




Experimento comportamental (humano)

Sesgo: Inclinación o preferencia hacia un significado específico de una palabra polisémica

- **Sesgo Base:** Sesgo que se tiene sobre una palabra dado que el contexto brindado es neutro
- **Sesgo Generado:** Sesgo que se tiene cuando se brinda un contexto sesgador

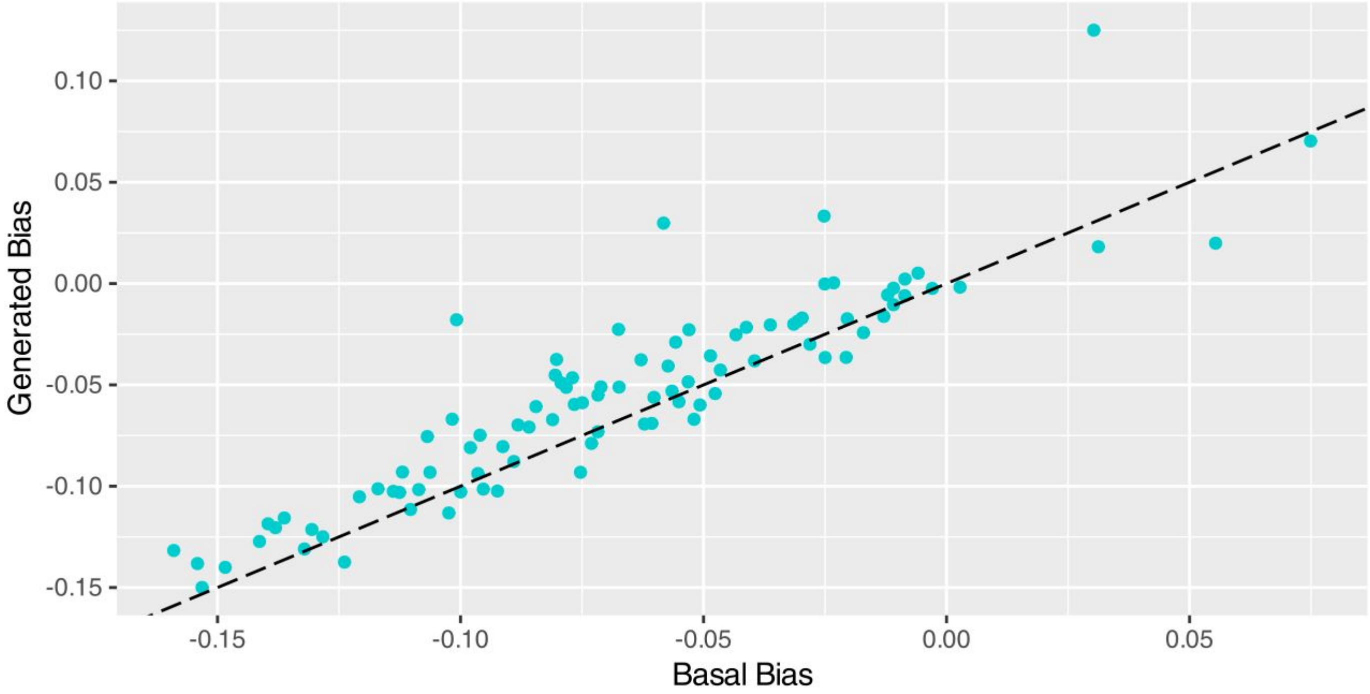


Experimento comportamental (computacional)

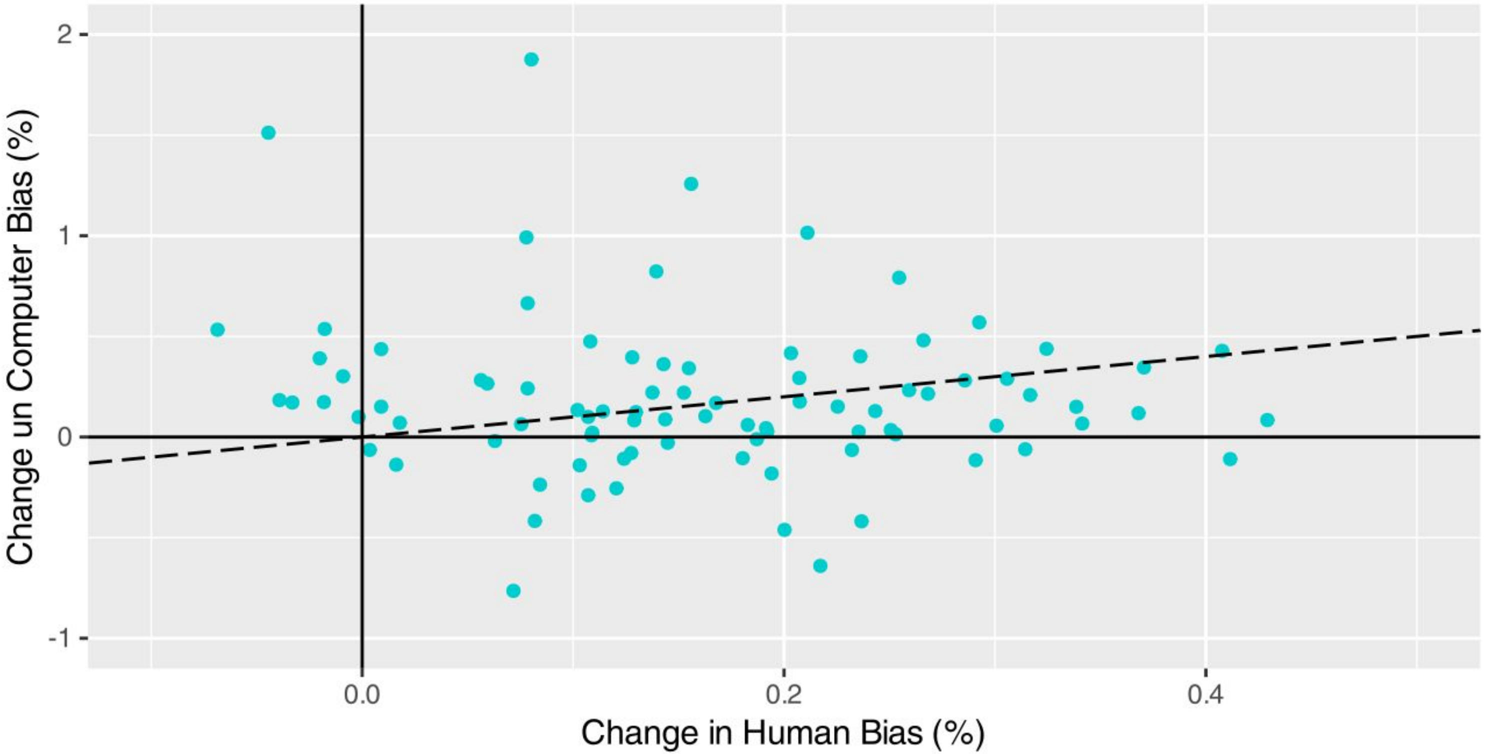
Se define:

- **Sesgo Base:** Distancia entre el embedding estático del significado y el embedding contextualizado de la palabra polisémica dado que el contexto es neutro
- **Sesgo Generado:** Distancia entre el embedding estático del significado y el embedding contextualizado de la palabra polisémica, dado que el contexto es sesgador.

Experimento comportamental (computacional)



Experimento comportamental (comparación)



Trabajo Futuro

- Análisis de embeddings capa a capa
- Experimentar con diferentes modelos
- Desarrollar nuevas métricas de desambiguación
- Comparar con neuroimágenes

Hasta mañana!