RESPUESTAS OBSERVABLES Y ESTADOS OCULTOS EN REDES NEURONALES ARTIFICIALES PARA RAZONAR SOBRE ASPECTOS COGNITIVOS DEL LENGUAJE

OBSERVABLE RESPONSES AND HIDDEN STATES IN ARTIFICIAL NEURAL NETWORKS TO REASON ABOUT COGNITIVE ASPECTS OF LANGUAGE

GUILLERMO JORGE-BOTANA[1],

JOSE ÁNGEL MARTÍNEZ-HUERTAS[2] Y
ALEJANDRO MARTÍNEZ-MINGO[2]

Cómo referenciar este artículo/How to reference this article:

Jorge-Botana, G., Martínez-Huertas, J. A y Martínez-Mingo, A. (2025). Respuestas observables y estados ocultos en redes neuronales artificiales para razonar sobre aspectos cognitivos del lenguaje. [Observable responses and hidden states in artificial neural networks to reason about cognitive aspects of language]. *Acción Psicológica, 22*(1), 41–56. https://doi.org/10.5944/ap.22.1.43347

Resumen

Para estudiar los procesos psicológicos involucrados en el lenguaje, la Ciencia Cognitiva indaga sobre las representaciones internas que se manejan a la hora de comprender o producir lenguaje. También postula las operaciones que modifican esas mismas representaciones dadas unas constricciones contextuales. Así, contexto y representación interactúan para crear significados. Con arreglo a esto, se ofrecen diferentes hipótesis de como el sistema cognitivo produce lenguaje. Al igual que existen metodologías experimentales para su estudio, distintas arquitecturas de redes neuronales artificiales permiten dotar a dichas hipótesis de un aparataje formal. En estos modelos, las representaciones y las operaciones

Correspondence address [Dirección para correspondencia]: Guillermo Jorge-Botana, Facultad de Psicología, Univesidad Complutense de Madrid, España.
Email: guijorge@ucm.es

ORCID: Guillermo Jorge-Botana (https://orcid.org/0000-0001-5879-6783), José Ángel Martínez-Huertas (https://orcid.org/0000-0002-6700-6832) y Alejandro Martínez-Mingo (https://orcid.org/0000-0002-8375-0952).

Agradecimientos: esta publicación es parte del Proyecto de I+D+i PID2022-136905OB-C22 financiado por el Ministerio de Ciencia e Innovación MCIN/ AEI/ 10.13039/501100011033/ FEDER, UE.

[1] Universidad Complutense de Madrid, España.
[2] Universidad Nacional de Educación a Distancia, España.

Recibido: 12 de noviembre de 2024.
Aceptado: 22 de enero de 2025.

participantes quedan exhaustivamente caracterizadas. Las redes neuronales recurrentes (RNNs) con mecanismos LSTM y los *Transformers* destacan como arquitecturas especialmente útiles para modelar la secuencialidad contextual presente en el lenguaje. Este número especial nos brinda la ocasión para explicar el uso de sus expresiones externas (sus salidas) como de sus representaciones internas (estados ocultos) para entender en términos cognitivos el efecto que tienen los cambios de expectativas en distintas marcas temporales de las frases. Para hacerlo, se ilustra la formalización mediante una RNN Secuencia-Secuencia con codificador y decodificador y se homologan sus mediciones a los experimentos de potenciales evento-relacionados (ERPs) en un tema nuclear en el lenguaje: la composicionalidad sistemática.

Palabras clave: Redes Neuronales Artificiales; Redes Recurrentes; LSTM; Estados Ocultos; Sorpresividad; Lenguaje; Potenciales Evento-Relacionados; Transformers.

Abstract

In order to study the psychological processes involved in language, Cognitive Science investigates the internal representations involved in understanding or producing language. It also postulates the operations that modify those representations given contextual constraints. Thus, context and representation interact to create meanings. Accordingly, there are different hypotheses about how the cognitive system produces language. Just as there are experimental methodologies for their study, different architectures of artificial neural networks make it possible to provide these hypotheses with a formal apparatus. In these models, the representations and operations involved are exhaustively characterized. Recurrent neural networks (RNNs) with LSTM mechanisms and Transformers stand out as particularly useful architectures for modeling the contextual sequentiality of language. This special issue gives us the opportunity to explain how to use their external expressions (outputs) as well as their internal representations (hidden states) to understand, in cognitive terms, the effect that changes of expectations have on different temporal markings of sentences. To do so, we illustrate such formalization using a Sequence-Sequence RNN with encoder and decoder and relate its measures with event-related potentials (ERPs) experiments on a nuclear issue in language: systematic compositionality.

Keywords: Artificial Neural Networks; Recurrent Networks; LSTM; Hidden States; Surprisal; Language; Event-Related Potentials; Transformers.

Respuestas observables y estados ocultos en Redes Neuronales Artificiales para razonar sobre aspectos cognitivos del lenguaje

Desde que se superaron las constricciones metodológicas de los modelos conductistas, la Psicología ha avanzado hacia el estudio de las representaciones internas que se generan en los razonamientos humanos (Neisser, 1967). En este contexto, la Ciencia Cognitiva propone que estas representaciones internas son clave para comprender los procesos mentales (Anderson, 2005; Pitt, 2022; Sterelny, 1990), y que su carácter emergente es el resultado de la interacción entre aspectos primitivos del entorno y la interpretación contextualizada del individuo. Estas representaciones internas se llaman emergentes porque son los símbolos que las personas manejamos que se generan a partir de aspectos primitivos de la realidad. Podríamos apelar al término emergente en tanto que han sido contextualizadas (i.e., sesgadas) a la situación interna o externa del individuo. No obstante, es ampliamente reconocido que, en general, se dedica un mayor esfuerzo a la recopilación de datos empíricos, que al desarrollo de modelos formales que describan las representaciones mentales y sus operaciones. Esto hace que las teorías psicológicas estén eminentemente sustentadas en lenguaje natural, manera legítima de describir fenómenos, pero mucho más ambigua que los modelos formales (e.g., Busemeyer et al., 2015; Farrell y Lewandowsky, 2010; Sun, 2023). Así, la propuesta de los modelos formales permite superar esta ambigüedad describiendo las representaciones mentales en términos formales y proponer qué operaciones las construyen y manejan.

Nos encontramos en el mismo escenario con los aspectos cognitivos del lenguaje. Rompiendo el marco del análisis del lenguaje como cadenas de conductas verbales del conductismo (Skinner, 1957), la Ciencia Cognitiva ha tratado de inferir qué tipo de representaciones se utilizan y qué tipo de operaciones son desplegadas sobre ellas en el momento de comprender o producir lenguaje (e.g., Anderson, 2005). Con carácter general, se trata de proponer el formato de las representaciones y qué tipo de información portan, además de estudiar su pervivencia y preeminencia en distintos formatos (e.g., modal y amodal). Así tenemos el debate entre el uso obligado o no de representaciones modales (De Vega et al., 2012), del uso de indicios emocionales y sus consecuencias (Lindquist, 2021), de la separación o no del sistema sintáctico y semántico (Kaan, 1999), etc. En este tipo de investigaciones, se aíslan propiedades que están presentes en ciertos estímulos, mientras que en otros quedan ausentes. Por ejemplo, se pueden utilizar palabras cuyo referente tiene contenido emocional frente a palabras neutras, palabras con referencia a aspectos manipulativos o sensoriomotores frente a las abstractas, frases cuya sintaxis es legítima pero con semántica confusa frente a la coincidencia de semántica plausible y sintaxis ilegítima, distintos tipos de dependencia sintáctica en términos de cercanía o lejanía, etc. (ver el manual de Belinchón et al., 2009 para entender la sutilidad de tales manipulaciones). Existen distintos paradigmas experimentales en Ciencia Cognitiva para estudiar los aspectos cognitivos del lenguaje. A modo de ejemplo, el paradigma conductual suele estudiar tiempos de reacción o registros de movimientos oculares. Otro paradigma clásico que ha madurado las últimas décadas es el de los Potenciales Relacionados con Eventos (Event Related Potencials, ERPs). Con el mismo control experimental, este paradigma trata de identificar momentos en la línea temporal que desvelen sensibilidades, entendidas como la capacidad de algunas localizaciones corticales de responder de manera diferente a esas manipulaciones. De ahí que se busque resolución temporal más que espacial, aunque las técnicas actuales puedan aunar ambos, como es el caso de la Magnetoencefalografía. Un ejemplo de ERP es el N400, actividad diferencial localizada normalmente en la zona centro-parietal que se asocia al cambio de expectativas al leer frases (Federmeier y Kutas, 1999; Kutas y Hillyard, 1980). Más tarde hablaremos de él.

En resumen, lo importante de estos paradigmas radica en su capacidad para captar los procesos cognitivos subyacentes a partir de las sensibilidades detectadas y de las representaciones mentales implicadas, así como de la información que estas contienen. En este artículo, presentamos y discutimos el aparataje que nos permite estudiar esas representaciones y sus propiedades a partir del modelado formal con ciertas arquitecturas de redes neuronales artificiales (RNAs). El objetivo de este texto no es la exhaustividad, sino fomentar el interés y la reflexión en torno a este enfoque de modelización formal.

Las Redes Neuronales Artificiales (RNAs) en Ciencia Cognitiva

Una posibilidad para estudiar las representaciones internas es su simulación con RNAs. Una RNA puede representarse mediante una serie de nodos (neuronas artificiales) y conexiones (con pesos asociados) que reciben señales de otros nodos, las procesan y envían una respuesta a través de una función de activación concreta. Los grupos de nodos suelen agruparse en capas y la señal viaja desde la primera capa (capa de entrada) hasta la última capa (capa de salida). Esta última capa es la expresión externa de la resolución de una tarea. Será en la capa de salida donde se emitirá la predicción que hace la red a partir de una entrada concreta en función de sus pesos, funciones de activación y estructura. Este paradigma tiene una gran tradición desde los primeros estudios del conexionismo (e.g., McClelland y Rumelhart, 1989; McClelland et al., 1987; Rumelhart et al., 1986). Sin ánimo de ser exhaustivos, podemos describir su aprovechamiento en el ámbito cognitivo en varios ejes de manipulación:

Las características de la propia red. Esto incluye el tipo de arquitectura de la red (el tipo de red y su funcionamiento estructural), su topología (el número de capas y nodos, además de posibles ensamblajes entre redes) y su parametrización (funciones, coeficientes de aprendizaje, optimizadores, ratios de dilución, etc.). La manipulación de estas características puede sustentar hipótesis sobre los mecanismos implicados en el procesamiento del lenguaje.

Las muestras con las que aprende la red. Es decir, la información con la que la red es entrenada, lo que incluiría el corpus textual (textos, frases, pares de frases, etc.) y sus características. También el tamaño es una cuestión clave, ya que puede generar un tipo de aprendizaje más estadístico o más emergente.

Las entradas puestas bajo escrutinio una vez entrenada la red. Consiste en aislar propiedades presentes y ausentes en ciertas entradas y comprobar el comportamiento de la red y el tipo de errores que comete. Sería una suerte de Psicología comparada persona-máquina.

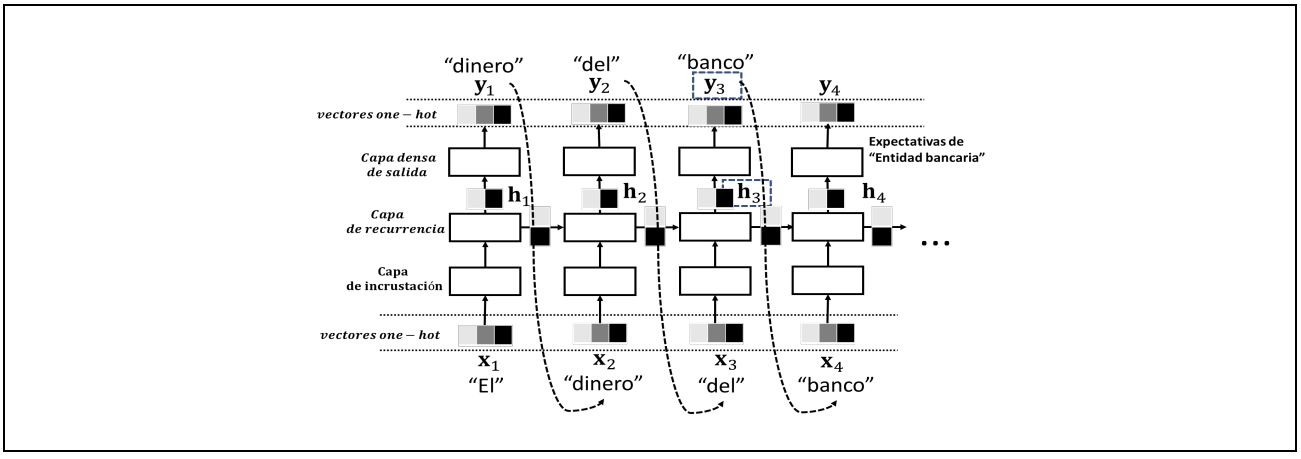
Son varias las topologías que se han propuesto para el estudio del lenguaje y aquí queremos destacar las redes neuronales recurrentes (RNN; e.g., Elman, 1990; Jordan, 1997) con mecanismos LSTM (RNN-LSTM; Hochreiter y Schmidhuber, 1997), así como los Transformers (Vaswani et al., 2017). En este artículo, ilustraremos distintas estrategias, observables y ocultas, de cómo se puede estudiar el lenguaje desde un punto de vista psicológico con una RNN.

La Figura 1 presenta un esquema del funcionamiento de una RNN. Cada palabra de una unidad de texto (normalmente, frases) será la entrada en cada marca de tiempo t (siendo t el orden de la palabra en la frase) como x_t. La tarea más comúnmente utilizada para entrenar este tipo de red es la predicción de la siguiente palabra en la secuencia (tarea autoregresiva). Introducida una palabra, la red tiene que predecir la siguiente palabra en la frase. Operativamente, la entrada x_t es un vector *one-hot* (una representación binaria que codifica las palabras del vocabulario del modelo), el cual es transformado en la capa de incrustación en un vector que captura propiedades lingüísticas relevantes (como aspectos semánticos y sintácticos). En otras palabras, esta capa convierte las entradas binarias en representaciones vectoriales densas dentro de un espacio vectorial. Posteriormente, esta representación densa se procesa en la capa de recurrencia, generando un estado oculto (h_t) en su salida. Así, la capa de recurrencia utiliza conexiones recurrentes que permiten que la señal generada en un instante sea enviada a la misma capa en el tiempo siguiente, funcionando, así como un mecanismo de contexto temporal. Como se ve en la Figura 1, esta salida está en función de la entrada actual (x_t) y el estado oculto anterior (h_{t-1}), que codifica la información de la frase ya procesada. De esta forma, h_t es tanto la salida de la capa de recurrencia, como la entrada de la misma capa en el momento inmediatamente posterior. Cada estado oculto h_t se utiliza finalmente como entrada para la capa de salida, que produce un vector y_t a partir de una función softmax (Jurafsky y Martin, 2023). Este vector representa la predicción del modelo en el punto temporal correspondiente para la siguiente palabra en la secuencia. La capa de salida está compuesta por tantos nodos como palabras tenga el vocabulario del modelo, y su señal puede ser expresada en forma de vector de probabilidad, indicando cada componente la probabilidad de una palabra concreta. La palabra con más probabilidad será seleccionada como la continuación más probable de la secuencia. Volviendo a la Figura 1, si estamos en el momento t = 3, la entrada es «del». La capa de recurrencia recibe información de la entrada «del» y del estado oculto anterior h_2. Ese estado oculto anterior lleva a su vez la información de la situación de la frase en ese punto ya que existe un contexto marcado por lo ya dicho (en este caso, «El dinero»). Por tanto, h_3 está en función de h_2 (información contextualizada por «El dinero») y la entrada actual «del» gracias a la capa de recurrencia. Es previsible que la secuencia sea continuada por un sustantivo masculino en un contexto de dinero. En este ejemplo, la continuación sería «banco» y la salida y_3 tendrá su mayor valor en el componente que corresponde a esta palabra. Es decir, el nodo de salida correspondiente a «banco» daría la señal (probabilidad) más grande.

Quizá lo más interesante del estado oculto h_t es su naturaleza recurrente, que le permite actuar como una memoria dinámica que se actualiza en cada marca temporal (t). Este estado no solo acumula información sobre la frase procesada hasta el momento, sino que también integra el contexto de las palabras anteriores para formar una representación contextualizada de la secuencia. Desde una perspectiva cognitiva, h_t puede interpretarse como una representación de las expectativas del modelo, es decir, una proyección de lo posible: qué palabras son más probables de aparecer a continuación según el patrón lingüístico observado.

Figura 1

Representación desenrollada en el tiempo de una RNN clásica donde se muestran tanto los estados ocultos h_t generados en cada momento como las salidas y_t de la red en forma de probabilidades asignadas a palabras. Por ejemplo, y_3 es la salida y h_3 es el estado oculto construido a partir de «El dinero del». h_3 participará en construir h_4 para predecir y_4.



Modelando la ruptura de expectativas en la frase

Cuando ya tenemos una RNN entrenada como la de la Figura 1, podemos emplear varias estrategias para comprobar las consecuencias de procesar frases manipuladas de distintas formas. Estas consecuencias se evalúan en las llamadas muestras de evaluación. La primera estrategia podemos llamarla «observable» y consiste en comparar las salidas de la red con patrones de comportamiento sujetos a hipótesis previas. Si la red predice palabras, como la del ejemplo de la Figura 1, la comparación será frente a palabras previamente establecidas y se analizará si el modelo es capaz de predecir las palabras adecuadas. Aquí entraría, por ejemplo, la constatación de que una red (manipulada convenientemente) produce lenguaje o comete errores similares a los de las personas.

Otra de las formas de análisis dentro de esta estrategia es una evaluación sobre la previsibilidad del lenguaje que produce una persona tomando como referencia el modelo. En este caso se calcula la «sorpresividad» del lenguaje[1] (Oh y Schuler, 2023) como medida de su predictibilidad. Por ejemplo, la producción verbal de una persona con afasia no fluyente divergirá del modelo normativo representado en una red (Cong et al., 2024). Dentro de una frase, cada palabra podrá ser sorpresiva si no es esperada normativamente. Concatenar tres determinantes seguidos no es normativo, y la divergencia con el modelo normativo representado en la red es muestra de ello. Es sorpresivo que después de «El dinero del» la persona haya producido un determinante como «este», ya que su probabilidad estimada será baja y su sorpresividad notable. La forma de calcular la sorpresividad es a partir de la salida de la red (y_t) representada en la Figura 1. La otra estrategia podríamos llamarla «oculta», y consistiría en aprovechar los estados ocultos que la red va generando para diseñar índices que consignen la situación interna (no observable) de la frase. Estos estados pueden homologarse con las representaciones mentales construidas dadas unas constricciones o, por lo menos, poner tal homologación bajo exploración. Como hemos visto, un estado oculto h_t no es más que la salida de los nodos de una capa de recurrencia (también

[1] En Psicología, la sorpresividad (o sorpresa) del estímulo es la magnitud en la que un estímulo es inesperado o incongruente con respecto a las expectativas previas del individuo. Es un concepto clásico de modelos como el de Rescorla-Wagner (Rescorla y Wagner, 1972).

llamada oculta). Es decir, es una representación interna en forma de valores numéricos. Si hay una sucesión de entradas en una línea temporal (una sucesión de palabras de una frase), se generarán tantos estados ocultos como palabras. Cada estado representa el estado de la situación de la red en el momento de introducir una nueva palabra. Como se intuye, esto es crucial en las redes neuronales que sirven para procesar lenguaje, ya que esos estados internos muestran las expectativas de la frase en cada momento.

Cabe destacar que el lenguaje tiene una clara secuenciación temporal. Las frases siguen una línea temporal que, aunque no sea monótona (ya que pueden anticiparse palabras posteriores), sí marca su orden de procesamiento tanto en comprensión como en producción (e.g., Rayner, 2012 contiene múltiples ejemplos del procesamiento secuencial del lenguaje con metodología de movimientos oculares). Topologías como las RNN-LSTM o los *Transformers* son útiles para modelar esa línea temporal implícita en las frases. Aunque estas topologías sean sensiblemente distintas, ambas van generando estados ocultos, los cuales participan de una u otra forma en generar la salida de la red. En el caso de las RNNs, los estados ocultos están en función de la palabra (entrada, x_t) de la marca de tiempo t de la frase y del estado oculto generado en el momento anterior (h_{t-1}). Así es cómo las RNNs hacen que la representación de la situación de la frase sea sensible a dos circunstancias: (a) la palabra procesada en el momento actual, y (b) un contexto que recoge la parte de la frase procesada anteriormente (implícito en el estado oculto anterior). En última instancia, la salida de la red y_t, que indicará las palabras más probables a suceder la secuencia, se instanciará como resultado de introducir estado oculto actual (h_t) como entrada en la capa de salida. Todo este proceso se llama de recurrencia: siempre hay un estado oculto anterior (h_{t-1}) que, junto con la entrada actual x_t, genera un nuevo estado oculto h_t, y éste una salida y_t. En el caso de los *Transformers*, esta recurrencia se omite y se hace uso de los llamados mecanismos de autoatención (Vaswani et al., 2017). No obstante, al igual que en las RNN, también se genera un estado oculto actual.

El hecho de que podamos tener acceso a los estados ocultos de la red es muy ventajoso, ya que podemos analizarlos para estudiar la representación que se tiene de la frase en cada momento. Es una representación interna de la situación de la frase y contiene información implícita tanto de las posibles relaciones gramaticales como de la semántica expresada (o incluso de indicios sensoriomotores o emocionales, si se ha construido el modelo con ellos). Pongamos como ejemplo la frase «Fui a un banco cercano a retirar dinero». Se intuye fácilmente que, si hemos leído solo «Fui a un banco cercano a…», las expectativas generadas son aún ambiguas. Esto significa que la situación de la frase es incierta, ya que la situación puede referirse a sentarse o sacar dinero. En cualquier caso, lo esperable es continuar la frase con un verbo. El estado oculto generado por el modelo en ese momento nos podrá informar de tal fenómeno y los cambios de ese estado en lo sucesivo nos pueden dar una métrica de certidumbre. Imaginemos que manejamos la frase: «Fui a un banco cercano a rastrojos». En este nuevo caso podemos cotejar el cambio del estado oculto entre el punto en que se ha leído «Fui a un banco cercano a…" (h_{t-1}) y el punto en que se completa la frase "Fui a un banco cercano a rastrojos» (h_t). El cambio entre h_t y h_{t-1} puede ser notable, y de ello extraerse una métrica de cambio o de ruptura de expectativas.

Tanto el cálculo de la sorpresividad (estrategia basada en la salida y_t; Oh y Schuler, 2023) como del cambio en el estado oculto (h_t) pueden considerarse como indicadores de cambio de expectativas tanto semánticas como gramaticales. Si la sorpresividad de una palabra producida por una persona es notable, se rompen las expectativas, al igual que se rompen si la representación de la situación de la frase inferida por el estado oculto cambia de un momento a otro. Este fenómeno es precisamente lo que buscan capturar ciertos potenciales relacionados con eventos (ERPs). Así, se establece una posible conexión entre algunos ERPs y los índices derivados de estas estrategias, como se ha demostrado en investigaciones previas (e.g., Rabovsky y McClelland, 2020; Rabovsky et al., 2018).

No obstante, aun pudiendo representar ambos índices una ruptura de expectativas, se han observado diferencias entre la sorpresividad y el cambio de estado oculto. En primer lugar, se ha observado que el cambio de estado oculto está más correlacionado con el ERP N400 (Rabovsky y McClelland, 2020; Rabovsky et al., 2018), planteándose como hipótesis principal que ambos están más relacionados con la semántica. Es un hallazgo interesante para razonar sobre el impacto de la plausibilidad temática de las

frases y su confrontación con el más fino procesamiento sintáctico. Hipotéticamente, tanto el N400 como el cambio de estado oculto serían mucho más sensibles a la frase «Fui a un banco cercano a rastrojos» por su semántica inesperada, que a «Fui un dinero retirar en banco a cercano» por encontrarse los distintos elementos semánticamente relacionados. De igual manera, frases empleadas en pruebas de dislexia fonológica como «El perro es perseguido por el gato», podrían generar un N400 menor y poco cambio en los estados ocultos de la red. En ambos ejemplos se puede crear una ilusión semántica donde se interpreta la frase por plausibilidad (Rabovsky y McClelland, 2020). Sin embargo, se ha observado que la sorpresividad correlaciona en mayor manera con el P600 (Rabovsky y McClelland, 2020). La sorpresividad, al igual que el P600, es sensible a las expectativas sintácticas, aunque también en cierta medida a las semánticas (ver el trabajo de Slaats, y Martin, 2023 para una reflexión sobre este amalgamiento de fuentes de variabilidad de la sorpresividad). De esta manera, es sensible a la agramaticalidad y es en cierto modo una alarma de que algo no cuadra en la frase.

En este texto ilustramos conceptualmente la forma de conseguir los estados ocultos que se generan en los distintos momentos de las frases, y la forma de medir la sorpresividad de las palabras de las frases. Así, mostraremos la utilidad tanto de los estados ocultos y sus cambios como de la sorpresividad, en un fenómeno que inunda todos los debates de Psicología del lenguaje: la composicionalidad sistemática (e.g., Liñán, 2009; Szabó, 2001, 2020). Esta es una propiedad de los sistemas cognitivos que explica cómo somos capaces de derivar el significado de expresiones complejas a partir del significado de sus partes y las reglas que las combinan. Este concepto nos dará la oportunidad de ilustrar cómo se utilizaría un modelo de RNNs sustentado en un diseño experimental con ERPs.

Cómo calcular índices en las estrategias observables y oculta

Vamos a analizar de manera pormenorizada una RNN con una topología meramente autogenerativa[2], asumiendo que el modelo ya ha sido entrenado. Esta topología se muestra en la Figura 1 (para más detalles se puede consultar Jorge-Botana, 2024). Hemos aludido antes a dos estrategias posibles para estudiar las expectativas lingüísticas del modelo: una observable y otra oculta.

La primera, cuyo índice es la sorpresividad (Oh y Schuler, 2023), consiste en operar en la capa de salida y sus predicciones, es decir, en la parte observable (y_t). Esto significa que se utilizan los vectores de salida generados por las distintas entradas de una secuencia. Si tenemos un conjunto de palabras dispuestas como secuencia $\{x_1, x_2, x_3, \ldots, x_t\}$, podemos tomar el vector y_t de salida como la predicción contextualizada de una entrada en la marca de tiempo t, cuyo vector es x_t. El vector y_t expresará la probabilidad que tiene cada palabra del vocabulario de ser salida en ese momento de la frase. De esa manera, y_t será un vector que contiene las probabilidades de cada una de las palabras del vocabulario de suceder a la secuencia que ya se ha procesado. Si se procesa una frase como "El dinero del banco piedra", se puede estimar en el vector y_4 que la probabilidad de «piedra» es muy baja, acaso por la temática o porque un sustantivo no es previsible detrás de otro. Esto es posible porque y_4 tiene tantos componentes como palabras tenga el vocabulario, siendo un listado de probabilidades para todas las palabras que conoce el modelo. El primer componente del vector y_4 representa la probabilidad de que la palabra indizada como primera en el vocabulario suceda a la frase «El dinero del banco». El segundo componente del vector y_4 representa la probabilidad de que la palabra indizada como segunda en el vocabulario suceda a la frase. Y así con todos los componentes del vector. Así, si «piedra» ocupase la posición 5 en el vocabulario, el componente 5 de ese vector y_4 sería la probabilidad de que ocurra piedra detrás de «El dinero del banco»:

[2] Se usa únicamente un decodificador. En las topologías llamadas codificador-decodificador, el codificador proporciona el contexto para que el decodificador empiece a autogenerar lenguaje de manera probabilística. Si el contexto es fuerte, la autogeneración estará instigada por él. De ahí que lo que emita el decodificador no sea arbitrario sino apegado a un tema (una pregunta, una imagen, una frase, etc.). No obstante, en alguna topología solo se requiere de decodificador (por ejemplo, cuando se le da un pie para que el decodificador simplemente lo autocomplete).

$$y_5 = P(\text{w}_{\text{piedra}} | \text{w}_{\text{el}}\ \text{w}_{\text{dinero}}\ \text{w}_{\text{del}}\ \text{w}_{\text{banco}}) \quad [1]$$

Aplicando el logaritmo en negativo a lo obtenido en y_5, obtenemos la medida de sorpresividad. Así pues, el cálculo de la sorpresividad (S) de una palabra en una posición t concreta (w_t) dadas una serie de n palabras previas quedaría definido como:

$$S(w_t) = -\log P(\text{w}_t | w_{t-1}\ w_{t-2},\ldots,w_{t-n}) \quad [2]$$

En nuestro ejemplo, la sorpresividad de la palabra "piedra" sería:

$$S(w_{piedra}) = -\log (y_5) = -\log \text{P}(w_{piedra} | w_{el}\ \ w_{di\text{-}nero}\ w_{del}\ w_{banco}) \quad [3]$$

Se puede calcular la sorpresividad de cualquier palabra que forme parte del vocabulario en cada momento t. Salta a la vista que esta estrategia es llamada observable por ser la materialización observable en la capa de salida del estado oculto h_4 generado en ese mimo momento. Así, y_4 sería la predicción a partir de la entrada x_4, pero contextualizada con el estado oculto h_3 (ver Figura 1). Tomando ese vector de salida y_4, podemos obtener las probabilidades asociadas a que cada palabra del vocabulario suceda después y con esto calcular su sorpresividad (en el Apéndice ofrecemos algunos enlaces a códigos que calculan la sorpresividad empleando modelos preentrenados). Esta estrategia asume que los nodos de la capa de salida tienen como tarea predecir palabras con la función *softmax* (Bridle, 1989; cuyo origen puede rastrearse hasta Boltzmann, 1868) como función de activación, dado que se trabaja con vectores one-hot para indexar las palabras del vocabulario.

No obstante, una cosa es dar una probabilidad a cada palabra y otra contar con la representación oculta de la situación de la frase en ese momento. Con la predicción (y_t) podemos obtener una distribución en la que algunas palabras tendrán más probabilidad de continuar la secuencia. Con la representación oculta (estado oculto h_t) tenemos una forma de representar la situación de la frase en el marco de un contexto. Esto permite, por ejemplo, representar dinámicamente la situación de la frase antes y después de introducir «banco» en el contexto previo de «No me queda mucho dinero en el…» o en el contexto de «Como estoy cansado me siento en el…».

Se suele formalizar la situación de la frase contextualizada en un momento concreto tomando el valor del estado oculto h_t que se genera en él. De esta forma, se toma h_t como la representación interna de la situación de la frase en el tiempo t. Si quisiéramos consignar el cambio en tal situación oculta antes y después de introducir la palabra «banco» (como en la Figura 1), podríamos calcular una simple distancia vectorial entre los estados ocultos antes y después de tal hecho:

$$Cambio\ («banco») = h_3 - h_4 \quad [4]$$

Este índice, llamado cambio de estado oculto (Rabovsky y McClelland, 2020), puede calcularse en todos los momentos de la frase como la diferencia entre el estado oculto antes y después de la palabra: h_t- h_{t-1}. Puede decirse que h_4 representa una expectativa a partir de x_4 y su contexto anterior. Así, de una entrada x_4, con un formato incierto, se consigue en h_4 una representación con expectativas de entidad bancaria. Piantadosi (2023) define esos estados internos como aspectos latentes de la sintaxis y la semántica que gobiernan la interpretación del texto (en el Apéndice ofrecemos algunos enlaces a códigos que ayudan a conseguir los estados ocultos en los diferentes momentos de la frase y de las diferentes capas ocultas).

En términos de plausibilidad cognitiva, es sugerente reflexionar sobre el concepto de expectativa en referencia a los estados ocultos. Cuando en una secuencia se produce una nueva entrada x_t y ésta interactúa con el estado oculto anterior h_{t-1}, el nuevo estado oculto h_t puede entenderse como las expectativas sobre qué palabras podrían acompañar a la entrada x_t en la siguiente marca temporal dado ese contexto. Este estado es pues una constelación de posibilidades y, así, cada momento conlleva una constelación de palabras diferentes. Esas expectativas son una potencia que se materializa en forma de palabras en la capa de salida mediante la distribución de probabilidad y_t.

Composicionalidad sistemática como núcleo del debate

El conocimiento humano no parece estar exclusivamente basado en la experiencia (o, al menos, en un mero cómputo probabilístico sobre ella). Parece haber habilidades *emergentes* que generalizan el uso de reglas sobre estructuras nunca vistas. Esto se relaciona con el conocido fenómeno de la «pobreza del estímulo» (e.g., Pearl, 2022). Este fenómeno presenta la paradoja de que, aunque las personas están expuestas a un conjunto relativamente pequeño de oraciones, sus expresiones suelen ser correctas formalmente y pueden llegar a inferir el significado de frases con palabras legítimamente combinadas, aunque nunca vistas en concurrencia. Esto significa que las personas pueden comprender y producir combinaciones lingüísticas que no han visto. Se postula que tal capacidad viene dada por la denominada composicionalidad sistemática del lenguaje (e.g., Liñán, 2009; Szabó, 2001, 2020).

Formalmente, en el concepto de composicionalidad sistemática se mantiene que la entrada no proporciona evidencia sobre todas las oraciones posibles, y que tampoco contiene reglas explícitas sobre las posibles combinatorias ni sus significados (Lasnik y Lidz, 2016). Sin embargo, las personas se desenvuelven relativamente bien en ambos casos, aunque no ven todo el lenguaje (y sus variedades) a lo largo de su vida, ni se les dice explícitamente cómo combinarlo.

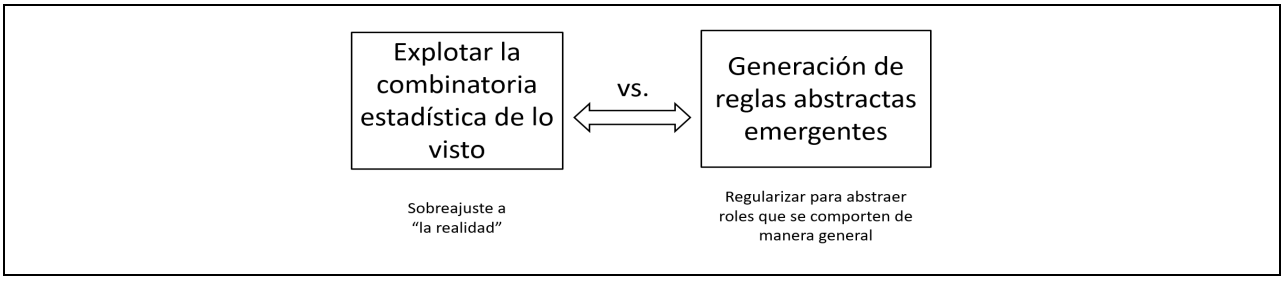
Tradicionalmente, se ha sugerido que la composicionalidad sistemática está relacionada con la sintaxis. Visto de esa forma, existirían dos sistemas, uno con representaciones semánticas de las palabras y otro que aplicaría reglas universales. Este último sistema está separado del significado de las palabras individuales (trabajo seminal de Chomsky, 1957). Es de resaltar también un componente modular introducido por Fodor y Pylyshyn (1988) a esta concepción composicional: las unidades del sistema semántica son módulos y las palabras participarían con la misma carga semántica independientemente de su rol sintáctico. En las frases «Julieta quiere a Romeo» y «Romeo quiere a Julieta», Julieta y Romeo participarían con la misma carga semántica en ambas frases. No obstante, parece complicado asumir estos supuestos (Rabovsky y McClelland, 2020; Rabovsky et al., 2018). Asumir ese componente modular implicaría que las representaciones de las palabras se recuperan antes de asignarles roles sintácticos, sin mediar ninguna contextualización más que la que posteriormente imponga la sintaxis. Sin embargo, la Julieta de la primera frase es una chica amante, y la de la segunda es amada (incluso pudiera ser que participen distintas Julietas en ambas frases), cada una con su constelación de posibilidades. Consecuentemente, algunos autores rebajan esa constricción de modularidad y hablan más bien de cuasicomposicionalidad (Rabovsky y McClelland, 2020). Esta nueva composicionalidad no asumiría que las palabras contribuyen de manera independiente a su rol sin-táctico.

El fenómeno de la composicionalidad puede evaluarse a través de índices como el cambio en los estados ocultos y la sorpresividad mencionados antes. Consideremos secuencias de palabras a las que hemos sido expuestos como, por ejemplo, «Julieta ama a Romeo». Hemos visto a chicas llamadas Julieta amar a chicos llamados Romeo y, por tanto, forma parte de lo plausible. Sin embargo, podemos abstraernos de la realidad e interpretar frases como “Ju-lieta ama el morado” o más implausibles como «Julieta ama la garrapata». La idea es jugar con la realidad (i.e., los estímulos que participan en el estudio) para confrontar su efecto en las personas y en los modelos. Tómese como ejemplo la frase que Chomsky (1957) introdujo como reto: «Las ideas verdes incoloras duermen furiosamente» (*Co-lorless green ideas sleep furiously*). Aunque nunca hemos sido expuestos a esa escena, podemos producir la frase y evocar su significado, incluso en contra de su plausibili-dad. De manera similar, también entenderemos la frase «Julieta ama la garrapata». Lo importante es que estas fra-ses pueden ser producidas y su significado se puede infe-rir, aunque la plausibilidad del lenguaje favorecerá las expectativas en algunas ocasiones y las penalizará en otras. Si un sistema no fuera capaz de abstraerse de la experiencia previa a la que ha sido expuesto y entender frases que no haya visto antes (e.g., «Julieta ama la garrapata»), estaría sobreajustado a la realidad.

Algunos antropólogos ya propusieron la existencia de algunos sistemas lingüísticos primitivos o protolenguas en el que no existía dicha generalización. Estos fueron llama-dos lenguajes libres de sintaxis, donde la capacidad sintác-

Figura 2

La composicionalidad sistemática del lenguaje tiene como resultado la generación de reglas abstractas emergentes más que simplemente explotar la combinatoria estadística de la información procesada. Podemos pues entender esto como resultado de un proceso de regularización durante el proceso de aprendizaje



tica se amalgama con la semántica de las palabras (Bickerton, 1995; Jackendoff, 1987, 1999). Aludimos a este caso extremo para hacer gráfico un caso en el que la composicionalidad sistemática es difícil o imposible de desplegar. En el otro extremo, podemos tener un sistema capaz de abstraer reglas aisladas de su experiencia con la realidad explotando los indicios sintácticos para dilucidar cuando un sustantivo es activo o es pasivo. En términos evolutivos, algunos autores afirman que el proceso de gramaticalización (es decir, la atribución de un papel gramatical o sintáctico a una palabra) es análogo al de la metaforización (Heine y Kuteva, 2007). Cuando dos palabras comparten algunas propiedades funcionales, el sistema cognitivo evoluciona y se apercibe de que ambas comparten un papel común y, por tanto, son intercambiables en determinados roles genéricos. Así, el sistema aprende que dos palabras comparten un papel común a través de esas propiedades funcionales compartidas a lo largo de su experiencia. Esto nos lleva también a pensar en términos piagetianos, ya que las palabras podrían ser asimiladas a una categoría abstracta como un rol sintáctico. Podríamos extender estos razonamientos también a otros ejemplos como el caso de los verbos inventados o las pseudopalabras donde, aunque no conozcamos las palabras, somos capaces de interpretarlas como sustantivos o verbos dependiendo de cómo se comportan sintácticamente.

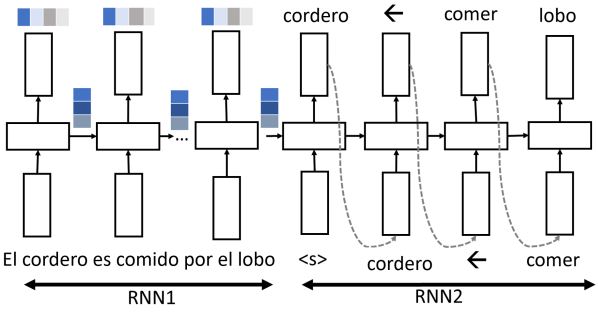
La semántica sería, según lo argumentado, una cosmología de expectativas del uso de las palabras, y estas expectativas pueden ser semánticas o incluso sintácticas en cuanto a continuidad de la frase. Este tipo de expectativas estarán constreñidas por las posibilidades de usos reales del lenguaje. Sin embargo, la composicionalidad es una manera de regularizar el conocimiento del sistema para que no se sobreajuste a la realidad (Figura 2). Tanto el término de sobreajuste como el de regularización tienen el sentido que se les da en el ámbito del aprendizaje automático. Regularizar es impedir que el modelo se sobreajuste a características irrelevantes de la muestra de entrenamiento y deje holgura para usar de manera implícita indicios mucho más genéricos como, por ejemplo, los roles sintácticos. Puede verse de manera clara esta lógica en trabajos que imponen a la función de coste usada para la actualización de los pesos una regularización basada en maximizar la independencia de los constituyentes de las frases (Nandi et al., 2024). Esto último puede considerarse una forma de instigar que los estados ocultos creados por las sucesivas palabras de una frase no se amalgamen en base a la combinatoria vista en la realidad, y el coste de la desambiguación haga que no abstraiga propiedades genéricas de la estructura composicional (Choi et al., 2017).

Un experimento tentativo sobre los efectos de la plausibilidad

En este apartado vamos a plantear un experimento mental donde se pretenden homologar los índices de ruptura de expectativas lingüísticas (plausibilidad) de los modelos de RNNs con las sensibilidades extraídas de algunos ERPs. El objetivo es mostrar uno de los posibles paradigmas que podrían emplearse para evaluar el comporta-

Figura 3

Entrenamiento del codificador-decodificador en la reconstrucción de una escena tanto en sus participantes como en sus roles pasivo-activo. La RNN1 (el codificador) proporciona a la RNN2 (el decodificador) la codificación de la frase en su último estado oculto, y la RNN2 autogenera la escena. Los pesos tanto de la RNN1 como de la RNN2 se modifican en función del éxito de la autogeneración de la escena. La RNN2 produce <s> como inicio de la escena y asigna el rol pasivo-activo a través de la dirección de una flecha (en este caso, ← indica que el verbo comer se produce de manera pasiva y que el cordero es comido por el lobo)



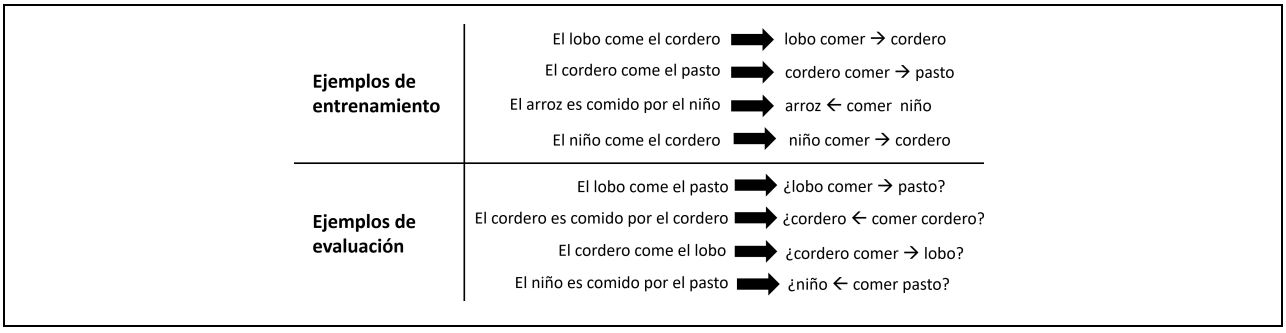
miento composicional y su correlación con ERPs (Rabovsky y McClelland, 2020 o Rabovsky et al., 2018 presentan algunos diseños completos, aunque con un modelo idiosincrático llamado *Sentence Gestalt* que no se corresponde con los modelos más empleados en el campo de Inteligencia Artificial Generativa). Este experimento mental también nos permitirá mostrar una forma tentativa de comprobar si el cambio de estado oculto en una RNN está relacionado con el cambio abrupto de la situación de la frase en términos eminentemente temáticos más que sintácticos. Se espera, según lo dicho previamente, que el cambio de estado oculto esté más alineado con el potencial N400 que con el P600.

Para ello, vamos a plantear una topología de RNN Secuencia-Secuencia con codificador y decodificador (consultar Jorge-Botana, 2024 para más detalles) como la que se presenta en la Figura 3. Durante la fase de entrenamiento, se codifica una frase en lenguaje natural en el codificador y se autogenera la reconstrucción de su escena en el decodificador, teniendo en cuenta tanto el aprendizaje de las palabras (sustantivos y verbos) como la asignación de roles pacientes y agentes. El codificador simula la transformación de la información de una frase en una representación interna, tal y como haría una persona que lee una frase e imagina una escena. Otros autores han sugerido aproximaciones similares con la intención de dotar a los modelos del lenguaje de conocimiento del mundo mediante diversas vías (véase Carta et al., 2023; Hernández et al., 2023; Ivanova et al., 2024).

La especificación de las características de la red es el primer eje de manipulación sobre el que podemos trabajar. El segundo eje de manipulación son las muestras con las que aprende la red. En este caso, se utilizaría un conjunto de pares frase-escena. Lo interesante de este corpus es que se trata de una muestra muy controlada en la que se limita el tamaño evitándose así el efecto de escala de los Grandes Modelos de Lenguaje. Pero lo más importante es que, en este conjunto, habrá sustantivos que actúan como agentes y pacientes, otros sólo como agentes, y el resto sólo como pacientes, tal y como podríamos encontrar en el mundo real (ver los ejemplos de entrenamiento en la Figura 4). Una frase plausible será, por ejemplo, «El lobo come el cordero», pues se ha presentado durante el entrenamiento y el modelo ha aprendido a generar esa escena. Así, el modelo estará sobreajustado a estas frases, pudiendo controlar la plausibilidad de las frases que, una vez entrenado, se evaluarán en el modelo. Las entradas puestas bajo escrutinio una vez entrenada la red son el tercer eje de manipulación. Dentro del conjunto de evaluación, existirán también frases no vistas cuya escena nunca se ha representado

Figura 4

Entrenamiento y evaluación del modelo con pares frases-escena. El entrenamiento se hace con frases plausibles con escenas sistemáticamente correctas. La evaluación se hace con frases plausibles y no plausibles para analizar la reconstrucción de las escenas por parte del modelo

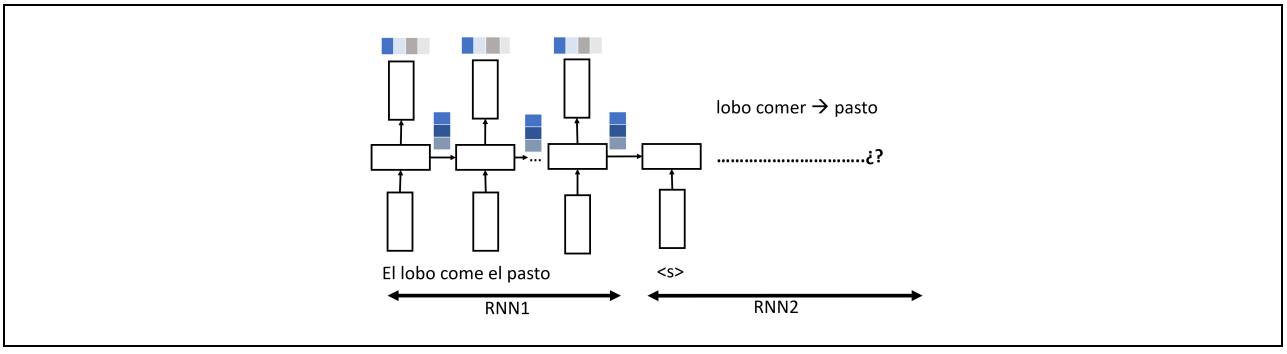


nuestro modelo (ver los ejemplos de evaluación con frases implausibles en la Figura 4) como, por ejemplo, «El cordero come el lobo» o «El niño es comido por el pasto». Por tanto, habrá distintos niveles de plausibilidad en los que solo algunos casos se podrán resolver por composicionalidad. También se podría manipular la legitimidad sintáctica de las frases de prueba cambiando el orden de algunas palabras e o intercambiando los roles pasivos y activos de los sustantivos para estudiar los roles sintácticos.

La clave de todo esto, además de ver si el modelo es capaz de reconstruir escenas de frases no vistas con distinto grado de plausibilidad (e.g., «El lobo come el pasto» o «El niño es comido por el pasto»), es medir el cambio de estado oculto en el momento de introducir en el codificador las palabras que rompen las expectativas tanto por plausibilidad como por legitimidad sintáctica (Figura 5). Con este paradigma podemos comprobar cómo la medida de cambio del estado oculto explica las representaciones del lenguaje de una manera controlada a través de esta topología de RNN. Resultados previos sugieren que el N400

Figura 5

Evaluación del codificador-decodificador a través de frases plausibles e implausibles. Los estados ocultos y sus diferencias permiten consignar el cambio como ruptura de las expectativas lingüísticas ante una frase nunca vista por el modelo. Los cambios en los estados ocultos pueden ser sensibles a la ruptura de expectativas y, consecuentemente, se pueden calcular índices que cuantifiquen dicha ruptura de expectativas como: $h_{(El\ lobo\ come\ el\ pasto)} - h_{(El\ lobo\ come\ el)}$. *En cada marca de tiempo, al introducir la siguiente palabra en el codificador, se genera un nuevo estado oculto que podría generar un cambio en las expectativas de la frase*



está más alineado con medidas como el cambio de estado oculto y que ambos son eminentemente temáticos y situacionales (Rabovsky y McClelland, 2020; Rabovsky et al., 2018). De esta manera, estos modelos nos permiten hacer predicciones sobre cómo funciona la composicionalidad y qué mecanismos hay implicados, pudiendo simular respuestas en las RNNs que son parecidas a los ERPs que se consignan en el cerebro. Además, los razonamientos sobre la plausibilidad de los modelos de redes neuronales pueden ayudar a mejorarlos y, con ayuda de las estrategias basadas en respuestas ocultas y observables, desplegar en ellos mecanismos basados en las mismas sensibilidades que se detectan en los experimentos de ERPs. Terra ignota.

Conclusión

Las nuevas arquitecturas de red neuronal (RNN-LSTM y *Transformers*) y sus distintas topologías ponen en nuestras manos herramientas muy poderosas para formalizar las teorías cognitivas del lenguaje que describimos en lenguaje natural. Además, su aparataje permite calcular índices sobre sus salidas y analizar los estados ocultos que se van generando en cada momento de la línea temporal de una frase. Esto permite confrontar modelos y experimentos para corregir tanto las teorías como los modelos, y aplicar esas correcciones a las arquitecturas que hoy día están por debajo de los Grandes Modelos del Lenguaje.

Este texto se focaliza en la arquitectura RNN-LSTM, puesto que sus mecanismos son muy interesantes en términos de plausibilidad cognitiva. El hecho de capturar las dependencias temporales del lenguaje con mecanismos de memoria de trabajo con diferente pervivencia de la información las hace muy interesantes a nivel psicológico (la memoria de trabajo a corto plazo, a largo plazo o los propios mecanismos de olvido y aportación). Puede decirse que las RNN-LSTM son muy intuitivas para entender un posible modelo situacional de las frases. Sin embargo, actualmente, los *Transformers* (Vaswani et al., 2017) han sustituido a las RNNs-LSTM en muchas aplicaciones porque presentan ciertas ventajas: procesamiento paralelo (procesan toda la entrada de una secuencia a la vez gracias a sus mecanismos de autoatención), mayor eficiencia (mayor escalabilidad y eficiencia al no requerir recurrencia), y menor restricción secuencial (ya que pueden acceder al contexto sin tener que recorrer las secuencias en orden). No obstante, aunque de forma diferente, los *Transformers* también generan estados ocultos en cada marca de tiempo a partir de la integración de la información de las palabras de la frase. Es por ello por lo que se pueden generalizar directamente los razonamientos que hemos mostrado en el texto a esta arquitectura. Además, la capa de salida no difiere de la de las RNN-LSTM y, por tanto, los cálculos de cambio en los estados ocultos y el índice de sorpresividad serán comunes.

Referencias

Anderson, J. R. (2005). *Cognitive Psychology and its Implications*. Macmillan.

Belinchón, M., Igoa, J. M. y Rivière, Á. (2009). *Psicología del Lenguaje. Investigación y Teoría* [Psychology of Language: Research and Theory]. Trotta.

Bickerton, D. (1995). *Language and Human Behavior.* University of Washington Press.

Boltzmann, L. (2012). Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten [Studies on the Equilibrium of Living Force Between Moving Material Points]. En F. Hasenöhrl (Ed.), *Wissenschaftliche Abhandlungen* (pp. 49–96). Cambridge University Press.

Bridle, J. (1989). *Training Stochastic Model Recognition Algorithms as Networks Can Lead to Maximum Mutual Information Estimation of Parameters*. In Proceedings of the 2nd International Conference on Neural Information Processing Systems (NIPS-89) (pp. 211–217). Morgan Kaufmann.

Busemeyer, J. R., Wang, Z., Townsend, J. T. y Eidels, A. (2015). *The Oxford Handbook of Computational and Mathematical Psychology.* Oxford University Press.

Carta, T., Romac, C., Wolf, T., Lamprier, S., Sigaud, O. y Oudeyer, P. Y. (2023, July). Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning. En *International Conference on Machine Learning* (pp. 3676–3713). PMLR.

Choi, H., Cho, K. y Bengio, Y. (2017). Context-Dependent Word Representation for Neural Machine Translation. *Computer Speech & Language, 45,* 149–160.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Cong, Y., LaCroix, A. N. y Lee, J. (2024). Clinical Efficacy of Pre-Trained Large Language Models through the Lens of Aphasia. *Scientific Reports, 14*(1), Artículo 15573. https://doi.org/10.1038/s41598-024-66576-y

De Vega, M., Glenberg, A. y Graesser, A. (2012). *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford University Press.

Elman, J.L. (1990). Finding Structure in Time. *Cognitive Science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Farrell, S. y Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology. *Current Directions in Psychological Science, 19*(5), 329–335. https://doi.org/10.1177/0963721410386677

Federmeier, K. D. y Kutas, M. (1999). Right Words and Left Words: Electrophysiological Evidence for Hemispheric Differences in Meaning Processing. *Cognitive Brain Research, 8*(3), 373–392. https://doi.org/10.1016/S0926-6410(99)00036-1

Fodor J. A. y Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition, 28*(1-2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Heine, B. y Kuteva, T. (2007). *The genesis of Grammar: A Reconstruction.* Oxford University Press.

Hernandez, E., Sen Sharma, A., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y. y Bau, D. (2023). Linearity of Relation Decoding in Transformer Language Models. *ArXiv,* https://doi.org/10.48550/arXiv.2308.09124

Hochreiter, S. y Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780.

Hubel, D. H. y Wiesel, T.N. (1968). Receptive Fields and Functional Architecture of Monkey Striate Cortex. *Journal of Physiology, 195,* 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Ivanova, A., Sathe, A., Lipkin, B., Kumar, U., Radkani, S., Clark, T., Kauf, C., Hu, J., Pramod, R., Grand, G., Paulun, V., Ryskina, M., Akyurek, E., Wilcox, E., Rashid, N., Choshen, L., Levy, R., Fedorenko, E., Tenenbaum, J. y Andreas, J. (2024). Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv* https://doi.org/10.48550/arXiv.2405.09605

Jackendoff, R. (1999). Possible Stages in the Evolution of the Language Capacity. *Trends in Cognitive Sciences, 3*(7), 272–279. https://doi.org/10.1016/S1364-6613(99)01333-9

Jackendoff, R. (1987). *Consciousness and the Computacional Mind*. MIT Press.

Jordan, M.I. (1997). Serial Order: A Parallel Distributed Processing Approach. *Advances in Psychology, 121,* 471–495. https://doi.org/10.1016/S0166-4115(97)80111-2

Jorge-Botana, G. (2024). *Redes neuronales recurrentes y Transformers para modelos cognitivos del lenguaje* [Recurrent Neural Networks and Transformers for Cognitive Language Models]. Ediciones Complutense.

Jurafsky, D. y Martin, J. H. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson.

Kaan, E. (1999). Syntax and Semantics? *Trends in Cognitive Sciences, 3*(9), Artículo 322. https://doi.org/10.1016/S1364-6613(99)01376-5

Kutas, M. y Federmeier, K.D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology, 62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M. y Hillyard, S. A. (1980). Reading Senseless Sentences: Brain potentials Reflect Semantic Incongruity. *Science, 207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Lasnik, H. y Lidz, J. (2016). The Argument from the Poverty of the Stimulus. En I. Roberts (Ed.), *The Oxford Handbook of Universal Grammar* (pp.221–248). Oxford Academic.

Lindquist, K. A. (2021). Language and Emotion: Introduction to the Special Issue. *Affective Science, 2*(2), 91–98. https://doi.org/10.1007/s42761-021-00049-7

Liñán, J. L. (2009). Sistematicidad, productividad y composicionalidad: Una aproximación pragmatista [Sistematicity, Productivity, and Compositionality: A Pragmatistic Approach]. *Revista de Filosofía, 34*(1), 51–75.

McClelland, J. L. y Rumelhart, D.E. (1989). *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*. MIT press.

McClelland, J. L., Rumelhart, D. E. y PDP Research Group. (1987). *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models* (Vol. 2). MIT press.

Nandi, A., Manning, C. D. y Murty, S. (2024). Sneaking Syntax into Transformer Language Models with Tree Regularization. arXiv preprint arXiv. https://doi.org/10.48550/arXiv.2411.18885

Neisser, U. (1967). *Cognitive Psychology*. Prentice-Hall.

Oh, B. D. y Schuler, W. (2023). Why Does Surprisal from Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics, 11*, 336–350.

Pearl, L. (2022). Poverty of the Stimulus Without Tears. *Language Learning and Development, 18*(4), 415–454. https://doi.org/10.1080/15475441.2021.1981908

Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. En E. Gibson y M. Poliak (Eds), *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett* (pp.353–414). Language Science Press.

Pitt, D. (2022). Mental Representation. En E. N. Zalta y U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition). https://plato.stanford.edu/archives/fall2022/entries/mental-representation/

Rabovsky, M., Hansen, S. S. y McClelland, J. L. (2018). Modelling the N400 Brain Potential as Change in a Probabilistic Representation of Meaning. *Nature Human Behaviour, 2*(9), 693–705. https://doi.org/10.1038/s41562-018-0406-4

Rabovsky, M. y McClelland, J. L. (2020). Quasi-Compositional Mapping from form to Meaning: A Neural Network-Based Approach to Capturing Neural Responses during human Language Comprehension. *Philosophical Transactions of the Royal Society B, 375*(1791), Artículo 20190313. https://doi.org/10.1098/rstb.2019.0313

Rayner, K. (Ed.). (2012). *Eye Movements in Reading: Perceptual and Language Processes*. Academic Press.

Rescorla, R. A. y Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. En A. H. Black y W. F. Prokasy (Eds.), *Classical Conditioning II* (pp. 64–99). Appleton-Century-Crofts.

Rumelhart, D. E., McClelland, J. L. y PDP Research Group. (1986). *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition*. Foundations. MIT press.

Skinner, B. F. (1957). *Verbal Behavior*. Prentice-Hall.

Slaats, S. y Martin, A. E. (2023). *What's Surprising about Surprisal*. https://osf.io/7pvau/download/

Sterelny, K. (1990). *The Representational Theory of Mind.* Basil Blackwell.

Sun, R. (2023). *The Cambridge Handbook of Computational Cognitive Sciences*. Cambridge University Press.

Szabó, Z. G. (2001). *Problems in Compositionality*. Garland.

Szabó, Z. G. (2020). Compositionality. En E. N. Zalta y U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab. http://seop.illc.uva.nl/entries/compositionality/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. y Polosukhin, I. (2017). Attention is all you need. En U. von Luxburg, I. Guyon y S. Bengio (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Curran.

Apéndice

Códigos que calculan la sorpresividad empleando modelos preentrenados:
https://github.com/tmalsburg/llm_surprisal
https://github.com/aalok-sathe/surprisal
https://github.com/simonepri/lm-scorer
https://github.com/TomSgrizzi/surprisal-with-psychformers
https://github.com/samer-noureddine/GPT-2-for-Psycholinguistic-Applications/blob/master/get_probabilities.py

Códigos para extraer estados ocultos:
https://stackoverflow.com/questions/48302810/whats-the-difference-between-hidden-and-output-in-pytorch-lstm
https://www.geeksforgeeks.org/difference-between-hidden-and-output-in-pytorch-lstm/