



PEDECIBA Biología - Subárea Biofísica
Tesis de Doctorado



Aproximaciones neurocomputacionales al análisis de la coherencia temática en el lenguaje.

Mag. Álvaro Cabana

Grupo de Modelización de Sistemas Cognitivos.
Sección Biofísica, Facultad de Ciencias.
Universidad de la República

Orientador: Dr. Eduardo Mizraji
Co-orientador: Dr. Juan C. Valle Lisboa
Tribunal: Dr. Ángel Caputi (Presidente)
Dr. Ruben Budelli
Dr. Diego Fernández Slezak

Índice general

Parte I Marco General	2
1 Introducción	3
1.1 El problema	3
1.2 Facultad del lenguaje	5
1.2.1 Gramática	6
1.2.2 Semántica	8
1.2.3 Tema y discurso	8
1.3 Cerebro, lenguaje y patologías	9
1.3.1 Afasias	10
1.3.2 Modelos de lectura y dislexia	10
1.4 Esquizofrenia y coherencia	13
2 Redes neuronales	14
2.1 La neurona abstracta	14
2.1.1 Pesos sinápticos y aprendizaje	15
2.2 Perceptrones y memorias lineales	16
2.2.1 Funcionamiento de la red	17
2.2.2 Memorias asociativas lineales	17
2.2.3 Algoritmos de aprendizaje	19
2.2.4 Entrenando memorias lineales	20
2.2.5 Propiedades de las memorias lineales	22
2.2.6 Separabilidad lineal	24
2.2.7 Memorias matriciales contexto-dependientes	25
2.3 Redes neuronales y lenguaje	28
2.3.1 Red recurrente simple de Elman	28
3 Modelos semánticos	32
3.1 Detección automática de tópicos	35
3.2 Espacios semánticos	35
3.2.1 Análisis Semántico Latente	37
3.2.2 Limitaciones	37
3.2.3 Temas en espacios semánticos	39
3.3 Modelos probabilísticos	40
3.4 Modelos de redes neuronales	42

3.4.1	Selector de tópicos	43
3.5	Herramientas	46
4	Esquizofrenia	47
4.1	Síntomas y dimensiones patológicas	48
4.2	Desorden del pensamiento	49
4.2.1	La importancia del lenguaje	50
4.2.2	Análisis automatizados	53
4.3	Redes neuronales en esquizofrenia	55
4.3.1	Alucinaciones auditivas y pérdida de conectividad	55
4.3.2	Modelo SPELT	56
5	Objetivos	59
5.1	Objetivo general	59
5.2	Objetivos específicos	59
	Parte II Resultados y Discusión	60
6	Redes neuronales	61
6.1	Módulo selector de tópicos	61
6.2	Categorización de textos	61
6.2.1	Medidas de clasificación	62
6.2.2	Métodos	63
6.2.3	Resultados	65
6.3	Red recurrente simple	66
6.3.1	Métodos	69
6.3.2	Resultados	73
6.4	Discusión	77
7	Cuantificación de coherencia	80
7.1	Entropías temáticas	80
7.1.1	Entropía de tema	81
7.1.2	Entropía de transición	84
7.1.3	Entropías y desorganización	86
7.2	Entropías y esquizofrenia	88
7.3	Asignación automática	90
7.3.1	Generación de grafos temáticos	90
7.3.2	Agrupamiento jerárquico de proyecciones	93
7.4	Discusión	99
8	Conclusiones y perspectivas	111
8.1	Validación de espacios	112
8.2	Entropías y redes	114

Parte III Apéndices	134
A Redes neuronales	135
A.1 Derivación de la regla delta	135
B Entropías temáticas	137
B.1 Máximo de entropía temática	137
B.1.1 Definición	137
B.1.2 Caso 2 temas	137
B.1.3 Caso general	139
B.2 Máximo de entropía de transición	141
B.2.1 Definición	141
B.2.2 Caso 3 tópicos	142
B.2.3 Caso general	142
C Espacios semánticos	144
C.1 Análisis Semántico Latente	144
C.1.1 <i>Corpus Wikipedia</i> en español	145
C.2 Búsquedas en LSA	145
C.3 Discursos artificiales	146
Bibliografía	148

Agradecimientos

- En primer lugar, agradezco la paciencia, el estímulo y los consejos de mis orientadores, Eduardo y Juan, quienes me han guiado en este camino.
- Muchos compañeros de viaje aportaron su granito de arena, desde la Sección Biofísica, como desde el CIBPsi. A todos ellos un saludo.
- Esta tesis fue posible gracias al apoyo económico de una Beca de Doctorado de la Agencia Nacional de Investigación en Innovación (ANII), y los fondos de alícuota para estudiantes del Programa de Desarrollo de las Ciencias Básicas (PEDECIBA Biología).
- También agradezco el apoyo nutritivo del bar Gran Sportman (18 de Julio y Tristán Narvaja), escenario de escritura de buena parte de esta tesis.
- Nada de esto hubiera sido posible sin el constante apoyo de Paula y las sonrisas de Josefina. A ellas dedico este trabajo.

Resumen

El lenguaje es un fenómeno cognitivo conspicuo y complejo, que comprende fenómenos que ocurren en un gran rango de escalas espaciales y temporales. Provee a su vez, una ventana a través de la cual intentar comprender los mecanismos neurales que subyacen a los fenómenos mentales. La disponibilidad de grandes colecciones de textos y de una siempre creciente capacidad de cómputo han permitido el desarrollo de aproximaciones computacionales al estudio de la relación entre cerebro y lenguaje. Los modelos de redes neuronales son un ejemplo de herramientas que han sido utilizadas exitosamente con este fin.

La esquizofrenia es una patología psiquiátrica que afecta en forma conspicua al lenguaje. Entre otras alteraciones, produce una pérdida en la coherencia temática del discurso. Este deterioro tiene valor diagnóstico y pronóstico, a la vez que su cuantificación resulta difícil de estandarizar y automatizar. Resulta de interés entonces, contar con una herramienta que permita evaluar en forma automática y estandarizada el grado de coherencia temática de un discurso.

Por otra parte, los modelos de redes neuronales han intentado evaluar hipótesis fisiopatológicas acerca de la esquizofrenia intentando emular los síntomas y signos de la enfermedad. Contar con, por un lado, un modelo de red neuronal capaz de representar significados, y por otro, con una herramienta para cuantificar la pérdida de la coherencia temática, permitiría explorar las condiciones y mecanismos por los cuales esta pérdida ocurre.

En esta tesis se exploran dos aproximaciones neurocomputacionales al problema de la coherencia temática en el discurso. Por un lado, se exploró la capacidad de modelos de redes neuronales de capturar relaciones semánticas en forma similar a la exhibida por métodos tales como los espacios semánticos. Por otro, se construyó y evaluó un instrumento para cuantificar la coherencia temática presente en un discurso en forma automática.

Los modelos de redes neuronales empleados en esta tesis lograron capturar relaciones semánticas presentes en las palabras a partir de los patrones de co-ocurrencia en colecciones de textos. En particular, una implementación no supervisada resulta escalable a colecciones de tamaños suficientemente grandes como para intentar emular el desempeño humano en algunas tareas psicolingüísticas, algo poco común en modelos de naturaleza neural.

El instrumento construido para cuantificar la coherencia temática de un discurso fue aplicado en forma promisorio al análisis de discursos de pacientes diagnosticados con esquizofrenia. En particular, el procedimiento completamente automático aquí propuesto muestra resultados interesantes, que sentarán las bases de futuros desarrollos.

Organización del trabajo

Esta tesis está organizada en dos partes. En la primera parte, se introduce el marco general y los objetivos de la tesis. En el primer capítulo se explica la motivación principal de el presente trabajo, se introducen brevemente algunos aspectos de la facultad del lenguaje, la importancia de las patologías y los modelos de redes neuronales en el estudio de las bases neurales de esta facultad, y la pérdida de coherencia en el discurso de pacientes esquizofrénicos. En los siguientes tres capítulos se presentan los modelos de redes neuronales, los modelos semánticos y de tópicos, y las alteraciones discursivas en la esquizofrenia, respectivamente. Luego en el capítulo 5 se explicitan los objetivos del presente trabajo de tesis.

En la segunda parte, se presentan los resultados obtenidos. En el capítulo 6 se muestran los resultados obtenidos empleando modelos de redes neuronales que son capaces de capturar información semántica a partir de colecciones de textos, mientras que en el capítulo 7 se muestran los resultados obtenidos el elaborar y evaluar una herramienta de cuantificación de la coherencia temática del discurso. Finalmente se elaboran conclusiones y perspectivas en el capítulo 8.

Parte I

Marco General

Capítulo 1

Introducción

1.1 El problema

Este trabajo de tesis puede ser considerado como parte del esfuerzo colectivo por entender cómo el cerebro da origen a la mente. Quizás una de los problemas más grandes a los que se enfrenta la ciencia, el estudio de la mente ha sido abordado desde hace mucho tiempo a través de múltiples enfoques. Nos centramos aquí entonces en el problema de cómo el cerebro sustenta el lenguaje. En particular, se trata de una contribución a intentar reducir la brecha entre el conocimiento del cerebro que aportan las neurociencias, y el que brindan las distintas ciencias cognitivas sobre el funcionamiento de la mente.

Los modelos de redes neuronales ofrecen un paradigma de investigación muy interesante que permite al mismo tiempo remitir a aspectos de la naturaleza funcional y anatómica de los circuitos neurales, y también funcionar como modelos cognitivos (Bechtel y Abrahamsen, 2002; Eliasmith y Anderson, 2004; Elman *et al.*, 1997; McClelland y Rumelhart, 1986). Desde los años 1980 en particular, los modelos de redes neuronales han influenciado el desarrollo de las ciencias cognitivas, aportando al desarrollo de implementaciones computacionales concretas de fenómenos cognitivos, algo que algunos proponen como un paso necesario para una comprensión cabal de estos fenómenos.

El lenguaje se presenta como un observable comportamental conspicuo, y provee un punto de acceso privilegiado al mundo cognitivo (Pinker, 2007; Spitzer, 1999), siendo objeto de estudio de la psicología y elemento diagnóstico en neuropsicología y psiquiatría (McKenna y Oh, 2005). Además, es en derecho propio un objeto de estudio en sí mismo, como lo es para la lingüística. En su relación con el cerebro que le da sustrato, ha sido estudiado por la neuropsicología y la neurología, entre otras. Y como corolario de la llegada de grandes capacidades de cómputo y la disponibilidad de grandes colecciones de bases de datos lingüísticas, el lenguaje ha sido objeto

de estudio de las ciencias de la computación y la información, en un rango de disciplinas que tienen como paraguas los términos de Procesamiento del Lenguaje Natural, o más en general, *information sciences*.

Por un lado, los modelos de redes neuronales han sido aplicados a fenómenos lingüísticos, habiendo contribuido en temas tales como el procesamiento fonológico (Kohonen, 1988), producción morfológica (Rumelhart y McClelland, 1986), la captura de las dependencias sintácticas (Christiansen y Chater, 1994; Elman, 1990), y la naturaleza de las representaciones semánticas (Ritter y Kohonen, 1989; Rogers y McClelland, 2004). Es en este último aspecto que esta tesis se enfoca, intentando contribuir al estudio de modelos de redes neuronales que logren implementar representaciones semánticas de palabras, oraciones y textos.

Por otro lado, el estudio de las patologías que afectan al lenguaje han aportado al conocimiento del funcionamiento del cerebro. Tal es el caso de las afasias, la dislexia y el trastorno específico del lenguaje, entre otras. Un fenómeno interesante y difícil de cuantificar es la pérdida de coherencia en el discurso de algunos pacientes esquizofrénicos (Elvevåg *et al.*, 2007; McKenna y Oh, 2005). Este fenómeno es informativo acerca del nivel de desarrollo de la patología del paciente, y tiene el potencial de ilustrar acerca de los mecanismos que sustentan la representación del significado y el control ejecutivo necesarios para la elaboración de un discurso organizado y coherente (Kuperberg, 2010a).

Por último, en el campo de las ciencias de la información han surgido herramientas que permiten dar sustento formal y computacional a la representación del significado, tanto de palabras como de fragmentos de texto o discursos. Estos métodos surgieron para dar respuesta a problemas relacionados a la búsqueda de información en grandes colecciones de datos, pero recientemente han sido aplicados al análisis y la modelación cognitiva (Foltz *et al.*, 1998; Jones *et al.*, 2006; Landauer y Dumais, 1997). De particular importancia para esta tesis son los modelos de espacios semánticos y los modelos probabilísticos de tópicos. Estos métodos por un lado han sido aplicados al estudio de la desorganización del discurso en la esquizofrenia (Elvevåg *et al.*, 2010, 2007), y además han sido vinculados con los modelos de redes neuronales (Blouw y Eliasmith, 2013; Mizraji *et al.*, 2009).

El panorama entonces puede representarse como se muestra en la figura 1.1. Los tres grandes temas que se intersectan en esta tesis son los modelos de redes neuronales, la identificación de tópicos o temas en textos y discursos, y las alteraciones discursivas en la esquizofrenia. El objetivo último del programa de investigación en el que se enmarca esta tesis se centra en la intersección de estos tres temas: la formulación de modelos de redes neuronales que emulen las alteraciones discursivas en la esquizofrenia, pudiendo ser exploradas estas últimas mediante instrumentos de diagnóstico que se basan en la noción de tópico. Un ejemplo concreto de este enfoque es el artículo de Valle-Lisboa *et al.* (2014) que incluimos en el capítulo 8 de Perspectivas, ya

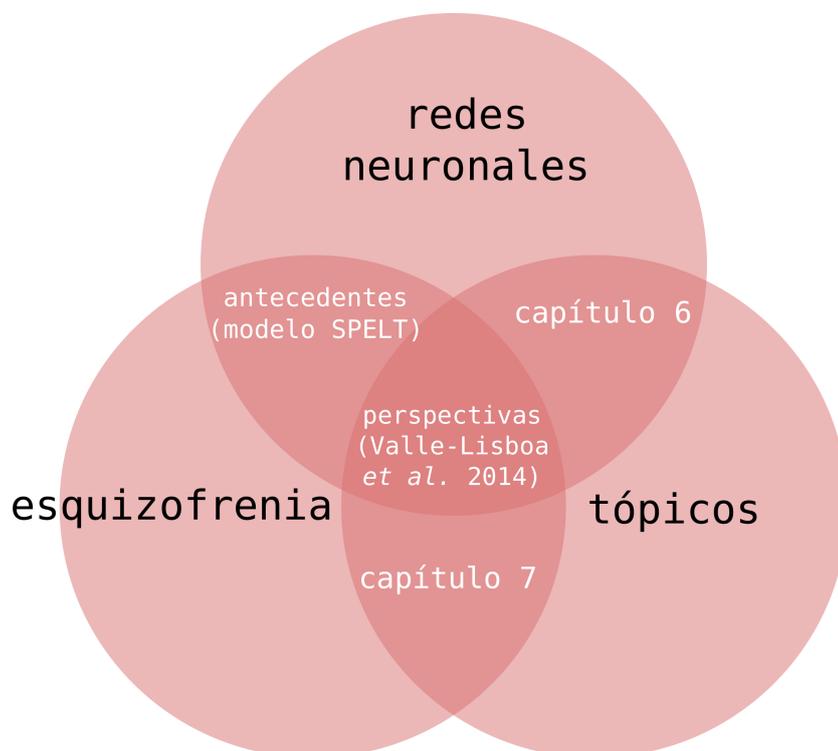


Figura 1.1: Representación de los temas que abarca esta tesis usando un diagrama de Venn.

que ahí se plantea una estrategia que será objeto de ulteriores desarrollos. En el capítulo 6 se presentan los resultados relacionados a los modelos de redes neuronales que sustentan nociones de tema y significado. En tanto, en el capítulo 7 y en el apéndice B.1 se presentan los resultados relacionados con la cuantificación de las alteraciones discursivas en pacientes con esquizofrenia, utilizando para ello representaciones de la estructura temática del discurso.

1.2 Facultad del lenguaje

El lenguaje es un fenómeno que abarca muchas escalas espaciales y temporales, desde los milisegundos y los milímetros observables en el estudio de la psicofísica y psicofisiología de la percepción y procesamiento del lenguaje, hasta los miles de años y kilómetros de la historia de los pueblos y el cambio diacrónico, además del cambio evolutivo que abarca centenas de miles de años.

De forma transversal a estas escalas, el lenguaje en tanto estructurada y compleja forma de comunicación y representación del mundo exterior, se compone de diversos niveles que interactúan entre sí. En cierto modo pueden considerarse como capas de complejidad en una dirección que va

desde la naturaleza física del mensaje emitido, hasta las implicaciones que tiene en el contexto social el contenido del mismo.

1.2.1 Gramática

El lenguaje hablado implica la emisión y percepción de combinaciones de sonidos (gestos en el caso del habla de señas) producidos por la vibración del aire en la laringe y las cavidades nasal y bucal. La señal acústica resultante es de una gran complejidad, y los recursos neurales destinados a la producción y el análisis de estos patrones parece ser sustancial (Pulvermüller, 2002; Tankus *et al.*, 2012).

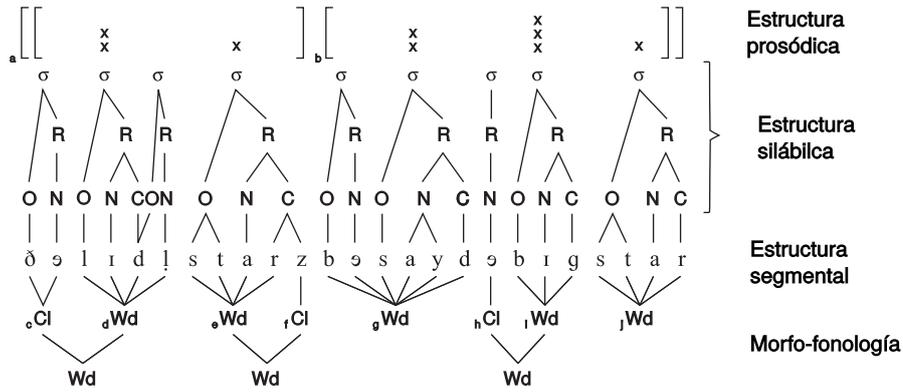
La forma en la que este canal sonoro es utilizado para transmitir un potencial infinito de mensajes, utilizando un conjunto finito de símbolos discretos (lo que permite almacenarlos en memoria), es a través del uso de reglas combinatorias a diferentes niveles. Se conoce como *gramática* al conjunto de reglas que operan para organizar la forma en que las diferentes unidades discretas se combinan. Son componentes de la gramática la *fonología*, la *morfología* y la *sintaxis*, que rigen la forma en la que se combinan sonidos (la unidad es el *fonema*), unidades subléxicas en palabras (la unidad es el *morfema*) y las palabras en frases y oraciones (la unidad es la frase o el *sintagma*).

En la figura 1.2 se ilustran estos niveles a través de esquemas que representan las estructuras que se generan al combinar los elementos o unidades básicas. Se puede apreciar la existencia de combinación de unidades básicas en unidades mayores, las cuales se combinan entre sí en varios niveles para engendrar estructuras ricas. La existencia de estas estructuras le imprime poder expresivo al lenguaje, por un lado, y potencia la ocurrencia de regularidades que facilitan la tarea de producción y procesamiento, por otro (Jackendoff, 2002; Pinker, 1999).

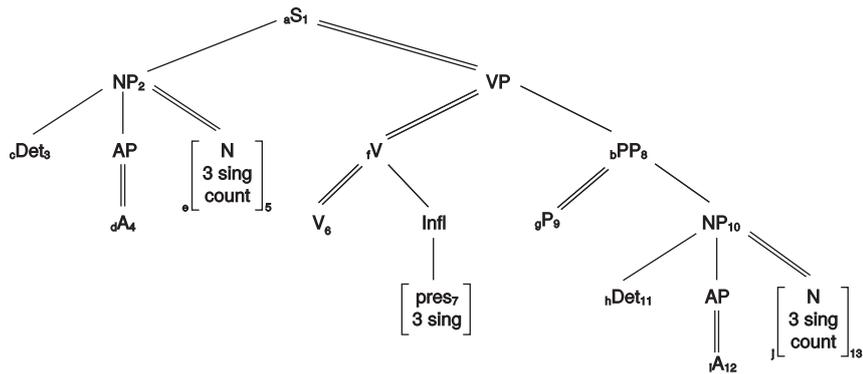
Los diferentes niveles gramaticales han sido descritos como clásicamente constituidos por una serie de reglas que operan sobre símbolos. Por ejemplo, para producir la forma plural de un nombre en español, una regla especifica el agregado del sufijo *ES*, *CES* o *S* dependiendo si la terminación de la forma singular es *s*, *z*, u otra.

En el seno de la lingüística se han desarrollado varios marcos teóricos que definen qué tipo de reglas y estructuras rigen los niveles gramaticales. Las gramáticas *generativas*, desde sus primeras formulaciones a mediados del siglo xx hasta sus versiones más recientes (Chomsky, 1993; Pinker, 1999; Pollard, 1994), mantienen a pesar de sus diferencias algunas características en común. Estas características (estructuras gramaticales jerárquicas y recursivas por ejemplo) son consideradas parte integral de las mismas. A este enfoque basado en reglas y la manipulación de símbolos discretos, suele denominarse el “enfoque clásico” de las ciencias cognitivas, y no sólo es aplicado en el dominio del lenguaje, sino que está muy desarrollado en

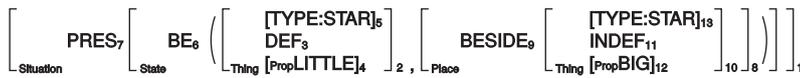
(a) Estructura fonológica



(b) Estructura sintáctica



(c) Estructura semántica/conceptual



(d) Estructura espacial

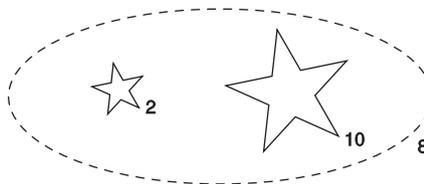


Figura 1.2: Estructuras lingüísticas. Adaptado de Jackendoff (2002).

el campo de la inteligencia artificial (véase por ejemplo [Russell y Norvig \(1995\)](#)), entre otras áreas ([Jackendoff, 2009](#)). Como se verá más adelante, algunos modelos de redes neuronales, en particular los amparados bajo el paradigma del *conexionismo*, desafían esta visión clásica argumentando que estos modelos pueden remedar el desempeño cognitivo sin la necesidad de implementar ningún tipo de reglas en forma explícita ([Bechtel y Abrahamson, 2002](#); [McClelland y Rumelhart, 1986](#); [Seidenberg y McClelland, 1989](#)). Esta discrepancia ha dado origen a un debate que lleva ya tres décadas, y aparentemente aún no ha sido resuelto ([Elman *et al.*, 1997](#); [Fodor y Pylyshyn, 1988](#); [Pinker y Prince, 1988](#)).

1.2.2 Semántica

Otro nivel estructural cuya descripción es menos consensuada ([Jackendoff, 2002](#)) es el de la representación del significado, tanto a nivel de cada palabra individual (y en su estructura interna), como la forma en que se codifican mediante la gramática las relaciones entre los significados de estas palabras. Así, en el ejemplo de la figura 1.2, a nivel léxico, las palabras *star*, *big*, *little*, *beside* tienen un significado más o menos definido en sí mismas, mas en su combinación se ponen en relación para crear un escenario nuevo.

Entre los enfoques simbólicos aplicados a la semántica se pueden contar al desarrollo de la lógica de predicados, hasta teorías de semántica generativa ([Lakoff, 1990](#)). Estos enfoques han permitido caracterizar algunos fenómenos frecuentes en las producciones lingüísticas como el manejo de referencias temporales, roles temáticos, etc ([Ingram, 2007](#)). Sin embargo, tanto a nivel “microscópico” (la representación del significado de cada palabra) como a nivel “macroscópico” (significado de un fragmento de discurso o texto), ningún marco teórico ha generado descripciones o modelos lo suficientemente potentes como para dar cuenta de muchos fenómenos de orden semántico ([Rogers y McClelland, 2004](#)).

1.2.3 Tema y discurso

El uso del habla se encuentra enmarcado naturalmente en un contexto funcional en la vida cotidiana de las personas. El habla se usa para comunicar algo a alguien, y esto implica por un lado el manejo de información acerca del contexto actual (por ejemplo, qué información es compartida por los hablantes). Por otro lado involucra una planificación de lo que se va a decir que va más allá de la oración actual, y tiene que ver con la estructuración del discurso para cumplir un objetivo.

Puede pensarse el discurso entonces como un acto motor, que tiene un objetivo, una planificación, un comienzo y un fin. El resultado es una secuencia ordenada de expresiones que mantienen un cierto grado de coherencia interna. A nivel de las oraciones, la coherencia está fuertemente domina-

da por la gramática, que impone restricciones acerca de la morfología y la posición de las palabras. A mayor escala (párrafos en el caso de un texto), pocas veces ocurren dependencias gramaticales, y por tanto la coherencia se exhibe fundamentalmente a nivel semántico. Las palabras que ocurren en una porción del discurso tienden a estar relacionadas semánticamente con las que ocurren en otra porción lejana (Doxas *et al.*, 2007; Foltz, 2007; Montemurro y Zanette, 2001). Esto refleja la existencia de unidades de cierta homogeneidad semántica, porciones del discurso que tienen un *tema* o tópico, que hacen referencia de un asunto o cosa.

La noción de tema (Kintsch, 2002; van Dijk, 1993) resulta de gran importancia en el campo de las ciencias de la información (Blei *et al.*, 2003). En los últimos años se han desarrollado una serie de herramientas cuyo objetivo es identificar el o los temas a los que hace referencia un fragmento de texto. De particular relevancia a esta tesis son los modelos de espacios semánticos y los modelos probabilísticos de tópicos (Blei, 2012; Griffiths *et al.*, 2007; Landauer y Dumais, 1997). Estos métodos han sido aplicados al estudio cognitivo del lenguaje en distintos contextos. Por ejemplo, Jones *et al.* (2006) usaron espacios semánticos para reproducir efectos de facilitación semántica en tareas de decisión léxica.

1.3 Cerebro, lenguaje y patologías

El cerebro es el sustrato de nuestras actividades cognitivas, entre las cuales siempre ha recibido especial atención el lenguaje articulado. A mediados del siglo XIX comenzaron a aparecer los primeros grandes avances en el entendimiento de las funciones cerebrales, y en particular cuáles son las bases anatómicas que subyacen al lenguaje. Mucho camino se ha recorrido desde entonces, y es mucho lo que sabemos hoy en día acerca de los qué, dónde y cómo del lenguaje en el cerebro. Esto ha sido posible gracias al desarrollo de numerosas técnicas experimentales, desarrolladas fundamentalmente en el siglo XX que permitieron abrir de a poco la ventana al cerebro, volviéndolo cada vez un poco menos opaco. Entre ellas se destacan las recientes técnicas de imagenología funcional¹.

Pero también, desde sus comienzos, el estudio de las patologías relacionadas al lenguaje sido muy valioso al aportar evidencia acerca de la localización cerebral y la naturaleza de algunas funciones lingüísticas. En esta sección repasaremos muy brevemente dos ejemplos de patologías han servido de fuente de información acerca de la implementación neural de las funciones lingüísticas, al tiempo que han dado origen a modelos cognitivos sobre distintos aspectos de la facultad de lenguaje (Ingram, 2007; Kuperberg, 2007; Pulvermüller, 2002; Stemmer y Whitaker, 2008).

¹Por ejemplo, resonancia magnética funcional, electroencefalografía, y magnetoencefalografía, entre otras

1.3.1 Afasias

Las afasias son síndromes que suelen aparecer en pacientes con daño cerebral, como hemorragias, stroke o resección quirúrgica, que afectan en varias maneras la producción o el procesamiento del habla. Suelen distinguirse las afasias en las que la producción está relativamente preservada (afasias receptivas o sensoriales), en contraste de aquellas en las que la recepción lo está (afasias de producción o motoras)². Esta disociación tiene orígenes históricos relacionados a los primeros reportes sobre la neurofisiología de las afasias.

Suele atribuirse a los trabajos de Paul Broca publicados a partir de 1861 con pacientes afásicos el comienzo del estudio de las bases neurales del lenguaje (Dronkers *et al.*, 2000). Broca encontró que los pacientes que tenían problemas para articular el habla (afasia de producción), exhibían *post-mortem* lesiones en el lóbulo frontal inferior izquierdo, dando un nuevo giro a la postura localizacionista de las funciones cerebrales propuesta por Franz Joseph Gall décadas antes. Pocos años después, Karl Wernicke publicaba sus trabajos con pacientes que presentaban afasias receptivas, los cuales exhibían lesiones en el tercio posterior del giro temporal superior izquierdo.

De estos descubrimientos surgió el primer modelo neurocognitivo del lenguaje: el modelo de Wernicke-Lichtheim, propuesto por Wernicke y Ludwig Lichtheim hacia los últimos años del siglo XIX. Este modelo fue extendido por Norman Geschwind para incorporar hallazgos acumulados hasta la primera mitad del siglo XX (Dronkers *et al.*, 2000). A pesar de sus limitaciones a la hora de explicar muchos fenómenos, es utilizado como referencia para organizar los esfuerzos de investigación y provee la base de un sistema una clasificación de las afasias que es usado hoy en día (Alexander, 1997).

1.3.2 Modelos de lectura y dislexia

La lectura es una habilidad perceptual y cognitiva adquirida recientemente en términos evolutivos, y esto se refleja en la dificultad que tiene la adquisición de esta habilidad, que es necesario entrenar a conciencia y que no surge espontáneamente (como la comprensión y producción del lenguaje oral). Por tanto resulta un fenómeno interesante en el que la plasticidad neuronal permite la aparición de nuevas estructuras y conexiones funcionales entre áreas corticales dedicadas al procesamiento visual y al lenguaje, en particular áreas que están funcional y anatómicamente relacionadas con otras dedicadas a la codificación fonológica y semántica de las palabras (Nakamura *et al.*, 2012).

La dislexia del desarrollo es un conjunto de trastornos que tienen como

²En realidad pueden distinguirse por lo menos 9 grandes tipos de afasia según la preservación o no de los siguientes tres rasgos: producción, comprensión y repetición (Alexander, 1997)

denominador común el afectar el desempeño lector desde el comienzo de la escolarización. El grado y la naturaleza de la dificultad en la lectura depende de la lengua y el sistema de escritura, en particular de la complejidad fonológica de las sílabas y la transparencia de la ortografía (Ziegler y Goswami, 2005). Aunque esto complica la comparación del trastorno en diferentes contextos culturales, la dislexia puede ser considerada el extremo bajo en la distribución de desempeño lector (1.5 desvío estándares por debajo de la media), afectando a un 7% de las personas a nivel mundial (Peterson y Pennington, 2012). Si bien las causas pueden ser múltiples, en muchas variantes la dislexia pareciera originarse en un déficit en el procesamiento fonológico debido a alteraciones neurológicas no del todo claras.

La dislexia ha tenido un papel importante en el desarrollo de los modelos cognitivos de lectura. En particular, en esta temática existe una interesante oposición entre dos visiones de cómo funciona la lectura desde un punto de vista cognitivo. Una variedad muy influyente son los modelos de doble ruta, que proponen la existencia de dos rutas paralelas por las que la información visual es procesada para lograr activar una representación de la palabra independiente de la modalidad (que incluye el contenido semántico, sintáctico) (Coltheart *et al.*, 2001). Como se ilustra en la figura 1.3 A, una de las rutas implica el acceso seriado a la representación fonológica de cada elemento constituyente de la palabra, componiendo estos elementos para obtener el sonido de la palabra entera. A esta ruta se la llama ruta indirecta, o fonológica, e involucra la ejecución de reglas que rigen el mapeo forma-visual/sonido para cada letra o sílaba de la palabra, lo que permite acceder al sonido de una palabra nueva para la que no se tiene experiencia previa.

Por otro lado, la ruta directa, o léxica, prescinde del acceso fonológico de las letras y las sílabas, para directamente acceder a la representación de la palabra a través de asociación visual. Es decir, recurre a la asociación entre forma visual y palabra que se ha almacenado en la memoria a partir de la experiencia previa, por lo que se necesita la exposición repetidas veces de una forma visual junto a una identidad de palabra. Sería esta vía también la que permite almacenar las excepciones a las reglas del mapeo forma visual-sonido, como por ejemplo en el caso del español, extranjerismos como *bijouterie* o *google*.

La existencia de estas dos rutas ha permitido acomodar buena parte de las evidencias experimentales en la literatura, y por eso ésta se ha establecido como la visión aceptada por buena parte de la comunidad científica (Coltheart, 2006).

En oposición a los modelos de doble ruta, Seidenberg y McClelland (1989) proponen mediante la construcción de modelos de redes neuronales, la existencia de un único proceso en el cual interactúan la forma ortográfica, la fonológica y la identidad semántica. La principal diferencia en su propuesta radical en que en esta propuesta, no se explicita en ningún lugar la existencia o ejecución de reglas de conversión entre la forma visual y la representación

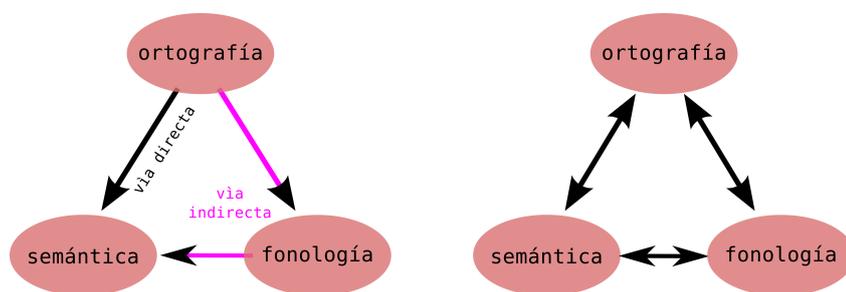


Figura 1.3: Diagrama de modelos de lectura. A. Modelo de doble ruta de Coltheart *et al.* (2001). B. Modelo de Seidenberg y McClelland (1989).

del sonido de las palabras. Más aun, estos autores aducen que en vez de palabras que respetan estas reglas y excepciones (como *casa* y *rouge* en español), existen formas cuasi-regulares, que respetan en mayor o menor medida algunas casi-reglas de conversión, es decir, tendencias probabilísticas (y no determinísticas) de conversión. Por ejemplo, en el idioma inglés que posee un mapeo grafo-fonológico muy irregular, existen palabras como *pint*, que riman con *paint* y no con *mint*, o *tint*. Como en otros contextos cognitivos, este modelo argumenta que lo que parecen reglas y excepciones en realidad son cuasi-regularidades estadísticas, y por tanto, el aprendizaje se convierte en un problema de inferencia estadística (Bechtel y Abrahamsen, 2002; Harm y Seidenberg, 2004).

En el debate entre estas visiones, las dislexias tanto adquiridas como del desarrollo ha jugado un papel importante. Primero, en muchos contextos las dislexias se clasifican en base a la visión de la existencia de dos rutas, una léxica y una fonológica; según qué vía esté afectada, se tendrán dislexias superficiales (vía léxica afectada), fonológicas (la vía fonológica es afectada), o profundas (ambas vías afectadas). Por otro lado, las formulaciones computacionales de los diferentes modelos permiten testear la validez de las hipótesis causales de la dislexia. Por ejemplo, Harm y Seidenberg (1999) simularon un déficit en el desarrollo del módulo fonológico, obteniendo un modelo cuyo desempeño era notoriamente más lento en general, y casi nulo para pseudopalabras.

El caso de los modelos cognitivos de lectura ilustra cómo se integran por un lado, modelos computacionales, y por otro, evidencia experimental acerca del comportamiento tanto en personas sanas como en individuos afectados por trastornos o patologías. En los últimos años, además, las técnicas de neuroimagen son una fuente extra de evidencia que complementa los enfoques antes mencionados (aunque hay quienes sostienen que un conocimiento más detallado de la implementación neural no necesariamente debe constreñir los modelos cognitivos; véase a Coltheart (2004) como un ejemplo de esta postura).

1.4 Esquizofrenia y coherencia

La esquizofrenia es un trastorno psiquiátrico que afecta muchos componentes cognitivos y afectivos (Frith, 1995; Kuperberg, 2010a). Es una psicosis que se caracteriza principalmente por la presencia de delirio, alucinaciones sensoriales complejas, así como de retraimiento social y aplanamiento afectivo, entre otros síntomas. Una condición presente en muchos pacientes es el llamado *desorden del pensamiento* (*thought disorder*) y se caracteriza por la presencia de razonamientos disfuncionales o aberrantes: las personas afectadas producen asociaciones extrañas, presentan dificultades en el razonamiento, y una pérdida de coherencia en el hilo de sus pensamientos.

En particular, el lenguaje de los pacientes esquizofrénicos, y en particular de aquellos que sufren desorden del pensamiento, ha sido estudiado desde hace décadas debido a que presenta varias particularidades. Una de las más conspicuas, es la pérdida de coherencia en el hilo narrativo (McKenna y Oh, 2005). Estas alteraciones de la organización temática del discurso, son importantes para el diagnóstico y la prognosis del tratamiento en pacientes (Elvevåg *et al.*, 2007), y por tanto se han desarrollado escalas que cuantifican el grado de deterioro de éste y otros aspectos del lenguaje (Andreasen, 1986). Sin embargo, la aplicación de estos métodos requiere la evaluación de expertos entrenados, y por tanto resulta costosa y difícil de estandarizar. Sería deseable por tanto, en especial para la implementación de estudios clínicos a mediana o gran escala, contar con herramientas automatizadas y estandarizadas que cuantifiquen el grado de coherencia de la producción discursiva.

Por otra parte, los modelos de redes neuronales han sido utilizados en el contexto neuropsiquiátrico para testear hipótesis fisiopatológicas y simular los efectos de posibles tratamientos (Gradin y Pomi, 2008; Siegle, 1999; Spitzer, 1999; Stein y Ludik, 1998a). En particular, la esquizofrenia ha sido el objetivo de varios esfuerzos de modelización, dada la amplitud de los síntomas y los diferentes componentes que parecen estar afectados, así como la diversidad de posibles mecanismos patológicos propuestos (Hoffman *et al.*, 2011; Hoffman y McGlashan, 2001). Por tanto, las redes neuronales resultan útiles para modelar algunos aspectos del funcionamiento cerebral, y en particular en el caso de patologías poco comprendidas como la esquizofrenia.

El tema principal de esta tesis resulta de la interacción entre la necesidad de generar una herramienta de cuantificación del desorden del discurso esquizofrénico con potencial uso en investigación clínica, utilizando herramientas surgidas del ámbito de las ciencias de la información (Foltz *et al.*, 1998) por un lado, y la construcción de modelos neuronales capaces de implementar la generación de discursos artificiales, con el fin de estudiar la fisiopatología de las alteraciones discursivas presentes en la esquizofrenia (Elvevåg *et al.*, 2007; Valle-Lisboa *et al.*, 2014).

Capítulo 2

Redes neuronales

Los modelos de redes neuronales pretenden capturar las propiedades esenciales de las neuronas reales para poder explorar matemáticamente sus capacidades computacionales. En este capítulo se introducirán algunos de los modelos de redes neuronales usados para comprender cómo el cerebro implementa aspectos del funcionamiento del lenguaje, además de presentar el tipo de modelo neural que fue utilizado en esta tesis.

2.1 La neurona abstracta

Los modelos de redes neuronales generalmente definen a sus unidades básicas, las neuronas¹, como unidades discretas capaces de procesar información en forma unidireccional. Reciben estímulos a través de sinapsis en sus “somas” o “dendritas”, y envían una señal a lo largo de su “axón” hacia otras neuronas. El primer modelo de neurona de este tipo fue propuesto por [McCulloch y Pitts \(1943\)](#), y hasta hoy en día muchos modelos se basan en esta primera conceptualización. Un esquema de una neurona típica en los modelos de redes neuronales se muestra en la figura 2.1. En algunos casos, como en el ya mencionado, las actividades de salida adquieren valores binarios. La neurona puede entonces estar o no activa, lo que pretende remedar la naturaleza “todo o nada” del potencial de acción de las neuronas reales. Cuando la actividad de las neuronas es binaria suele existir un “umbral” de actividad entrante que tiene que superarse para producirse una salida. En otros modelos, las actividades adquieren valores continuos, muchas veces valores reales, pudiendo ser interpretados como valores de frecuencia relativa de disparo. El hecho de que una neurona tenga mayor o menor actividad suele depender de una función de conversión que asigna la magnitud de la salida según el grado de activación de entrada. Típicamente, el efecto que las actividades de entrada tienen sobre una neurona está ponderado a

¹En algunos casos las unidades básicas pueden ser asimilables a conjuntos de neuronas. Por ejemplo, véase el trabajo de [Amari \(1972\)](#).

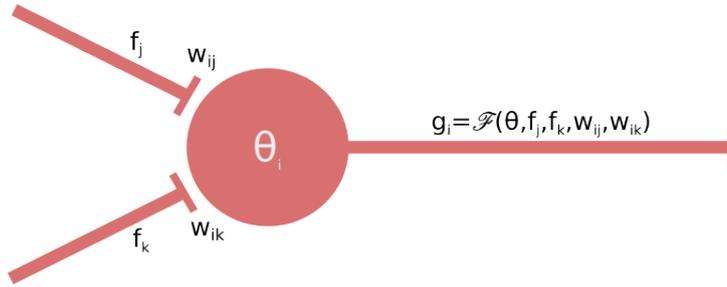


Figura 2.1: Esquema de una neurona abstracta. Los axones aferentes tienen actividades f_j , y sus contactos están modulados por pesos sinápticos w_{ij} . La actividad de salida estará dada por la función de transferencia \mathcal{F} , que tiene como parámetros al umbral θ_i , y a los pesos, y a las actividades f_j como variables.

través de los *pesos sinápticos* w_{ij} , que reflejan el diferente grado de eficacia (y la naturaleza excitatoria o inhibitoria) de las diferentes sinapsis sobre la actividad de salida de la neurona post-sináptica.

La actividad de salida de una neurona entonces, está determinada por una función \mathcal{F} de las entradas f_i , los pesos w_{ij} y el umbral θ_i :

$$g_i = \mathcal{F}(w_{ij}, f_j, \theta_i) \quad (2.1)$$

La función \mathcal{F} se denomina *función de transferencia*. En el caso del modelo de McCulloch-Pitts, esta función es:

$$\mathcal{F} = \mathcal{H} \left(\sum_j w_{ij} f_j - \theta_i \right) \quad (2.2)$$

donde \mathcal{H} es la función escalón de Heaviside ([Bose y Liang \(1996\)](#)):

$$\mathcal{H}(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}$$

2.1.1 Pesos sinápticos y aprendizaje

Los coeficientes de los pesos sinápticos son los principales determinantes del funcionamiento de la red: de hecho, una parte central de todos los modelos de redes neuronales consiste en la prescripción de un método de selección de sus valores. En muchos casos se prescribe un método de ajuste progresivo de los valores para los pesos sinápticos, denominado “algoritmo de aprendizaje”. Típicamente estos procedimientos producen un cambio desde una configuración de valores de los pesos que no realiza la tarea deseada (por ejemplo una configuración aleatoria) a una configuración óptima o sub-óptima que es capaz de realizarla.

En la siguiente sección se ilustrarán estos conceptos con la presentación de los modelos lineales que fueron utilizados en este trabajo de tesis, luego

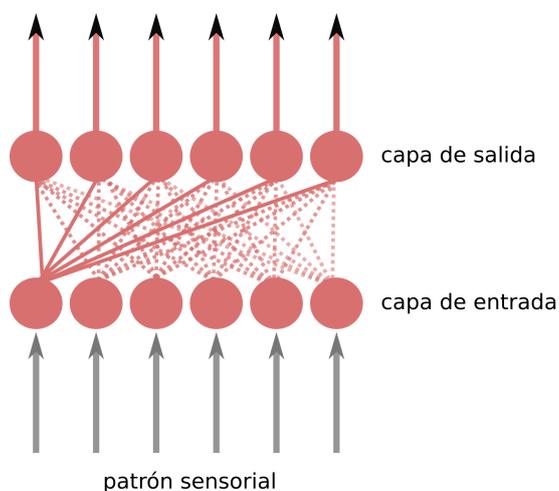


Figura 2.2: Diagrama de un perceptrón de una capa. La información sensorial llega a la capa sensorial, o capa de entrada, que se encarga de relevarla a la capa de salida. Son estas últimas conexiones las que tienen pesos sinápticos ajustables.

de una breve introducción a un muy influyente modelo de red neuronal: el perceptrón.

2.2 Perceptrones y memorias lineales

El perceptrón fue propuesto por F. Rosenblatt (1958) como modelo de procesamiento visual. En su versión más simple consiste de una capa de neuronas McCulloch-Pitts que reciben aferencias de un conjunto de neuronas sensoriales. Estas neuronas sensoriales envían sus eferencias a un conjunto de neuronas "motoras" o "efectores". En la figura 2.2 se ilustra un perceptrón sencillo de una capa.

La tarea del perceptrón es la de discriminar patrones en los estímulos sensoriales, activando las unidades efectoras "apropiadas" para cada patrón. Estos mapeos entrada/salida son definidos arbitrariamente, y en base a algún mecanismo de aprendizaje, los pesos sinápticos y umbrales de la capa efectora se pueden ajustar de modo que el mapeo adecuado sea realizado.

En los perceptrones simples, entonces, hay una capa de entrada que recibe aferencias sensoriales, cuyos pesos sinápticos y umbrales están fijos de antemano, y una capa de salida, cuyos pesos sinápticos y umbrales pueden modificarse según algún procedimiento de aprendizaje. En general, este tipo de arquitecturas se denomina "feedforward", y en este caso particular se considera como de una sola capa, dado que hay sólo una capa de neuronas cuyos parámetros son modificables.

2.2.1 Funcionamiento de la red

Consideremos una red “feedforward” de una capa, que recibe aferencias f_j , cada una pesada por su peso sináptico w_{ij} . La suma ponderada de estas aferencias será igual a la *actividad* de entrada de una neurona. Por simplicidad, asumiremos que la actividad de salida no es una expresión binaria como en el caso de los perceptrones, sino es igual a la actividad de entrada. En este caso entonces, la función de transferencia \mathcal{F} es la identidad²:

$$g_i = \sum_j w_{ij} f_j \quad (2.3)$$

El funcionamiento de la red entonces puede plantearse en los siguientes términos: Definimos que $f^k = \{f_j\}$ es un vector cuyas coordenadas son las actividades de las neuronas entrantes, es decir, el *patrón de entrada*, y $g^k = \{g_i\}$ es un vector cuyas coordenadas son las actividades de salida de las neuronas de la red, o sea, el *patrón de salida*. Para cada patrón de entrada f^k , se puede asignar una patrón de salida deseado y^k deseado, estableciendo la asociación:

$$\mathcal{G} : f^k \rightarrow y^k \quad (2.4)$$

La red entonces realiza la tarea en buena forma cuando para todos los k se tiene que $g^k = y^k$, esto es, la actividad de salida de la red es la deseada para cada patrón de entrada f^k . La existencia de un conjunto de pesos w_{ij} que cumpla con esta condición no siempre está garantizada, dependerá de la arquitectura de la red y de los patrones particulares que se quiera asociar.

Muchos de los modelos de redes neurales que vinieron después se basaron en gran medida en el desarrollo de los perceptrones. En particular, consideraremos a continuación el caso de las memorias asociativas lineales.

2.2.2 Memorias asociativas lineales

Introducidos en la década de 1970 en forma independiente por Teuvo Kohonen, James A. Anderson y Shun-Ichi Amari ([Amari, 1972](#); [Anderson, 1972](#); [Kohonen, 1972](#)), estos modelos son capaces de almacenar la asociación entre pares de vectores entrada y salida. El principal aporte de estos autores fue plantear el problema en términos matriciales, vinculando a los modelos de redes neuronales con problemas del álgebra lineal. En el caso de las memorias asociativas lineales, la tarea de la red es almacenar la asociación de vectores de entrada y salida, es decir, especificar un mapeo \mathcal{G} como se muestra en la ecuación 2.4. Para simplificar la notación en esta sección, el vector salida deseado y^k lo anotaremos sencillamente g^k :

²En este caso estamos ignorando los umbrales de cada neurona θ_j , aunque los resultados son esencialmente los mismos: puede incluirse una entrada extra constante e igual a 1 para toda la capa, y en tal caso cada peso sináptico para esa entrada extra equivaldrá $-\theta_j$.

Dado que cada neurona es un integrador lineal, cada actividad de salida equivale a la suma ponderada por los pesos de las actividades entrantes (ecuación 2.3). Entonces la red puede ser descrita mediante la siguiente ecuación:

$$g = Mf \quad (2.5)$$

donde f y g son los vectores de patrones de entrada y salida, respectivamente, y $M = w_{ij}$ es la matriz de pesos sinápticos. La asunción de linealidad en la función de transferencia encuentra una justificación en el análisis de Nass y Cooper (1975), quienes encuentran que esta función es prácticamente lineal para un rango amplio de niveles de actividad de entrada en algunos tipos de neuronas reales.

Lo que Kohonen y Anderson encontraron fue que, bajo ciertas condiciones, el problema de encontrar la matriz M que almacenara las asociaciones podía solucionarse algebraicamente. En particular, si los vectores de entrada f^k son ortonormales, la matriz que resuelve el problema en forma exacta puede calcularse como:

$$M = \sum_k g^k f^{kT} \quad (2.6)$$

Si F y G son las matrices donde cada columna k son los vectores de entrada y salida respectivos, entonces puede escribirse como:

$$M = GF^T \quad (2.7)$$

Puede observarse que ante la presentación de una entrada concreta f^k , la respuesta de la red será la respuesta deseada g^k :

$$g = Mf = \sum_j g^j f^{jT} f^k = \sum_{j \neq k} g^j f^{jT} f^k + g^k f^{kT} f^k = g^k$$

dado que para todos los $j \neq k$,

$$f^{jT} f^k = \langle f^j, f^k \rangle = 0,$$

mientras que

$$f^{kT} f^k = \langle f^k, f^k \rangle = \|f^k\|^2 = 1.$$

Si se relajan un poco las restricciones, y sólo se exige al conjunto de vectores de entrada $\{f^k\}$ ser linealmente independiente, entonces la mejor solución³ estará dada por la pseudoinversa de Moore-Penrose de la matriz F , denotada F^+ :

$$M = GF^+ \quad (2.8)$$

³La pseudoinversa de Moore-Penrose garantiza que la solución encontrada es la que minimiza el error cuadrático: $E(M) = \sum_k \|g^k - Mf^k\|^2$ y al mismo tiempo tiene norma $\|M\|^2$ mínima. (Meyer, 2000)

Con respecto a la pseudoinversa, vale la pena observar que se calcula de la siguiente manera⁴:

$$A^+ = (A^T A)^{-1} A^T$$

2.2.3 Algoritmos de aprendizaje

Como se describe en la sección anterior, existe un método analítico para encontrar los pesos sinápticos óptimos para almacenar las asociaciones. En cambio, y en un espíritu que intenta remedar la realidad neurobiológica, existen algoritmos de aprendizaje que especifican procedimientos de actualización iterativa de los pesos sinápticos de manera que en algunos casos puede garantizarse que converge a la solución deseada. Analizamos una categoría de estos mecanismos, y en particular el método propuesto por [Widrow y Hoff \(1960\)](#).

Los algoritmos de aprendizaje suelen especificar un método a través del cual se realizan pequeños ajustes a los pesos sinápticos. A través de la iteración continua del algoritmo, por lo general variando las entradas y las salidas, se especifica la variación que tendrán los pesos sinápticos:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij} \quad (2.9)$$

Un componente primordial de estos algoritmos es, entonces la especificación de la variación Δw_{ij} en los pesos sinápticos.

Existe una gran diversidad de algoritmos de aprendizaje, reflejando en parte la gran diversidad de arquitecturas de modelos neurales desarrollados con distintos fines y en diferentes contextos. A grandes rasgos, estos algoritmos pueden clasificarse en tres grandes categorías, de acuerdo con la existencia de una “señal de error” externa, y el rol que ésta juega en el aprendizaje.

Una clase particular de algoritmos de aprendizaje implica la existencia de un conjunto predefinido de pares de f^k y y^k asociados, denominado “conjunto de aprendizaje”, que guiará el proceso de aprendizaje de la red. A estos casos se los agrupa bajo la denominación de “aprendizaje supervisado”, por la existencia de un “supervisor” o “tutor” que indica a la red durante su entrenamiento cuál es la respuesta correcta. Esta noción de respuesta ideal o esperada permite cuantificar el grado de “error” presente en el desempeño de la red y resulta crucial para guiar el cambio o actualización de los pesos sinápticos, de modo de aproximarse iterativamente a el objetivo.

Para muchos otros algoritmos (o más bien situaciones de aprendizaje) esta información no está disponible en ningún momento, y la actualización de los pesos se hace únicamente en función de los patrones de entrada. A este tipo de aprendizaje se le denomina “aprendizaje no supervisado”.

⁴Se considera el caso de matrices con componentes reales. Para el caso complejo, en lugar de la transposición se debe operar el conjugado complejo.

Por último, la tercera clase la constituyen los algoritmos que realizan “aprendizaje por refuerzos”. En estos casos sí existe una señal del exterior en función de la respuesta actual, pero no es de carácter específico (véase [Bechtel y Abrahamsen \(2002\)](#); [Bose y Liang \(1996\)](#)).

2.2.4 Entrenando memorias lineales

Los primeros procedimientos ideados para lograr entrenar una red para realizar una tarea determinada estaban inspirados en la máxima enunciada por el neurobiólogo Donald O. Hebb (1949). Su hipótesis fue que si una neurona A descarga sobre otra neurona B cuando esta última está cursando un potencial de acción, entonces la fuerza de conexión entre A y B se ve aumentada. Para esto, supuso que debían ocurrir cambios en la fisiología de la sinapsis que alteren su funcionamiento. Esta idea fue tomada por muchos algoritmos de aprendizaje, habiendo hoy todo un repertorio de procedimientos denominados “hebbianos”, “anti-hebbianos”, entre otros. Hoy en día, con el conocimiento adquirido acerca de la fisiología de las sinapsis en varias décadas de investigación neurobiológica, se han incorporado nociones más detalladas y realistas acerca de la plasticidad sináptica en los algoritmos de aprendizaje. De todos modos, la hipótesis hebbiana continúa siendo influyente en esta área.

Implementando la regla de Hebb, puede entrenarse una memoria lineal como la descrita en la sección 2.2.2. La variación en los pesos sinápticos sería entonces:

$$\Delta w_{ij} = \eta g_i f_j \quad (2.10)$$

donde η es un parámetro llamado *tasa de aprendizaje*, y f_i y g_j son las actividades de la neurona de entrada i y la neurona de salida j , respectivamente. Esta ecuación en formato matricial es igual a:

$$\Delta W = \eta g f^T \quad (2.11)$$

Por lo tanto, si se comienza con una matriz de pesos nula, la regla de Hebb genera la matriz W después de N_k presentaciones de los diferentes pares asociados (siempre y cuando se tenga que $\eta = 1/N_k$):

$$W = \sum_k^{N_k} g^k f^{kT} = GF^T \quad (2.12)$$

y esta matriz es la establecida en la ecuación 2.7 para el caso del conjunto de entradas f^k ortonormal.

Gradiente descendente

Muchos algoritmos de aprendizaje se enfocan en minimizar la discrepancia entre los patrones de salida de la red g^k y las salidas deseadas y^k . Esta

discrepancia suele cuantificarse mediante el cálculo del error cuadrático medio sobre el de conjunto de pares k :

$$E(W) = \frac{1}{2N} \sum_k (y^k - g^k)^2 \quad (2.13)$$

El efecto del aprendizaje sobre los pesos sinápticos se vería reflejado en una disminución del error: pueden concebirse algoritmos de aprendizaje que se basen en minimizar la función $E(W)$ al modificar los pesos.

Un tipo de algoritmos de aprendizaje utilizan la técnica de *gradiente descendente* para tal fin, utilizando para ello el gradiente de la función error:

$$\nabla E = \frac{\partial E}{\partial W}.$$

El gradiente ∇E es un vector que señala la dirección en el espacio de pesos W en la que la función crece más rápido. Por tanto, si en cada iteración se actualizan los pesos en la dirección contraria al gradiente:

$$W(t+1) = W(t) + \Delta W(t) = W(t) - \eta \nabla E \quad (2.14)$$

es posible encontrar un mínimo local de la función error. Si la función error resulta ser monótona, entonces se puede encontrar un mínimo global, encontrándose un conjunto de pesos que realiza la tarea de asociación lo mejor posible.

La regla delta

Una implementación de esta idea es la desarrollada por [Widrow y Hoff \(1960\)](#), también denominada *regla delta* por [Hinton, McClelland y Rumelhart \(Rumelhart et al., 1986\)](#). La misma utiliza una variante de la ecuación 2.13. En vez de trabajar con la función error para todos los pares k , define la función error para cada uno de los pares k :

$$e^k = \frac{1}{2} (y^k - g^k)^2 \quad (2.15)$$

cuyo gradiente es:

$$\nabla e^k = (y^k - g^k) f^{kT} \quad (2.16)$$

y por lo tanto, la forma que toma la variación en los pesos sinápticos para cada par k es:

$$\Delta W^k = -\eta \nabla e^k = -\eta (y^k - g^k) f^{kT} \quad (2.17)$$

Este método de aprendizaje fue el utilizado en esta tesis para entrenar memorias asociativas lineales. Una deducción detallada de la ecuación 2.16 se presenta en el apéndice A.1.

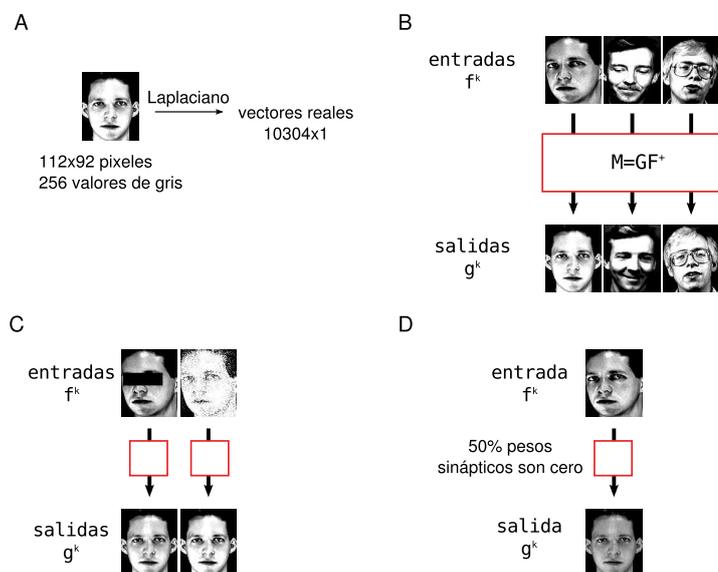


Figura 2.3: Ilustración del funcionamiento de una memoria asociativa lineal. Se asociaron vectores que corresponden a fotografías de personas tomadas de la base de datos de [Samaria y Harter \(1994\)](#). A. Partiendo de las imágenes de las caras, se calculó el Laplaciano para reforzar los bordes y volver menos correlacionadas las imágenes. Después se convirtió la imagen en un vector anexando en una sola todas las columnas de la matriz original. B) Se construyó una matriz asociando pares de vectores que pertenecen a distintas fotografías de la misma persona, utilizando la ecuación 2.8. El desempeño de la memoria es satisfactorio tanto ante entradas incompletas y ruidosas (C), como ante una destrucción de la mitad de las sinapsis (D), mostrando tolerancia ante el deterioro tanto en las entradas como en la integridad de la red.

2.2.5 Propiedades de las memorias lineales

En la figura 2.3 se ilustran dos de las principales propiedades que exhiben los modelos de memorias asociativas lineales. Cabe destacar que las mismas no son exclusivas de este tipo particular de red neuronal, sino que son compartidas por muchos otros modelos. Éstas y otras propiedades han sido señaladas como importantes a la hora de utilizar los modelos de redes neuronales en contraposición con otras aproximaciones al modelado cognitivo, como por ejemplo enfoques simbólicos o probabilísticos ([Bechtel y Abrahamsen, 2002](#)).

Accesible por contenido

En primer lugar, este tipo de memorias son accesibles por contenido, en oposición a las memorias accesibles mediante índices. Por ejemplo, en las memorias implementadas en las computadoras, cada elemento se almacena en un lugar específico. Para acceder a esta información, es preciso contar con un índice, en este caso la dirección física donde se encuentra almacenada.

En contraste, el acceso por contenido permite acceder a la información con una clave que no especifica dirección o instrucción alguna, sino que consiste, en el caso de las memorias en cuestión, en un vector de entrada f . En el caso particular de una memoria autoasociativa, en los que se almacenan los vectores asociados a sí mismos ($g^k = f^k$), es posible acceder a un vector f^k , usando como clave de búsqueda el propio f^k .

La utilidad de las memorias accesibles por contenido se ilustra en la figura 2.3 C. Si se presenta a una memoria lineal construida usando la ecuación 2.7 con una versión levemente alterada de un vector almacenado $f^{*l} = f^l + \epsilon$, donde $\|\epsilon\| \ll 1$ la respuesta de la red será:

$$\begin{aligned} g &= Mf^{*l} = Mf^l + M\epsilon \\ &= g^l + \sum_{k \neq l} g^k \langle f^{kT}, \epsilon \rangle \\ &= g^l + \delta \end{aligned} \quad (2.18)$$

donde entonces δ es una combinación lineal del resto de patrones almacenados g^k , cuyos coeficientes serán $\langle f^k, \epsilon \rangle \ll 1$. Si la perturbación no es muy grande, la respuesta de la memoria será correcta. En implementaciones concretas, y en particular si la dimensión de las memorias es suficientemente grande, incluso utilizando vectores de entrada sustancialmente perturbados (incompletos, por ejemplo), el desempeño de la red es aceptable. Es más, si la correlación lineal entre la perturbación y el vector almacenado es muy pequeña o cero, la respuesta de la memoria tiene el efecto de limpiar la señal del ruido.

Resistencia al deterioro

Una propiedad muy común en muchos sistemas biológicos es la resistencia frente a perturbaciones, incluido el deterioro estructural. El tejido neural no escapa a esta regla. A partir de la evidencia de la clínica neurológica y de los estudios de lesiones en roedores (Lashley, 1950), se ha demostrado que el cerebro es capaz de tolerar cierto grado de deterioro o destrucción, perdiendo gradualmente la funcionalidad a medida que el grado de daño se hace mayor. Esta resistencia al deterioro estructural está también presente en los modelos de memorias lineales (Kohonen, 1972), como se ilustra en la figura 2.3 D. Allí se muestra que cuando se destruyen hasta el 50% de las sinapsis (igualando a 0 el valor de los pesos sinápticos), la salida de la red no sufre grandes cambios.

Aprendizaje y generalización

Otra interesante propiedad de los modelos de redes neuronales, es la incorporación de un mecanismo de aprendizaje que sea capaz de llevar la

x	y	xORY	xXORY
1	1	1	0
1	0	1	1
0	1	1	1
0	0	0	0

Tabla 2.1: tablas de verdad par las funciones OR y XOR.

red desde un estado de pobre desempeño hasta uno de alto desempeño, ya sea incorporando mecanismos de plasticidad sináptica inspirados en la neurobiología de las neuronas reales, o agregando información proveniente de un monitor externo del desempeño (véase la sección 2.2.3).

La mayoría de los modelos logran aprender a través de la repetición un número muy grande de veces de todo un conjunto de patrones a asociar. No obstante, en muchas situaciones el cerebro es capaz de almacenar información en una memoria después de tan sólo una información. A este fenómeno se le suele llamar *one-shot learning*, y si bien existen modelos de redes neuronales que implementan estas estrategias, se trata de modelos desarrollados para tareas específicas que no se integran fácilmente en el marco de una teoría general.

2.2.6 Separabilidad lineal

En un libro que tuvo gran repercusión en su momento, Marvin Minsky y Seymour Papert (1969) señalaron una importante limitación de los perceptrones, y que comparten todos los modelos de neuronas lineales: su incapacidad para clasificar patrones que no son linealmente separables. Un ejemplo de tal tarea, por ejemplo, es el cálculo de la función lógica “O Exclusivo” (XOR, por *exclusive or*). En la figura 2.4 se ilustra el concepto de separabilidad lineal. En suma, los perceptrones son capaces de particionar el espacio de entrada usando hiperplanos, lo que limita enormemente sus capacidades computacionales. Y de forma más importante, operaciones como el XOR se pueden realizar de forma trivial por sistemas simbólicos. La tabla 2.1 muestra la tabla de verdad para las funciones OR y XOR, donde 1 equivale a *verdadero*, y 0 a *falso*.

Perceptrones multicapa

La crítica de Minsky y Papert en cierta forma tiñó de un aura de incertidumbre no sólo a los perceptrones, sino en buena medida al conjunto de modelos de redes neuronales. Algunos autores argumentan que durante algunos años, el desarrollo de los modelos de redes se desaceleró a causa de esta crítica, que hacía parecer a este enfoque poco propicio para emular las capacidades cognitivas humanas.

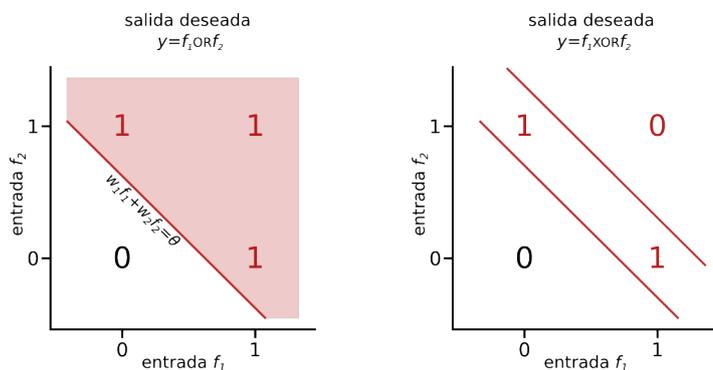


Figura 2.4: Diagrama de entradas y salida de una neurona tipo McCulloch-Pitts, en los que se plantea la implementación de las funciones lógicas binarias A) OR, y B) XOR. Para la primera función es posible encontrar un conjunto de pesos y umbral que implemente la función ya que los valores iguales de la función pueden separarse por una línea en el diagrama de entradas. En cambio, para el XOR esto no es posible, dado que ninguna línea podrá separar los unos de los ceros.

Una solución al problema de la separabilidad lineal consiste en añadir una capa (o varias) extra de neuronas, interpuesta entre las entradas y la capa de salida, denominada *capa oculta*. En su crítica, Minsky y Papert reconocen que con la inclusión de capas ocultas el problema de la separabilidad lineal se solucionaba; sin embargo, señalaron que no existían hasta el momento algoritmos que permitieran entrenar ese tipo de arquitecturas.

La solución a este problema tardó varios años en llegar. A mediados de la década de 1980, Rumelhart y McClelland publicaron una obra que sería muy importante para el resurgimiento del desarrollo de las redes neuronales, fundando el paradigma *conexionista*. En su obra *Parallel Distributed Processing* (McClelland y Rumelhart, 1986), compilan varios trabajos, entre los que se destaca la solución al problema del entrenamiento de perceptrones multicapa, el algoritmo de *backpropagation*.

Otra propuesta, dada en el contexto de las memorias lineales, fue planteada por la misma época en Montevideo, y es el tema de la siguiente sección.

2.2.7 Memorias matriciales contexto-dependientes

Una solución alternativa a la restricción impuesta por la necesaria separabilidad lineal del conjunto de entrenamiento fue dada por Eduardo Mizraji hacia finales de la década de 1980 (Mizraji, 1989, 1992). En vez de agregar capas ocultas, plantea una extensión de las memorias asociativas lineales que incorpora la presencia de dos entradas, f y c , donde la segunda actúa como contexto de la primera. De esta forma, es posible almacenar diferentes respuestas g para una misma entrada f , utilizando diferentes contextos c . La clave está en el uso del producto de Kronecker entre la entrada y el

contexto $f \otimes c$:

$$g = M(f \otimes c)$$

La memoria entonces almacena una única respuesta para cada producto $f \otimes c$. Esto implica o bien una capa de preprocesamiento que realice el producto entre entrada y contexto, o bien, que el producto entre f y c se compute en la misma capa de salida (Mizraji *et al.*, 1994).

En el caso en que las entradas y los contextos sean conjuntos ortonormales, la memoria funciona en forma perfecta si la matriz adquiere la forma:

$$M = \sum_{k,j} g^{kj} (f^k \otimes c^j)^T \quad (2.19)$$

Cuando se introduce una dupla f^m y c^l en la memoria, la salida es la deseada (g^{ml}):

$$\begin{aligned} g &= M(f^m \otimes c^l) = \sum_{k,j} g^{kj} (f^k \otimes c^j)^T (f^m \otimes c^l) \\ &= \sum_{k,j} g^{kj} (f^{kT} \otimes c^{jT}) (f^m \otimes c^l) \\ &= \sum_{k,j} g^{kj} (f^{kT} f^m \otimes c^{jT} c^l) \\ &= \sum_{k \neq m, j \neq l} g^{kj} (f^{kT} f^m \otimes c^{jT} c^l) + g^{ml} (f^{mT} f^m \otimes c^{lT} c^l) \\ &= \sum_{k \neq m, j \neq l} g^{kj} \langle f^k, f^m \rangle \langle c^j, c^l \rangle + g^{ml} \langle f^m, f^m \rangle \langle c^l, c^l \rangle \\ &= g^{ml} \end{aligned} \quad (2.20)$$

Este formalismo permite construir modelos capaces de implementar una versión vectorial de la lógica proposicional y modal (Mizraji, 1992). En particular, es capaz de resolver problemas en los que no existe separabilidad lineal de las entradas, como en el caso del O Exclusivo. En algunos casos puede emular el funcionamiento de una red de capa oculta entrenada con *backpropagation*, con la ventaja de que existe un método analítico para construir la matriz de pesos (ecuación 2.19), y además puede entrenarse usando la regla delta de Widrow-Hoff (véase la sección 2.2.4). En esta tesis hemos utilizado estos modelos para investigar aspectos de la organización semántica, basándonos en trabajos previos del laboratorio, uno de los cuales introduciremos en la siguiente sección.

Implementación del XOR

Para ilustrar el poder de este formalismo, presentamos a continuación una implementación en una única capa de neuronas de la función XOR. La misma se enmarca en la teoría de la *lógica vectorial* desarrollada por Mizraji (1992).

x	y	$xXORy$
s	s	n
s	n	s
n	s	s
n	n	n

Tabla 2.2: tabla de verdad para la función XOR con los vectores s y n tomando el rol de “verdadero” y “falso”.

Para ello, supongamos la existencia de un conjunto ortonormal de dos vectores s y n , asimilables a los valores “verdadero” y “falso”:

$$\{s, n\} \in \mathbb{R}^n$$

En este contexto, la tabla de verdad de la función XOR puede escribirse como se ilustra en la tabla 2.2. Puede pensarse en los vectores x e y como un par de vectores de entrada y contexto que están asociados a sus correspondientes patrones de salida $xXORy$. Entonces puede construirse la matriz X que almacene estas asociaciones, utilizando la ecuación 2.19, que adquiere la siguiente forma:

$$X = n(s \otimes s)^T + s(s \otimes n)^T + s(n \otimes s)^T + n(n \otimes n)^T \quad (2.21)$$

La función XOR puede obtenerse entonces calculando la salida g ante la entrada de una dupla de vectores tomados del conjunto s, n . Puede verificarse entonces que:

$$\begin{aligned} X(s \otimes s) &= n(s \otimes s)^T(s \otimes s) + s(s \otimes n)^T(s \otimes s) + s(n \otimes s)^T(s \otimes s) + n(n \otimes n)^T(s \otimes s) \\ &= n\langle s, s \rangle \langle s, s \rangle + s\langle s, s \rangle \langle n, s \rangle + s\langle n, s \rangle \langle s, s \rangle + n\langle n, s \rangle \langle n, s \rangle \\ &= n \\ X(s \otimes n) &= s \\ X(n \otimes s) &= s \\ X(n \otimes n) &= n \end{aligned} \quad (2.22)$$

lo que se corresponde con la tabla 2.2.

Este mismo procedimiento puede utilizarse para definir operadores monádicos y diádicos de la lógica proposicional, pudiendo desarrollarse a partir de los mismos, toda una lógica formulada en torno a vectores y matrices denominada “lógica vectorial” (Mizraji, 1992; Mizraji *et al.*, 2009)

2.3 Redes neuronales y lenguaje

Uno de los campos cognitivos en los que las redes neuronales han servido como motor impulsor de los esfuerzos de modelado en general, es el del lenguaje. En esta sección revisaremos uno de los modelos de red desarrollado para explicar fenómenos lingüísticos que ha tenido mayor impacto en la literatura, y en el que se basan los modelos desarrollados en esta tesis. Junto a otras propuestas, ha contribuido a generar intensos debates y potenciado numerosos desarrollos posteriores que han ilustrado la potencia de las redes neuronales como modelos de la cognición.

2.3.1 Red recurrente simple de Elman

Una de las principales limitaciones de las redes de arquitectura “feedforward” es que el procesamiento de cada entrada no está influenciada por la entrada anterior. Es decir, no son capaces de identificar o procesar dependencias temporales o de orden en las entradas. El lenguaje, sin embargo, se caracteriza por la presencia de fuertes dependencias temporales, incluso a largas distancias. Por ejemplo, en español la forma del verbo debe coordinarse con la naturaleza del sujeto que lo precede en persona y número. Incluso, si entre el sujeto y el verbo se interpone una cláusula relativa, como en la siguiente oración:

1. El *perro*_{1s}, que se había quedado dormido en la alfombra junto al calor de la estufa, *bostezó*_{1s} al oír su voz.

en la que el verbo debe tener conjugación en primera persona del singular al igual que el sujeto. Este tipo de dependencias son obvias en la sintaxis, pero operan también a nivel fonológico, semántico y pragmático. Es más, la presencia de correlaciones y dependencias temporales es una característica intrínseca de las funciones cognitivas en general.

Para lograr manipular este tipo de dependencias es necesario contar con una suerte de memoria de trabajo, que almacene la historia de las entradas, al menos durante un tiempo. Una arquitectura que implementa esta noción es la *red recurrente simple* (SRN por *simple recurrent network*), y fue propuesta originalmente por Jordan (1986). Sin embargo, la versión introducida por Jeffrey Elman en 1990 es la que tuvo mayor repercusión, y es utilizada usualmente en tareas lingüísticas.

La arquitectura de la red puede verse en la figura 2.5: es una red “feedforward” de capa oculta, a la que se le agrega una capa de contexto que recibe una copia de la salida de la capa oculta. Esta capa de contexto releva esta señal y la incorpora como entrada nuevamente en la capa oculta formando un bucle de retroalimentación. Esto determina que la capa oculta no sólo recibe aferencias de la capa de entrada, sino que recibe una copia de su actividad en la iteración anterior. De esta forma, la red procesa no sólo la

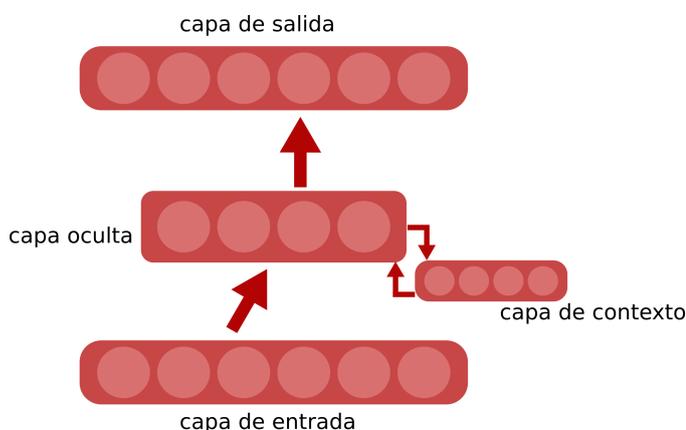


Figura 2.5: Arquitectura de una red recurrente simple (Elman, 1990). Se trata de una red de capa oculta, con una capa extra (capa de contexto) que se encarga de ingresar a la capa oculta la actividad de la misma en el tiempo anterior.

actividad de entrada en un momento dado, sino que retiene la historia de los procesamientos pasados.

Aprendiendo categorías gramaticales

En su trabajo original, Elman (1990) utilizó una red recurrente simple para predecir la siguiente palabra en un conjunto de oraciones. Las mismas fueron originadas a partir de “moldes” que especificaban qué tipo de constituyentes podían ocurrir en cada posición. Algunos ejemplos de los “moldes” son los siguientes:

- NOMBRE_{humano} VERBO_{intransitivo}
- NOMBRE_{animado} VERBO_{transitivo} NOMBRE_{animado}
- NOMBRE_{animado} VERBO_{COMER} NOMBRE_{comida}

Estos “moldes” los rellenó aleatoriamente con un conjunto de palabras (nombres y verbos) que pertenecían a diferentes categorías gramaticales con distintos grados de inclusión. Por ejemplo, la categoría de nombre “animado” incluía nombres de criaturas míticas, animales y personas, mientras que la categoría “humano” sólo incluía nombres de personas.

Luego de entrenar a la red, el desempeño de la red se aproximó al óptimo: las salidas consistían en combinaciones lineales de las palabras que podían ocurrir, ponderadas por la frecuencia de ocurrencia en el corpus de oraciones de entrenamiento. Es decir, la red podía aprender las probabilidades de transición entre las palabras que se desprendían de la estructura de los “moldes” y las categorías gramaticales utilizadas.

Un análisis de las activaciones de la capa oculta frente la aparición de las distintas palabras muestra cómo la red puede lograr realizar la tarea con tal eficiencia. Como puede verse en la figura 2.6, las categorías gramaticales de las palabras se reflejan en la actividad de la capa oculta: palabras que pertenecen a la misma categoría producirán activaciones similares, transformando a la representación de palabra en una representación aproximada de categoría gramatical, la cual es la única determinante de la distribución de las palabras en las oraciones.

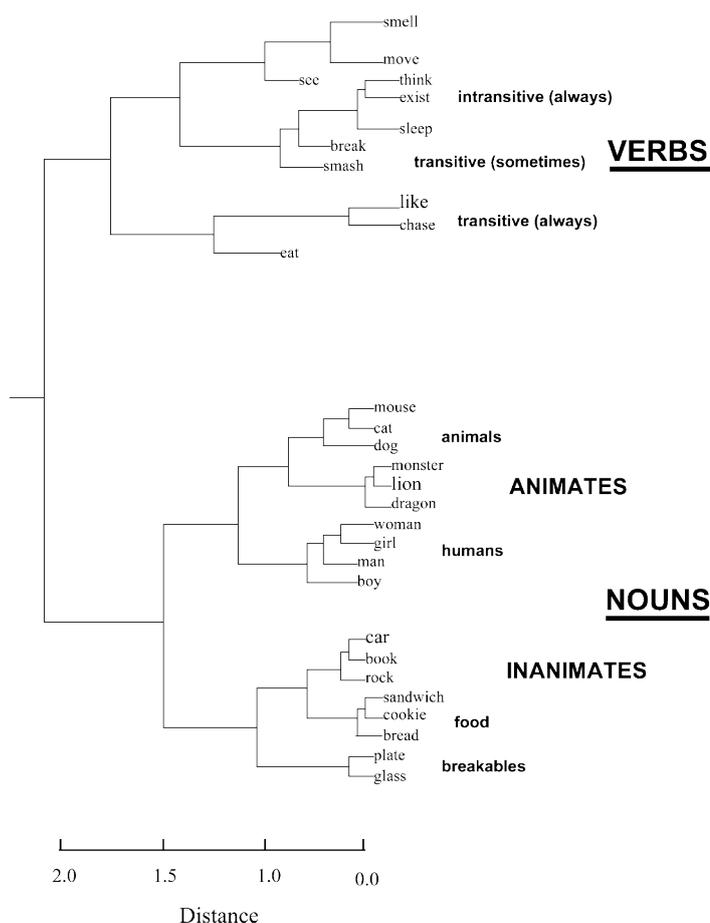


Figura 2.6: Dendrograma de agrupamiento jerárquico de las activaciones de la capa oculta obtenido por Elman (1990) ante las palabras del corpus. Puede verse que emergen de este análisis las diferentes categorías gramaticales, desde las más globales (verbos y nombres), hasta las más refinadas (animales y humanos).

En trabajos posteriores, algunos autores lograron capturar dependencias más sofisticadas, como las generadas por gramáticas más complejas, que incorporan por ejemplo cláusulas relativas (Elman, 1993, 1991; Lewis y Elman, 2001). Algunos trabajos plantean que las limitaciones exhibidas por estos modelos remedan las limitaciones presentes en hablantes, permitien-

do considerar a las redes recursivas simples como un potente modelo de procesamiento de oraciones (Christiansen y Chater, 1994).

El propio Elman (2009) ha llevado la relevancia de estos modelos un paso más allá, al argumentar que ilustran la posible existencia de conocimiento léxico sin la existencia de un léxico propiamente dicho. El argumento incorpora el hecho de que el procesamiento de cada palabra se ve influenciado (en el modelo, y posiblemente en la realidad) por las palabras que la anteceden. En tal sentido, el significado de cada palabra es función del significado del contexto precedente, y por tanto, la noción de léxico como almacén (estático) de información semántica (así como sintáctica y fonológica) de una palabra no resulta adecuada.

Más allá de este planteo polémico, las redes recurrentes simples son una herramienta utilizada regularmente en el estudio del lenguaje normal y patológico (Botvinick y Plaut, 2004; Reali y Christiansen, 2005).

En el presente trabajo de tesis (capítulo 6), utilizamos dos variaciones de redes recurrentes simples, que a diferencia de ser redes multicapa entrenadas mediante *backpropagation*, son memorias asociativas contexto-dependientes, donde la salida actúa como contexto de la siguiente entrada (véanse las secciones 3.4.1 y 4.3.2).

Capítulo 3

Modelos semánticos y la noción de tema

Cuando uno habla produce una serie ordenada de palabras que transmiten una o varias ideas, e invariablemente, todas ellas tienen algo en común, están insertas en un contexto, se está hablando de un tema, de un tópico. El tópico o tema caracteriza el contenido de un fragmento de discurso o documento, el “espíritu”, de “lo que habla” el texto. Por ejemplo, podemos decir que una persona está hablando del precio del carbón en los últimos 20 años, o un texto trata de la independencia de Angola.

Según algunos autores (Kintsch, 2002; van Dijk, 1993), el tema o tópico puede concebirse como una suerte de “resumen” de un segmento de texto. En el modelo de comprensión lectora denominado “construcción-integración” (Kintsch, 1988; van Dijk, 1993) se propone que de cada párrafo o sección de un texto se obtienen “macro-oraciones” que capturan el rasgo principal de ese segmento. Estas macro-oraciones se obtienen través de ciertas reglas de derivación (“macro-reglas”). Estas macro-oraciones pueden servir a modo de resumen del texto, ya que son una descripción concisa del texto original. A su vez, a partir de estas macro-oraciones puede construirse una estructura de orden superior, que logra abarcar o resumir el contenido o significado que comparten, dando lugar a una descripción aún más sintética del texto. Este proceso de abstracción y síntesis, puede repetirse hasta lograr una única descripción del texto, o sección, y podremos decir entonces que ése es el “tema” o “tópico” del que trata el texto.

Por ejemplo, considérese el siguiente fragmento de texto¹, analizado en (van Dijk, 1993):

Fairview estaba muriéndose. En el pasado, había sido una pequeña ciudad emprendedora y próspera y sus dos grandes

¹“Just the way it is”, novela policíaca de James Headley Chase, publicada en 1944.

fábricas, especializadas en herramientas, habían sido una fuente lucrativa de riqueza.

Ahora Fairview había pasado ya su edad de oro. La producción masiva se había encargado de eso. Los métodos de producción de la pequeña ciudad no podían competir con las fábricas modernas que habían surgido de la noche a la mañana en los distritos vecinos.

La producción masiva y Bentonville habían acabado con Fairview. Bentonville era una ciudad fabril de rápida expansión a unas treinta millas. Era una ciudad que surgió como las setas. Una ciudad para las generaciones jóvenes con tiendas brillantemente pintadas, limpias casitas de campo baratas, rápidos tranvías y un joven, vigoroso y palpitante espíritu comercial.

La juventud de Fairview se había marchado o a Bentonville o más al norte; algunos incluso se habían ido a Nueva York. Los negocios más progresivos se habían transferido a Bentonville tan pronto como el escrito apareció en la pared. Sólo quedaban allí las tiendas menos emprendedoras, defendiéndose lo mejor que podían.

Fairview estaba derrotada. Se podía ver en las casas en ruina, las carreteras sin arreglar y la calidad de los productos en los escaparates. Se podía ver en la digna ruindad de la pequeña colonia de hombres de negocios retirados que habían prosperado en la edad dorada y se contentaban con terminar sus días en esta triste y estancada pequeña ciudad. Y se podía ver particularmente en el número de parados que se reunían en las esquinas, indiferentes y apáticos.

En su análisis, [van Dijk \(1993\)](#) argumenta que si un hablante nativo es cuestionado acerca de cuál es el tema o tópico de este pasaje, entre las posibles respuestas se encuentran “La decadencia de Fairview”, “Fairview, una pequeña ciudad”, o “La decadencia de Fairview, debido a la producción masiva y a la competencia de una ciudad vecina, Bentonville”. De hecho, algunas frases del propio fragmento expresan estas ideas:

- *Fairview se estaba muriendo.*
- *Fairview había pasado ya su edad de oro.*
- *La producción masiva y Bentonville habían acabado con Fairview.*
- *Fairview estaba derrotada.*

De alguna manera, estas frases capturan algo especial del contenido del todo. En particular, todas hacen referencia a *Fairview*, la ciudad que es objeto

de la mayoría de las oraciones del fragmento. Puede pensarse, además, que los conceptos “derrotada”, “se estaba muriendo”, “pasado ya su edad de oro” y “acabado con”, son parte de “decadencia”. Por tanto, “la decadencia de Fairview” es una proposición que está relacionada con casi todo el fragmento. Estas relaciones, muchas veces, dependen en los modelos mentales detrás de expresiones como “decadencia”: en este caso, la transición desde un tiempo de prosperidad a un tiempo de no prosperidad, probablemente causado por uno o más factores. En el caso de una ciudad, eso se traduce en un cambio en la apariencia y estado físico de las estructuras que componen la ciudad, sus comercios, sus habitantes, etc.

Podemos ver a partir de este ejemplo, que la definición de tópico depende en cierta medida del grado de granularidad con que se enfoque al texto o discurso. Así, en los dos primeros párrafos se habla de “Fairview, su pasado próspero y su decadencia”, en los dos siguientes se habla de que “Fairview no puede competir con Bentonville”. Sin embargo, como vimos antes, bien puede considerarse que “La decadencia de Fairview” representa el tópico de todo el fragmento.

A mayor escala, el libro “El origen de las especies” de Charles Darwin trata sin duda del problema de la evolución de las especies, y podríamos describir su “tema” como “evolución”, o “evolución de las especies” (suele argumentarse que el libro no trata, a pesar de su título sobre el origen de las especies). Sin embargo, cada uno de los capítulos trata de un tema particular, a veces “incluido” dentro de éste gran tema de la evolución: por ejemplo, el primer capítulo trata de los criadores y la mejora de razas.

Además de la identificación de un tópico, un hablante es capaz de detectar cambios en el tópico, o producir estos mismos cambios al hablar. En algún momento, las proposiciones que se introducen en el discurso comienzan a no poder integrarse al tópico vigente. Así cuando se introduce a Bentonville en el ejemplo anterior, no está claro si se continuará dentro del tópico vigente (“la decadencia de Fairview”) o si pasará a enfocar la atención en Bentonville. Luego queda clara la relación entre Bentonville y el tópico vigente (una relación causal). En contraste, en el fragmento siguiente se introduce un concepto que dará origen a un nuevo tópico:

Pero todavía se podía encontrar una chispa de vida en Fairview (...) cuando Fairview estaba en la cima de su prosperidad, Harman había fundado un periódico para la ciudad (...)

La aparición de “una chispa de vida” es tan contrastante con la noción de “decadencia de Fairview” que parece dar inicio a un nuevo tópico, a la vez que introduce un nuevo elemento, el periódico. De pasajes ulteriores, se desprenderá en este caso que el periódico es parte del tópico subsiguiente.

3.1 Detección automática de tópicos

Si se quisiera contar con un método automático mediante el cual detectar el tópico de un fragmento de texto o discurso, se deben resolver por lo menos dos problemas. El primero, es el de la naturaleza jerárquica de los temas, pudiendo asignársele un tema a todo un discurso, a la vez que se pueden reconocer sub-temas en diferentes porciones del mismo. Así, la asignación de fragmentos de texto a un tópico o a otro dependerá del grado de granularidad elegido para esta asignación.

Sumado a este problema de granularidad, se encuentra el de la delimitación, es decir, cuándo empieza un tema y termina otro. Muchas veces los textos suelen estar organizados en párrafos, secciones, capítulos, etc, de modo que es posible guiarse por esas marcas para establecer los límites entre temas (aunque esto no es necesariamente algo deseable). En el discurso oral², sin embargo, estas marcas no se encuentran, y el problema de dónde trazar estos límites carece de guías estructurales. En cualquier caso, este problema es conocido en las ciencias de la información como el de *text segmentation*, y existen diferentes algoritmos destinados a tal fin (Choi, 2000).

En lo que resta del capítulo se presentarán algunas metodologías que permiten detectar automáticamente la estructura de tópicos de un texto o discurso, sentando las bases de lo que se presentará en los capítulos de resultados acerca de la evaluación de la coherencia discursiva en pacientes con esquizofrenia.

3.2 Espacios semánticos

Se conocen como espacios semánticos a un tipo de representación del significado de expresiones lingüísticas³ que consiste en el mapeo de estas expresiones a coordenadas en un espacio vectorial, usualmente de dimensión en el orden de los cientos (típicamente cercano a 300). En estos espacios, la similitud semántica está dada por la distancia entre distintos puntos en el espacio.

En su primera formulación, el espacio vectorial surgía con naturalidad atribuyendo a cada dimensión del mismo un *rasgo* semántico (*semantic feature*) (Sahlgren, 2006). Osgood (1952) fue el primero en proponer la representación del significado a partir de la obtención de normas de rasgos semánticos. A

²Puede pensarse en que la transcripción de un discurso oral es un texto, y aquí no se distinguirá en principio entre uno y otro. En todo caso, en las transcripciones no se cuentan con demarcaciones de estructura como párrafos o secciones, a menos que sean introducidas *a posteriori* por quien transcribe, utilizando algún criterio de delimitación que justamente es lo que se pretende emular

³En realidad, en muchos casos no se representa el significado en sí mismo sino la relación de similitud entre significados. Es decir, las dimensiones de este espacio no siempre son directamente interpretables.

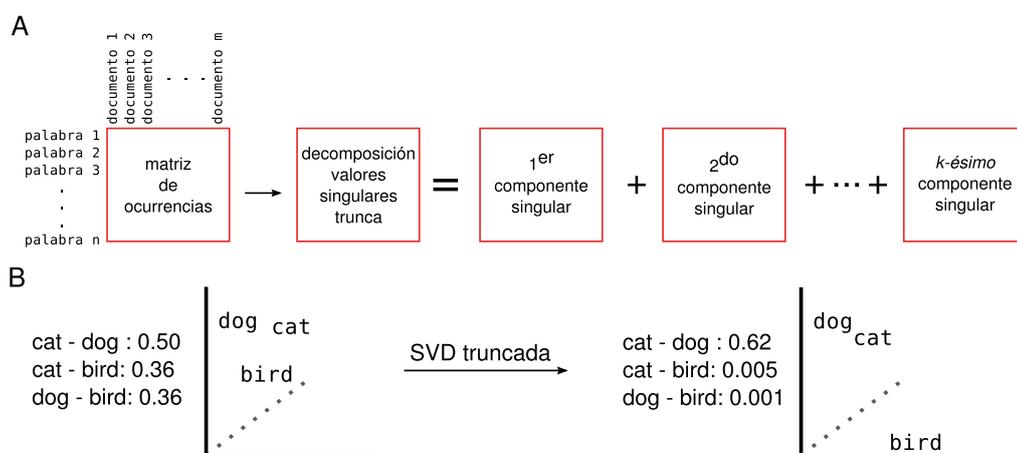


Figura 3.1: Obtención de un espacio semántico mediante Análisis Semántico Latente. A. Se construye una matriz de ocurrencias de palabras en documentos. Generalmente se aplica una transformación a los valores de frecuencia. Se realiza la descomposición en valores singulares (svd), reteniendo sólo los k primeros (pero véase [Valle-Lisboa y Mizraji \(2007\)](#) para una propuesta alternativa). B. Ilustración del efecto del truncado del svd sobre el comportamiento del espacio. Las distancias entre las palabras (coseno del ángulo entre los vectores) se ajustan a su relación semántica.

partir de llevar cada rasgo a una dimensión en un espacio vectorial, cada palabra se vuelve un punto en dicho espacio (una combinación lineal de las distintos rasgos). El uso de rasgos semánticos es particularmente útil para representar palabras que muestran un alto grado de concreitud, como por ejemplo nombres que denotan objetos concretos, o acciones de naturaleza física ([Barsalou et al., 2008](#); [Riordan y Jones, 2011](#)).

En décadas recientes, sin embargo, se han utilizado espacios semánticos derivados no de información explícita acerca del significado de las palabras, sino de información acerca de su co-ocurrencia o asociación. En ellas se incorporan ideas propuestas por lingüistas como Zellig Harris ([1968](#)) acerca de la posibilidad de caracterizar a las palabras por sus estadísticas de distribución. En particular, palabras que co-ocuran en un mismo contexto tendrán características similares⁴. En cuanto al significado, es esperable que tanto “auto” como “coche” sean utilizadas en contextos muy similares, a diferencia de estas palabras y “monje”, por ejemplo. Este tipo de información se suele reunir en matrices de ocurrencia de palabras en contextos: típicamente, los contextos suelen ser documentos o párrafos. El uso de la información distribucional puede ser usado como complemento a aproximaciones basadas en rasgos semánticos para generar espacios semánticos de mayor desempeño ([Recchia y Jones, 2009](#)).

⁴Esta noción puede trazarse a las relaciones sintagmáticas y paradigmáticas de Saussure ([de Saussure, 1990](#); [Dennis y Harrington, 2001](#))

3.2.1 Análisis Semántico Latente

Entre los métodos más utilizados para generar espacios semánticos a partir de las estadísticas de co-ocurrencia de palabras en grandes colecciones de texto se destaca el Análisis Semántico Latente (LSA por sus siglas en inglés) (Deerwester *et al.*, 1990; Landauer y Dumais, 1997). Esta técnica consiste en realizar la descomposición en valores singulares de la matriz de ocurrencias, reteniendo sólo algunos de los vectores singulares de mayor valor propio. La matriz aproximada define un espacio semántico en el cual dos palabras pueden tener coordenadas similares a pesar de no compartir directamente contexto alguno (en este caso seguro comparten contexto con otras palabras que sí co-ocurren entre sí, y el método captura así relaciones indirectas o “latentes” Landauer y Dumais (1997); Valle-Lisboa (2007)). En la figura 3.1 se esquematiza el procedimiento de obtención de un espacio semántico utilizando este método.

La descomposición en valores singulares de la matriz de ocurrencias X , tiene la forma:

$$X = \sum_{i=1}^r u_i \sigma_i v_i^T = U \Sigma V^T \quad (3.1)$$

Esta descomposición es única, y las matrices U , Σ y V son únicas y de rango r , el rango de la matriz X . Las matrices U y V retienen los vectores singulares u_i y v_i y Σ los valores singulares σ_i .⁵ Para obtener un espacio semántico, es necesario retener sólo los primeros k valores y vectores singulares:

$$\tilde{X} = \sum_{i=1}^k u_i \sigma_i v_i^T = U_k \Sigma_k V_k^T. \quad (3.2)$$

3.2.2 Limitaciones

Los espacios semánticos han demostrado emular el desempeño humano en algunas tareas psicolingüísticas, tales como la evaluación de similitud semántica, facilitación (*priming*) semántico, y patrones de asociación libre, entre otros (Dumais, 2003; Jones *et al.*, 2006; Landauer y Dumais, 1997; Lund *et al.*, 1995; Steyvers *et al.*, 2004; Steyvers y Tenenbaum, 2005).

Sin embargo, los espacios semánticos tienen algunas limitaciones importantes. Una de las más importantes es la naturaleza métrica de la similitud semántica (Tversky, 1977). En un espacio semántico, cualquier métrica D que se defina necesariamente debe respetar tres propiedades:

- La distancia de un elemento consigo mismo debe ser cero: $D(a, a) = 0$.
- La distancia es simétrica: $D(a, b) = D(b, a)$.

⁵De modo que si u y v son vectores singulares para un mismo valor singular σ (o sea que corresponden a la misma columna de U y V respectivamente), entonces $u^T X = \sigma v$, y $X v = \sigma u$.

- Las distancias cumplen la desigualdad triangular:

$$D(a, b) + D(b, c) \geq D(a, c).$$

La validez de las propiedades de simetría y desigualdad triangular han sido cuestionadas en varias oportunidades (Griffiths *et al.*, 2007; Tversky y Gati, 1982). Los juicios de similitud semántica suelen ser fuertemente asimétricos; por ejemplo, un ornitorrinco se percibe como más similar a un gato de lo que un gato a un ornitorrinco, o juzgamos que un retrato se parece más a la persona, que la persona se parece al retrato. Entre otros aspectos, la tipicalidad y la saliencia de los diferentes conceptos pueden subyacer estas asimetrías. Algo similar ocurre en las normas de asociación libre de palabras (Nelson *et al.*, 1998; Steyvers *et al.*, 2004), donde por ejemplo, *estufa* es una respuesta asociativa más frecuente ante *invierno* que a la inversa.

La desigualdad triangular es más difícil de interpretar, pero en todo caso implica que si a y b están relacionadas, así como b y c , entonces la distancia entre a y c no puede ser muy grande. Algunos análisis de juicios semánticos y asociación de palabras sugieren que esto no siempre es así (Steyvers *et al.*, 2004; Tversky, 1977; Tversky y Gati, 1982).

Es cierto que las nociones de distancia o similitud semántica son implementadas en los espacios semánticos mediante instrumentos que no califican como métricas, como por ejemplo la “distancia” coseno, esto es, el coseno del ángulo que forman dos palabras o fragmentos. De todas maneras, los anteriores cuestionamientos referidos a las métricas siguen siendo relevantes. En particular, el coseno de un ángulo entre dos vectores es igual sin importar el orden en que se los considere; $\cos(\theta) = \cos(180 + \theta)$ ⁶, y también respeta la desigualdad triangular.

Otro problema que se ha señalado refiere a la estructura topológica de similitudes semánticas o de asociación entre palabras. Si se analiza la topología de las asociaciones libres (Griffiths *et al.*, 2007; Steyvers y Tenenbaum, 2005), puede verse entre otras cosas que el número de asociados que presenta una palabra dada sigue una distribución de tipo ley de potencia⁷. Para analizar el número de palabras relacionadas en un espacio semántico es necesario convertir la noción continua de similitud semántica en una relación discreta (asociados o no). Esto puede hacerse, por ejemplo utilizando un criterio de umbral; si la similitud semántica es superior a ese umbral, entonces se las considera relacionadas. Usando este y otros métodos, Steyvers *et al.* (2004) encontraron que la distribución de asociados que emergía de un espacio semántico no se correspondía con la ley de potencia observada en los datos empíricos.

A pesar de todas estas observaciones, sin embargo, algunos autores plantean que estos inconvenientes pueden solucionarse. Jones *et al.* (2011)

⁶Más específicamente, el coseno es una función monótonica del ángulo, y el ángulo tiene propiedades de métrica.

⁷Esto es, la probabilidad de que una palabra tenga k asociados es proporcional a $k^{-\gamma}$

plantean el uso de un mecanismo de decisión que medie entre las medidas de similitud o distancia semántica provenientes del espacio, y la decisión de considerar como relacionadas o asociadas a dos palabras. Los autores logran, a partir de diferentes espacios semánticos, evadir las restricciones antes descritas y obtener patrones de asociación con propiedades más realistas (asimétricos, que no respetan la desigualdad triangular, y con distribuciones tipo ley de potencia).

3.2.3 Temas en espacios semánticos

La mayoría de los modelos de espacios semánticos no incorporan explícitamente la noción de tópico⁸, y dada la naturaleza continua de sus representaciones espaciales, no es trivial cómo pueden utilizarse para representar el o los tópicos de un fragmento discursivo.

En el contexto del enfoque de “construcción-integración” previamente mencionado, [Kintsch \(2002\)](#) propone usar espacios semánticos (en particular LSA) para estimar cuáles oraciones son las más típicas de un párrafo, o de un documento, y de esa manera obtener una aproximación a la macro-oración a la que puede ser asimilable el tópico subyacente. Una forma de hacer esto es calculando las distancias entre las diferentes oraciones y todo el segmento: una oración típica será la que guarde menor distancia con el todo. Volviendo al ejemplo de la novela de Chase usado al comienzo del capítulo, si se calculan las distancias entre cada oración y todo el fragmento, seguramente las oraciones “La producción masiva y Bentonville habían acabado con Fairview” y “Fairview estaba derrotada” estén entre las más similares al todo, es decir, las más típicas. En función de esto, podría utilizarse alguna de estas oraciones para caracterizar el tópico de todo el texto.

En cuanto a la detección de cambios en el tópico vigente, [Kintsch \(2002\)](#) sugiere evaluar la “conectividad” entre oraciones, es decir, la distancia entre oraciones sucesivas. Como es muy probable que cuando se producen cambios de tópico la naturaleza semántica de los componentes de las oraciones cambien, puede pensarse que una caída en la similitud entre oraciones (o párrafos) consecutivos corresponda a un cambio de tema. Esta aproximación a la detección de cambio de tópico fue analizado en la tesis de maestría de [Cabana \(2009\)](#), con resultados dispares. En el capítulo 7 se presentará una forma diferente de evaluar los cambios de tópicos utilizando espacios semánticos aplicada al problema del análisis del discurso esquizofrénico.

⁸aunque algunos análisis permitan suponer que las dimensiones de los espacios cumplan ese rol; véase [Blei et al. \(2003\)](#); [Hofmann \(1999\)](#); [Papadimitriou et al. \(2000\)](#); [Valle-Lisboa y Mizraji \(2007\)](#) por ejemplo.

3.3 Modelos probabilísticos

A partir del análisis del LSA en términos estadísticos realizado por Hofmann (1999), Papadimitriou *et al.* (2000) establecen las bases de una clase de modelos conocidos bajo el nombre de “modelos probabilísticos de tópicos” (*probabilistic topic models*) (Blei, 2012; Blei *et al.*, 2003; Griffiths y Steyvers, 2011; Griffiths *et al.*, 2007; Steyvers y Griffiths, 2007).

En este enfoque, la noción de tópico adquiere una definición precisa: un tópico z es una distribución de probabilidad sobre palabras w : $P(w|z)$. De esta manera, convierten el problema de la identificación de tópicos en un problema de inferencia estadística. Así, por ejemplo, en un tópico que refiere a “finanzas”, las palabras “banco”, “interés” y “cambio” tendrán altas probabilidades, mientras que en un tópico sobre “deportes”, esas mismas palabras tendrán probabilidad baja, mientras que serán altas las de “ganar”, “torneo” y “jugador”, por ejemplo.

La variable tópico es una variable latente, no directamente observable, que condiciona la ocurrencia de las variables observables, las palabras. Así, la probabilidad de observar una palabra dada w_i estará dada por la suma de las probabilidades condicionales en los distintos tópicos, multiplicadas por las probabilidades a priori de cada tópico.

$$P(w_i) = \sum_{j=1}^T P(w_i|z_i = j)P(z_i = j) \quad (3.3)$$

Sin embargo, la noción de tópico z que se maneja aquí no es directamente asociable a la noción de tópico manejada por Kintsch (2002) y van Dijk (1993). En el caso de los modelos probabilísticos, los tópicos son como dimensiones de significado en las que se inscribe cada fragmento de texto dado, mientras que el tópico según estos últimos autores se correspondería con el *gist* o la idea o espíritu central del texto, como lo denominan (Griffiths *et al.*, 2007). El *gist* g de un documento, en este contexto, es una distribución de probabilidad sobre tópicos: $P(z|g)$. Entonces, la probabilidad de observar una palabra w_i dado un *gist* g se vuelve:

$$P(w_i|g) = \sum_{z_j=1}^T P(w_i|z_j)P(z_j|g). \quad (3.4)$$

La especificación del *gist* de un documento, por tanto, determina un modelo generativo capaz de generar documentos nuevos. Este modelo generativo se resume en la figura 3.2.

A partir de una colección de textos, se pueden calcular las secuencias de tópicos $\mathbf{z} = (z_1, z_2 \dots z_n)$ que maximiza la probabilidad de ocurrir las secuencias de palabras $\mathbf{w} = (w_1, w_2 \dots w_n)$, usando la regla de Bayes:

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}|\mathbf{z})P(\mathbf{z})}{P(\mathbf{w})} \quad (3.5)$$

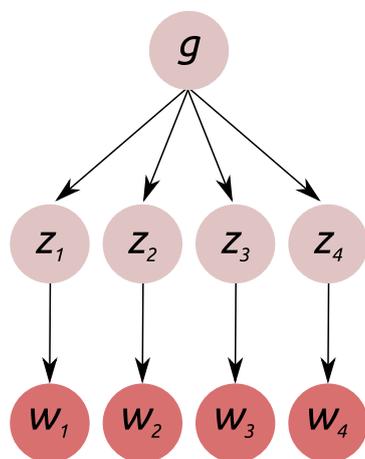


Figura 3.2: Modelo gráfico detrás del modelo probabilístico de tópicos en la versión de Griffiths *et al.* (2007).

En un caso sencillo, en el que cada documento tiene un *gist* que consiste en un sólo tópico z_k , la expresión anterior se vuelve:

$$P(z|\mathbf{w}) = \frac{P(\mathbf{w}|z_k)P(z_k)}{P(\mathbf{w})} \quad (3.6)$$

$$= \frac{\prod_{i=1}^n P(w_i|z)P(z)}{\sum_{j \neq k}^T \prod_{i=1}^n P(w_i|z_j)P(z_j)}, \quad (3.7)$$

donde el producto en el numerador hace que aquellos tópicos que dan probabilidades muy bajas a las palabras observadas tengan baja probabilidad. Dicho de otro modo, un tópico z probable de la secuencia de palabras \mathbf{w} es aquel que asigna altas probabilidades a cada una de ellas. Cuando se observa la palabra “banco”, por ejemplo, tanto los tópicos “finanzas” como “espacios públicos” tienen alta probabilidad; sin embargo, si la siguiente palabra es “fuente”, el tópico “finanzas” tendrá una probabilidad menor.

Usando este tipo de planteos, Griffiths *et al.* (2007) muestran cómo utilizar estos modelos para abordar tres problemas cognitivos diferentes que se presume resolvemos cuando procesamos el lenguaje:

Predicción Predecir la siguiente palabra w_{n+1} dado \mathbf{w} , calculando $P(w_{n+1}|\mathbf{w})$.

Desambiguación Identificar los distintos significados de las palabras. Dada una palabra ambigua w , detectar el tópico z , $P(z|w)$. En este planteo, cada palabra tiene un único “significado” para cada tópico.

Extracción del *gist* Detectar el *gist* del conjunto de palabras, mediante el cálculo de $P(g|\mathbf{w})$

El planteo detallado del cálculo de las probabilidades es bastante complejo, e involucra la especificación de distribuciones *a priori* tanto para los $P(w)$ como para los $P(z)$, así como el muestreo mediante procedimientos sofisticados de cálculo numérico de las complejas distribuciones de probabilidad⁹.

Usando estos métodos, los modelos probabilísticos de tópicos han sido utilizados como modelo de memoria semántica. Por ejemplo, [Griffiths et al. \(2007\)](#) los utilizaron para reproducir patrones de asociación libre, obteniendo mejores resultados que con métodos como el LSA. Entre otras propiedades de estos modelos, los autores señalan que la similitud semántica, calculada mediante la relación $P(w_2|w_1)$ bajo la asunción de un sólo tópico z (“inducido” por w_1):

$$P(w_2|w_1) = \sum_z P(w_2|z)P(z|w_1), \quad (3.8)$$

no es necesariamente simétrica, ni tiene por qué cumplir con la desigualdad triangular, dos de los principales objetivos de las críticas hacia los modelos semánticos espaciales (véase la sección 3.2.2). Mediante el cálculo de estas distancias, los autores también pudieron construir redes semánticas con una topología de distribución de ley de potencia similar a la observada en redes construidas a partir de asociación libre.

3.4 Modelos de redes neuronales

Existen variadas implementaciones de modelos de redes neuronales que pretenden dar cuenta del desempeño humano en tareas de tipo semántico. Buena parte de ellas, se basan en el manejo de características perceptuales y motoras de conceptos concretos y las categorías superordinales que los contienen. En [Rogers y McClelland \(2004\)](#) se pueden encontrar muchos ejemplos en este sentido, orientados a dar cuenta de los fenómenos de categorización, como la adquisición, uso y eventual deterioro de las categorías semánticas.

Por un lado, como se mencionó anteriormente (sección 3.2), bien puede pensarse que los *features* perceptuales y de acción son importantes en la codificación de palabras concretas, y menos relevantes para palabras abstractas. Por otro, estos modelos no pretenden dar cuenta de otros fenómenos tales como la asociación de palabras o la detección de tópicos en discursos, a los que buscan dar respuesta los modelos revisados en este capítulo. En la siguiente sección se presentará un modelo de red neuronal que se enmarca este último enfoque, en el que se basa parte del presente trabajo de tesis (véase la sección 6.1).

⁹Entre otros métodos, se utilizan métodos Monte Carlo de Cadenas de Markov. Los detalles pueden verse en [Blei et al. \(2003\)](#); [Griffiths y Steyvers \(2011\)](#); [Griffiths et al. \(2007\)](#).

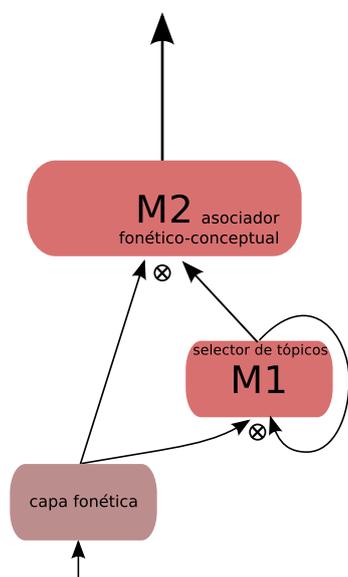


Figura 3.3: Topología modular de la red propuesta por Valle-Lisboa (2007). Un módulo encargado de suministrar las representaciones “fonéticas” de las palabras alimenta a dos módulos, M1 y M2. El primero está encargado de seleccionar el tópico vigente, y recibe una conexión recurrente de su propia salida, a la vez que alimenta al segundo módulo. Este último recibe las salidas de los otros dos módulos, y es el encargado de seleccionar el significado adecuado de la palabra dado el tópico vigente.

3.4.1 Selector de tópicos

En su tesis de doctorado, Juan C. Valle Lisboa introdujo un modelo neural de selector de tópicos Valle-Lisboa (2007). Este modelo pretende funcionar de manera análoga a los modelos probabilísticos de tópicos y los espacios semánticos.

El modelo se inserta en una red cuya tarea podría decirse que es la “desambiguación” de una palabra, es decir, identificar su significado dado el contexto semántico. Para ello, el modelo cuenta con dos módulos neurales constituidos por memorias asociativas dependientes de contexto (véase la sección 2.2.7), como se ilustra en la figura 3.3. Un patrón de activación conteniendo la identidad “fonética” de una palabra llega como entrada a dos módulos, M1 y M2. El módulo M1 es el encargado de identificar el tópico vigente, y enviar esta información al módulo M2, encargado de asociar a la entrada original su interpretación “conceptual”. Ambos módulos utilizan la salida del módulo M1 como contexto en sus entradas: así M2 puede asociar varias interpretaciones para una misma entrada (para distintos tópicos), mientras que M1 es capaz de asignar distintos tópicos a una palabra, dependiendo de la asignación realizada a la palabra anterior.

El módulo selector de tópicos, M_1 , utiliza una conexión recurrente de manera análoga a las redes recurrentes simples de [Elman \(1990\)](#) (véase la sección 2.3.1). Para que este módulo funcione correctamente identificando el tópico de un documento, [Valle-Lisboa \(2007\)](#) propone configurar la matriz de memoria de la siguiente forma:

$$M_1 = \sum_{i=1}^n t_i \left(t_i \otimes \sum_{j=1}^m \Omega_j(i) p_j \right)^T, \quad (3.9)$$

donde los t_i son los vectores que codifican los diferentes tópicos, los $\Omega_j(i)$ cuantifican la importancia de la palabra j en el tópico i , y p_j es el vector conteniendo la representación “fonética” de la palabra j .

La salida de este módulo frente a la entrada k puede plantearse en dos diferentes modalidades: o bien,

$$t(k) = M_1(t(k-1) \otimes p(s)), \quad (3.10)$$

la salida es estrictamente el producto de las entradas combinadas por el producto de Kronecker multiplicadas por la matriz de memoria, o bien:

$$t(k) = M_1(t(k-1) \otimes p(s)) + \beta t(k-1), \quad (3.11)$$

la salida es el producto de las entradas por la matriz, más la salida frente a la palabra $s-1$ multiplicada por un parámetro $\beta \ll 1$.

Para la primera palabra en un documento, la salida de la matriz M_1 no existe. En ese caso, puede incluirse como contexto un vector “blanco” w consistente en una combinación lineal de todos los tópicos, donde los coeficientes ϵ_i dan cuenta de la frecuencia de los mismos en el *corpus*:

$$w = \sum_{i=1}^n \epsilon_i t_i, \quad (3.12)$$

Entonces, la salida del módulo para la palabra k en una secuencia, según la ecuación 3.10 será:

$$t(k) = \sum_{i=1}^n \epsilon_i t_i \prod_{j=1}^k \Omega_j(i), \quad (3.13)$$

esto es, una combinación lineal de todos los tópicos, pesados por sus frecuencias ϵ_i , y por el producto de los “pesos específicos” de cada palabra en cada tópico. En cambio, si se retienen las salidas anteriores (ecuación 3.11), la salida será:

$$t(k) = \sum_{i=1}^n \epsilon_i t_i \left(\prod_{j=1}^k \Omega_j(i) + \beta \right). \quad (3.14)$$

La principal diferencia entre estos dos modos puede ilustrarse fácilmente en el caso en que uno de los $\Omega_j(i) = 0$. Por ejemplo, estos coeficientes pueden definirse como la frecuencia con que ocurre la palabra j en el tópico i :

$$\Omega_j(i) = \frac{n_j(i)}{\sum_{s=1}^n n_s(i)} \quad (3.15)$$

donde $n_s(i)$ es el número de veces en la que aparece la palabra i en el tópico s , y

$$\epsilon_i = \sum_j \Omega_j(i) \quad (3.16)$$

puede interpretarse como la relativa importancia de cada tópico i .

En tal caso, si la palabra j no ocurre nunca en un tópico i , entonces $\Omega_j(i) = 0$, y el producto en j en la ecuación 3.13 se vuelve 0, pero no en la 3.14. En el primer modo de operación, por tanto, en la combinación lineal esté ausente el término t_i . De esta forma, el primer modo sólo retiene en su salida los tópicos para los que todas las k palabras introducidas en la memoria tienen frecuencia definida, mientras que el segundo modo suaviza esa restricción.

El funcionamiento en primer modo remeda la ecuación 3.6 de los modelos probabilísticos de tópicos: el término ϵ_i tiene un rol similar a la probabilidad *a priori* $P(z)$ (si bien en esa ecuación está dentro del producto), y $\Omega_j(i)$ puede considerarse igual a $P(w_i|z)$. De hecho, Valle-Lisboa (2007) mostró que bajo ciertas restricciones, el módulo detector de tópicos puede funcionar como un clasificador bayesiano, vinculando así modelos basados en redes neuronales con modelos probabilísticos.

Otro modo de funcionamiento del módulo propuesto por Valle-Lisboa (2007) consiste en calcular la salida en el caso en que todas las k palabras hayan ingresado al mismo tiempo, “sumadas”. La salida en tal caso sería:

$$t(k) = \sum_{i=1}^n \epsilon_i t_i \left(\sum_j^k \Omega_j(i) \right) \quad (3.17)$$

En este caso, no existe el riesgo de que algunos coeficientes que multiplican a cada $\epsilon_i t_i$ se vuelvan cero porque uno de los $\Omega_j(i)$ sea cero: la sumatoria asegura que sólo se anulan los coeficientes si todos los $\Omega_j(i)$ son cero. A este modo de funcionamiento le llamaremos “modo paralelo”, haciendo referencia a que simula la presentación en paralelo de todas las palabras del documento. Por oposición, a los modos anteriores les llamaremos “modo serial”.

En la figura 3.4 se esquematizan los tres modos de funcionamiento del selector de tópicos. La salida del modo serial puro equivale a la intersección de los tópicos i que comparten todas las palabras j , mientras que el modo paralelo corresponde a algo similar a una “unión pesada” de los tópicos

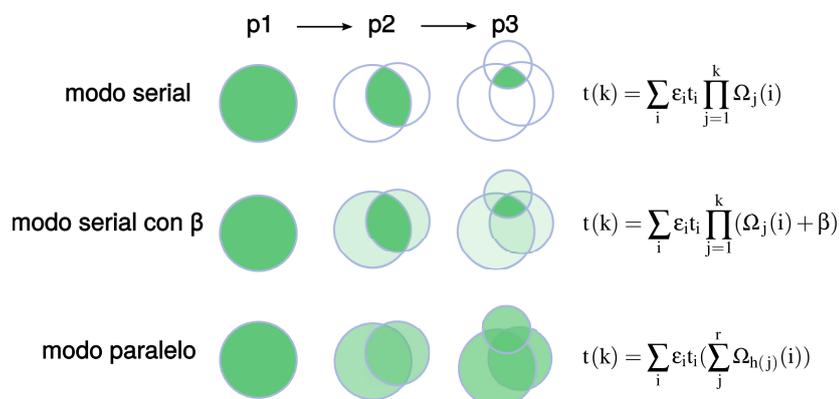


Figura 3.4: Ilustración de los distintos modos de funcionamiento del selector de tópicos. El “modo serial” (ecuación 3.13) corresponde a la intersección de los tópicos j para los cuales los $\Omega_j(i)$ son no nulos. El “modo paralelo” (ecuación 3.17) corresponde a una “unión pesada” de todos los tópicos, de acuerdo a la sumatoria de los $\Omega_j(i)$. El “modo serial con β ” (ecuación 3.14) corresponde a una situación intermedia.

de las palabras. El modo serial con la adición del parámetro β tiene un comportamiento intermedio, dado que la salida está a medio camino entre la intersección y la unión de tópicos.

En cuanto al desempeño del módulo, en su tesis de doctorado [Valle-Lisboa \(2007\)](#) demostró que podía utilizarse para desambiguar palabras utilizando *corpora* relativamente pequeños. Sin embargo, hasta el momento no se han realizado tests exigentes utilizando colecciones de gran tamaño utilizadas normalmente en tareas de categorización de textos. En el presente trabajo de tesis, se realizó una evaluación de este tipo en el módulo selector de tópicos, con el objetivo de evaluar la viabilidad de incluir un módulo de este tipo en una red dedicada al procesamiento y producción del discurso. Los resultados obtenidos se presentan en el capítulo 6.

3.5 Modelos semánticos como herramientas

Los tipos de modelos semánticos aquí revisados, además de ser interesantes como modelos del desempeño humano, resultan útiles como herramientas para enfocar diversos problemas prácticos. En el contexto de esta tesis, es de destacar la posible utilización de estos enfoques al problema de la caracterización objetiva de la coherencia en el discurso de pacientes con esquizofrenia. En el siguiente capítulo, se presentarán algunos aspectos de este problema, mientras que en el capítulo 7 se presentan algunos resultados que intentan aportar una posible vía de resolución que involucra el uso de modelos de espacios semánticos, entre otras herramientas.

Capítulo 4

El desafío de la esquizofrenia

La esquizofrenia es una enfermedad mental de gran incidencia¹, y que ha sido reconocida por la Organización Mundial de la Salud como una de las 10 enfermedades que más contribuyen a la carga mundial de enfermedades (Murray *et al.*, 1996). Incluso con medicación, en muchos casos llega a ser una enfermedad deshabilitante para el paciente, alterando notablemente la capacidad de llevar una vida normal.

Descrita por primera vez como una enfermedad psiquiátrica específica por Emil Kraepelin y Eugen Bleuler a principios del siglo xx (Andreasen, 2011), hoy parece claro que no constituye una enfermedad con una fisiopatología común sino que se trata de un conjunto de desórdenes estrechamente asociados por sus síntomas y signos pero no necesariamente sus causas (Ritsner y Gottesman, 2011). Tradicionalmente se han creado sub-categorías que dan cuenta de la heterogeneidad en la presentación del trastorno (catatónica, paranoide, desorganizada o hebefrénica, simple, indiferenciada), así como perfiles de trastorno relacionados a la esquizofrenia, como desorden esquizo-afectivo, desórdenes de personalidad (esquizoide, paranoide, y esquizotípico), etc. Sin embargo, la efectividad clínica y la relevancia en la investigación de las sub-categorías no han sido satisfactorias, y por tanto nuevos enfoques tratan a la esquizofrenia como un complejo o espectro, dentro del cual se pueden caracterizar algunas dimensiones patológicas. Así, cada paciente es incluido dentro de estas dimensiones a lo largo del tiempo, permitiendo la individualización del tratamiento basado en información clínica (Tandon y Bruijnzeel, 2014).

La esquizofrenia puede describirse como un trastorno progresivo del pensamiento, caracterizado por la aparición de psicosis, episodios en los que la persona pierde contacto con la realidad, así como mostrar pensamientos o emociones alteradas e ideas extrañas. El transcurso de la enfermedad incluye una fase premórbida en la que aparecen trastornos cognitivos, sociales y

¹La incidencia de la esquizofrenia parece variar en distintas partes del globo, sin embargo varios estudios la sitúan entre 10 y 30 por cada 100.000. Véase por ejemplo Menezes (2009)

motores sutiles, seguida de una fase de *prodromos* en la que éstos se agravan. En los últimos años ha cobrado importancia la caracterización de este estadio, de forma de poder identificar a pacientes con riesgo de desarrollar la enfermedad y así iniciar tratamiento en forma preventiva (Cannon, 2009). El primer episodio psicótico, por lo general ocurre en la juventud temprana, y se toma a éste como el comienzo de la enfermedad, que suele ser de carácter crónico. En esta etapa, al comienzo se suceden varios episodios psicóticos puntuados por episodios de remisión, que por lo general se vuelven más largos con el paso del tiempo (Janicak *et al.*, 2014a). Finalmente, en la fase de estabilidad posterior prevalecen los síntomas negativos y cognitivos, ante la presencia de algún brote psicótico aislado. Si bien no existe cura, muchos pacientes logran variados niveles de recuperación a partir de los tratamientos.

4.1 Síntomas y dimensiones patológicas

Bajo el nombre de esquizofrenia se presenta un complejo síndrome que incluye una amplio rango de síntomas que abarcan todos los dominios mentales. Pueden distinguirse varias dimensiones psicopatológicas (síntomas positivos, negativos, cognitivos, emocionales y afectivos, motores, etc.), muchas veces asociadas a distintas alteraciones neurobiológicas y a diferentes patrones de respuesta al tratamiento (Janicak *et al.*, 2014a).

Los episodios psicóticos son un componente principal de estos síntomas, conformando el grupo de los llamados *síntomas positivos*. La mayoría de las personas diagnosticadas han sufrido en algún momento de algún tipo de alucinación; son muy frecuentes las alucinaciones auditivas, desde ruidos simples hasta voces que dialogan, ordenan o comentan las acciones de la persona. Buena parte de los pacientes también manifiesta un pensamiento desorganizado, así como un habla muy difícil de entender por su falta de coherencia. Son muy frecuentes los delirios, muchas veces relacionados a la ilusión de que su mente está controlada o es transparente a terceros, o la sensación de persecución constante por parte de una entidad poderosa (Mamah y Barch, 2011).

También son importantes los denominados *síntomas negativos*, que incluyen falta de voluntad e iniciativa, pérdida de interés por el relacionamiento social y retraimiento. Otras dimensiones reconocidas son las alteraciones emocionales (como por ejemplo respuestas emocionales inapropiadas para el contexto), los síntomas motores, incluyendo catatonía o hiperactividad (Arango y Carpenter, 2011), los déficits cognitivos (problemas de atención, memoria de trabajo, solución de problemas y cognición social, entre otras), y la desorganización del pensamiento, tema de la siguiente sección (Tandon y Bruijnzeel, 2014).

Estas dimensiones patológicas, sin embargo, no son enteramente inde-

pendientes: a través del análisis factorial de cada uno de los síntomas en pacientes diagnosticados con esquizofrenia, primero Liddle (1987) y luego muchos otros autores identificaron tres grandes factores que dan cuenta de la mayoría de la varianza observada. Las “cargas” de los diferentes síntomas en cada uno de los factores permiten caracterizar a éstos como correspondiendo a síntomas negativos, positivos, y de “desorganización”. Esta separación lleva a la descripción de los “tres síndromes de Liddle” (McKenna y Oh, 2005). Si bien la relativa independencia de los síntomas positivos y negativos estaba bien establecida, estos estudios permiten identificar a la “desorganización” (del que la “desorganización del pensamiento” puede ser un gran componente) como un síndrome independiente y caracterizable mediante un análisis de los síntomas que lo componen.

4.2 Desorden del pensamiento y el comportamiento

En buena parte de los pacientes con esquizofrenia, los procesos que conllevan el pensamiento se ven alterados en mayor o menor medida. Esto es aparente ya sea por la propia introspección (meta-cognición) del paciente acerca de sus patrones de pensamiento aberrantes, o, de manera más importante, por el análisis de las manifestaciones presentes en el habla espontánea.

Ya Emil Kraepelin, en la descripción original de la enfermedad, hacía notar que las “asociaciones mentales” del paciente pueden ser muy peculiares, rayando lo incomprensible, y que esto constituye el fundamento de la “incoherencia del pensamiento” (descrito en McKenna y Oh (2005)). Según este autor, la severidad del desorden puede ir desde una “distractibilidad aumentada”, con el paciente “pasando de un tema a otro sin ninguna conexión”, hasta la pérdida completa de conexión de las ideas. Un ejemplo brindado por el mismo Kraepelin, ante la pregunta “¿Está usted enfermo?”, un paciente le respondió:

You see as soon as the skull is smashed and one still has flowers [laughs] with difficulty, so it will not leak out constantly. I have a sort of silver bullet which held me by my leg, that one cannot jump in, where one wants, and that ends beautifully like the stars. Former service, then she puts it on her head and will soon be respectable, I say, O God, but one must have eyes. Sits himself and eats it. Quite excited, I was quite beside myself and say that therefore there should be meanness and there is a merry growth over. It was the stars. I, and that is also so curious, the nun consequently did not know me any more, I should come from M. because something always happens, a broken leg or something, they've had a quarrel with each other, the clergyman and she; a leg has just been broken. I believe that it is caused by this that such a misfortune happens, such a reparation for

damages. I have also said I shall then come in the end last, with the sun and the moon, and too much excitement, and all that makes still a great deal of trouble. Kings do not collect the money, in this way the letters have been taken away from me, as I that at last of those that particularly believe as I at last specially think, and all are burned. You can imagine that comes always from one to the other.

A pesar de la evidente naturaleza incomprensible y fragmentaria de la respuesta, Kraepelin observó que en algunas partes se pueden encontrar conexiones entre las ideas. En otros casos, las vinculaciones eran de naturaleza fonológica: en alemán, “octavo” rima con “carnear” (*achten* vs *schlachten*). Kraepelin usó varios términos para describir este fenómeno, que se tradujo al inglés y otros idiomas como “descarrilamiento” (McKenna y Oh, 2005).

Además de un “descarrilamiento” en la asociaciones entre ideas, la falta de planificación u objetivos en el habla, en muchos pacientes es evidente una afectación de naturaleza intrínsecamente lingüística. Son comunes las alteraciones gramaticales, los circunloquios, neologismos, parafasias, incluso el habla telegráfica o agramática presente en algunas afasias y disfasias, o la “ensalada de palabras” o “esquizofasia” en el que el habla se vuelve ininteligible al nivel de las palabras individuales (Kuperberg, 2010a).

4.2.1 La importancia del lenguaje

Como se mencionó anteriormente, el habla del paciente esquizofrénico tiene el doble rol de ser, por un lado, ventana al mundo interior del paciente a través de su relato acerca de las alteraciones del pensamiento, así como los delirios o alucinaciones que éste pueda experimentar (como sucede por lo general en medicina, es el relato del paciente el que da cuenta de los síntomas presentes). Por otro lado, la naturaleza de la construcción del habla en sí mismo permite caracterizar las alteraciones del pensamiento (DeLisi, 2001).

Con tal fin, Norman Cameron (1938) fue el pionero en realizar entrevistas estructuradas y analizar las respuestas otorgadas por pacientes con esquizofrenia. Así, identificó tres clases de distorsiones características presentes en el habla de los pacientes: “pensamiento asindético”, “distorsión metonímica”, e “interpenetración de temas”. La primera refiere a que las respuestas solían contener elementos relacionados y relevantes a la pregunta, pero el orden lógico entre ellos se pierde, mientras que la segunda describe el uso de circunloquios o aproximaciones indirectas en lugar de palabras concretas. Estas distorsiones encuentran correlato en algunos conceptos que manejaban ya Kraepelin y Bleuler, como “descarrilamiento” y “asociaciones laxas”, aunque encontraron en Cameron un nivel de detalle y rigor mayores.

La “interpenetración de temas” refiere como su nombre lo indica, a la entremezcla de temas en el discurso, y resulta uno de los principales aportes

de Cameron. En algunos pacientes el tema de la pregunta se mezcla con un tema recurrente, en general el tema central de sus delirios. En este ejemplo, el paciente tenía ambiciones no realizadas acerca de ser ingeniero, y se le pidió que continuara la frase “Mi pelo es marrón porque...”:

[My hair is brown because...] Because it is a sort of hydraulic evering. [What does that mean?] It means that it gives you some sort of color-blindness because it works through the roots of the hair and hydrasee –that is a study of the growth of plants, a sort of human barometer, hydraulic hydrosceenic method.

Puede apreciarse que el tema (“hair color”) está correctamente relacionado con algunas palabras como “roots” y “growth”, pero el color lleva a “color-blindness” y “growth” al crecimiento de las plantas. La frase siguiente, “... a sort of human barometer, hydraulic hydrosceenic method” parece referirse al tema sujeto del delirio del paciente (McKenna y Oh, 2005).

Hacia la segunda mitad del siglo xx, surgen nuevos aportes a la caracterización del lenguaje presente en el trastorno del pensamiento y el lenguaje en la esquizofrenia. Así los trabajos de Wing (Wing *et al.*, 1974) y de Harrow (Harrow y Quinlan, 1985) enfatizan la pobreza en el contenido del habla y la aparición de pensamientos bizarros e idiosincrásicos, respectivamente. La noción de pobreza de contenido en el habla comenzó a disociarse entonces de otro concepto relacionado: la pobreza del habla. Esta última refiere a la observación ya hecha por Kraepelin y Bleuler, acerca de que en muchos pacientes crónicos, el habla espontánea casi desaparece, las respuestas son lacónicas o directamente inexistentes. Sin embargo, la pobreza en el contenido del habla refiere a un habla más o menos fluida, donde es notoria la carencia de palabras y oraciones con contenido, como en el siguiente ejemplo (McKenna y Oh, 2005):

— How do you like it in hospital? — Well, er... not quite the same as, er... don't know quite how to say it. It isn't the same, being in hospital as, er... working. Er... the job isn't quite the same, er... very much the same but, of course, it isn't exactly the same.

Como puede apreciarse en base a los ejemplos mencionados, son varias las alteraciones del habla características del trastorno del pensamiento, que van desde la casi ausencia de palabras (o incluso directamente mutismo), a la verborrea llena de expresiones bizarras en las que se entrecruzan los temas de conversación y se utilizan giros y expresiones infrecuentes reemplazando palabras comunes. Ante este panorama, se desarrollaron varios instrumentos que intentaron cuantificar cada una de estas dimensiones, con el fin de poder monitorizar, caracterizar y comparar poblaciones y pacientes

en particular ante el desarrollo de la enfermedad (McKenna y Oh, 2005). Entre esos instrumentos se destaca la escala de Andreasen, una de las más utilizadas hoy en día.

La escala de Andreasen

Hacia finales de la década de 1970, Nancy Andreasen desarrolló y evaluó un conjunto de dimensiones en las cuales poder cuantificar el nivel de deterioro de las habilidades de pensamiento, lenguaje y comunicación. Poniendo orden y refinando definiciones un tanto vagas que existían en la literatura, elaboró un conjunto de 18 dimensiones en las que analizar el habla en búsqueda de anormalidades, entre los que se destacan (Andreasen, 1986):

Pobreza del habla Restricción en la *cantidad* del habla espontánea, las respuestas son concretas, breves y sin elaboración.

Pobreza del contenido del habla Respuestas de largo adecuado, pero que conllevan poca información. Lenguaje vago, sobreabstracto o sobreconcreto, repetitivo, estereotipado.

Tangencialidad Responder a una pregunta en una forma oblicua, tangencial, o irrelevante. Relacionada poco o nada a la pregunta.

Descarrilamiento Las ideas descarrilan, en sucesión clara u oblicuamente relacionadas, si acaso. Yuxtaposición de conceptos sin significado aparente.

Incoherencia El habla se vuelve incomprensible. Porciones de oraciones coherentes en medio de oraciones globalmente incoherentes. O palabras semánticamente no relacionadas en sustitución de las apropiadas, de manera que la oración pierde sentido.

Ilogicalidad *Non-sequiturs*, inferencias lógicas inválidas. Llegar a conclusiones basadas en premisas falsas, sin presencia de pensamiento delirante.

Clanging Elección de palabras en base al sonido y no al contenido. Aparición de rimas y asociaciones basadas en palabras que suenan similares.

Circunstancialidad Habla muy indirecta que se demora en alcanzar su idea objetivo. Al explicar algo, se incorporan muchos detalles tediosos, incluyendo acotaciones “entre paréntesis”.

Pérdida de objetivo Falla en seguir una cadena de pensamiento hasta su conclusión natural. Habla que comienza con un tópico, comienza a perderse en otros, para no volver al original.

Perseveración Repetición persistente de palabras, ideas o temas, que permanecen en el habla una vez utilizadas.

Ecolalia El paciente repite las palabras del entrevistador. Típicamente repetitiva y persistente.

Para analizar una muestra del habla de un paciente, cada uno de estos ítems se evalúan por separado por dos o más jueces en una escala entre 0 y 3 o 4 dependiendo del ítem. Las puntuaciones de cada ítem se promedian (algunos ítems tienen más peso que otros, por ejemplo “descarrilamiento” vs. “pérdida de objetivo”) para obtener una puntuación global (TLC global score).

Desde su aparición, la escala de Andreasen pasó a ser el instrumento de medida estándar para todos los estudios del desorden del pensamiento en pacientes psicóticos (Andreasen y Grove, 1986; McKenna y Oh, 2005). De hecho, la mayoría de los ítems desarrollados en el planteamiento original de la escala (Andreasen, 1979) fueron incorporados en la tercera edición del Manual de Diagnóstico y Estadística de la Asociación Psiquiátrica Americana (DSM-III), que representó un hito en la caracterización del desorden del pensamiento en particular, y de las psicosis en general (American Psychiatric Association, 1980).

4.2.2 Análisis automatizados

Si bien la escala de Andreasen es de gran utilidad tanto en la práctica clínica (para diagnóstico y prognosis) como en la investigación, varios autores señalan la necesidad de contar con métodos de evaluación automatizados. La aplicación de la escala requiere de mucho esfuerzo humano: es necesario entrenar a quienes han de aplicarla; y cada aplicación lleva un tiempo considerable. Por otra parte, el hecho de que involucre juicios subjetivos dificulta la comparación entre distintos evaluadores. Esto dificulta la aplicación masiva y regular de la escala, así como la comparación directa de los resultados obtenidos en diferentes centros, o incluso en un mismo centro clínico si difieren mucho los criterios empleados por los jueces involucrados (Elvevåg *et al.*, 2007).

En los últimos años se han incrementado los esfuerzos por producir métodos de análisis automático del discurso que sean capaces de cuantificar algunos aspectos de las alteraciones del habla de pacientes con esquizofrenia que son consideradas en la escala de Andreasen.

El análisis cuantitativo del discurso esquizofrénico puede rastrearse al menos hasta un trabajo de George K. Zipf (Whitehorn y Zipf, 1943), en el que se realiza un análisis de la distribución de frecuencias de las palabras. El mismo Zipf ya había mostrado que una propiedad robusta de los lenguajes naturales es exhibir una relación de ley de potencia entre el rango y la frecuencia de las palabras; más precisamente el rango que ocupa una palabra es aproximadamente proporcional al inverso de la frecuencia² (Ferrer i Cancho

²Si se ordenan las palabras según mayor a menor frecuencia, el rango es la posición que

y Solé, 2003; Zipf, 1949). En el trabajo, los autores analizan la distribución de frecuencias de palabras en la correspondencia entre una mujer diagnosticada con “trastorno esquizoafectivo”, concluyendo que muestra propiedades similares a las del habla presente en niños. Específicamente, muestra que existe un rango de palabras con frecuencia intermedia que están sobrerrepresentadas en su frecuencia de uso (se usan más de lo que predice la ley) (Whitehorn y Zipf, 1943). A pesar de lo promisorio del enfoque, este tipo de análisis cuantitativo de la producción lingüística aplicado a la psiquiatría aparentemente no generó muchas repercusiones en la literatura relacionada a las patologías psiquiátricas (sin embargo véase Pennebaker *et al.* (2003) para una revisión de enfoques de análisis del discurso en el contexto psicológico).

Con el desarrollo de técnicas y recursos computacionales de las últimas décadas han surgido nuevas aproximaciones al análisis automático del discurso en el contexto psiquiátrico. Por ejemplo, Elvevåg *et al.* (2007) utilizaron Análisis Semántico Latente para cuantificar la pérdida de coherencia en transcripciones de habla de pacientes con esquizofrenia. En ese estudio lograron correlacionar medidas obtenidas en forma automática utilizando esta técnica con evaluaciones de tangencialidad y contenido realizados por expertos entrenados. Utilizando este tipo de medidas, los autores lograron discriminar a los pacientes con alto *versus* bajo grado de desorden del pensamiento; sin embargo, no lograron discriminar a este último grupo de los sujetos utilizados como grupo control. En un estudio subsiguiente, los autores utilizaron análisis basados en LSA en pacientes, así como en sus parientes cosanguíneos en primer grado, logrando un grado de discriminación similar Elvevåg *et al.* (2010).

Otro enfoque que ha ganado impulso en los últimos años es el de utilizar elementos de teoría de grafos para analizar el discurso. Mizraji y Valle-Lisboa (2007) propusieron el uso de estas herramientas a partir de la hipótesis de que las alteraciones discursivas podrían ser el resultado de que las redes semánticas exhiben un deterioro tal que origina el colapso entre distintos “clusters” o temas. Esto podría evidenciarse entonces reconstruyendo la topología de las redes de conceptos presentes en el discurso³. Por ejemplo, Mota *et al.* (2014; 2012) construyeron grafos a partir de relatos de pacientes diagnosticados con esquizofrenia y trastorno bipolar. En base a los grafos, lograron discriminar a los pacientes según su patología utilizando varios descriptores topológicos (como el número de ciclos, la densidad de conexiones, entre otros).

En el presente trabajo de tesis, presentamos una herramienta para cuan-

ocupan en la esta lista ordenada. La denominada “Ley de Zipf” establece que el rango de la palabra es proporcional al inverso de la frecuencia, $r \propto \frac{1}{f}$. Esta ley es un ejemplo de las denominadas “leyes de potencia”, o libres de escala: el cociente entre dos rangos de dos palabras cualesquiera es igual al cociente de sus frecuencias.

³véase por ejemplo Baronchelli *et al.* (2013) para una revisión sobre la teoría de redes en ciencias cognitivas

tificar el deterioro en la coherencia discursiva, utilizando grafos y espacios semánticos, que se presenta en el capítulo 7.

4.3 Redes neuronales en esquizofrenia

A partir de la década de 1990, enmarcados en el auge que el conexionismo estaba adquiriendo en ciencias cognitivas, surgieron varios esfuerzos por modelar fenómenos psicopatológicos utilizando redes neuronales (véase por ejemplo [Qi et al. \(2011\)](#); [Stein y Ludik \(1998a\)](#)). Los fenómenos relacionados a la esquizofrenia no escaparon a este movimiento, al contrario: varias propuestas se centraron en explorar estos modelos como campo de prueba para testear hipótesis fisiopatológicas e intentar explicar el origen de algunos síntomas ([Spitzer, 1999](#)).

Actualmente se consideran como fundacionales al uso de modelos de redes neuronales en esquizofrenia a los trabajos de [Cohen y Servan-Schreiber \(1992\)](#) sobre el surgimiento del desorden del pensamiento como consecuencia de una reducción en la transmisión dopaminérgica cortical ([Cohen y Servan-Schreiber, 1992](#)), y los de Ralph Hoffman sobre el origen de las alucinaciones auditivas y otros fenómenos como consecuencia de un sobregiro de la “poda” sináptica ([Hoffman y Dobscha, 1989](#); [Hoffman et al., 1995](#); [Siekmeier y Hoffman, 2002](#)).

Estos últimos dieron origen a una implementación de la topología de Elman (véase la sección 2.3.1) utilizando memorias matriciales dependientes de contexto (sección 2.2.7) que fue utilizada en este trabajo de tesis, y que es introducido en la siguiente sección. Además, constituyen la base de aproximaciones más recientes al uso de redes neuronales en esquizofrenia (y en psiquiatría en general, véase [Hoffman et al. \(2011\)](#)).

4.3.1 Alucinaciones auditivas y pérdida de conectividad

Los trabajos de Ralph Hoffman exploran el origen de uno de los síntomas más conspicuos de la esquizofrenia: la presencia de alucinaciones auditivas (véase la sección 4.1). En una serie de trabajos utilizando redes recurrentes simples ([Hoffman y McGlashan, 1997](#); [Hoffman et al., 1995](#); [Hoffman y McGlashan, 2001](#)), exploró la aparición de “alucinaciones” auditivas (falsos reconocimientos de palabras frente a estímulos neutros) realizando manipulaciones que alteran la estructura y/o funcionamiento de las redes. Una de las manipulaciones consistía en eliminar progresivamente las conexiones sinápticas.

Esta manipulación de “podado” sináptico se plantea como forma de testear la validez de un mecanismo fisiopatológico propuesto para el desenlace de la esquizofrenia ([Hollis y Rapoport, 2011](#)). Durante la adolescencia, ocurre normalmente que una fracción importante de las sinapsis presentes en la

corteza son eliminadas. Este fenómeno está vinculado a una “maduración” conectiva del cerebro. Sin embargo, varios autores plantean que un exceso en la poda de conexiones (o incluso una poda insuficiente), puede ser un factor que contribuye a las alteraciones morfológicas observadas en la corteza en pacientes con esquizofrenia, así como también ser el origen de algunos de los síntomas psicóticos (Lawrie y Pantelis, 2011).

En sus simulaciones, los trabajos de Hoffman plantean un vínculo entre la poda excesiva de sinapsis y la presencia de alucinaciones auditivas, aunque la validez de esta vinculación aún no ha sido comprobada en estudios con pacientes.

4.3.2 Modelo SPELT

Una variante del enfoque de Hoffman fue explorada por Florencia Reali y Juan C. Valle Lisboa en sus tesis de maestría y doctorado respectivamente (Reali, 2002; Valle-Lisboa, 2007; Valle-Lisboa *et al.*, 2005). En el trabajo se utilizaron memorias matriciales dependientes de contexto (sección 2.2.7) para implementar la topología de Elman, lo que redundaba en un procedimiento más sencillo (dado que el algoritmo de Widrow-Hoff resulta más sencillo de implementar que el de *backpropagation*, y el número de iteraciones necesarias para el entrenamiento es menor).

La topología del modelo SPELT (acrónimo de “Sigma Pi Elman Topology”, que hace referencia a la suma y la multiplicación presentes en el modelo de neurona) puede apreciarse en la figura 4.1. Un módulo de entrada se encarga de suministrar las representaciones fonéticas a la capa de salida, el asociador fonético-conceptual. La salida de este módulo es la correcta identificación de la palabra en base a su representación fonética. La capa de salida alimenta también una memoria de trabajo, a través de la cual recibe alimentación recurrente de su propia salida en el tiempo anterior (las unidades de la memoria de trabajo funcionan replicando la actividad de las unidades de la capa de salida). La actividad recurrente suministrada por la memoria de trabajo sirve de contexto a la entrada fonética. Esta topología es esencialmente la misma que la del módulo selector de tópicos, aunque el contexto de funcionamiento sea diferente (véase la sección 3.4.1).

En la simulación que se presenta en Valle-Lisboa *et al.* (2005), entrenaron a la red para predecir la siguiente palabra en una oración, utilizando un conjunto de oraciones sencillas formadas usando un vocabulario reducido de 28 palabras comprendiendo 3 categorías gramaticales; verbos, adverbios y sustantivos. En este contexto, cada oración está formada de 3 palabras, donde la primera y segunda son un nombre de persona o animal y la segunda un verbo congruente (“hablar” sólo puede seguir a un nombre de persona, por

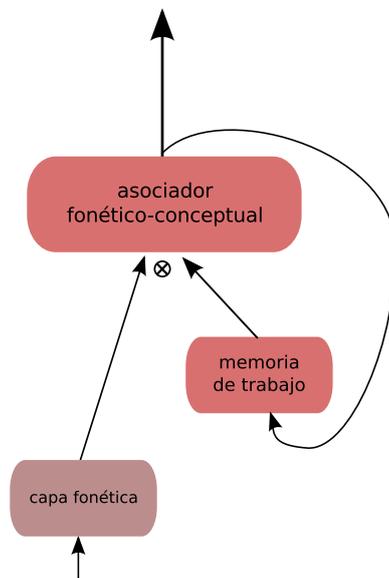


Figura 4.1: Topología de Elman implementada usando memorias matriciales dependientes de contexto.

ejemplo), mientras que la tercera puede ser un nombre o adverbio congruente con las primeras dos palabras. Después de entrenar la red, evaluaron su desempeño presentando todas las oraciones del *corpus* separadas por “silencios”, patrones fonéticos aleatorios que no corresponden a ninguna palabra conocida. Si la red reconoce una de estas ocurrencias de “silencios” como palabras, son consideradas como “alucinaciones”.

Para simular los efectos psicopatológicos presentes en la esquizofrenia, los autores presentaron dos estrategias. Primero, y siguiendo a [Hoffman et al. \(1995\)](#), simulamos el efecto de un exceso de eliminación de sinapsis (synaptic pruning) utilizando el criterio “Darwiniano” de neurodesarrollo esgrimido por estos autores, según el cual se eliminan antes las sinapsis más débiles. En este contexto, eliminar una sinapsis equivale a cambiar el valor de un peso sináptico en la matriz de asociación por 0. Segundo, simulamos la pérdida de funcionalidad de la memoria de trabajo a través de la eliminación de unidades del módulo de memoria de trabajo, volviendo 0 algunas unidades del vector contexto. Otro aspecto importante que exploran en el modelo es el efecto neuromodulatorio de los antipsicóticos, a través de la multiplicación de los valores del contexto proporcionados por la memoria de trabajo por una constante menor a 1.

El desempeño de la red sin alteraciones es el esperado para una red recurrente simple de Elman, alcanzando un alto grado de reconocimiento a la vez que genera expectativas correctas acerca de la categoría gramatical de las palabras subsiguientes. Esto demuestra que esta implementación basada en memorias matriciales dependientes de contexto representa una

alternativa a la formulación clásica de esta topología basada en perceptrones de capa oculta entrenados usando *backpropagation*.

Los efectos del deterioro de la red fueron mayores para la pérdida de funcionalidad de la memoria de trabajo que para el deterioro sináptico selectivo: en este último procedimiento, la red identifica correctamente un 90% de las palabras aun cuando se han eliminado el 80% de las sinapsis (la eliminación del 40% de unidades de memoria de trabajo produce un efecto similar).

Las “alucinaciones” auditivas, es decir, el reconocimiento de silencios como palabras, acompañan el deterioro en el desempeño de la red, y son por tanto también más sensibles a la eliminación de unidades de memoria de trabajo que a la poda selectiva de sinapsis. Es importante destacar que no en todas las simulaciones las redes produjeron alucinaciones: esto es efecto de las diferentes realizaciones particulares de la codificación aleatoria utilizada. La inclusión de “neuromodulación” inducida por antipsicóticos, expresada en la reducción de la actividad de las unidades de la memoria de trabajo, eliminó completamente la ocurrencia de “alucinaciones” en todas las simulaciones, al mismo tiempo que mejoró el reconocimiento de las palabras.

Si bien algunos de estos resultados no están alineados completamente a los de Hoffman (Hoffman *et al.*, 1995; Hoffman y McGlashan, 2001), contribuyen a establecer el uso de modelos de redes neuronales como herramientas válidas para explorar los mecanismos fisiopatológicos que subyacen a los trastornos psiquiátricos, vinculando fenómenos neurales con observables comportamentales, así como simulando el posible efecto de las terapias existentes (Hoffman *et al.*, 2011; Stein y Ludik, 1998a).

Por tanto, la importancia de este trabajo en el marco de esta tesis es doble: por un lado, establece la validez de la implementación de la topología SPELT, utilizada en esta tesis como herramienta para obtener espacios semánticos (véase el capítulo 6). Por otro, ilustra cómo los modelos de redes neuronales (y en particular los que utilizan las memorias matriciales dependientes de contexto) pueden ser utilizadas para explorar hipótesis fisiopatológicas en el contexto psiquiátrico, marcando un posible rumbo de perspectiva para los resultados obtenidos relacionados con la cuantificación de la alteraciones del discurso (presentadas en el capítulo 7, véanse la perspectivas en el capítulo 8).

Capítulo 5

Objetivos

5.1 Objetivo general

- Contribuir al desarrollo de modelos de redes neuronales capaces de generar discursos, de forma de servir de campo de prueba para entender mejor los mecanismos fisiopatológicos que subyacen a la pérdida de Coherencia en el discurso.

5.2 Objetivos específicos

- Desarrollar una herramienta/procedimiento que permita de cuantificar las alteraciones en la coherencia discursiva.
- Avanzar hacia la automatización del proceso de cuantificación, de modo de poder contar con una herramienta aplicable en situaciones clínicas en forma eficiente.
- Evaluar el desempeño de una red neuronal capaz de implementar la noción de tópico.
- Construir modelos de redes neuronales capaces de extraer de forma no supervisada relaciones semánticas, que logren desempeños similares a los obtenidos mediante espacios semánticos.

Parte II

Resultados y Discusión

Capítulo 6

Redes neuronales

En este capítulo se presentan los resultados de esta tesis concernientes al uso de modelos de redes neuronales capaces de representar el significado, de manera similar a los modelos semánticos discutidos en el capítulo 3. En particular, se presentarán los resultados obtenidos utilizando el modelo selector de tópicos (introducido en la sección 3.4.1) como un clasificador de documentos, y los obtenidos utilizando la arquitectura SPELT (presentado en la sección 4.3.2) para derivar representaciones semánticas similares a las obtenidas aplicando métodos de construcción de espacios semánticos, tales como el Análisis Semántico Latente (ver sección 3.2.1).

6.1 Módulo selector de tópicos

Como se vio en el capítulo 3, el módulo selector de tópicos desarrollado por Valle-Lisboa (2007) es un buen candidato para explorar su capacidad para implementar la noción de tópico. Para emplear este módulo es necesario contar con una colección de textos en las cuales se conozca previamente la distribución de tópicos, ya que el aprendizaje del módulo es supervisado. Esto restringe el tipo de tareas en las cuales se puede realizar una evaluación de desempeño, ya que son pocas las situaciones en las que este requerimiento es satisfecho.

A partir de esta restricción, la tarea elegida para evaluar el desempeño del módulo fue la categorización de textos (*text categorization*) (Manning y Schütze, 1999). Esta tarea resulta apropiada para este tipo de evaluaciones, ya que cuenta con una larga tradición y vasta literatura, y por ello existen abundantes recursos disponibles para implementarla.

6.2 Categorización de textos

En la categorización de textos (o clasificación o indexado de documentos), el objetivo es asignar a cada documento una o varias categorías definidas. En

la categorización supervisada, esta asignación es conocida para un conjunto grande de documentos, y el objetivo es lograr en base a esta información, asignar correctamente nuevos documentos. Esta colección de textos suele segmentarse en dos partes; una parte es usada como conjunto de entrenamiento para estimar los parámetros del clasificador, y la otra es usada como conjunto de testeo para evaluar su desempeño.

6.2.1 Medidas de clasificación

Las medidas de clasificación más frecuentemente utilizadas son *precisión* y *cobertura* (o *recall*), y sus combinaciones. La precisión P de una clasificación es la fracción de asignaciones realizadas que son correctas:

$$P = \frac{\# \text{ de asignaciones correctas realizadas}}{\# \text{ de asignaciones realizadas}} \quad (6.1)$$

mientras que la cobertura R es la fracción de posibles asignaciones correctas que fueron efectivamente realizadas:

$$R = \frac{\# \text{ de asignaciones correctas realizadas}}{\# \text{ de asignaciones correctas posibles}}. \quad (6.2)$$

Para cada documento, un algoritmo de clasificación puede producir un rango continuo de puntajes de pertenencia para cada categoría, o bien asignaciones binarias de pertenencia. En el primer caso, la precisión y la cobertura pueden medirse teniendo en cuenta qué categorías son relevantes para cada documento, esto es, qué categorías logran puntajes por encima de cierto umbral. Para umbrales altos, la precisión puede ser alta y la cobertura baja, mientras que lo opuesto podrá suceder para umbrales bajos. Una medida que integra el desempeño del clasificador para todo el conjunto de documentos teniendo en cuenta el compromiso entre precisión y cobertura es la “precisión promedio interpolada en 11 puntos” (*interpolated 11-point average precision*, 11pt-avgP) (Yang, 1999). Esta medida es el promedio de precisión obtenido a 11 niveles de cobertura (entre 0 y 1 a intervalos de 0,1) obtenidos al variar el umbral, y por tanto logra resumir en una cifra el comportamiento general del clasificador.

En el caso de los clasificadores binarios, precisión y cobertura suelen evaluarse con respecto a los documentos considerados relevantes para cada categoría. Los valores resultantes para cada categoría pueden ser integrados a través del *macro*- o *micro*-promediado. El primer método consiste en promediar directamente los valores de precisión y cobertura para cada categoría, y el segundo consiste en calcular medidas globales considerando todas las categorías y todos los documentos. El primer método otorga igual peso a cada categoría, mientras que el segundo otorga igual peso a cada documento, y por tanto mayor peso a las categorías más grandes.

Para este último tipo de análisis, la precisión y la cobertura suelen combinarse usando la métrica F_β (Manning y Schütze, 1999; Yang, 1999):

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad (6.3)$$

siendo ampliamente más usado el caso particular F_1 , que corresponde a la media armónica entre precisión y cobertura:

$$F_1 = \frac{2PR}{P + R}. \quad (6.4)$$

Los macro- F_β y micro- F_β se calculan utilizando los valores de precisión y cobertura macro- y micro-promediados, respectivamente.

6.2.2 Métodos

Para la tarea de clasificación se utilizaron dos *corpora* estándar para la tarea (Joachims, 1998; Li y Yang, 2003; Yang, 1996, 1999; Yoon *et al.*, 2006): la colección Reuters-2178-mod-Apté, y la colección de clasificación OHSUMED-87-91. Ambas han sido utilizadas para evaluar diferentes clasificadores, presentan un nivel de dificultad intermedia, y es posible comparar los resultados obtenidos con la literatura. Además, su tamaño intermedio permite utilizarlos en computadoras de escritorio sin grandes problemas.

Un procedimiento común para aumentar la efectividad de la clasificación es la selección de palabras (*feature selection*), según varios criterios. El primero es la tradicional remoción de las palabras funcionales y de escaso contenido semántico, conocidas como “stop words” (Salton y Lesk, 1965) (véase el apéndice C.1). Para refinar la selección de palabras, se suelen calcular métricas de especificidad que intentan medir qué tan específicas de una categoría son las palabras, como por ejemplo la información mutua, el estadístico χ^2 , o el índice Gini (Aggarwal y Zhai, 2012; Manning y Schütze, 1999; Yang, 1999). Aquí utilizamos el estadístico χ^2 ya que es el más utilizado en la literatura, en la que se reportan buenos resultados con esta métrica (Yang, 1999).

Una vez elegidas las palabras a retener, se calculan los valores de $\Omega_j(i)$ que cuantifican la importancia de la palabra j en el tópico i (ecuación 3.15) en base a los documentos de entrenamiento. También en base a estos documentos se calcularon los valores de ϵ_i , que cuantifica la importancia relativa (en número de palabras) de cada tópico (3.16). Los vectores de tópico t_i fueron vectores canónicos (coordenada i igual a 1, el resto nulas) con una dimensionalidad igual al número de categorías existentes.

Algunos resultados preliminares habían mostrado que de los tres modos de funcionamiento presentados en la sección 3.4.1, el modo “serial” estricto,

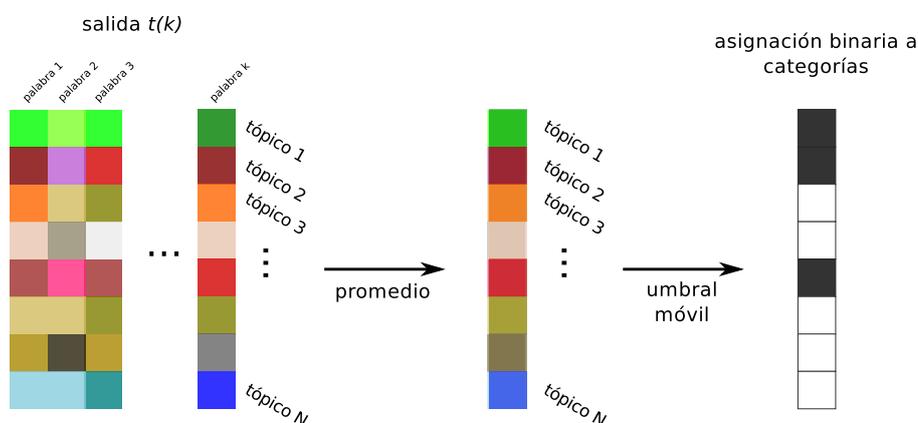


Figura 6.1: Ilustración del procedimiento utilizado para obtener asignaciones binarias a categorías a partir de los vectores de salida $t(k)$ del módulo selector de tópicos. Para estos vectores cada dimensión corresponde a una categoría (o tópico) diferente. Primero se obtiene el promedio del conjunto de vectores $t(k)$ obtenidos para cada una de las k palabras del documento (excluyendo el segmento inicial, ver texto). A este vector promedio se lo convierte en un vector binario utilizando un umbral variable, de una forma que depende del modo de evaluación (clasificación por documento o por categoría, ver texto).

sin parámetro β (ecuación 3.13) no es capaz de funcionar adecuadamente en este tipo de tareas, dada la ocurrencia de palabras con $\Omega_j(i) = 0$ en el conjunto de entrenamiento, que aparecen sin embargo en documentos del tópico i en el conjunto de testeo. Esto genera que el coeficiente del vector canónico t_i sea cero, ya que $\prod_j \Omega_j(i) = 0$, haciendo que el selector no elija este tópico como relevante. Por lo tanto, se decidió descartar este funcionamiento, y sólo evaluar los modos “serial con parámetro β ” (de ahora en adelante referido simplemente como “modo serial”, ecuación 3.14, y el modo “paralelo” (ecuación 3.17).

Para cada palabra de un documento puede obtenerse una salida del módulo de tópicos, que por construcción es una combinación lineal de la base canónica t_i . Por tanto, cada coeficiente del vector de salida $t(k)$ es interpretable directamente como el “peso” específico de cada tópico en el documento (después de procesadas k palabras). Estas salidas fueron normalizadas en cada paso utilizando la norma $\| \cdot \|_1$ (la suma de las coordenadas es igual a 1).

Para cada documento, los vectores obtenidos en cada palabra se promediaron (excluyendo la quinta parte inicial de cada documento, en el caso de que éste constara de más de 10 palabras), y este vector promedio se consideró como un ranking de puntajes de asignaciones de categoría. En la figura 6.1 se ilustra este procedimiento. A partir de estos puntajes se obtuvieron los valores de de precisión y cobertura. Para el caso de la categorización por documento, simplemente se varió el valor del umbral de clasificación, y se

computó el valor de $11pt-avgP$ global. Para la clasificación por categorías, se utilizó el método $Pcut$ (Yang, 1999) para producir asignaciones binarias de documentos para cada categoría. En base a los valores de precisión y cobertura obtenidos se calcularon los valores máximos de las medidas macro- y micro- F_1 .

Reuters 2178

Existen varios *corpora* para clasificación de textos elaborados a partir de artículos periodísticos de la agencia Reuters (Lewis y Ringuette, 1994; Lewis *et al.*, 2004; Yang, 1999). El primero fue introducido por Hayes y Weinstein (1991), al que le surgieron varias modificaciones (Yang, 1999). Una de las más utilizadas es la proporcionada por Apté *et al.* (1994), en particular en su versión de 90 categorías. Esta versión tiene 7768 documentos de entrenamiento, y 3018 documentos de test, etiquetados en una o más de 90 categorías según el tema de la noticia (algunos nombres de categorías son “hierro-acero”, “adquisiciones”, “petróleo”, “soja”, “maíz”, “café”, y reflejan que los artículos tratan principalmente con el mercado de valores, en particular sobre *commodities*).

Después de eliminar las palabras “stop” se retuvieron las 10.000 de mayor frecuencia, y de éstas, las 9000 que mostraron mayor valor para el estadístico χ^2 .

OHSUMED HD-119

La colección OHSUMED consiste en abstracts de publicaciones médicas de Medline. Cada abstract está etiquetado con categorías de la ontología MeSH (*medical subject headings*), organizada por la Ohio State University (Hersh *et al.*, 1994). El conjunto utilizado aquí es la subdivisión de artículos publicados entre 1987 y 1991 que cuentan con categorizaciones provenientes del ramal “Heart Diseases” de MeSH, más específicamente, la que cuenta con 119 categorías MeSH (OHSUMED HD-119) (Yang, 1996). Si se consideran sólo las categorías que tienen al menos un documento en el conjunto de entrenamiento, este número se reduce a 103 categorías.

Esta versión cuenta con 12823 documentos de entrenamiento y 3758 documentos de test. Después de eliminar palabras “stop” se retuvieron las 14.000 palabras de mayor frecuencia, y de éstas se eligieron las 5000 con mayor valor de χ^2 .

6.2.3 Resultados

En la figura 6.2 se muestra una comparación entre la salida del módulo de tópicos para los modos serial y paralelo (ecuaciones 3.14 y 3.17), en dos documentos de la colección Reuters. Cada trazo corresponde al coeficiente

Reuters-2178 ModApte				OHSUMED
	11pt-AvgP	macro- F_1	micro- F_1	
kNN	0.93 ¹	0.59 ²	0.85 ²	
NaiveBayes		0.47 ²	0.80 ²	
SVM		0.64 ²	0.88 ²	
Selector de tópicos	0.91	0.48	0.79	
HD119				
	11pt-AvgP	macro- F_1	micro- F_1	
LLSF	0.78 ³	0.55 ³		
NaiveBayes				
SVM		0.58 ⁴		
Selector de tópicos	0.67	0.47	0.51	

Tabla 6.1: Comparación de los puntajes 11-avgP y macro-F1 obtenidos mediante el selector de tópicos con ejemplos de la literatura. Referencias: 1: Yang (1999); 2: Li y Yang (2003); 3: Yang (1996); 4: Yoon *et al.* (2006). Acrónimos: kNN: k-nearest neighbors, LLSF: linear least squares fit, SVM: support vector machines.

de cada una de las 90 categorías. Como se ilustra en estos dos ejemplos, los coeficientes que alcanzan valores más altos corresponden a las categorías correctas para cada documento sólo en el modo serial. En el modo paralelo no se logran diferenciar demasiado los coeficientes entre sí, creando asignaciones temáticas incorrectas. Esta situación se cumple para prácticamente todos los documentos de ambas colecciones, y por lo tanto, en adelante sólo se presentarán resultados del modo serial.

El desempeño del selector de tópicos en la colección Reuters se presenta en la figura 6.3. Para valores de β en el rango $[10^{-7}, 10^{-5}]$ se obtuvieron los mayor valores de F_1 y 11pt-avgP (cada máximo en diferentes valores del parámetro). Como puede apreciarse en la tabla 6.1, los niveles de desempeño obtenidos son comparables a los reportados en la literatura para los clasificadores bayesianos ingenuos (Naive Bayes, (Li y Yang, 2003)), y están algo por debajo de lo reportado para técnicas de referencia (por ejemplo, máquinas de soporte vectorial, SVM, (Li y Yang, 2003)).

En la figura 6.4 se muestran los valores de los indicadores de desempeño alcanzados en la colección OHSUMED HD-119. Los valores del parámetro β que lograron el mejor desempeño también estuvieron en el rango $[10^{-7}, 10^{-5}]$. Como se muestra en la tabla 6.1, los valores obtenidos son algo inferiores a los resultados reportados en la literatura para clasificadores de referencia.

6.3 Red recurrente simple

El módulo selector de tópicos evaluado en la sección precedente ilustra el potencial de este tipo de arquitecturas para implementar la noción de tópicos

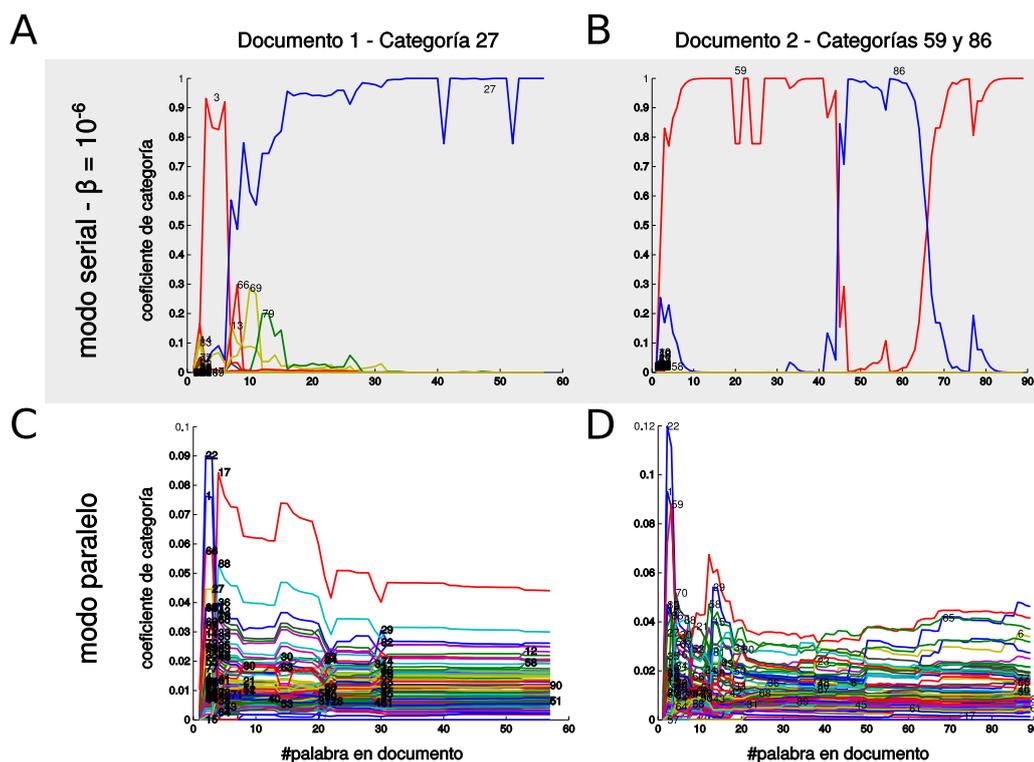


Figura 6.2: Ejemplo de salida del módulo selector de tópicos para dos documentos de la colección Reuters. Cada línea corresponde valor de la coordenada i del vector salida, para las sucesivas palabras del documento. Los números en cada línea corresponden a la dimensión i de cada trazo. Puede verse que en el modo serial, las dimensiones que resultan con coeficientes altos coinciden con las categorías a las que pertenecen. A. Modo serial para el Documento 1, categoría 27. El coeficiente de la dimensión 27 alcanza grandes valores luego de que las primeras 5 palabras ingresan al módulo. B. Modo serial para el Documento 27, categorías 59 y 86. A lo largo de todo el documento, se alternan en dominancia la dimensión 59 y la 86. C. Modo paralelo para el Documento 1.

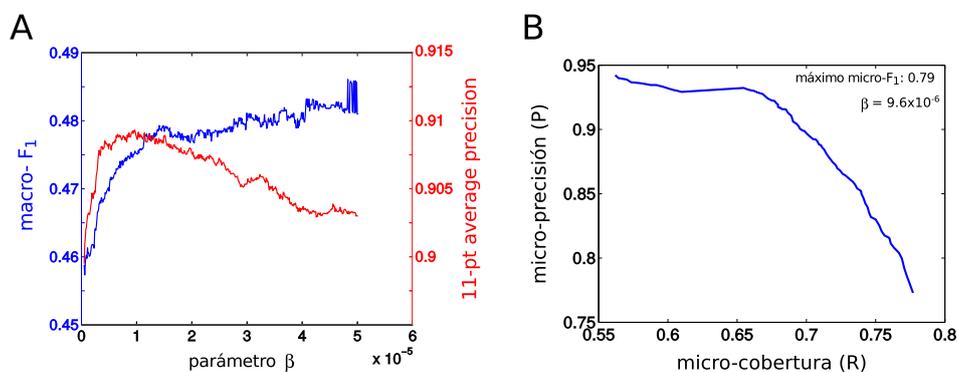


Figura 6.3: Desempeño del selector de tópicos en la colección Reuters-2178. A. Macro- F_1 y 11pt-avgP en función del parámetro β . B. Perfil de precisión y cobertura micro-promediados para el valor de β donde se alcanza el máximo valor de micro- F_1 .

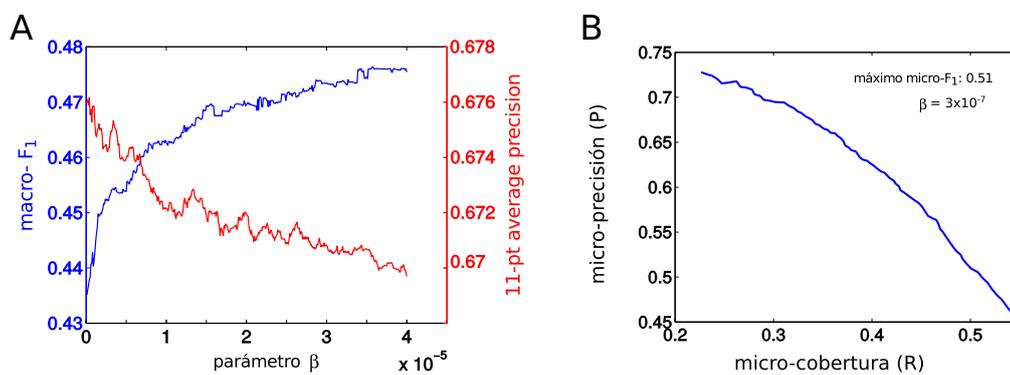


Figura 6.4: Desempeño del selector de tópicos en la colección OHSUMED-HD-119. A. Macro- F_1 y 11pt-avgP en función del parámetro β . B. Perfil de precisión y cobertura micro-promediados para el valor de β donde se alcanza el máximo valor de micro- F_1 .

o tema. Sin embargo, el hecho de que el entrenamiento deba ser supervisado en un conjunto de documentos cuyos temas ya estén identificados *a priori* va en detrimento de su posible en situaciones psicolingüísticas más realistas, que implican el análisis de un conjunto de información cuya estructura temática es desconocida.

Para intentar sortear esta dificultad planteamos la posibilidad de adaptar una arquitectura similar, la arquitectura SPELT (presentada en el capítulo 4.3.2), al procesamiento de grandes *corpora* no estructurados, con el fin de obtener una representación del significado similar a la que generan los espacios semánticos.

Para tal fin, llevamos adelante una estrategia en dos pasos. Primero probamos la funcionalidad de la arquitectura a pequeña escala, utilizando para ello un *corpus* de textos de poco volumen, más concretamente las primeras 1000 oraciones tomadas de 126 artículos de la enciclopedia en línea *Wikipedia* (la misma versión utilizada para desarrollar un espacio semántico en base a LSA, véase el apéndice C.1).

En esta colección se evaluó la capacidad de la red recurrente simple para obtener representaciones que reflejen aspectos gramaticales de las palabras, algo en lo que este tipo de topologías se especializa (Christiansen y Chater, 1994; Elman, 1991).

Una vez evaluada esta funcionalidad, se decidió probarla en una colección de textos de mayor tamaño y gran cobertura temática. Para ello se utilizó un *corpus* basado en la enciclopedia en línea *Wikipedia* en su versión en español. Los detalles acerca de la obtención de esta colección de textos se encuentran descritos en el apéndice C.1.1.

6.3.1 Métodos

Como se describe en la sección 4.3.2, la topología SPELT es una implementación de una red recurrente simple usando memorias matriciales dependientes de contexto (Valle-Lisboa *et al.*, 2005). La ventaja de esta arquitectura es que logra capturar relaciones de orden de palabras (un rasgo predominantemente sintáctico), lo que sumado a la expectativa de que desarrolle representaciones semánticas, tiene el potencial de lograr una funcionalidad avanzada (véase Jones *et al.* (2006) para argumentos en este sentido).

Sin embargo, como el objetivo es capturar también relaciones semánticas, se introdujo una variante en el funcionamiento de esta arquitectura. En la arquitectura general, la matriz de memoria M asocia a un vector de entrada que representa a la palabra (su representación “fonética”), f , con un vector de salida g que corresponde a la representación “conceptual” de la palabra que ocupa la siguiente posición en el discurso. Al entrar a la matriz, el vector f está modulado por el contexto p , que corresponde a la salida g del paso anterior. Este funcionamiento puede resumirse en la siguiente expresión

para la salida en el tiempo t :

$$g(t) = Mf(t) \otimes g(t-1) \quad (6.5)$$

La naturaleza de los vectores f y g es en buena medida arbitraria, si bien, como se vio en la sección 2.2.7, este tipo de memorias alcanza buenos desempeños cuando el conjunto de vectores de entrada es linealmente independiente. En las implementaciones realizadas en este trabajo, se eligieron vectores canónicos para el conjunto de vectores f .

La modificación introducida en este trabajo involucra la naturaleza de las representaciones de salida, g . Más específicamente, se introdujo un método de “agregación” de los vectores de salida. Esto es, para cada iteración k de la fase de aprendizaje, cada vector de salida que representa a la palabra w , $g_k(w)$ es actualizado de la siguiente manera:

$$g_{k+1}(w) = g_k(w) + \epsilon \left[\bar{g}_k(w) + \epsilon \sum_j \bar{g}_k(w_j) \right] \quad (6.6)$$

donde $\bar{g}_k(w)$ es el promedio de las salidas correspondientes a todas las ocurrencias de la palabra w , y la sumatoria en j es la suma de las salidas promedio de las palabras w_j que conforman las frases en las que la palabra w ocurre. El parámetro $\epsilon \ll 1$ controla la influencia de estos promedios.

La idea detrás de esta modificación es generar un acercamiento de los vectores de salida $g(w)$, de modo que si dos palabras w_i y w_j comparten el contexto de ocurrencia, por ejemplo, tienden a ocurrir después de las mismas palabras:

$$w_1 w_2 w_3 w_i, \quad w_2 w_3 w_3 w_j,$$

las predicciones de la red, a medida que va aprendiendo, serán una combinación de los vectores de $g(w_i)$ y $g(w_j)$. Por tanto, en el promedio $\bar{g}_k(w_i)$, habrá un peso importante de $g_k(w_j)$ y viceversa.

Esta idea está inspirada en modelos de espacios semánticos obtenidos por procedimientos de aglomeración de representaciones vectoriales, como por ejemplo BEAGLE (Jones *et al.*, 2006). En estos métodos las representaciones de las palabras comienzan como vectores completamente aleatorios, y se actualizan mediante un procedimiento que incluye sumar para cada palabra, los vectores de las palabras que co-ocurren en un mismo contexto, reflejando también información acerca del orden en que ocurren las palabras.

Producto de Kronecker incompleto

Si se pretende utilizar un *corpus* con 20.000 palabras, y una dimensión del espacio de salida de 300, por ejemplo, la naturaleza del producto de Kronecker impone que la matriz de memoria tenga 6 millones de columnas y 300 filas, lo que equivale a $1,8 \times 10^9$ elementos. Almacenar estos elementos

en doble precisión numérica implica unos 64 bits, u 8 bytes por elemento, es decir, $1,4 \times 10^{10}$ bytes, o 13.4 Gigabytes de memoria. Esto plantea una limitación importante en el número de palabras a incluir, ya que la memoria (y el tiempo de ejecución) son limitados (se emplearon en este trabajo equipos con 16Gb de RAM total). Para sortear esta dificultad se utilizó un resultado conocido para las memorias dependientes de contexto (Reali, 2002), que muestra que se pueden obtener buenos desempeños cuando el producto de Kronecker entre entrada y contexto es incompleto, y sólo se toman en cuenta una fracción de los elementos del producto. En las simulaciones efectuadas en este trabajo, se utilizó 1/3 de las dimensiones elegidas en forma aleatoria (fijas para cada simulación, diferentes para cada palabra) de las 300 o 400 presentes en el contexto. De esta forma, el tamaño de la matriz en la memoria se reduce a un tercio, y además se reduce el tiempo de cálculo.

Evaluación: categoría gramatical

Para evaluar el grado en que las representaciones obtenidas reflejan información gramatical, se tomaron las representaciones de las palabras más frecuentes, y se las etiquetó de acuerdo a la categoría gramatical más frecuente. Luego, se evaluó la capacidad de categorizar gramaticalmente a cada palabra en base a las representaciones y categorías de las palabras restantes. Para eso se utilizó el algoritmo de k vecinos más cercanos (kNN, *k-Nearest-Neighbors*), empleando como conjunto de entrenamiento a las $n - 1$ palabras restantes, para cada una de las n palabras (lo que se conoce como "validación cruzada dejando uno afuera", o *leave-one-out cross-validation*). La significación estadística de esta validación se estimó realizando 1000 réplicas de la misma, en las que los vectores de las representaciones se permutaron aleatoriamente, sin reposición (Phipson y Smyth, 2010).

Evaluación: predicción de asociados

Una forma de evaluar la incorporación de relaciones semánticas en un espacio semántico es compararlos con indicadores del desempeño humano, como por ejemplo las normas de asociación libre de palabras (Griffiths *et al.*, 2007; Kiss *et al.*, 1973; Nelson *et al.*, 1998). Por ejemplo, puede evaluarse en qué medida las palabras que ocurren con mayor frecuencia como asociados a una palabra dada en las normas de asociación libre (los primeros o segundos asociados) se encuentran entre las palabras más cercanas a esa palabra en el espacio semántico.

Otro recurso utilizado en la literatura para evaluar espacios semánticos son los experimentos de facilitación semántica (Hare *et al.*, 2009; Jones *et al.*, 2006; Lund *et al.*, 1995), donde la reducción en los tiempos de reacción a una palabra es causada por otra presentada previamente es evidencia de una relación semántica entre ellas.

Con la intención de realizar este tipo de validaciones, y al mismo tiempo generar material de interés psicolingüístico adaptado a la realidad local, decidimos coleccionar datos de asociación libre en español rioplatense. En colaboración con Simon De Deyne de la Universidad Católica de Lovena (KU Leuven), en Bélgica, coleccionamos datos de asociación libre de hablantes nativos del español rioplatense para 1.000 palabras de alta frecuencia (utilizando la plataforma web y el método descrito por De Deyne *et al.* (2013)). Las palabras estímulo fueron elegidas a partir de las normas de frecuencia LEXESP (Sebastián-Gallés, 2000) y SUBTLEX-ESP (Cuetos *et al.*, 2011), excluyendo palabras de la categoría cerrada. Se coleccionaron respuestas de más de 2.500 personas, cada palabra fue respondida por un promedio de 40 personas. Se descartaron las respuestas de más de una palabra, y se contabilizó la frecuencia de las respuestas. Para cada palabra estímulo, las respuestas se ordenaron en frecuencia descendiente, siendo la más frecuente el primer asociado, la segunda más frecuente el segundo asociado, etc.

Para evaluar la correlación entre los resultados de asociación libre y los espacios semánticos obtenidos, se calcularon las distancias entre cada palabra estímulo y todas las palabras del espacio. En base a estas distancias, para cada palabra estímulo se ordenan las palabras según similitud descendiente, y se contabiliza el rango que ostenta el primer asociado en esta lista.

Para comparar el desempeño de la red recurrente simple con otros métodos basados en co-ocurrencia léxica, se obtuvo un espacio semántico utilizando LSA en base a la misma colección de textos. Es necesario notar, sin embargo, que el tratamiento de los documentos es diferente. Por ejemplo, para el LSA es necesario remover gran parte de las palabras funcionales, lo que determina que la naturaleza de los textos en los que se apoyan los diferentes métodos evaluados no sean exactamente idénticos (véase el apéndice C.1.1 para más detalles).

Corpora y parámetros de simulación

Para la primera simulación se utilizó el *corpus* pequeño, que contó con 4331 palabras y 18789 ocurrencias, se utilizó un espacio de salida con 300 dimensiones, y se siguió la evolución de las representaciones de salida durante 990 iteraciones. La constante de aprendizaje utilizada fue $\alpha = 0,2$, y la constante de aglomeración de representaciones $\epsilon = 0,001$.

En la simulación en base al *corpus* Wikipedia en español, que contó con 20287 palabras y 374361 ocurrencias en 39709 oraciones provenientes de 20217 documentos, se utilizó un espacio de salida de 400 dimensiones, y se siguió la evolución de las representaciones de salida durante 390 iteraciones. Este número resulta bastante bajo, pero es un resultado del extenso tiempo que lleva cada iteración. La constante de aprendizaje fue $\alpha = 0,2$, y se fue reduciendo en un factor de 5/7 cada vez que el error entre dos iteraciones se reducía en un valor menor a un umbral fijado en 0.0001. La constante de

aglomeración fue de $\epsilon = 0,001$.

En ambos casos, las representaciones de las palabras de entrada fueron vectores canónicos, lo que favorece la perpendicularidad de la entrada a la matriz de memoria a la vez que agiliza los cálculos (en cada iteración sólo se modifica una porción de la matriz de pesos sinápticos). La simulación fue implementada utilizando el lenguaje Java, utilizando una librería de álgebra lineal específicamente diseñada para aprovechar la disponibilidad de varios núcleos en los procesadores actuales (“Parallel Colt”, [Wendykier y Nagy \(2010\)](#)). Se utilizó una computadora de escritorio con procesador Intel i7 (3770 Ivy Bridge), con 16 Gb de memoria RAM.

6.3.2 Resultados

Corpus Wikipedia en miniatura

En la figura 6.5 se muestra un dendrograma que ilustra el agrupamiento jerárquico obtenido utilizando las representaciones de salida para 46 palabras de alta frecuencia (distancia promedio sin peso (UPGMA), distancia coseno). Para cada palabra se calculó su categoría gramatical más probable, en base a las ocurrencias de las palabras en la sección J del *corpus Brown* en su versión Semcor 3.0 ([Miller et al., 1994](#)). Esta colección cuenta con anotaciones gramaticales de “part of speech” según la codificación del Penn Treebank [Marcus et al. \(1994\)](#). En el dendrograma puede verse un cierto agrupamiento de los verbos, los determinantes y las preposiciones, que dista de ser perfecto.

La significación estadística de este agrupamiento puede inferirse de la figura 6.6. En la figura 6.6 A se muestra en número de categorizaciones correctas utilizando kNN en base a las representaciones de salida obtenidas, en función de la iteración alcanzada. A partir de la iteración 300 se clasifican correctamente más de 20 palabras, alcanzando un máximo de 28 en las 420 iteraciones, para estabilizarse relativamente en el entorno de las 22 clasificaciones correctas hacia el final de la simulación. El histograma que se muestra en la figura 6.6 B permite estimar la probabilidad de clasificar correctamente n palabras en forma aleatoria (bajo la hipótesis nula de que las representaciones no informan sobre la categoría gramatical de las palabras). Dado que de las 1000 réplicas realizadas ninguna superó las 20 correctas, es razonable pensar que la ocurrencia de $n > 20$ clasificaciones correctas es un hecho estadísticamente significativo.

Wikipedia en español

En la figura 6.7 se muestra un dendrograma obtenido a partir de la agrupamiento jerárquica (distancia promedio sin peso (UPGMA), distancia coseno) de las representaciones de salida para 195 palabras muy frecuentes en el *corpus Wikipedia* utilizado. Cada palabra está categorizada según su categoría gramatical más probable, según el número de acepciones presentes en el

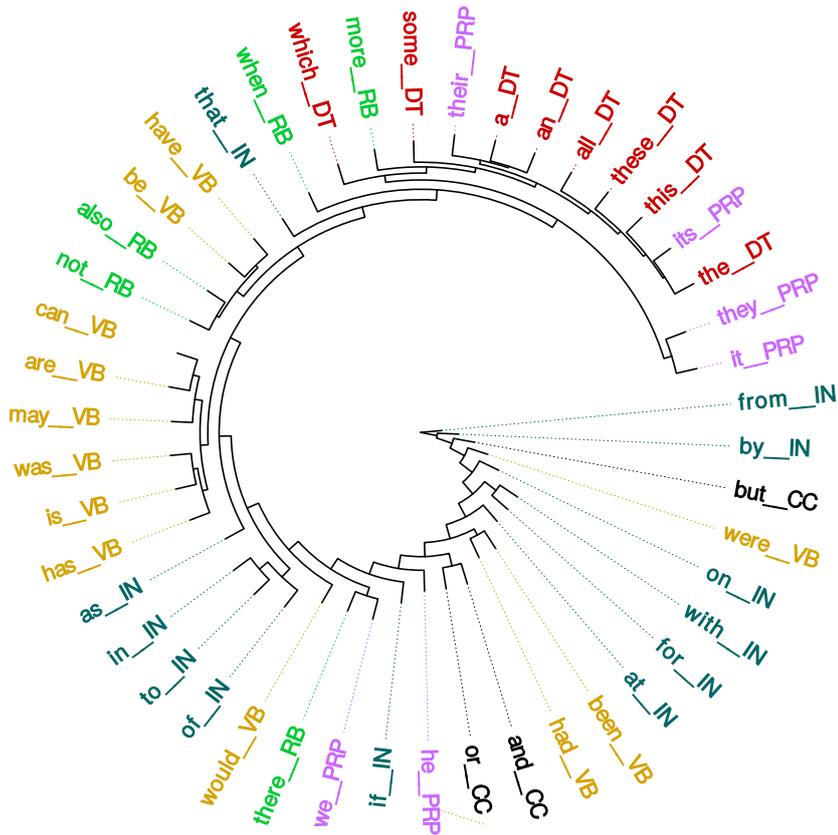


Figura 6.5: Dendrograma obtenido a partir de las representaciones de salida para 46 palabras muy frecuentes realizadas por el módulo de red recurrente simple aplicado a una muestra del *Wikipedia* en inglés. Las marcas en mayúsculas en las hojas del dendrograma, al igual que los colores, corresponden a la etiqueta gramatical más probable de cada palabra, adaptadas de la notación del Penn Treebank (Marcus *et al.*, 1994). vb: verbo, rb: adverbio, prp: pronombre personal, in: preposición, dt: determinante, cc: conjunción.

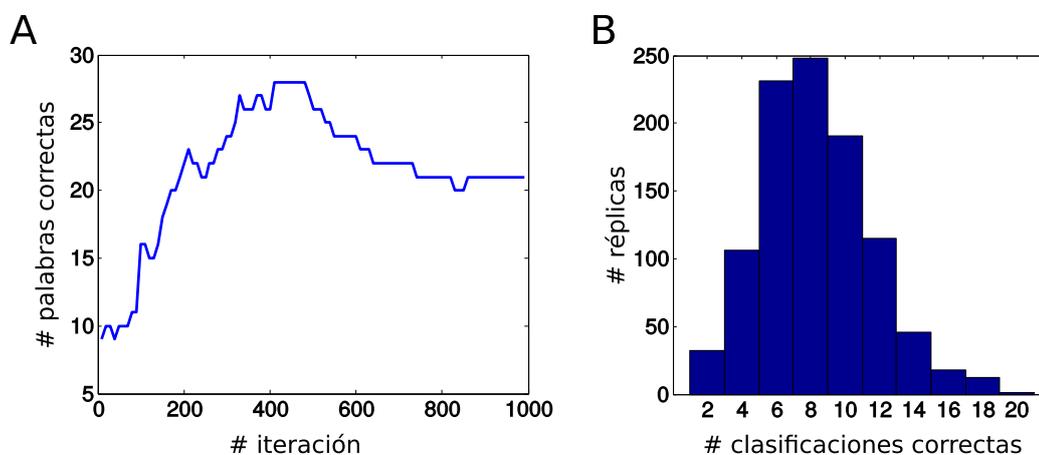


Figura 6.6: A. Número de palabras correctamente clasificadas de acuerdo a su categoría gramatical más probable, en base a las representaciones obtenidas por la red recurrente simple utilizando el algoritmo de k -Nearest-Neighbors, con $k = 5$. B. Estimación de la probabilidad de clasificación correcta de n palabras en base al algoritmo kNN , con $k = 5$, 420 iteraciones. El histograma refleja el número de veces que se obtuvieron n clasificaciones correctas, de un total de 1000 corridas con los vectores de salida permutados aleatoriamente. Ninguna de las permutaciones logró más de 20 clasificaciones correctas.

Diccionario de la Real Academia Española ([Real Academia Española, 2014](#)) y en base a inspección y evaluación subjetiva. Pueden identificarse en el árbol una región donde se agrupan principalmente los sustantivos masculinos, y otra región donde preponderantemente aparecen sustantivos femeninos. También aparecen agrupados algunos adverbios y algunas preposiciones.

En la figura 6.8 A se muestra el desempeño de un clasificador kNN en el conjunto de las 195 palabras, en función de la iteración. A partir de la iteración 250 aproximadamente, se clasifican correctamente más de 70 palabras, alcanzándose un máximo de 88 categorizaciones correctas a las 320 iteraciones. El análisis de permutaciones aleatorias que se muestra en la figura 6.8 B sugiere que la clasificación correcta de más de 70 palabras es estadísticamente significativa.

Los resultados de la comparación con el experimento de asociación libre pueden verse en la figura 6.9. En la figura 6.9 A se observa el desempeño de la red recurrente simple en comparación con una matriz de ocurrencia palabra-documento (que oficia de línea de base), y dos versiones de LSA utilizando 50 y 100 dimensiones. Puede verse que la probabilidad de incluir a un primer asociado en las m palabras más cercanas en función del tamaño m es mucho menor para la red recurrente simple (incluso está debajo de la línea de base propuesta en base a la matriz de ocurrencias documento-palabra). Sin embargo, como puede verse en la figura 6.9 B, el desempeño de la red no es debido al azar, al menos no a partir de la iteración 180. Es

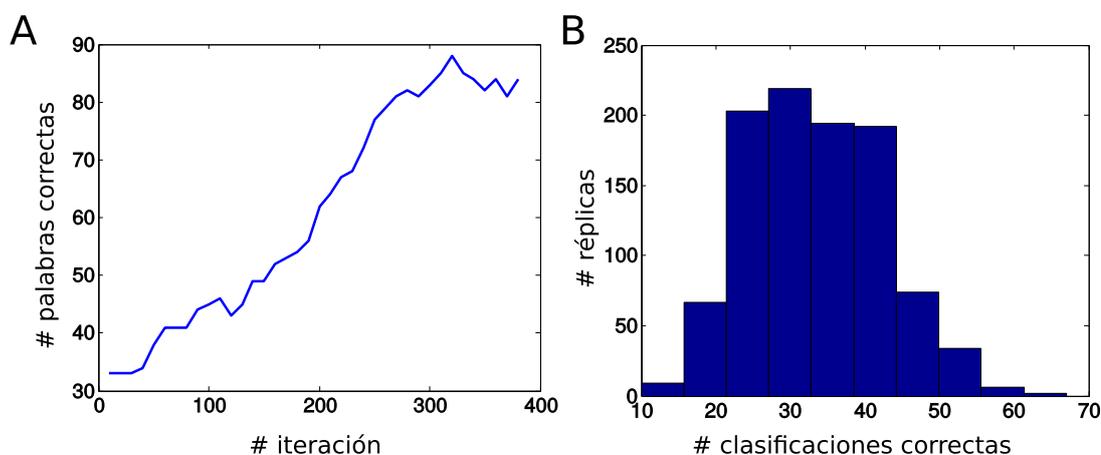


Figura 6.8: A. Número de palabras correctamente clasificadas de acuerdo a su categoría gramatical más probable, en base a las representaciones obtenidas por la red recurrente simple utilizando el algoritmo de k -Nearest-Neighbors, con $k = 5$, en función del número de iteraciones alcanzadas por la red. B. Estimación de la probabilidad de clasificación correcta de n palabras en base al algoritmo kNN , con $k = 5$, para 320 iteraciones. El histograma refleja el número de veces que se obtuvieron n clasificaciones correctas, de un total de 1000 corridas con los vectores de salida permutados aleatoriamente. Ninguna de permutaciones logró más de 71 clasificaciones correctas.

más, el desempeño de la red mejora en forma monotónica al aumentar las iteraciones. A partir de la iteración 180 supera el desempeño mínimo posible en base al orden aleatorio de las palabras, en particular cuando se consideran conjuntos de palabras superiores a 200.

6.4 Discusión

El módulo selector de tópicos logró un desempeño comparable a otras implementaciones no neurales encontradas en la literatura, en particular, el desempeño es casi idéntico al de los clasificadores bayesianos ingenuos. Esto último es entendible, dado que el selector de tópicos funciona, en esencia, como un clasificador de este tipo (Valle-Lisboa, 2007).

En suma, el selector de tópicos implementa un espacio semántico donde las dimensiones son directamente interpretables en términos de tópicos. Si bien podría utilizarse este tipo de módulos para implementar un espacio semántico de naturaleza general, que emule el desempeño cognitivo humano, es necesario para ello contar con un *corpus* etiquetado, en el que los tópicos de cada documento sean conocidos a la hora de realizar el entrenamiento. Esta situación dista mucho de ser la común para grandes colecciones de texto como las empleadas usualmente para derivar espacios semánticos de este tipo. Este hecho motiva la búsqueda de arquitecturas alternativas, para las

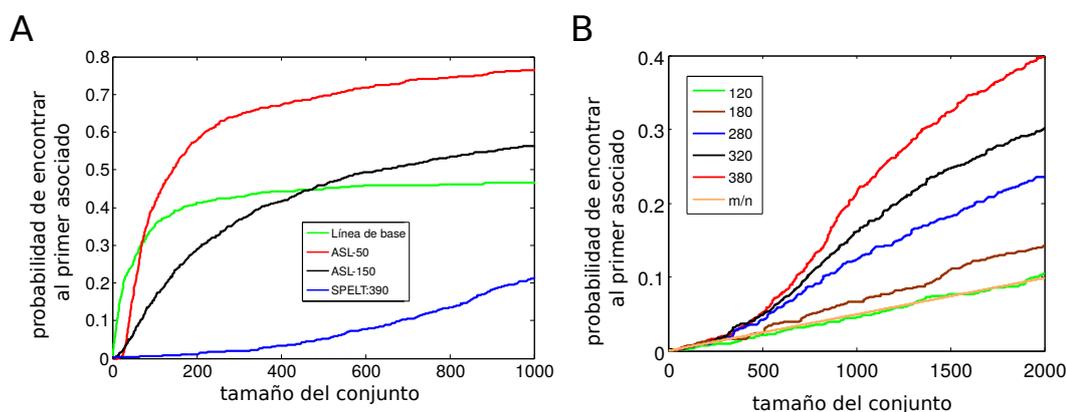


Figura 6.9: Probabilidad de encontrar al primer asociado de una palabra en el conjunto de las m palabras más cercanas, en función de m . A. Comparación entre los resultados obtenidos con LSA (con $k = 50$ y $k = 150$ dimensiones) y la red recurrente simple (iteración 390). La línea de base resulta de utilizar la matriz documento-palabra utilizada en el LSA sin reducción de dimensionalidad (con k igual al rango de la matriz). El desempeño de la red recurrente simple resulta muy inferior incluso a la línea de base. B. Comparación entre los resultados obtenidos en distintas iteraciones de la red recurrente simple. Puede verse que el desempeño es mayor para iteraciones posteriores. Además, el desempeño para la iteración 120 no supera el desempeño mínimo debido al azar, igual a una probabilidad de m/n de encontrar al primer asociado entre las primeras m palabras, de un total de n palabras.

cuales el desarrollo de las representaciones semánticas se realice en forma no supervisada.

El modelo de red recurrente simple empleado en esta tesis, logra capturar información de carácter gramatical, y codificarla en las representaciones de los vectores de salida. Las implementaciones de redes recurrentes simples basadas en perceptrones multicapa entrenadas con *backpropagation* codifican información gramatical en las activaciones de la capa oculta [Elman \(1991\)](#) (véase sin embargo a [Borensztajn et al. \(2014\)](#) para una posición contrastante). En suma, puede verse a la arquitectura aquí presentada como funcionalmente equivalente a las redes recurrentes simples clásicas.

Sin embargo, la implementación aquí utilizada (propuesta en esencia por [Reali \(2002\)](#)) se basa en memorias matriciales dependientes de contexto, lo que resulta una gran ventaja desde el punto de vista de los requerimientos computacionales. Esto permite escalar el funcionamiento de este tipo de redes y exponerlas a *corpora* realistas y de gran volumen. En este trabajo de tesis se entrena por primera vez una red recurrente simple en colecciones de textos de estas características.

En cuanto a la incorporación de información gramatical en las representaciones de salida, puede verse en ambos *corpora* que parece alcanzarse durante el entrenamiento de la red un máximo en el desempeño de la red. En iteraciones posteriores el contenido de información presente pareciera dismi-

nuir. Esto puede deberse a un fenómeno de sobre-ajuste a las peculiaridades del *corpus* empleado.

Sin embargo, es relevante recordar que el procedimiento de agregación de las representaciones de salida fue propuesto con la intención de no sólo capturar relaciones de tipo gramatical, sino también relaciones de carácter semántico. En el *corpus* donde el contenido semántico de las representaciones fue evaluado, el desempeño de la red en esta dimensión parece ir en aumento con las iteraciones, incluso después de alcanzar el máximo de contenido gramatical. Pareciera entonces que en las representaciones de salida aumenta el contenido semántico, quizás en detrimento del contenido gramatical, si bien las iteraciones alcanzadas en esta simulación son muy pocas para poder afirmar esto con seguridad.

Muy probablemente a causa de esto último, el desempeño de la red en la tarea de predicción de primeros asociados fue muy inferior al mostrado por el Análisis Semántico Latente, incluso por una simple matriz de ocurrencia documento-palabra. Sin embargo, a partir de la iteración 180 por lo menos, el desempeño está por encima del esperado debido completamente al azar, lo que muestra efectivamente que la red codifica información de naturaleza semántica en las representaciones vectoriales.

Puede argumentarse entonces que esta arquitectura es una implementación de naturaleza neural de un espacio semántico, entrenado a partir de colecciones de textos de naturaleza similar a los utilizados en implementaciones no neurales como el Análisis Semántico Latente, BEAGLE, los modelos probabilísticos de tópicos, etc. Forma parte, por tanto, de una categoría poco poblada, de la que es un ejemplo el modelo de [Blouw y Eliasmith \(2013\)](#) (que en realidad es una propuesta de implementación neural, más que una implementación en sí misma), entre otros ([Mizraji et al., 2009](#); [Shankar et al., 2009](#)).

Capítulo 7

Cuantificación de coherencia del discurso

Como se argumentó en el capítulo 4, la cuantificación de la coherencia en el discurso puede ser una herramienta importante en el ámbito de la investigación en esquizofrenia; en particular, si esta cuantificación se logra realizar en forma automática y estandarizada. En este capítulo se presentan los esfuerzos en esa dirección que se realizaron en el marco de esta tesis. En primer lugar se introduce un nuevo tipo de medida de la desorganización de la sucesión de temas, las entropías temáticas. Luego se presenta la aplicación de estas medidas a discursos de pacientes con esquizofrenia. Finalmente se aborda una propuesta de automatización del proceso de identificación de tópicos y la subsiguiente aplicación de las entropías, también aplicado al análisis del discurso de pacientes con esquizofrenia.

7.1 Entropías temáticas

La entropía es un concepto surgido en el marco de la termodinámica clásica, propuesto por Rudolf Clausius, relacionado con la reversibilidad de los procesos y la disipación de energía. La definición posterior, en términos estadísticos fue realizada por Ludwig Boltzmann, y reencontrada en el contexto de la transmisión de la información por Claude Shannon (Shannon, 1949), en lo que daría origen a la moderna teoría de la información. En esta última formulación, la entropía H de una fuente x de símbolos toma la siguiente forma:

$$H(x) = - \sum p_i \log(p_i). \quad (7.1)$$

Los p_i son las probabilidades de emisión de los diferentes símbolos, de modo que una distribución de p_i sesgada fuertemente hacia un valor muy probable y otros muy poco probables genera una entropía menor, mientras que una distribución totalmente homogénea ($p_i = 1/N$, con N el número de símbolos diferentes existentes) alcanza una entropía máxima en $H = \log(N)$.

En el contexto del análisis del discurso, utilizamos una herramienta inspirada en esta formulación de la entropía para cuantificar el nivel de desorden presente en la sucesión de temas que muestra un texto o discurso. Un texto o discurso ordenado en extremo, en este contexto, es aquel en el que cada tema aparece en un solo bloque contiguo, sin estar interrumpido por porciones de otros temas. También aquel en el que la sucesión entre temas es fija: después del tema α siempre le sigue el tema β , por ejemplo¹. En cambio, en un discurso desordenado los temas se interrumpen unos a otros reaparecen, varias veces, y sin una sucesión clara. En la figura 7.1 se ilustran estos dos casos paradigmáticos.

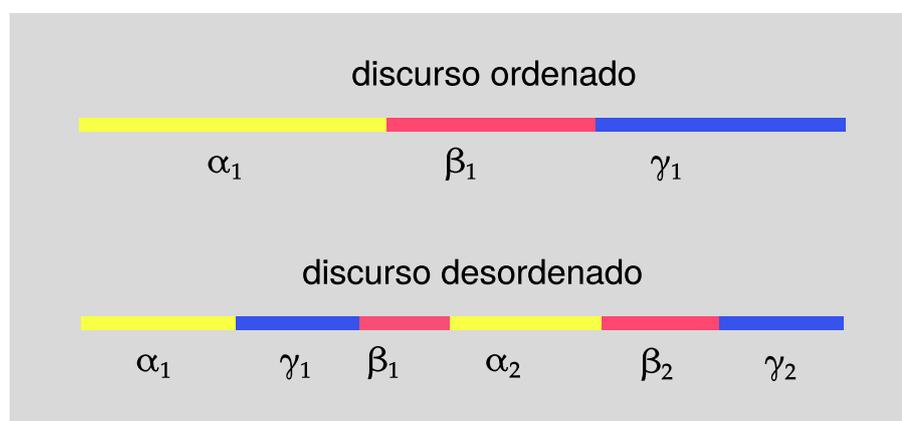


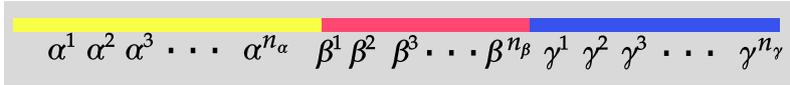
Figura 7.1: Esquema de sucesión de temas en un discurso hipotético, ordenado vs. desordenado.

Para caracterizar estas dos nociones de orden y desorden (por un lado grado de particionamiento o interrupción, por otro el grado de regularidad en las transiciones), formulamos las nociones de *entropía de tema* y *entropía de transición*, descritas en la sección siguiente.

7.1.1 Entropía de tema

La entropía de tema refleja el grado de particionamiento que tiene un tema en el discurso. Por ejemplo, como se mostró anteriormente, si el discurso trata de los temas α , β , y γ , un discurso ordenado será uno en el que cada uno de éstos se transiten en su totalidad en forma compacta y continuada, sin intercalarse. A modo de ilustración, si a cada palabra j se le puede asignar un tema y representarla entonces como α^j , β^j o γ^j , un discurso muy ordenado se vería como una sucesión similar a ésta:

¹Esto fuerza que si un texto es ordenado en el primer sentido, también lo sea en el segundo, dado que sólo existe una sola transición entre α y β .



donde n_α, n_β y n_γ son el número de palabras asignadas a cada tema.

En el otro extremo de la distribución de fragmentos de temas, un texto muy desordenado podría ser uno donde se intercalen fragmentos de texto de diferentes temas repetidamente:



En este caso, los temas están distribuidos en diferentes segmentos, que en adelante llamaremos “bloques temáticos” y denotaremos con subíndices, $\alpha_i, \beta_i, \gamma_i$, etc. Para el ejemplo anterior, los bloques temáticos son los siguientes:



Para cada bloque temático α_i puede definirse la fracción del tamaño total que representa. Este tamaño relativo de cada bloque, que denotaremos aquí como $p(\alpha_i)$, se define como:

$$\begin{aligned}
 p(\alpha_i) &= \frac{\text{número de palabras del bloque } i}{\text{número total de palabras del tópico } \alpha} \\
 &= \frac{n(\alpha_i)}{n_\alpha}
 \end{aligned}
 \tag{7.2}$$

La entropía de tema usa estas frecuencias relativas de forma análoga a las “probabilidades” en la entropía de Shannon. La entropía de un tema $\alpha, S(\alpha)$ se define mediante la siguiente expresión:

$$S(\alpha) = - \sum_{i=1}^{N_\alpha} p(\alpha_i) \log p(\alpha_i)
 \tag{7.3}$$

donde N_α es el número de bloques en que está particionado el tema α .

En el primero de los ejemplos anteriores todos los temas están presentes en un solo bloque; solo existe entonces un p igual a 1, y la entropía de tema se vuelve cero ($S(\alpha) = S(\beta) = S(\gamma) = 0$). En cambio, en el segundo ejemplo, se tiene que:

$$\begin{aligned}
 p(\alpha_1) &= 2/5 & p(\alpha_2) &= p(\alpha_3) = p(\alpha_4) = 1/5, \\
 p(\beta_1) &= 3/6 & p(\beta_2) &= 1/6 & p(\beta_3) &= 2/6, \\
 p(\gamma_1) &= 1/3 & p(\gamma_2) &= 2/3
 \end{aligned}$$

y por tanto las entropías se vuelven:

$$S(\alpha) = \frac{2}{5} \log \frac{5}{2} + \frac{3}{5} \log 5$$

$$S(\beta) = \frac{3}{6} \log \frac{6}{3} + \frac{1}{6} \log 6 + \frac{2}{6} \log \frac{6}{2}$$

$$S(\gamma) = \frac{1}{3} \log 3 + \frac{2}{3} \log \frac{3}{2}$$

La entropía de un discurso \mathcal{D} , $S(\mathcal{D})$, es simplemente la suma de las entropías para todos los temas presentes,

$$S(\mathcal{D}) = - \sum_{\tau} S(\tau), \quad (7.4)$$

mientras que la entropía promedio del discurso es el promedio de las entropías de los temas:

$$\overline{S(\mathcal{D})} = - \frac{1}{N_T} \sum_{\tau} S(\tau), \quad (7.5)$$

donde N_T es el número de temas presentes en el discurso.

Entropía de tema normalizada

A la hora de comparar valores de entropía de tema de diferentes discursos, con distinto largo y distinto número de temas, puede ser necesario calcular el valor máximo de entropía a fin de contar con una métrica normalizada. Una posible normalización implica considerar, dado un discurso con n_T palabras, y una partición en temas α, β, \dots , la entropía máxima que puede alcanzar cada tema depende de los tamaños relativos de los temas $n_{\alpha}, n_{\beta}, \dots$, etc.

Si el tamaño del tema de mayor, digamos n_{α} , es menor o igual a la suma de los otros tamaños más uno, es decir:

$$n_{\alpha} \leq s = 1 + \sum_{\tau \neq \alpha} n_{\tau}$$

entonces la entropía máxima de todos los temas será:

$$S^{max}(\tau) = \log n_{\tau}. \quad (7.6)$$

De lo contrario, la expresión es la siguiente:

$$S^{max}(\tau) = (a + b)n_{\tau} - a \log c - b \log(c - 1) \quad (7.7)$$

donde

$$a = \frac{rc}{n_{\tau}}, \quad b = \frac{(s-r)(c-1)}{n_{\tau}}, \quad c = \left\lceil \frac{n_{\tau}}{s} \right\rceil, \quad r = \frac{\text{resto}\{n_{\tau}/s\}}{c-1}$$

En el Apéndice B.1 se encuentran los detalles de la derivación de estas cantidades.

La entropía máxima de un discurso, la definimos aquí como la suma de las entropías máximas de cada tema:

$$S^{max}(\mathcal{D}) = \sum_{\alpha} S^{max}(\alpha)$$

y la entropía de temas normalizada, es el cociente entre la entropía del discurso y la entropía máxima:

$$S^{norm}(\mathcal{D}) = \frac{S(\mathcal{D})}{S^{max}(\mathcal{D})} \tag{7.8}$$

7.1.2 Entropía de transición

La entropía de transición refleja la regularidad con la que un tema es precedido por otros. Para definirla, se recurre a la noción de “probabilidad de transición”, $p_{\alpha}(\tau)$:

$$p_{\alpha}(\tau) = \frac{\text{número de transiciones del tema } \alpha \text{ al tema } \tau}{\text{número de transiciones del tema } \alpha \text{ a otros temas}} \tag{7.9}$$

La expresión para la entropía de transición de un tema α , $T(\alpha)$, es la siguiente:

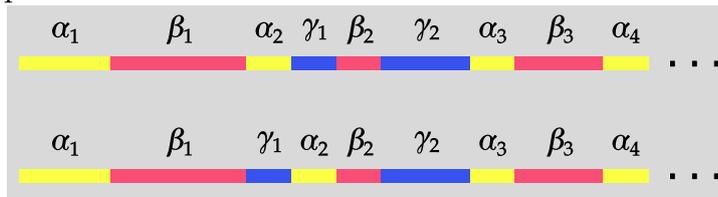
$$T(\alpha) = - \sum_{\tau \neq \alpha} p_{\alpha}(\tau) \log p_{\alpha}(\tau), \tag{7.10}$$

La entropía de transición de un discurso \mathcal{D} , $T(\mathcal{D})$ es la suma de las entropías de los temas del discurso, y la entropía promedio es el promedio de las entropías de los temas:

$$T(\mathcal{D}) = \sum_{\tau}^{N_T} T(\tau) \tag{7.11}$$

$$\overline{T(\mathcal{D})} = \frac{1}{N_T} \sum_{\tau}^{N_T} T(\tau) \tag{7.12}$$

En el siguiente ejemplo, los dos discursos tienen el mismo grado de partición de los temas, y por lo tanto la misma entropía de tema. Sin embargo las probabilidades de transición son diferentes:



En el primer discurso, las probabilidades de transición son:

$$\begin{aligned} p_\alpha(\beta) &= 2/3 & p_\alpha(\gamma) &= 1/3 \\ p_\beta(\alpha) &= 2/3 & p_\beta(\gamma) &= 1/3 \\ p_\gamma(\alpha) &= 1 & p_\gamma(\beta) &= 0 \end{aligned} \tag{7.13}$$

mientras que para el segundo son:

$$\begin{aligned} p_\alpha(\beta) &= 1 & p_\alpha(\gamma) &= 0 \\ p_\beta(\alpha) &= 1/3 & p_\beta(\gamma) &= 2/3 \\ p_\gamma(\alpha) &= 1 & p_\gamma(\beta) &= 0 \end{aligned} \tag{7.14}$$

y por tanto las entropías se vuelven:

$$T(\alpha) = \frac{2}{3} \log \frac{3}{2} + \frac{1}{3} \log 3$$

$$T(\beta) = \frac{2}{3} \log \frac{3}{2} + \frac{1}{3} \log 3$$

$$T(\gamma) = 0$$

para el primer ejemplo, y para el segundo los valores son:

$$T(\alpha) = 0$$

$$T(\beta) = \frac{1}{3} \log 3 + \frac{2}{3} \log \frac{3}{2}$$

$$T(\gamma) = 0$$

En el segundo ejemplo, la entropía de transición del tema α es cero, ya que siempre después de este tema sigue el tema β , y por tanto la entropía de transición del discurso es menor que en el primer ejemplo, si bien, como ya se mencionó, las entropías temáticas son las mismas.

Puede verse entonces que para un discurso periódico en el que cada vez que termina el tópico α le sigue siempre el tópico β , entonces $p_\alpha(\beta) = 1$, y todos los otros $p_\alpha(\tau) = 0$, haciendo que la entropía de transición de α sea cero: $T(\alpha) = 0$. Contrariamente, cuando después de un tema α siguen los otros temas con igual probabilidad, la entropía para el tema alcanza su máximo (véase el Apéndice B.1).

Entropía de transición normalizada

La entropía de transición alcanza su máximo cuando las transiciones hacia los otros temas están igualmente representados, y por tanto, la entropía máxima de transición se vuelve:

$$T^{max}(\alpha) = \log(N_T - 1), \quad (7.15)$$

Sin embargo, puede suceder que el tema conste de menos palabras que el número total de temas:

$$n_\alpha < N_T - 1$$

y en ese caso el máximo de entropía de transición se vuelve:

$$T^{max}(\alpha) = \log(n_\alpha),$$

Tomando en cuenta esta posibilidad, la expresión general queda:

$$T^{max}(\alpha) = \log \min(N_T - 1, n_\alpha). \quad (7.16)$$

La entropía máxima de transición para un discurso es la suma de las entropías máximas de cada tema:

$$T^{max}(\mathcal{D}) = \sum_{\alpha} T^{max}(\alpha),$$

y la entropía normalizada del discurso se vuelve:

$$T^{norm}(\mathcal{D}) = \frac{T(\mathcal{D})}{T^{max}(\mathcal{D})} \quad (7.17)$$

En el Apéndice B.1 pueden verse algunos detalles más específicos sobre estas derivaciones.

7.1.3 Entropías y desorganización

Por definición, las entropías serán bajas para discursos cuyas progresiones temáticas sean ordenadas, y altas para los casos en los que sean desordenadas. Para ilustrar este comportamiento se realizó el siguiente test.

Partiendo de una asignación temática manual realizada a un fragmento de la obra “A Study in Scarlet” de Arthur Conan Doyle (Cabana, 2009), se realizó una mezcla aleatoria progresiva de los fragmentos del texto como se ilustra en la figura 7.2 A: de esta manera se obtienen versiones cada vez más desordenadas del fragmento original (figura 7.2 B). Tanto la entropía de tema como la de transición aumentan a medida que el texto es desordenado (figura 7.2 C y D). La entropía de transición alcanza un valor estable mucho antes que la entropía de tema, mostrando una variación menor frente al desorden. Este valor estable está muy cerca del valor máximo de entropía de transición, el cual en general puede alcanzarse mucho antes que el máximo de la de tema (véase Apéndice B.1).

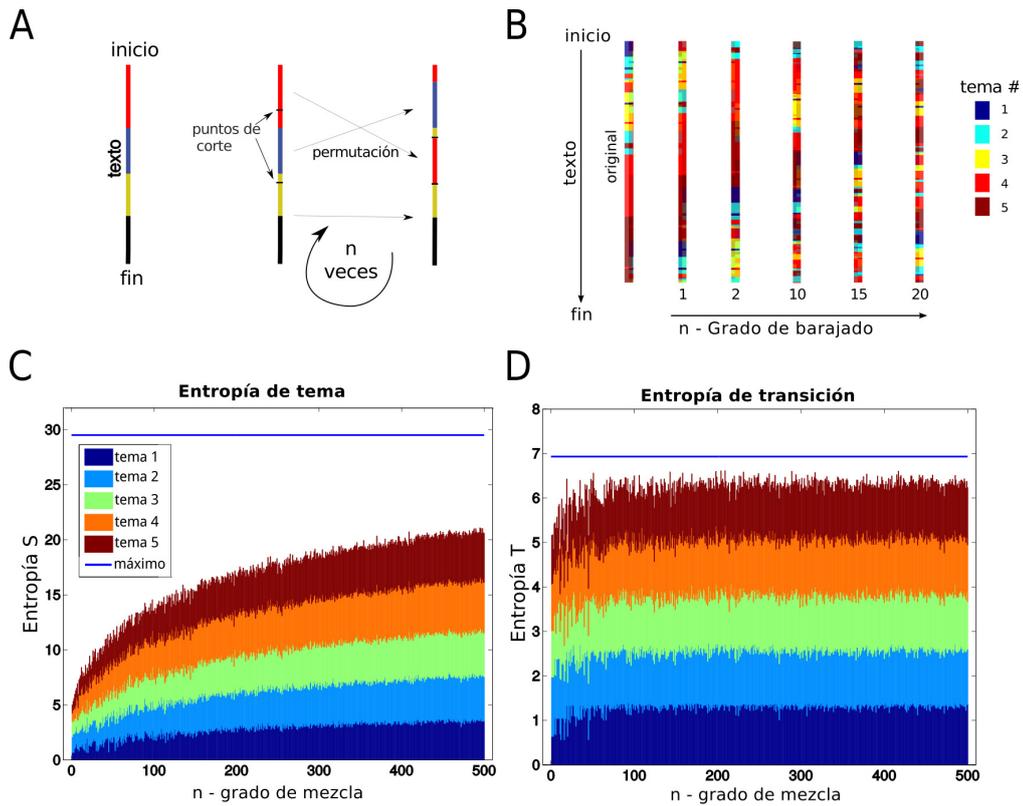


Figura 7.2: Respuesta de las entropías temáticas frente al desorden temático. A. Procedimiento de barajado de un texto. B. Ilustración del efecto del barajado en una asignación temática del fragmento de “A Study in Scarlet” usado en (Cabana, 2009). C. Aumento de la entropía de tema y D. de la entropía de transición a medida que el texto se desordena.

7.2 Aplicación a discurso de pacientes con esquizofrenia

Con el fin de evaluar la aptitud de las entropías para cuantificar el grado de desorden temático del discurso, aplicamos las entropías al análisis del discurso de pacientes esquizofrénicos. Para eso contamos con la colaboración de la Dra. Brita Elvevåg, en ese entonces en el National Institute of Mental Health, Maryland, EEUU (actualmente en la Universidad de Tromsø, Noruega). Los resultados de esta colaboración se publicaron en un artículo que se adjunta al final de esta sección (Cabana *et al.*, 2011).

En base a una técnica desarrollada en mi tesis de maestría (Cabana, 2009), construimos grafos temáticos para diferentes textos, a partir de los cuales se obtuvieron asignaciones temáticas para cada uno. Los textos correspondieron a un fragmento de “A Study in Scarlet”² que narraba en primera persona un monólogo de Sherlock Holmes, y un fragmento de transcripción de un paciente con esquizofrenia reportado en Andreasen (1986). Ambos fragmentos tiene una longitud comparable, y el carácter metafórico del primero hace que en éste se encuentren palabras que en un otro tipo de texto, como un relato, es más difícil que co-ocuran.

Una mejor comparación entre textos se realizó utilizando otros tres fragmentos de transcripciones de entrevistas con pacientes con esquizofrenia, y otro fragmento de entrevista con un participante sano como control (Discursos 2 a 5). Estas transcripciones corresponden a las repuestas obtenidas frente a la pregunta “¿Qué actividades hacen generalmente las personas en el transcurso del día?”, y fueron tomadas de un trabajo anterior (Elvevåg *et al.*, 2007). Para las mismas se contó con puntajes de Coherencia y Tangencialidad de la escala de Andreasen (Andreasen, 1986) asignados por un experto (véase la sección 4.2.1).

El procedimiento utilizado para construir los grafos temáticos fue el siguiente (Cabana, 2009): Primero se delimitan bloques de texto, de no menos de 20 palabras, en las que se desarrolla una unidad temática. A cada uno de estos bloques temáticos se les asignaron entre 3 y 5 palabras a modo de “claves” que hacen referencia al contenido del bloque. Con todas las etiquetas obtenidas para un texto, se construye un grafo: las etiquetas corresponden a los nodos, mientras que las conexiones entre éstos denotan la co-ocurrencia de las etiquetas en un mismo bloque de texto.

Una vez construidos los grafos para cada transcripción se aplicaron las entropías temáticas descritas anteriormente de la siguiente forma: Dado un grafo, se consideró como “tema” a cada uno de sus componentes conexos. De esta manera se obtiene una asignación de un tema a cada bloque de texto, permitiendo esto calcular las entropías correspondientes. En la figura 7.3 se muestran los grafos temáticos elaborados para cada discurso, mientras

²Conocido en español como “Estudio en escarlata”, de Arthur Conan Doyle.

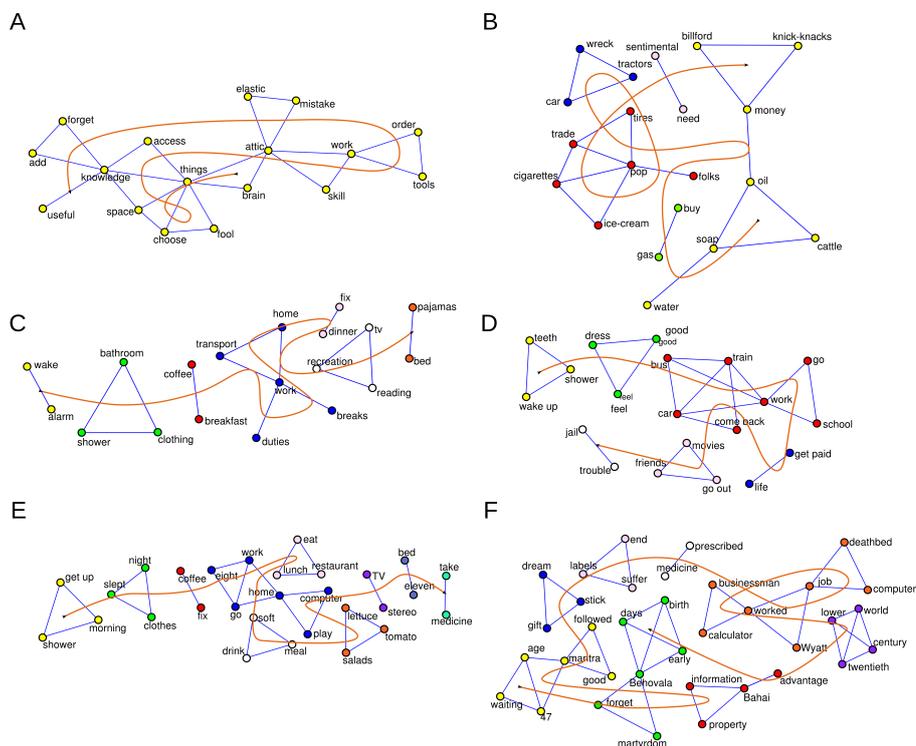


Figura 7.3: Grafos temáticos obtenidos mediante la técnica de etiquetado manual presentada en (Cabana, 2009), aplicada a: A. un fragmento de “A Study in Scarlet”; B. transcripción de un paciente con esquizofrenia tomado de Andreasen (1986); C. transcripción de un sujeto control; y de otros tres pacientes con esquizofrenia (D., E. y F.), tomados de (Elvevåg et al., 2007).

que en la tabla 7.1 pueden verse los valores de entropías calculadas sobre las asignaciones temáticas en base a grafos para las seis muestras. Puede verse que para el fragmento de texto de Sherlock Holmes (Texto), todas las entropías son cero. Esto se debe a que el grafo temático solo cuenta con un único componente conexo, y por tanto comprende a un solo tema. En cambio, el fragmento del paciente con esquizofrenia, (Discurso 1) tiene 4 temas, y las entropías son diferentes a cero.

En cuanto a las respuestas de la pregunta acerca de la rutina diaria de una persona cualquiera (Discursos 3 a 5), las entropías de tema, entropías de tema promedio, así como las entropías de transición promedio son más altas para los pacientes que para el control (Discurso 2). Por otra parte, el discurso calificado con mayor incoherencia (Discurso 5) muestra las mayores entropías de tema y de transición.

	# temas	$S(\mathcal{D})$	$\overline{S(\mathcal{D})}$	$T(\mathcal{D})$	$\overline{T(\mathcal{D})}$	COH	TAN
Texto (Holmes)	1	0	0	0	0	–	–
Disc. 1 (Paciente)	4	1.94	0.38	1.79	0.35	–	–
Disc. 2 (Control)	7	0.3	0.043	0.69	0.099	1	1
Disc. 3 (Paciente)	10	1.08	0.11	1.10	0.11	1	4
Disc. 4 (Paciente)	6	0.67	0.11	0.69	0.12	3.5	4
Disc. 5 (Paciente)	8	2.05	0.26	1.39	0.17	6	6

Tabla 7.1: Medidas de entropías para seis textos con asignación temática manual (Cabana *et al.*, 2011). Las columnas COH y TAN corresponden a los puntajes de Coherencia y Tangencialidad.

7.3 Asignación automática de estructura temática

Para automatizar completamente el proceso del cálculo de las entropías, resta encontrar una manera de identificar la estructura temática (en particular la sucesión de temas) automáticamente. En esta sección se describen algunos intentos en esta dirección, utilizando un espacio semántico obtenido mediante de Análisis Semántico Latente a partir de un corpus de artículos de Wikipedia (Cabana, 2009).

Se exploraron dos estrategias de identificación de tópicos usando un espacio semántico; ambas involucran la proyección de “ventanas” de texto en el espacio. Una “ventana” es un conjunto de palabras contiguas, que son usadas para construir vectores de búsqueda, o “query”, los cuales son proyectados en el espacio semántico (véase el Apéndice C.2 para más detalles) En la figura 7.4 se ilustran los primeros pasos comunes a las estrategias aquí descritas.

7.3.1 Generación de grafos temáticos

La primera estrategia abordada fue la emulación del proceso manual de generación de grafos temáticos, en particular, el proceso de selección de etiquetas temáticas para los segmentos de texto. Dado que este proceso involucra un análisis del contenido temático de un segmento de texto, similar a la identificación de tópico (véase la sección 3.2.3), se decidió emular este paso usando un espacio semántico.

El procedimiento usado fue el siguiente. Para cada ventana de texto, se construyó el vector de búsqueda q_i , y se calcularon las proyecciones p_i sobre el espacio definido por la matriz U (véase el apéndice C.1 para más detalles). Para cada proyección, se calcularon las 3 palabras del espacio semántico cuya proyección (las filas de la matriz U) resulta más cercana a la proyección de la ventana p_i (utilizando como criterio de “distancia” el coseno del ángulo formado por los vectores). Se eligieron estas palabras “cercanas” como etiquetas de la ventana.

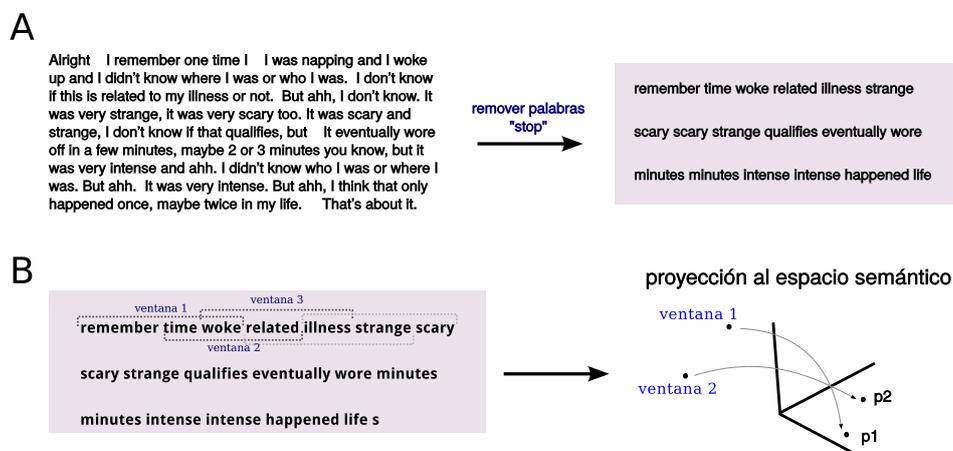


Figura 7.4: Procedimiento general para el procesamiento automático de las transcripciones. A. Primero se remueven las palabras de escaso contenido semántico (palabras de la categoría cerrada, adverbios, preposiciones, etc.), en base a un listado de palabras “stop” usadas comúnmente en la literatura [Salton y Lesk \(1965\)](#) (véase el Apéndice C.1), y las palabras que no están presentes en el espacio semántico a utilizar. B. Se construyen “ventanas” sobre el texto que abarcan en general de 3, 4 o 5 palabras de largo. En base a estas ventanas se construyen “queries” o vectores de búsqueda que se proyectan al espacio semántico (ver apéndice C.2).

Con todas las etiquetas generadas (3 para cada ventana de texto), se construyeron grafos temáticos siguiendo el mismo criterio utilizado en [Cabana \(2009\)](#): cada etiqueta es un nodo, y los nodos se conectan si las etiquetas correspondientes co-ocurren en una ventana de texto (véase la figura 1 de [Cabana et al. \(2011\)](#), adjunto al fin de este capítulo para más detalles).

A partir del grafo temático se calcularon las entropías temáticas, utilizando el mismo criterio que para los grafos construidos a mano: se consideró a cada componente conexo del grafo como un tema diferente. Este procedimiento se aplicó a las 4 transcripciones de [Elvevåg et al. \(2007\)](#) utilizadas en el caso del etiquetado manual, que corresponden a un sujeto control y 3 transcripciones de pacientes con esquizofrenia. Las transcripciones además cuentan con puntajes asignados para las categorías de “coherencia” y “tangencialidad” de la escala de [Andreasen \(1986\)](#).

Como se aprecia en la figura 7.5, ninguna de las entropías calculadas logra reflejar el aumento en los puntajes de coherencia y tangencialidad de las transcripciones de los pacientes. En base a este resultado adverso, se elaboró una estrategia alternativa de cálculo automatizado de entropías, descrito en la siguiente sección.

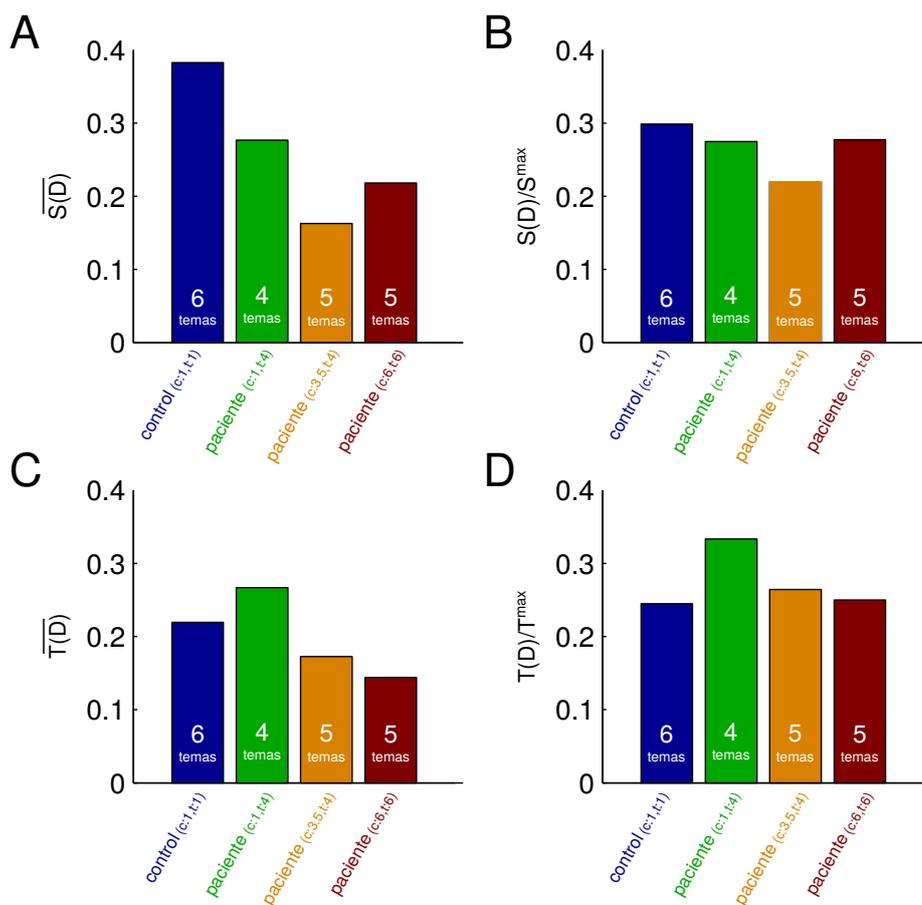


Figura 7.5: Entropías obtenidas mediante la generación automática de grafos temáticos para las 4 transcripciones de Elvevåg *et al.* (2007) utilizadas anteriormente, y descritas en Cabana *et al.* (2011). De izquierda a derecha, los puntajes de coherencia y tangencialidad aumentan, reflejando una mayor desorganización. A. Entropía de tema promedio. B. Entropía de tema normalizada. C. Entropía de transición promedio. D. Entropía de transición normalizada.

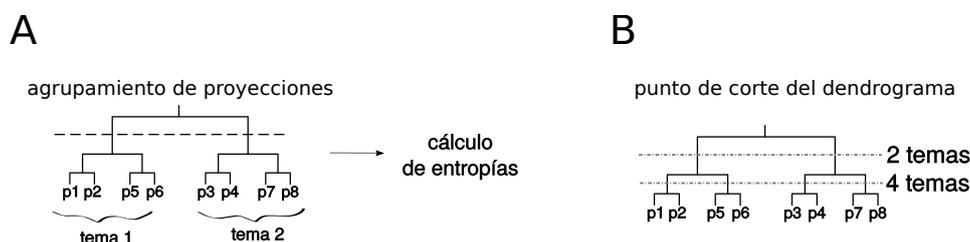


Figura 7.6: Ilustración del proceso asignación temática mediante el agrupamiento jerárquico de las proyecciones. A. A partir del dendrograma obtenido, se elige un punto de corte, y las ventanas cuyas proyecciones quedan en diferentes ramas del árbol son asignadas a diferentes tópicos. B. El número de temas obtenidos depende de la posición del árbol en donde se decidió cortar.

7.3.2 Agrupamiento jerárquico de proyecciones

La estrategia alternativa para calcular las entropías en forma automática consistió en realizar un agrupamiento jerárquico de las proyecciones de las ventanas. La idea detrás de este procedimiento es la siguiente: si una sección del discurso trata de un tema, entonces las proyecciones de las ventanas en el espacio semántico estarán más cercanas entre sí que con las ventanas que corresponden a otros temas. Este procedimiento fue explorado en situaciones artificiales en Cabana (2009) con buenos resultados.

El procedimiento utilizado se ilustra en la figura 7.6. Se realizó un agrupamiento jerárquico de tipo aglomerativo (distancia promedio sin pesos (UPGMA)), la métrica utilizada fue la “distancia” coseno (véase 3.2.3). El resultado del agrupamiento jerárquico es un dendrograma, o árbol, que tiene como hojas a las proyecciones, y cuyos nodos internos son grupos de proyecciones. Cada uno de los nodos internos es un potencial tema, ya que reúne proyecciones que son similares entre sí.

Si se realiza un corte al dendrograma en algún punto, las ramas obtenidas (que corresponden a nodos internos en el dendrograma intacto) pueden identificarse como temas, lo que permitiría el cálculo de las entropías (figura 7.6 A). Por tanto, diferentes puntos de corte en el árbol generarían diferentes asignaciones temáticas, lo que conllevaría a distintos valores de entropías (figura 7.6 B). Una forma de resolver este problema es, en vez de elegir un punto de corte según un criterio dado, evaluar la asignación temática dada por diferentes cortes del dendrograma, obteniendo así diferentes valores de entropía para diferentes números de temas asignados.

Discurso artificial

Para evaluar la eficacia de las proyecciones y el agrupamiento jerárquico para identificar temas, se construyó un discurso artificial usando el mismo espacio semántico utilizado para realizar las proyecciones. Este discurso se

genera partiendo de una palabra inicial, que funciona a modo de “semilla” (en el ejemplo utilizado aquí es “science”), hasta llegar a una palabra objetivo, o final (aquí es “team”). Primero se buscan las n palabras del espacio cuyas proyecciones son más próximas a la semilla. De esta lista, se elige la que está más cerca de la palabra objetivo, y se la utiliza como semilla en una nueva búsqueda, y así sucesivamente hasta llegar a la palabra de destino. Este procedimiento se describe en detalle en el Apéndice C.3.

La sucesión de palabras obtenidas utilizando la palabra “play” como semilla se muestra en la figura 7.7 A. Se realizó una segmentación subjetiva del discurso en 6 temas, que se refleja en el código de color de la figura 7.7. Sobre el discurso artificial, se crearon ventanas de texto de largo 3, y se calcularon las proyecciones sobre el espacio semántico, como se describe en figura 7.4 B.

Las proyecciones obtenidas fueron sometidas a un análisis de agrupamiento jerárquico como se describió anteriormente, cuyo resultado se muestra en la figura 7.7 C. El código de color indica el tema al que corresponden las proyecciones de las ventanas. Puede verse que las proyecciones correspondientes a un mismo tema se agrupan casi todas en una rama, aunque la correspondencia entre ramas del dendrograma y temas no es exacta. En particular, no existe ningún punto de corte que separe a las proyecciones en los temas identificados.

También se realizó un análisis de segmentación basado en la construcción de un mapa autoorganizado de Kohonen (1982), una red neuronal capaz de construir una representación bidimensional de los vectores de entrada. Como se aprecia en la figura 7.7 B, el mapa organiza a las proyecciones de las ventanas de un mismo tema en regiones contiguas.

Ambos análisis ilustran la posibilidad de identificar temas en un discurso con agrupaciones de las proyecciones de las ventanas de texto.

Aplicación a transcripciones de pacientes con esquizofrenia

Del mismo modo que se realizó para el método ilustrado en la sección 7.3.1, se aplicó el método de agrupación jerárquica de proyecciones a las transcripciones provenientes de Elvevåg *et al.* (2007), utilizadas en la ilustración inicial de las entropías presentada en la sección 7.2 y en Cabana *et al.* (2011).

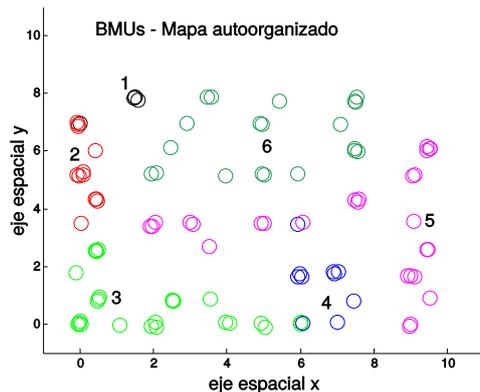
Para cada transcripción se obtuvieron las ventanas y sus respectivas proyecciones, de las cuales se obtuvo un árbol de agrupamiento jerárquico. Para cada árbol, se evaluaron los cortes que generan sucesivamente 3, 4, etc., hasta 22 particiones temáticas. De todas las entropías surgidas de las diferentes particiones temáticas de un mismo discurso, se eligieron las entropías máximas como representativas del grado de desorganización del discurso.

Los resultados del análisis se muestran en la figura 7.8 para la entropía de tema y en la figura 7.9 para la entropía de transición. Puede apreciarse que las máximas entropías promedio de tema $\overline{S(\mathcal{D})}$, así como las entropías

A

science fiction computer doctor fantasy technology novels dr stories books author story book
 published writer written wrote writing edition work songs works working project album released track
 albums listing release tracks bonus live records label record recorded recording recordings artists jazz
 studio sound art sounds blues artist gallery instrument instruments painter painting paintings musical
 played music playing role appeared play character characters actor fictional main anime actress
 manga japanese series japan tokyo television episode episodes channel cable show season network
 aired shows networks league seasons baseball football premier soccer sports club clubs golf
 basketball championship tournament men teams women team

B



C

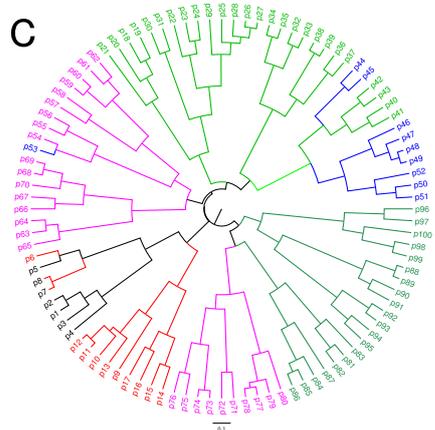


Figura 7.7: A. Discurso artificial obtenido empleando el procedimiento descrito en el Apéndice C.3. Los colores corresponden a los diferentes temas identificados manualmente; B. Posición de las unidades más activas para cada ventana del discurso artificial de un mapa de Kohonen (1982), entrenado con las mismas proyecciones de las ventanas. Puede verse que las ventanas que corresponden a diferentes temas activan unidades neurales presentes en diferentes secciones del mapa. C. Agrupamiento jerárquico de las proyecciones de las ventanas del discurso artificial. Las ventanas pertenecientes a un mismo tema se agrupan casi totalmente en ramas disjuntas, si bien no existe ningún punto de corte del dendrograma que resulte en ramas que reflejen exactamente la partición temática manual.

de tema normalizadas $S(\mathcal{D})/S^{max}(\mathcal{D})$ son mayores para los discursos de los pacientes con mayor puntaje de coherencia y tangencialidad (figura 7.8 A y C). El perfil de entropías para las particiones temáticas de entre 3 y 22 temas (figura 7.8 B y D) no permite distinguir claramente entre los discursos.

Para las entropías de transición la situación es similar, con la salvedad que sólo para la entropía de transición normalizada $T(\mathcal{D})/T^{max}(\mathcal{D})$ los valores aumentan con las puntuaciones de coherencia y tangencialidad (figuras 7.9 A y C).

Es de notar que para ambos tipos de entropía, los máximos de los distintos discursos se obtienen para cortes del dendrograma que originan diferente número de temas.

Correlación con puntuaciones de “Pobreza de contenido”

Los resultados auspiciosos de la sección precedente motivaron un análisis más riguroso de la efectividad del método de agrupación jerárquica de proyecciones. Para eso contamos con 130 transcripciones de respuestas realizadas por pacientes y controles ante tres preguntas que integran el denominado “Cuestionario de experiencias extrañas” (utilizado en [Elvevåg et al. \(2007\)](#)), cedidos gentilmente por la Dra. Elvevåg.

El cuestionario cuenta de tres preguntas:

1. “¿Cómo describirías el color a una persona ciega?” (COLOR)
2. “¿Cómo describirías el sonido a una persona sorda?” (SOUND)
3. “Cuenta un suceso extraño que te haya sucedido ” (STRANGE)

Las respuestas fueron evaluadas en el ítem “Pobreza de contenido del habla” de la escala de [Andreasen \(1986\)](#).

A las 130 respuestas se les aplicó el análisis por agrupación jerárquica, obteniéndose las respectivas entropías temáticas. Se calcularon los valores de las correlaciones entre las diferentes entropías obtenidas, y los puntajes de contenido. Las únicas correlaciones significativas fueron para las puntuaciones del ítem STRANGE y las entropías promedio $\overline{S(\mathcal{D})}$, como se muestra en la figura 7.10 A y B. Sin embargo, para este ítem se observa una correlación significativa entre la puntuación de contenido y la longitud del discurso (figura 7.10 C). Dado que además se observa una muy alta correlación entre el largo del discurso y la entropía media obtenida automáticamente (figura 7.10 D), es muy probable que la correlación entre la entropía media y las puntuaciones sea un reflejo de la correlación entre las puntuaciones y el largo del discurso.

En efecto, cuando se utilizan las entropías de tema normalizadas $S(\mathcal{D})/S^{max}(\mathcal{D})$ (la normalización implica ajustar por un factor proporcional al logaritmo del largo del discurso, véase el Apéndice B.1), la correlación con el largo del

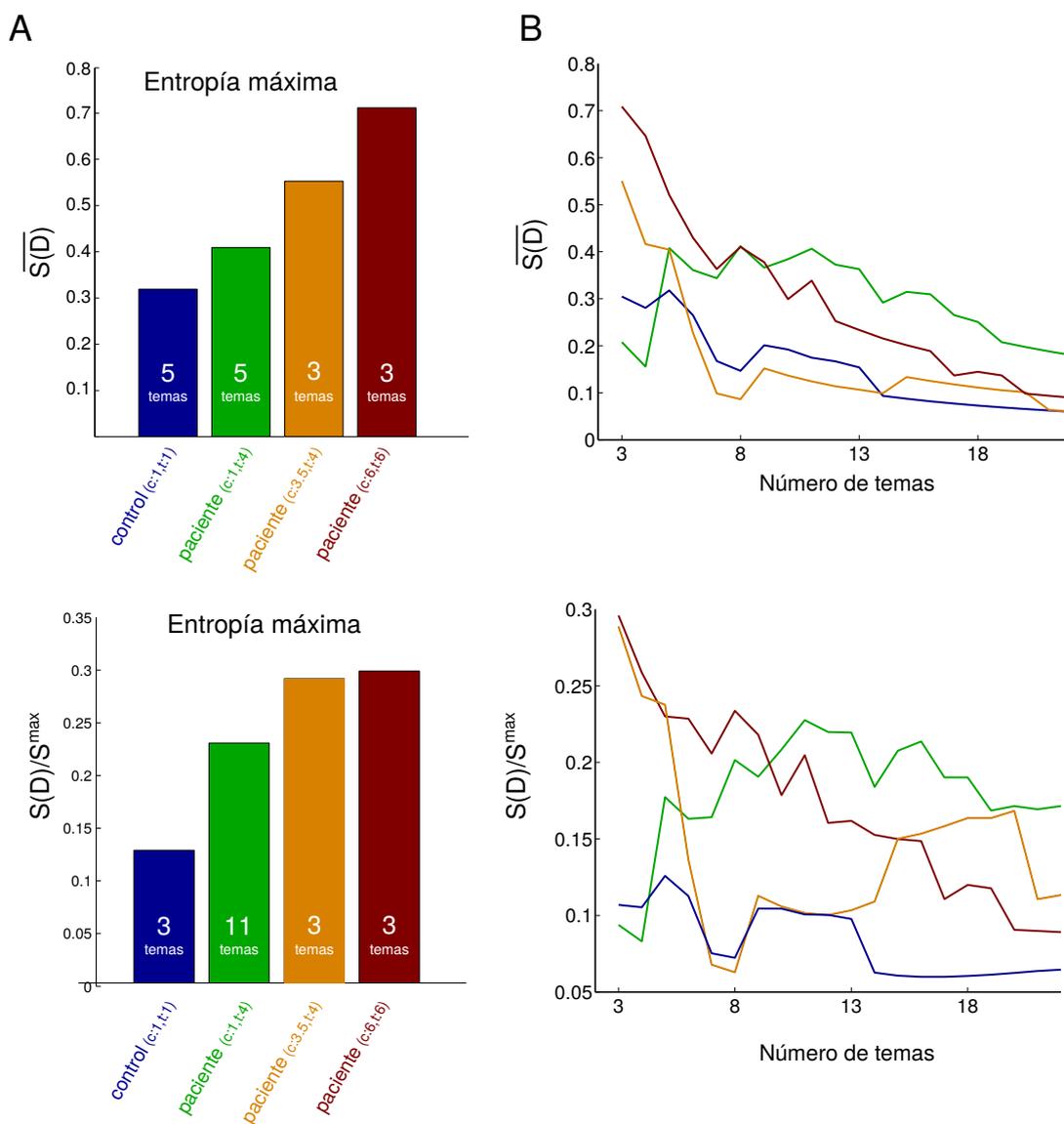


Figura 7.8: Entropías de tema obtenidas mediante el procedimiento de agrupamiento jerárquico de proyecciones para las cuatro transcripciones de [Elvevåg et al. \(2007\)](#) descritas anteriormente. A. Máximos valores de entropía de tema promedio $\overline{S(D)}$. B. Perfil de entropía promedio para particiones entre 3 y 22 temas. C. Máximos valores de entropía de tema normalizada $S(D)/S^{max}(D)$. D. Perfil de entropía normalizada para particiones entre 3 y 22 temas.

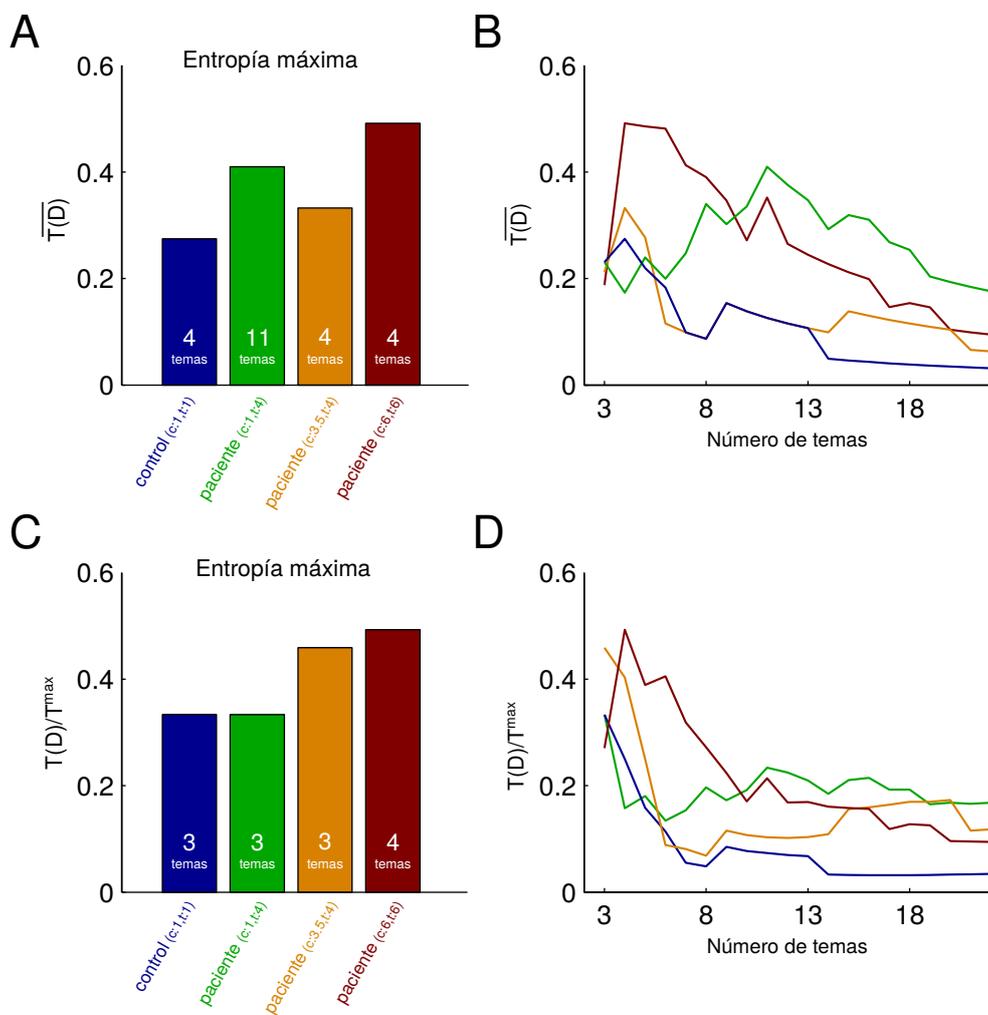


Figura 7.9: Entropías de transición obtenidas mediante el procedimiento de agrupamiento jerárquico de proyecciones para las cuatro transcripciones de [Elvevåg et al. \(2007\)](#) descritas anteriormente. A. Máximos valores de entropía de transición promedio $\overline{T(D)}$. B. Perfil de entropía promedio para particiones entre 3 y 22 temas. C. Máximos valores de entropía de transición normalizada $T(D)/T^{max}(D)$. D. Perfil de entropía normalizada para particiones entre 3 y 22 temas.

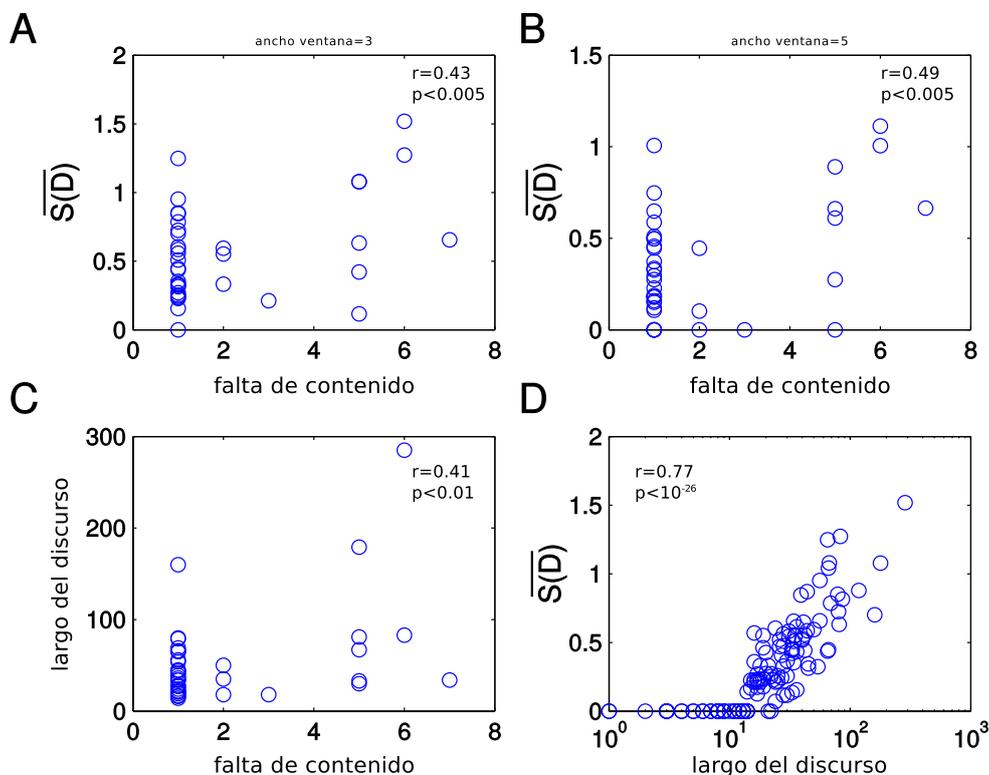


Figura 7.10: Correlaciones entre los valores de entropía de tema promedio obtenidos automáticamente mediante el método de agrupación jerárquica y las puntuaciones de “pobreza de contenido” para el ítem *STRANGE* del cuestionario de “Experiencias extrañas”, para una ventana de 3 palabras (A), y para una ventana de 5 palabras (B). C. Correlación entre las puntuaciones de “pobreza de contenido” y el largo del discurso. D. Correlación entre el largo del discurso y los valores de entropía obtenidos automáticamente. Nótese la escala logarítmica en el eje de las x .

discurso desaparece (figura 7.11 D), así como la correlación con los puntajes de contenido para todos los ítems (figura 7.11 A,B y C).

7.4 Discusión

Las entropías aquí presentadas son sensibles a la desorganización temática del discurso, y resultan herramientas potencialmente valiosas para utilizar tanto en el contexto clínico como experimental. Las evaluaciones realizadas en esta tesis permiten ilustrar su funcionamiento, y elaborar una suerte de “prueba de concepto”, más que representar una evaluación exhaustiva de su utilidad en un contexto dado. Es necesario notar que las entropías temáticas ofrecen posibilidades de variación; pueden usarse las entropías por tema, las entropías promedio, normalizadas, etc. De hecho, la normalización aquí utilizada no es la única posible.

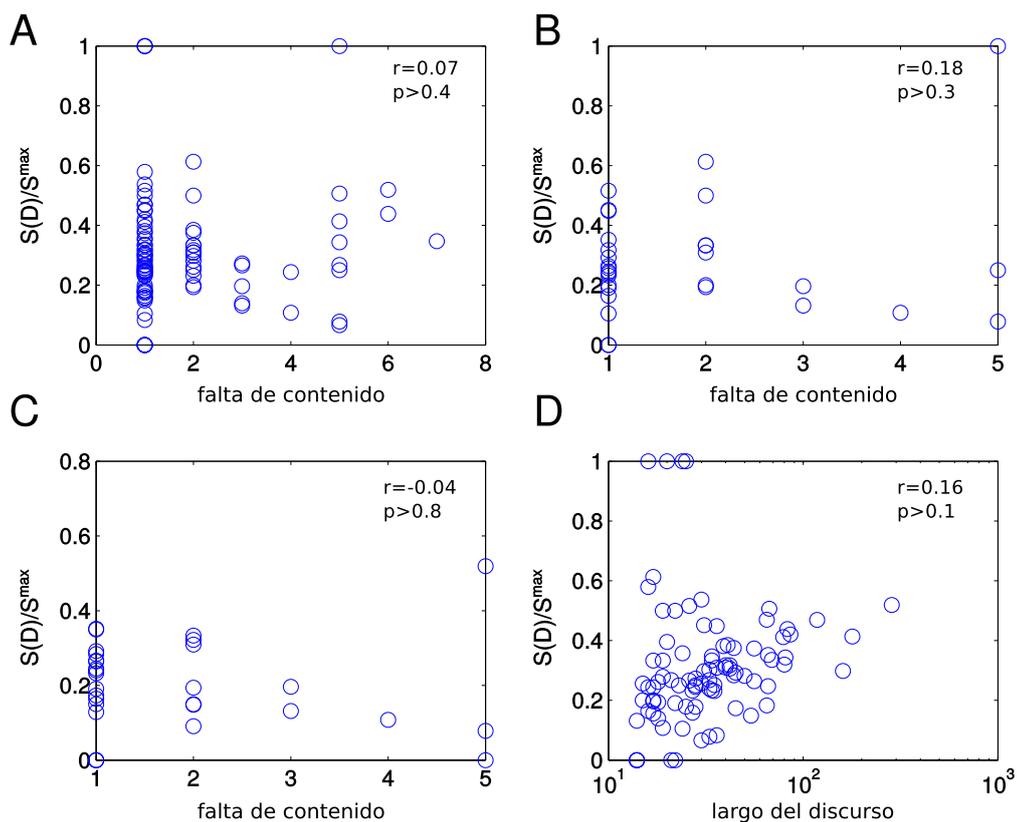


Figura 7.11: Correlaciones entre los valores de entropía de tema normalizada obtenidos automáticamente mediante el método de agrupación jerárquica y las puntuaciones de “pobreza de contenido” para el cuestionario de “Experiencias extrañas”. A. Todas las respuestas. B. Las respuestas al ítem STRANGE. C. Las respuestas al ítem STRANGE, restringiendo el análisis a las particiones entre 3 y 5 temas. D. Correlación entre el largo del discurso y la entropía normalizada. Nótese la escala logarítmica en el eje de las x .

En cuanto a los métodos automáticos, el método de construcción de grafos temáticos a partir de un espacio semántico no arrojó resultados auspiciosos. Si bien la evaluación propuesta no permite obtener una respuesta definitiva acerca de la utilidad de este procedimiento, se decidió no seguir explorando su potencial.

Por el contrario, los resultados obtenidos mediante el método de agrupamiento jerárquico de proyecciones resultan promisorios, ya que se obtuvieron entropías progresivamente mayores para los discursos con mayores puntajes de coherencia y tangencialidad, mostrando sintonía con los juicios subjetivos de los expertos. En base a estos resultados se decidió someter al método a una prueba estadísticamente más rigurosa, y evaluar la correlación entre los valores de entropía obtenidos automáticamente y puntajes de coherencia realizados por expertos. Sin embargo, el conjunto de datos con que se tuvo acceso (Cuestionario de experiencias extrañas) no contaba con puntuaciones de coherencia (o tangencialidad), sino de falta de contenido, lo que no representa una situación de evaluación ideal.

Si bien se obtuvieron correlaciones positivas significativas entre las puntuaciones de contenido y las entropías de tema promedio en uno de los 3 ítems del cuestionario (STRANGE), seguramente se trate del reflejo de la correlación entre el largo del discurso y la puntuación de falta de contenido, mediada por una correlación entre el largo del discurso y la entropía de tema promedio obtenida automáticamente. De hecho, la entropía de tema normalizada no presenta ninguna de las dos correlaciones, posiblemente debido a que el término de normalización incorpora el largo del discurso (véase el apéndice B.1).

Para poder validar las herramientas aquí desarrolladas como métricas de la pérdida de coherencia o desorganización temática es necesario realizar estudios normativos o de correlación con medidas preexistentes en contextos clínicos o experimentales. Es posible que este tipo de estudios arroje luz acerca de cuáles variantes resultan las que mejor se adecúan cada situación particular.

Artículo publicado en
Schizophrenia research



Detecting order–disorder transitions in discourse: Implications for schizophrenia

Álvaro Cabana ^a, Juan C. Valle-Lisboa ^a, Brita Elvevåg ^{b,*}, Eduardo Mizraji ^a

^a Group of Cognitive Systems Modeling, Biophysical Section, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay

^b Psychiatry Research Group, Department of Clinical Medicine, University of Tromsø, & Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway

ARTICLE INFO

Article history:

Received 3 October 2010

Received in revised form 15 April 2011

Accepted 18 April 2011

Available online 2 June 2011

Keywords:

Discourse trajectory

Incoherent speech

Complex networks

Topic entropy

Schizophrenia

ABSTRACT

Several psychiatric and neurological conditions affect the semantic organization and content of a patient's speech. Specifically, the discourse of patients with schizophrenia is frequently characterized as lacking coherence. The evaluation of disturbances in discourse is often used in diagnosis and in assessing treatment efficacy, and is an important factor in prognosis. Measuring these deviations, such as “loss of meaning” and incoherence, is difficult and requires substantial human effort. Computational procedures can be employed to characterize the nature of the anomalies in discourse. We present a set of new tools derived from network theory and information science that may assist in empirical and clinical studies of communication patterns in patients, and provide the foundation for future automatic procedures. First we review information science and complex network approaches to measuring semantic coherence, and then we introduce a representation of discourse that allows for the computation of measures of disorganization. Finally we apply these tools to speech transcriptions from patients and a healthy participant, illustrating the implications and potential of this novel framework.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Language production and comprehension provide a window into the cognitive and neural architecture underlying complex information processing in the brain (Pinker, 2000). They are high-level cognitive functions that reflect the state of numerous cognitive processes. The pattern and content of the communication can be traced back to individuals' cognitive abilities, knowledge, affective state and consequently their overall mental state. Disturbances in the domain of language, especially in speech, occur in a variety of psychiatric and neurological conditions, and their neural substrates are likely to be related to the pathophysiology of the disorder (DeLisi, 2001), and hence are a fundamental aspect in diagnosis and assessing treatment responsiveness and prognosis (Andreasen and Grove, 1986; Andreasen and Black, 2005; McKenna and Oh, 2005).

Indexing language comprehension and production disturbances has been conducted using a variety of neuropsychological measures and tests (Hodges et al., 1992; Tamlyn et al., 1992; McKenna et al., 1994). We focus on speech, which traditionally has been quantified for predictability and variability using a variety of manual (and labor-intensive) techniques, such as cloze analysis, type:token ratios, analysis of lexical and syntactic structure, and also discourse structure using cohesion analysis (for a review, Kuperberg, 2010). There are a variety of fine-grained rating scales of the coherence of speech and

communication, such as the Scale for the Assessment of Thought, Language and Communication (TLC; Andreasen, 1986), the Communication Disturbances Index (Docherty, 2005), and the Thought Disorder Index (TDI) (Solovay et al., 1987; Niznikiewicz et al., 2002), use of which requires extensive training but nonetheless remains open to variance across raters. In some sense these are probing “communication efficiency”, which can be assessed by a range of computational linguistic techniques (Jurafsky and Martin, 2000). Indeed, such studies – using Latent Semantic Analysis (LSA) which models and matches discourse content (Landauer and Dumais, 1997; Landauer et al., 2007) – have demonstrated that it is possible to evaluate patients with schizophrenia based on open-ended verbalizations. These automatically derived language scores have distinguished patients from controls accurately (and patients from other patients, and also from their family members), using both large discourse samples as well as responses consisting of only a few words (Elvevåg et al., 2007, 2010).

Our goal here is to present some tools derived from recent developments in network theory and information science that enable one to capture and index “meaning” in a quantifiable and biologically relevant manner. This is because there are statistical properties in expressed language that provide a rich source of information regarding “meaningful communication”. Specifically, we present measurements of disorganization of discourse based on topic randomness and semantic graph measures. Thus, in the next section we describe methods based on information science and complex network approaches to language, and we introduce a particular representation of discourse, and present ways to measure its

* Corresponding author at: UNN-Åsgård, 9291 Tromsø, Norway.
E-mail address: brita@elvevaag.net (B. Elvevåg).

disorganization. Then we apply our framework to speech samples from patients with schizophrenia and a healthy participant to illustrate the potential of the method.

2. Methods

2.1. Semantic graphs and complex networks

Our method requires a simple but rich representation of meaning. One approach to achieve this is graph representation, with roots in semantic network theory (Collins and Quillian, 1969; Collins and Loftus, 1975; Steyvers and Tenenbaum, 2005). Graphs are mathematical objects consisting of sets of nodes and sets of edges connecting the nodes. Traditional semantic networks are “graphs” with labeled connections that instantiate different relationships between entities (e.g., “a robin is a bird” is represented by a particular type of link (the IS-A link) between the “robin” node and the “bird” node, or the HAS linking “a bird has feathers”, that together support the inference that “a robin has feathers” (Quillian, 1968)). Semantic graphs (i.e., “stripped down” versions of semantic networks) can be used to capture associative and conceptual relationships by automatically analyzing large portions of text, usually linking together nodes that represent words that co-occur within a small range in a large corpus (graphs built this way are referred to here as “lexical graphs”; Ferrer i Cancho and Solé, 2001; Dorogovtsev and Mendes, 2001; Steyvers and Tenenbaum, 2005).

Recent developments in graph theory applied to the study of complex systems have shown that many natural and artificial complex networks show the small world and scale free properties (Albert and Barabási, 2002). The former means that networks tend to have high clustering coefficients¹ while keeping low path lengths (Watts and Strogatz, 1998), and the latter implies that link distribution is frequently characterized by a power law, enabling highly connected nodes to appear relatively often. These two characteristics confer interesting properties to networks, like fast transmission and failure tolerance (Motter et al., 2002; Ferrer i Cancho et al., 2005; Steyvers and Tenenbaum, 2005). In order to represent meaning, we use graphs that turn out to satisfy some of these properties.

As our goal here is to capture the thematic structure of a single instance of a linguistic expression (a relatively small sample of text, discourse or dialog) we will represent words as nodes in a semantic graph, and consider discourse as a trajectory in such a graph. “Goal directed” discourse would show an ordered and organized trajectory, whereas thought disordered discourse would appear as a disordered trajectory due to the disorganization of the semantic structure, or of the mechanism that searches through it. We propose that to measure this disorganization, semantic structures be represented using networks and characterized using measures inspired by information theory. However, in order to derive useful tools two methodological challenges have to be addressed: First, find a suitable representation of discourse, including a topic graph and a trajectory. Second, devise measures of disorganization, sensitive enough to detect subtle deviations.

2.2. The representation of discourse

To represent discourse we analyzed textual transcriptions of speech samples. The texts were subjectively evaluated by delimiting

¹ The clustering coefficient measures the degree of “socialization” of a node: it is the fraction of a node’s neighbors that are themselves neighbors. A mean clustering coefficient of 0 means that nodes are statistically solitary (in social terms, your friends are not friends of each other), and a coefficient of 1 implies the highest degree of socialization (your friends are all friends of each other). The characteristic path length of a graph is the mean of the lengths of the shortest path between each pair of nodes.

small blocks of text of just one theme or idea, and labeling each block in the text with a set of words representing that theme (Cabana, 2009).

As a calibration procedure we analyzed the first two chapters of “A study in scarlet” by Arthur Conan Doyle, as the descriptive nature of the text is devoid of complex metaphors or other literary devices that could complicate analysis. For this large text sample, we added the additional criteria that block size should be between two sentences and four paragraphs and that the selected theme be distinguishable from the previous and next blocks. We show this calibration example in Fig. 1. As the discourse advances, consecutive paragraphs share some labels, allowing the construction of a graph, whereby each label is a node and each pair of labels that co-occur in one block of text is linked by an edge (Fig. 1B). The resulting graph is the topic graph (Fig. 1C).

When the topic graph is displayed bi-dimensionally, the discourse trajectory can be represented as a line drawn over each block in the order of their appearance in the discourse (Fig. 1B, D). If this trajectory is drawn over the whole topic graph, the line appears convoluted and folded, as a result of the text “re-visiting” central topics of the story (Fig. 1D). These “long range interactions” are what we seek to capture by measuring the entropies (see Section 2.3). The rationale for this *a priori* expectation is that sentence sequence in discourse is not random when language is organized (see Kintsch, 1988; Foltz, 2007), and loss of this higher-level order would result in disorganization. From visual inspection of the topic graph and the trajectory line, at least five major sets of labels can be identified that delimit five major topics in the graph (Table 1). In a larger graph, this delimitation could be automatically performed by identifying connected components (sets of connected nodes disconnected from others) or communities (sets of nodes statistically more connected to each other) as representing major topics (Palla et al., 2005; see Section 4).

2.3. Measures of disorganization

The central hypothesis of this work is that loss of goal, tangentiality and incoherence frequently observed in schizophrenia are based in part on problems “following” an ordered trajectory among different topics. In coherent discourse adjacent words refer to connected topics. In contrast, in incoherent discourse a certain degree of “shuffling” of the topics occurs such that adjacent words may belong to different topics. This does not imply a “word salad” as the discourse may respect syntax, word order and even word similarity, but nevertheless reveal a high degree of disorder (i.e., semantic shuffling) in terms of meaning. To detect and measure this disorganization, we developed topic and transition entropy measures (closely related to that used in statistical mechanics).

2.3.1. Topic entropy

Topic entropy $S(\alpha)$, is defined as

$$S(\alpha) = - \sum_{i=1}^{n(\alpha)} p(\alpha_i) \log p(\alpha_i) \quad (1)$$

α being a particular topic in the text, $n(\alpha)$ the number of continuous stretches of text attributable to this topic and $p(\alpha_i)$ the ratio of length of the stretch i to total number of words attributable to this topic. This Eq. (1) measures the level of discontinuity of words belonging to the same topic. When discourse is organized in a perfect sequence of topics each consisting of an uninterrupted stretch of text, all topic entropies will be zero, and will grow in those cases where topics are more interspersed. To illustrate, we shuffled the text of our example, “A study in scarlet”, and present a visual representation of this shuffling (Fig. 2).

Fig. 3A shows how the entropy of each of the topics increases as a result of the shuffling. The first instances of shuffling disrupt the

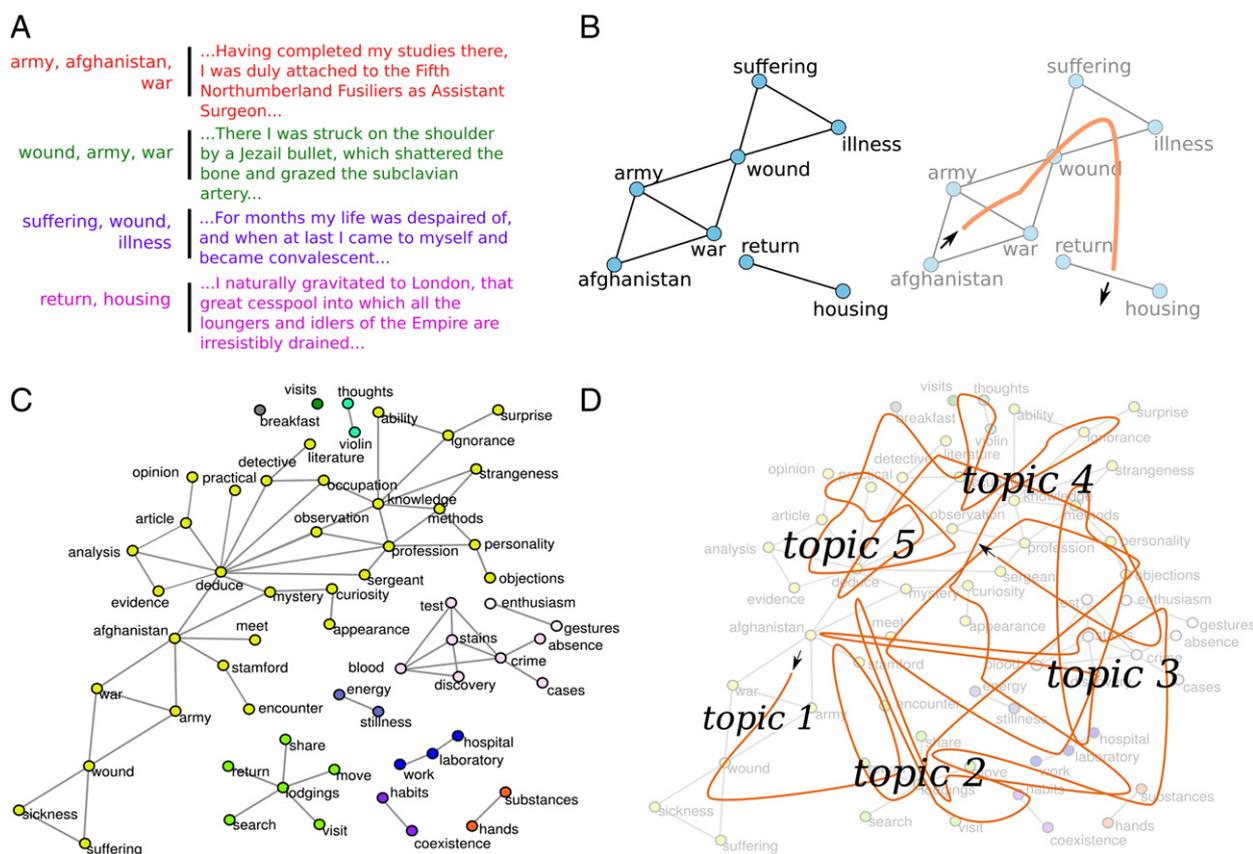


Fig. 1. To illustrate our approach, we analyzed the first two chapters of “A study in scarlet” by Arthur Conan Doyle, which features the first appearance of detective Sherlock Holmes (Conan Doyle, 2005). The first paragraphs and an illustration of the labeling process are shown. A) Four thematic blocks with labels are delimited. B) The resulting graph after assigning a node to each label, and linking labels that occur together in a block assignment, called topic graph. C) The resulting topic graph of the two chapters. D) The discourse trajectory is drawn over the topic graph.

original order only moderately, and a great deal of shuffling has to be imposed to result in some “randomness”, but the greatest increases in entropy occur in the first shuffling instances. Given that every topic’s entropy increases, but to different degrees, a good measure of disorganization is mean entropy across topics, estimated as

$$\overline{S(D)} = \frac{1}{N_T} \sum_{\kappa} S(\kappa) \tag{2}$$

where $\overline{S(D)}$ refers to the entropy of the discourse and N_T is the number of topics that are expressed.

2.3.2. Transition entropy

Transition entropy, is defined as

$$T(\alpha) = - \sum_{\tau \neq \alpha} p_{\alpha}(\tau) \log p_{\alpha}(\tau) \tag{3}$$

Table 1
Representative topic labels used to categorize each sentence of the text.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Afghanistan	return	crime	personality	mystery
war	search	cases	strangeness	deduce
wound	lodgings	blood	methods	evidence
sickness	move	laboratory	occupation	analysis
suffering	coexistence	substances	knowledge	detective

$p_{\alpha}(\tau)$ is the fraction of transitions from topic α to topic τ . Discourse can have a large topic entropy (calculated with Eq. (1)) but zero transition entropy (calculated with Eq. (3)). If the discourse were perfectly periodic (e.g., a repetition of sequence $\alpha\beta\gamma\alpha\beta\gamma\dots$), then transition frequencies would be $p_{\alpha}(\beta) = 1$ and $p_{\alpha}(\gamma) = 0$, and the entropy defined by Eq. (3) would be equal to 0. With reference to Fig. 3B illustrating transition entropy for the original and shuffled versions of “A study in scarlet”, the increase in transition entropy is apparent, and it is more subtle than with topic entropy. This is probably because in the original text the entropy is already high since topics are relatively independent, or because the small sample of transitions cannot be used to detect inter-topic structure. As with topic entropy, mean transition entropy can be defined as:

$$\overline{T(D)} = \frac{1}{N_T} \sum_{\kappa} T(\kappa) \tag{4}$$

where the sum is performed over all topics.

It remains to be established whether subtle disorganization in semantic structure of discourse can be detected reliably and reproducibly using this approach. We present below examples of its potential usefulness.

3. The topology of speech in schizophrenia

Clearly our method would benefit from further refinements, but we nonetheless illustrate the potential usefulness of the whole approach and demonstrate its “proof of concept”.

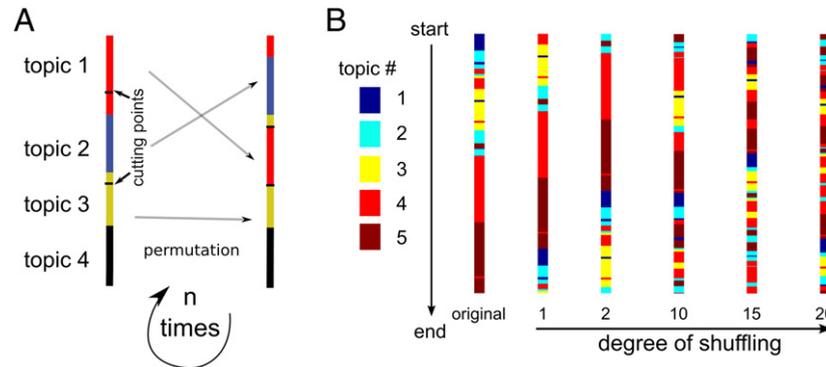


Fig. 2. A) Schematic diagram of the shuffling procedure. Two “cutting points” are randomly assigned in the text, and then the remaining three portions of text are randomly permuted. B) Visualization of the effectiveness of the shuffling procedure on the topic assignment of the first two chapters of “A study in scarlet”, based on the 5 topics identified in the thematic graph (Fig. 1D). Note how the mixture develops as the shuffling is iterated.

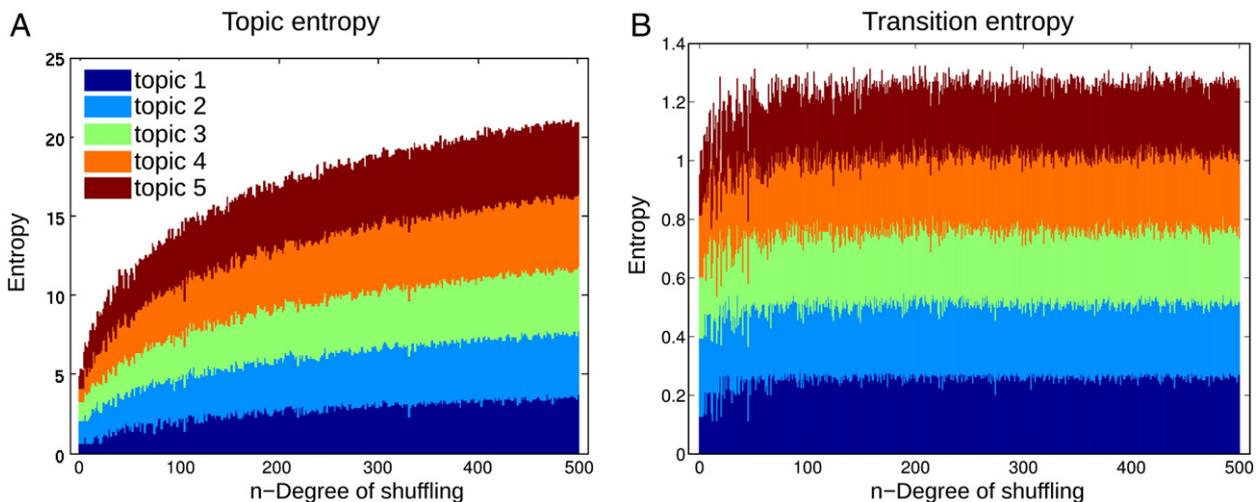


Fig. 3. Topic and transition entropy increase when the text of the first two paragraphs of “A study in scarlet” is shuffled. A) The topic entropies, calculated using Eq. (1). B) Transition entropy of each topic, according to Eq. (3).

One important difference between short speech transcriptions and the example we used to calibrate the procedure concerns size; to evaluate the effect of size we analyzed a paragraph of “A study in scarlet” (Text example), and compared it with a somewhat incoherent speech sample from a patient with schizophrenia (Sample 1). We selected the text example which has evident metaphorical character in stark contrast to Sample 1. In order to establish an even better comparison and analyze further, we examined speech samples generated in response to the question “What activities do people generally do during the course of the day?”, from a healthy participant (Sample 2), and three patients with schizophrenia (Samples 3 to 5) (from Elvevåg et al., 2007). The responses were rated by two human raters for coherence (a score of 1 = very coherent versus 7 = very incoherent) and tangentiality (a score of 1 = very incisively related to question versus 7 = completely unrelated to question).

For all samples we built lexical graphs (see below) and calculated the topological graph parameters (see Section 2.1). We also built topic graphs and calculated the topic and transition entropies (topological graph parameters were not estimated for these graphs because of their small size). Stuttering and repetitions were omitted from speech transcriptions.

We constructed lexical graphs by linking together words co-occurring in a text at a distance of three words (Ferrer i Cancho, 2005), but removing all function words (e.g., articles, prepositions). We

calculated the main graph parameters (clustering coefficient, characteristic path length) of the lexical graphs (see Table 2). Notice that all samples show similar measures, but the smallness of the graphs precludes us from concluding anything further. As discussed below, future studies using this approach should employ bigger speech samples. Next we focus on the topic graphs.

To obtain the topic graph, we performed a manual labeling procedure (as in the topic graph of “A study in scarlet”), selecting blocks of about one sentence in length. Once the topic graph was built (Fig. 4), each connected component was assigned a different topic, enabling the calculation of the topic and transition entropies.

After calculating the entropies, we detected important differences between the patient samples and the controls of comparable length. With reference to Table 3, the patient’s discourse (Sample 1) results in higher topic and transition entropies than the text example (Holmes). Regarding the responses to the question “What activities do people generally do during the course of the day?” (Samples 2 to 5), it can be seen that the healthy participant’s response (Sample 2) results in lower topic entropy than the patients’ responses. Within the patients’ responses, the one with the lowest coherence (Sample 5) has much higher topic entropy than the others. However, transition entropy was lowest in the healthy participant (Sample 2) and in one of the responses from a patient (Sample 4).

BOX 1

Text example

"I consider that a man's brain originally is like a little empty attic, and you have to stock it with such furniture as you choose. A fool takes in all the lumber of every sort that he comes across, so that the knowledge which might be useful to him gets crowded out, or at best is jumbled up with a lot of other things so that he has a difficulty in laying his hands upon it. Now the skilful workman is very careful indeed as to what he takes into his brain-attic. He will have nothing but the tools which may help him in doing his work, but of these he has a large assortment, and all in the most perfect order. It is a mistake to think that that little room has elastic walls and can distend to any extent. Depend upon it there comes a time when for every addition of knowledge you forget something that you knew before. It is of the highest importance, therefore, not to have useless facts elbowing out the useful ones." (p.16; "A study in scarlet" by Arthur Conan Doyle).

Sample 1

"They're destroying too many cattle and oil just to make soap. If we need soap when you can jump into a pool of water, and then when you go to buy your gasoline, my folks always thought they should, get pop but the best thing to get, is motor oil, and, money. May as well go there and, trade in some, pop caps and, tires, and tractors to grup, car garages, so they can pull cars away from wrecks, is what I believed in. So I didn't go there to get no more pop when my folks said it. I just went there to get a ice-cream cone, and some pop, in cans, or we can go over there to get a cigarette. And it was the largest thing you do to get cigarettes 'cause then you could trade off, what you owned, and go for something new, it was sentimental, and that's the only thing I needed was something sentimental, and there wasn't anything else more sentimental than that, except for knick-knacks and most knick-knacks, these cost 30 or 40 dollars to get, a good billfold, or a little stand to put on your desk." (p. 477; Andreasen, 1986).

Sample 2

"Get up, maybe the alarm clock would wake you up, turn off the alarm clock, or press the snooze bar or something like that, then use the bathroom, brush your teeth, take a shower, maybe shave if you're a man, then you do your hair, put on clothing, get some breakfast, some people just have coffee or something like that, then go wherever it is you go, school or work, so you might drive yourself, or take the bus or train whatever, to get to where you're supposed to be for the day, and do whatever it is you're responsible for doing, working or taking classes, or taking care of your children, or whatever you do during the day, taking breaks during the day for lunch and maybe coffee breaks, or bathroom breaks and at the end of the regular weekday, you go on home by whatever method you came, fixing dinner, or buying something for dinner and eating it, maybe doing some housework or running errands, maybe watching TV or doing something else for recreation, like reading, like a book or a magazine, then get ready for bed, brushing your teeth, putting on pajamas, and getting in the bed." (Coherence score: 1, Tangentiality score: 1).

Sample 3

"I'd get up. Usually I take a shower in the morning. Put on clean clothes 'cause I usually slept in the clothes I had on the night before. Eat some kind of breakfast like toast or something. Fix coffee. When I was working, I'd then go to work. Try to get

to work by eight. Or go to lunch and eat lunch usually at a restaurant. Then go back to work and work 'til five. Then go home. Then I'd a lot of times go out and have an O'Dooles or a some kind of soft drink, usually a soft drink like Diet Pepsi or eat a meal like a lot of times I just had salads, but I because I had a hard time with cholesterol so I'd just ate a salad, like lettuce, you know a side-salad, like lettuce, tomato, onion. Then I'd go home and play on my computer until it was time to go to sleep. I'd usually have the TV on and play the stereo with the usually have the sound up and the TV turned down and on ESPN or whatever sports that was best. And, then I would usually go to bed about eleven. I'd always take my medicine when I was supposed to would usually take my medicine when I got up in the morning and a lot of times if I had medicine I was supposed to take, I'd take it about lunch time but usually I didn't lately have that then one when I went to bed." (Coherence score: 1, Tangentiality score: 4).

Sample 4

"Ok, a person usually wakes up, at night time you brush your teeth, in the morning you take a shower or bath and you get dressed you feel good, you take a car a cab or a bus or a train to work, and you either go to school or work, and you get something out of it, you get paid, have a good life, and do what suits you. You come back on the bus or train or the cab, or your car an you go to have a good time you go home, go out to eat, go out with friends, and watch movies, go out to movies and stuff, you stay out of trouble, if you don't stay out of trouble, you go to jail, the worst place to be." (Coherence score: 3.5, Tangentiality score: 4).

Sample 5

"Well, at age forty-seven, I'm waiting around for age forty-eight, to be quite honest. I'm trying to bear martyrdom of the supreme families, Behovala. Why did you make me forget all those things? I didn't like that. I withhold information. I believe in private property. That's the Bahai faith. At age forty-seven, I'm following Dennis the Menace's mantra. He followed mantra I want you to be just as good as I am. Stick to your dream and now the LSA can make a responsible decision as to whether you would like another very expensive gift because all labels are one and we have to face that... Is that true or false? Do you want him to suffer again? The end has reached this. That was the medicine I prescribed. Yes, I have many jobs. I've worked with the Wyatt company, I've worked... It's a computer deathbed. It was a good job I worked at a calculator as a businessman. I've worked at nine to five. In the lower world... Twentieth century AD where I was raised. I was born in eleven-seventy BC. That's an unfair advantage knowing that the Bahai faith comes next. In the early days, they didn't know that. They had to decide if Imagine it is the birth of Buhevaloh, not the birth of Mary and God." (Coherence score: 6, Tangentiality score: 6).

Table 2
Characteristics of lexical graphs obtained from the samples.

Network	Clustering coefficient	Path length	No. of nodes	Links
Text example (Holmes)	0.55	3.63	49	109
Sample 1 (Patient)	0.51	4.04	45	95
Sample 2 (Control)	0.54	4.39	56	121
Sample 3 (Patient)	0.50	3.02	49	124
Sample 4 (Patient)	0.52	3.34	34	72
Sample 5 (Patient)	0.51	4.41	61	128

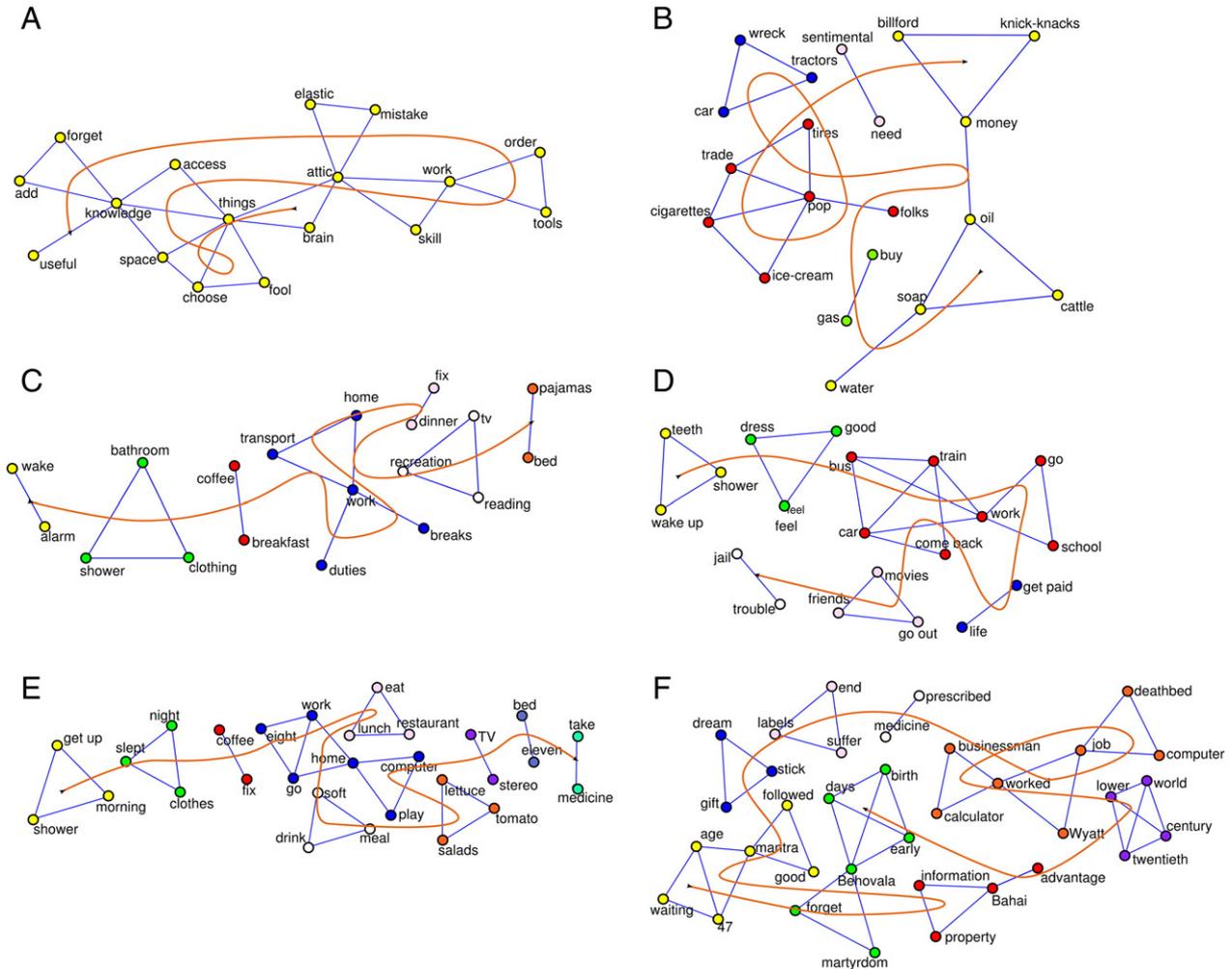


Fig. 4. Topic graphs obtained from manual label assignment for A) Sherlock Holmes' "speech", B) a patient with schizophrenia (Sample 1), C) a healthy participant (Sample 2), D, E and F) patients with schizophrenia (Samples 3 to 5). Discourse trajectories are shown as lines over the graphs.

Although preliminary, these results clearly demonstrate the possibility of applying this novel methodological framework to assay the nature of the disorder that is readily apparent in this discourse.

4. Prospects for an automated topic graph construction

One promising direction to automatically obtain the topic graphs is to employ multidimensional semantic spaces. In these, each concept is associated with a vector, a set of concepts is represented as a vector space, and semantic relatedness is gauged as the proximity of the corresponding vectors (e.g., Latent Semantic Analysis (LSA, Deerwester et al., 1990), BEAGLE (Jones et al., 2006)). Semantic spaces are usually built using information on how words co-occur with different frequencies in different contexts. If a large enough corpus (on the order of thousands of documents, each having hundreds of terms) is used, the resulting space can simulate human behavior on a variety of tasks (Landauer and Dumais, 1997; Jones et al., 2006). In order to devise an illustrative automatic procedure, we built an LSA space using 53,956 documents and 56,108 terms obtained from Wikipedia², applying standard methods (Landauer et al., 2007). The 390-

dimensional semantic space performed comparably well on the TOEFL synonym test (64.65%, versus the 'gold-standard' of 64.38% (Landauer and Dumais, 1997))³. The automatic labeling procedure was as follows: First, we projected each paragraph of text onto the semantic space, generating paragraph vectors representing their semantic content. Since each of the terms used to build the word-document matrix can also be represented as a vector in that space, we selected the three terms that were closest to each paragraph and used them as "automatic labels", to build a thematic graph (similar to Fig. 1C). We performed the dot product between every word vector and each paragraph vector to determine which word vectors were closest to each paragraph. A pre-selection of words was made by projecting "windows of words" of length 8, and selecting 3 labels for each. Then, when computing dot products for the whole paragraphs, the words were chosen from the previously obtained set of labels, not from the full 56,108 terms (Fig. 5).

Although the semantic space method produced noisy labels (Fig. 5), the results are nonetheless encouraging at least when applied to large portions of text. We discuss the potential of this and other methods in Section 5.

² <http://download.wikimedia.org/backup-index.html>.

³ Kindly provided by Prof. T. Landauer. Despite this good score we nonetheless validated all our results using the LSA space available at <http://lsa.colorado.edu>.

concerning verbal hallucinations, using different models of neural networks (Mizraji, 1989). Our long term goal is to apply neural models to the production of incoherent discourse. If the measurements presented here can be applied generally, and the translation of these procedures to neural models can be achieved, the modeling of language production deviances on a large scale will be possible, and thus provide much needed insight into the neural and cognitive processes underlying speech production in schizophrenia.

Role of funding source

EM and JCVL were supported by PEDECIBA and CSIC-UDELAR. AC was supported by PEDECIBA. BE was supported by the Northern Norwegian Regional Health Authority (Helse Nord RHF). None of the funding agencies (PEDECIBA, CSIC or Helse Nord RHF) had any further role in study design, data analysis and interpretation, or in the writing of the report and decision to submit for publication.

Contributors

EM and JVL conceived the original ideas and theoretical framework. AC designed the procedures to create and analyze the lexical and topic graphs, corrected and improved the LSA-Wikipedia previously created by JVL and wrote all the programs for text and network analysis. BE motivated the clinical application of the theoretical framework and tools. All authors discussed the methods and results and contributed to the writing of the manuscript. All authors have read and approved the final manuscript.

Conflict of interest

None of the authors have any potential conflicts of interest or biomedical financial interests.

Acknowledgements

We thank Dr. Andrés Pomi for the valuable discussions and encouragement.

References

- Albert, R., Barabási, A.L., 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74 (1), 47–97.
- Andreasen, N.C., 1986. Scale for the assessment of thought, language and communication (TLC). *Schizophr. Bull.* 12, 474–482.
- Andreasen, N.C., Black, D.W., 2005. *Introductory textbook of psychiatry*, 4th edition. American Psychiatric Association, Washington DC.
- Andreasen, N.C., Grove, W.M., 1986. Thought, language and communication in schizophrenia: diagnosis and prognosis. *Schizophr. Bull.* 12, 348–359.
- Cabana, A. 2009. Representación de la estructura del lenguaje escrito mediante grafos y espacios semánticos. [Representation of the structure of written language using graphs and semantic spaces] MSc Thesis. PEDECIBA-Universidad de la República, Uruguay.
- Chen, E.Y., 1994. A neural network model of cortical information processing in schizophrenia. I: interaction between biological and social factors in symptom formation. *Can. J. Psychiatry* 39, 362–367.
- Collins, A.M., Loftus, E.F., 1975. A spreading-activation theory of semantic processing. *Psychol. Rev.* 85 (6), 407–428.
- Collins, A.M., Quillian, M.R., 1969. Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* 8, 240–247.
- Dayan, P., Abbott, L., 2001. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. The MIT Press, Boston.
- Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R., 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, 391–407.
- DeLisi, L.E., 2001. Speech disorder in schizophrenia: review of the literature and exploration of its relation to uniquely human capacity for language. *Schizophr. Bull.* 27, 481–496.
- Docherty, N.M., 2005. Cognitive impairments and disordered speech in schizophrenia: thought disorder, disorganization, and communication failure perspectives. *J. Abnorm. Psychol.* 114, 269–278.
- Dorogovtsev, S.N., Mendes, J.F.F., 2001. Language as an evolving word web. *Proc. Roy. Soc. B-Biol. Sci.* 268, 2603–2606.
- Doyle, A.C., 2005. A study in Scarlet. In: *The complete Sherlock Holmes by Sir Arthur Conan Doyle*. Collector's Library Editions, CRW Publishing Limited, UK. pp. 13–52 [First published in 1887, Ward Lock & Co, UK].
- Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr. Res.* 93, 304–316.
- Elvevåg, B., Foltz, P.W., Rosenstein, M., DeLisi, L.E., 2010. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J. Neurolinguist.* 23, 270–284.
- Ferrer i Cancho, R., 2005. The structure of syntactic dependency networks: insights from recent advances in network theory. In: Altmann, G., Levickij, V., Perebyinis, V. (Eds.), *The problems of quantitative linguistics*. Ruta, Chernivtsi, pp. 60–75.
- Ferrer i Cancho, R., Solé, R.V., 2001. The small world of human language. *Proc. Roy. Soc. B-Biol. Sci.* 268 (1482), 2261–2265.
- Ferrer i Cancho, R., Riordan, O., Bollobás, B., 2005. The consequences of Zipf's law for syntax and symbolic reference. *Proc. Roy. Soc. B-Biol. Sci.* 272 (1562), 561–565.
- Foltz, P.W., 2007. Discourse coherence and LSA. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (Eds.), *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum, New York, pp. 167–184.
- Halligan, P.W., David, A.S., 2001. Cognitive neuropsychiatry: towards a scientific psychopathology. *Nat. Rev. Neurosci.* 2, 209–215.
- Hodges, J.R., Salmon, D.P., Butters, N., 1992. Semantic memory impairment in Alzheimer's disease: failure of access or degraded knowledge? *Neuropsychologia* 30, 301–314.
- Hoffman, R.E., 1987. Computer simulations of neural information processing and the schizophrenia–mania dichotomy. *Arch. Gen. Psychiatry* 44, 178–188.
- Hoffman, R.E., McGlashan, T.H., 1997. Synaptic elimination, neurodevelopment, and the mechanism of hallucinated "Voices" in Schizophrenia. *Am. J. Psychiatry* 154, 1683–1689.
- Hoffman, R.E., McGlashan, T.H., 1998. Reduced corticocortical connectivity can induce speech perception pathology and hallucinated 'voices'. *Schizophr. Res.* 30, 137–141.
- Hoffman, R.E., Rapaport, J., Ameli, R., McGlashan, T.H., Harcherik, D., Servan-Schreiber, D., 1995. A neural network simulation of hallucinated "voices" and associated speech perception impairments in schizophrenia patients. *J. Cogn. Neurosci.* 7, 479–497.
- Jones, M., Kintsch, W., Mewhort, D., 2006. High-dimensional semantic space accounts of priming. *J. Mem. Lang.* 55 (4), 534–552.
- Jurafsky, D., Martin, J.H., 2000. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, Upper Saddle River, NJ.
- Kintsch, W., 1988. The role of knowledge in discourse comprehension: a construction–integration model. *Psychol. Rev.* 2, 164–182.
- Kuperberg, G., 2010. Language in schizophrenia. Part 1: an Introduction. *Lang. Linguist. Compass* 4 (8), 576–589.
- Landauer, T., Dumais, S., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.* 104, 211–240.
- Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (Eds.), 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum, New York.
- McKenna, P., Oh, T., 2005. Schizophrenic speech: making sense of bathroofs and ponds that fall in doorways. Cambridge University Press, Cambridge, UK.
- McKenna, P.J., Mortimer, A.M., Hodges, J.R., 1994. Semantic memory and schizophrenia. In: David, A.S., Cutting, J.C. (Eds.), *The neuropsychology of schizophrenia*. Lawrence Erlbaum Associates, Hove, pp. 163–178.
- Mizraji, E., 1989. Context-dependent associations in linear distributed memories. *Bull. Math. Biol.* 51, 195–205.
- Mizraji, E., Pomi, A., Valle-Lisboa, J.C., 2009. Dynamic searching in the brain. *Cogn. Neurodyn.* 3, 401–414.
- Motter, A.E., de Moura, A.P.S., Lai, Y.C., Dasgupta, P., 2002. Topology of the conceptual network of language. *Phys. Rev. E* 65 (6) 065102(R).
- Niznikiewicz, M.A., Shenton, M.E., Voglmaier, M., Nestor, P.G., Dickey, C.C., Frumin, M., Seidman, L.J., Allen, C.G., McCarley, R.W., 2002. Semantic dysfunction in women with schizotypal personality disorder. *Am. J. Psychiatry* 159, 1767–1774.
- Palla, G., Derenyi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435 (7043), 814–818.
- Pinker, S., 2000. *The language instinct. 2nd reimpression*. Harper Perennial Classics, New York.
- Quillian, M., 1968. Semantic memory. In: Minsky, M. (Ed.), *Semantic Information Processing*. The MIT Press, Cambridge, Massachusetts, pp. 227–270.
- Solovay, M.R., Shenton, M.E., Holzman, P.S., 1987. Comparative studies of thought disorder. I. Mania and schizophrenia. *Arch. Gen. Psychiatry* 44, 13–20.
- Steyvers, M., Tenenbaum, J.B., 2005. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci.* 29 (1), 41–78.
- Talamini, L.M., Meeter, M., Elvevåg, B., Murre, J.M., Goldberg, T.E., 2005. Reduced parahippocampal connectivity produces schizophrenia-like memory deficits in simulated neural circuits with reduced parahippocampal connectivity. *Arch. Gen. Psychiatry* 62, 485–493.
- Tamlyn, D., McKenna, P.J., Mortimer, A.M., Lund, C.E., Hammond, S., Baddeley, A.D., 1992. Memory impairment in schizophrenia: its extent, affiliations and neuropsychological character. *Psychol. Med.* 22, 101–115.
- Utsumi, A., 2011. Computational exploration of metaphor comprehension processes using a semantic space model. *Cogn. Sci.* 35, 251–296.
- Valle-Lisboa, J., Mizraji, E., 2007. The uncovering of hidden structures by Latent Semantic Analysis. *Inf. Sci.* 177, 4122–4147.
- Valle-Lisboa, J., Reali, F., Anastasia, H., Mizraji, E., 2005. Elman topology with sigma-pi units: an application to the modeling of verbal hallucinations in schizophrenia. *Neural Netw.* 18, 863–877.
- Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of small-world networks. *Nature* 393, 440–443.

Capítulo 8

Conclusiones y perspectivas

En este trabajo de tesis se presentan resultados obtenidos utilizando aproximaciones neurocomputacionales al estudio de la coherencia temática en el lenguaje. En particular, son parte de un esfuerzo por entender los mecanismos que subyacen a la identificación de la estructura temática, y la pérdida de coherencia en la producción patológica del discurso, como sucede en el caso de la esquizofrenia.

La primera tanda de resultados explora, utilizando modelos de redes neuronales computacionalmente simples pero potentes, posibles arquitecturas capaces de aprender relaciones semánticas presentes en el ambiente lingüístico al que se expone un hablante, de modo similar a como lo hacen los modelos de espacios semánticos. Este tipo de modelos puede convertirse en una herramienta interesante para dos objetivos relacionados pero algo diferentes.

En primer lugar, resultan importantes en tanto implementaciones neuralmente plausibles de espacios semánticos. Como se revisó en el capítulo 3.2, los procedimientos utilizados para capturar información semántica y/o de tópico a partir de textos no hacen referencia a los posibles mecanismos neurales que estarían implementando esta funcionalidad en el cerebro. Esto último puede verse como una limitación si se está interesado en comprender los mecanismos neurales del procesamiento y producción del discurso tanto normal como patológico. Algunos otros esfuerzos en esta dirección, son por ejemplo el trabajo de [Mizraji et al. \(2009\)](#) que plantea una similitud formal entre el LSA y las memorias asociativas lineales (sección 2.2), el planteo de [Blouw y Eliasmith \(2013\)](#) de una posible implementación neural de la codificación holográfica propuesta por [Jones y Mewhort \(2007\)](#) y [Kanerva \(2009\)](#), y el modelo de memoria de contexto temporal de [Shankar et al. \(2009\)](#).

En segundo lugar, su carácter de modelo neural permite exploraciones que las implementaciones no neurales no permiten. Como se ha mostrado en los capítulos introductorios de esta tesis, los modelos de redes neuronales han sido utilizados para simular el desempeño en condiciones patológicas,

tanto en afecciones propias del desarrollo de la estructura neural o cognitiva, como en el caso de las dislexias del desarrollo (Harm y Seidenberg, 1999), así como también en el deterioro ocurrido en la madurez funcional, como en el caso de las alexias, afasias (Dominey, 2003), y la esquizofrenia (Hoffman y McGlashan, 2001). Estos modelos permiten evaluar hipótesis fisiopatológicas, así como evaluar la eficacia de posibles tratamientos (Hoffman *et al.*, 2011; Stein y Ludik, 1998a).

Los modelos neurales que implementan espacios semánticos podrían ser utilizados, entonces, para explorar el desempeño en condiciones que emulen patologías como la demencia semántica, la enfermedad de Alzheimer, la esquizofrenia, etc. El desempeño de los pacientes afectados con estas patologías parece alterado en algunas tareas, como la asociación libre de palabras, la facilitación semántica en tareas de decisión léxica, entre otras (Kuperberg, 2010b; Laisney *et al.*, 2011; Merck *et al.*, 2014; Vistoli *et al.*, 2011). Estas alteraciones observadas pueden proveer datos de validación para los modelos neurales que incorporen mecanismos fisiopatológicos presuntamente implicados.

8.1 Validación de espacios semánticos

Existen varias formas de comparar el desempeño de un espacio semántico al desempeño humano. Entre otras fuentes de información, las tareas de asociación libre y de facilitación semántica han sido utilizadas para tal fin (Griffiths y Steyvers, 2011; Hare *et al.*, 2009; Jones *et al.*, 2006; Lund *et al.*, 1995; Recchia y Jones, 2009; Shankar *et al.*, 2009).

Como se mostró en el capítulo 6, actualmente estamos colectando normas de asociación libre del español rioplatense para, entre otros fines, evaluar el desempeño de los modelos de espacio semántico desarrollados. El sitio web donde se lleva a cabo la recolección es: <http://www.smallworldofwords.com/platense>.

En forma complementaria, estamos realizando experimentos de facilitación semántica en tareas de decisión léxica. Camila Zugarramurdi, en el marco de sus Tesis de Maestría, replicó los experimentos de Hare *et al.* (2009) de facilitación semántica de palabras relacionadas por eventos. En un experimento, colectó los tiempos de reacción a palabras *target* cuando eran precedidas por palabras *prime* que denotaban eventos relacionados o no relacionados al *target*. En otro experimento, observó la modulación del potencial evocado N400 (relacionado con el procesamiento de la expectación semántica, (Kutas y Hillyard, 1980)) en la misma tarea. En la figura 8.1 A se ilustra el protocolo utilizado. Los resultados preliminares obtenidos por Zugarramurdi muestran una facilitación de unos 30 milisegundos en los tiempos de reacción causada por los eventos relacionados, y una modulación del componente N400 entre ambas condiciones (figura 8.1 B y C).

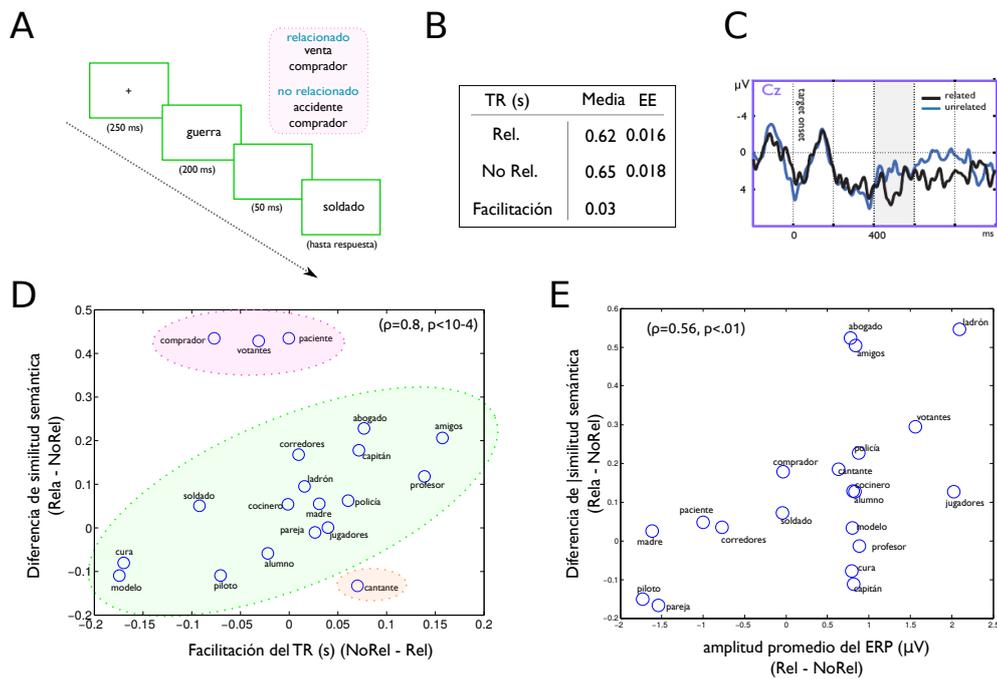


Figura 8.1: Resultados preliminares de los experimentos realizados por Camila Zugarramurdi. A. Diseño de la tarea. B. Tiempos de reacción por condición. C. Modulación del componente N400 en el electrodo Cz. D. Correlación parcial entre la diferencia de tiempos de reacción y la diferencia de similitud semántica para cada par *prime-target* en las dos condiciones. Nótese que se excluyen 4 ítems de la correlación. E. Correlación entre la diferencia de amplitud promedio N400 entre condiciones y la diferencia de similitud semántica.

Estos datos fueron utilizados para evaluar el desempeño de un espacio semántico obtenido mediante *ASL* en base a un *corpus Wikipedia* en español (similar al utilizado en el capítulo 6). Para cada par de palabras *prime-target* utilizadas en los experimentos, se calculó la similitud semántica en el espacio semántico obtenido. Luego, se calcularon las correlaciones entre la diferencia de tiempos de reacción y de amplitud promedio del componente N400 entre condiciones, y la diferencia de similitud semántica. Según se puede apreciar en la figura 8.1 C y D, se obtuvieron correlaciones significativas en ambos casos.

Estos resultados preliminares ilustran la posibilidad de obtener espacios semánticos cuyo desempeño se correlaciona con medidas comportamentales y electrofisiológicas de similitud semántica. Esto es particularmente importante si se repara que existen pocos ejemplos de este tipo de enfoques para el español, y menos aún para el español rioplatense.

8.2 Entropías y modelos de redes neuronales

Las entropías temáticas aquí presentadas pueden convertirse en herramientas útiles para cuantificar la coherencia del discurso, siempre y cuando se logre comprobar la confiabilidad y validez como instrumento de medida. Una potencial aplicación de estas entropías es la que motivó su formulación, la cuantificación de coherencia en discursos de pacientes con esquizofrenia, en particular con fines de investigación clínica (Cabana *et al.*, 2011; Elvevåg *et al.*, 2007).

Más en general, podrían utilizarse para cuantificar pérdida de coherencia en situaciones experimentales, por ejemplo, en experimentos en que se evalúa el desempeño de la función ejecutiva en diferentes condiciones de carga atencional (Michael Jones, comunicación personal). Dada la siempre creciente disponibilidad de textos en una escala masiva, las entropías podrían ser utilizadas para explorar la estructura temática de diferentes tipos de texto, a diferentes escalas (local *vs.* global, por ejemplo).

Además de ser aplicadas a discursos producidos obtenidos por humanos, las entropías temáticas pueden aplicarse a los “discursos” producidos por modelos de producción del lenguaje. Un ejemplo de esto es el trabajo de Valle-Lisboa *et al.* (2014), quienes utilizan un modelo de producción basado en memorias asociativas dependientes de contexto (sección 2.2.7) para emular el deterioro de la coherencia en pacientes con esquizofrenia.

El planteo de Valle-Lisboa *et al.* (2014) intenta resolver la aparente paradoja de la “desconexión–hiper-conexión” que implica la hipótesis patofisiológica de la reducción excesiva de sinapsis (Hoffman y McGlashan, 1997; Hoffman *et al.*, 1995) (véase la sección 4.3.1). Según esta hipótesis, varios de los síntomas de la esquizofrenia se deben a un sobregiro en el podado de las sinapsis que ocurre normalmente en la corteza durante la transición a

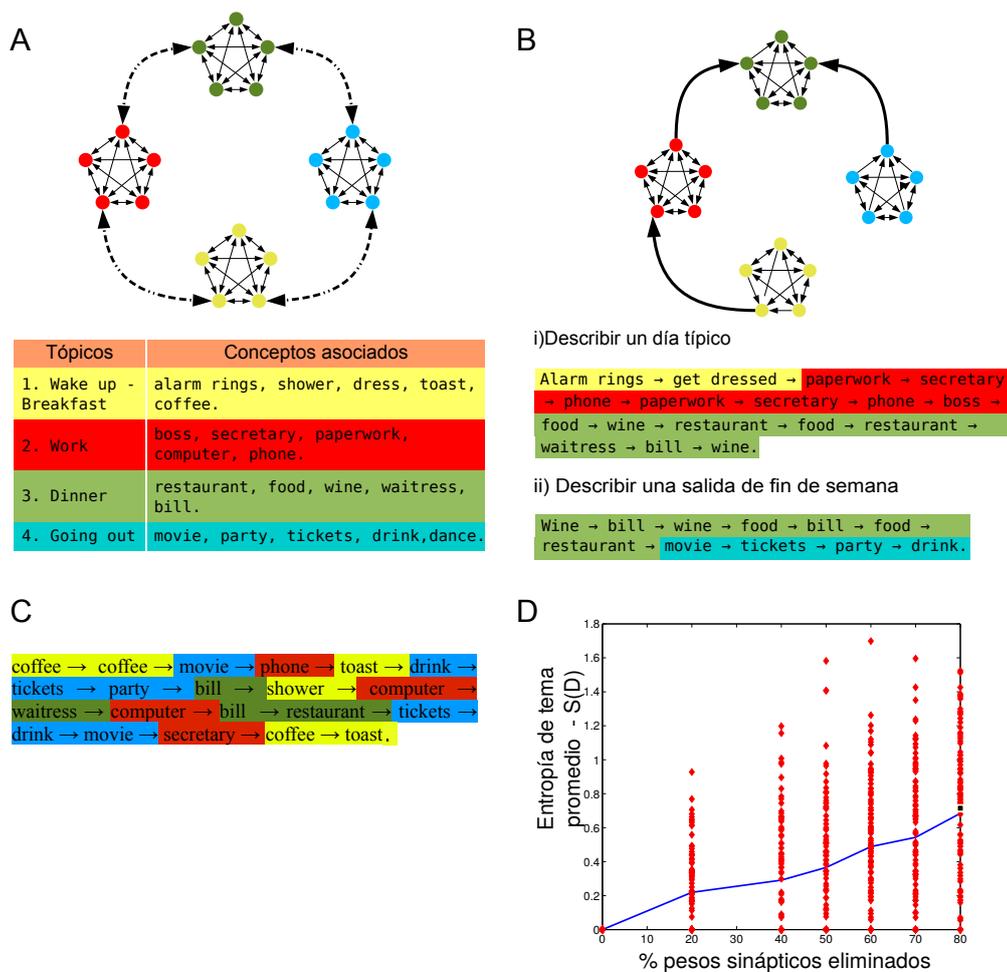


Figura 8.2: Modelo empleado en [Valle-Lisboa et al. \(2014\)](#). A. Red de asociación de conceptos almacenada. B. Asociaciones inducidas por un contexto específico, y ejemplos de discursos producidos en este (i) y otro (ii) contexto. C. Ejemplo de discurso patológico obtenido eliminando pesos sinápticos. D. Aumento de entropía de tema promedio al incrementarse el porcentaje de sinapsis eliminadas.

la vida adulta. Por tanto, esta “desconexión” excesiva sería la causante de fenómenos semánticos como la pérdida de coherencia en el habla, y las asociaciones extrañas (entre palabras que normalmente no parecen asociadas). Estos fenómenos que parecieran ocasionarse por una “hiper-conexión” de los significados de las palabras (en particular en situaciones en las que por ejemplo, se asocian o combinan dos conceptos diferentes debido al colapso de diferentes significados de una palabra).

El modelo de producción utilizado en ese trabajo almacena asociaciones entre conceptos, y su funcionamiento se esquematiza en la figura 8.2 A. Los conceptos pertenecen a distintos tópicos, y las asociaciones entre éstos dependen de un marcador de tópico que oficia como contexto. Durante la producción, el contexto genera en el modelo asociaciones específicas entre palabras, e impone una direccionalidad a las posibles transiciones entre tópicos. El funcionamiento del modelo intacto se muestra en la figura 8.2 B. Dado un contexto que oficia de objetivo, y una palabra inicial, el modelo genera una sucesión de palabras que recorre los tópicos en forma ordenada, hasta llegar al objetivo fijado.

Los autores del trabajo exploraron los efectos de una eliminación sináptica masiva en el desempeño de modelo. En la figura 8.2 C se muestra que en estas condiciones, el modelo tiende a generar sucesiones de palabra que recorren los tópicos en forma desordenada. Esta aparente pérdida de coherencia en el discurso queda evidenciada en la figura 8.2 D, que muestra cómo la entropía de tema de los discursos producidos aumenta al incrementarse el porcentaje de sinapsis eliminadas. Por tanto, un modelo que sufre la “desconexión” exacerbada de sus unidades neurales es capaz de producir, empero, una aparente “hiper-conexión” entre conceptos no asociados previamente, lo que genera un discurso con menor coherencia, a juzgar por el incremento en las entropías temáticas.

Esta propuesta resolución de la paradoja antes mencionada, si bien promisoriosa, está basada en un modelo muy elemental de producción del discurso. Esto último puede intentar resolverse utilizando modelos neurales de espacios semánticos de desempeño comparable al humano en algunas tareas. Los modelos de redes neuronales planteados en este trabajo de tesis constituyen un punto de partida en esta dirección. Futuros esfuerzos en este tema, por tanto, requerirán de modelos de espacios semánticos cuyo desempeño esté validado por evidencia comportamental, como las que se presentaron en la sección precedente.

En suma, el desafío planteado por la articulación de modelos de redes neuronales que implementen espacios semánticos con las medidas de desorganización del discurso, es un reflejo de la importancia y utilidad del lenguaje como ventana al mundo cognitivo, y a la naturaleza de la implementación neural subyacente.

Artículo publicado en *Cortex*

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue: Research report

A modular approach to language production: Models and facts



Juan C. Valle-Lisboa^a, Andrés Pomi^a, Álvaro Cabana^a, Brita Elvevåg^{b,c}
and Eduardo Mizraji^{a,*}

^a Group of Cognitive Systems Modeling, Biophysics Section, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

^b Psychiatry Research Group, Department of Clinical Medicine, University of Tromsø, Tromsø, Norway

^c Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway

ARTICLE INFO

Article history:

Received 11 July 2012

Reviewed 10 October 2012

Revised 19 December 2012

Accepted 7 February 2013

Published online 19 February 2013

Keywords:

Language

Neural models

Model neuroimaging

Schizophrenia

Discourse trajectories

ABSTRACT

Numerous cortical disorders affect language. We explore the connection between the observed language behavior and the underlying substrates by adopting a neurocomputational approach. To represent the observed trajectories of the discourse in patients with disorganized speech and in healthy participants, we design a graphical representation for the discourse as a trajectory that allows us to visualize and measure the degree of order in the discourse as a function of the disorder of the trajectories. Our work assumes that many of the properties of language production and comprehension can be understood in terms of the dynamics of modular networks of neural associative memories. Based upon this assumption, we connect three theoretical and empirical domains: (1) neural models of language processing and production, (2) statistical methods used in the construction of functional brain images, and (3) corpus linguistic tools, such as Latent Semantic Analysis (henceforth LSA), that are used to discover the topic organization of language. We show how the neurocomputational models intertwine with LSA and the mathematical basis of functional neuroimaging. Within this framework we describe the properties of a context-dependent neural model, based on matrix associative memories, that performs goal-oriented linguistic behavior. We link these matrix associative memory models with the mathematics that underlie functional neuroimaging techniques and present the “functional brain images” emerging from the model. This provides us with a completely “transparent box” with which to analyze the implication of some statistical images. Finally, we use these models to explore the possibility that functional synaptic disconnection can lead to an increase in connectivity between the representations of concepts that could explain some of the alterations in discourse displayed by patients with schizophrenia.

© 2013 Elsevier Ltd. All rights reserved.

* Corresponding author. Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay.

E-mail addresses: mizraj@fcien.edu.uy, emizraji@gmail.com (E. Mizraji).

<http://dx.doi.org/10.1016/j.cortex.2013.02.005>

0010-9452/© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Research investigating the neural substrates of language production and comprehension is a complex and rapidly evolving field (Pulvermüller, 2010, 2012). A clear understanding of these neurocognitive processes is also relevant for a variety of psychiatric and neurological conditions where language is often a crucial operational indicator of cognitive function, as well as the clinical course and state of the illness (Andreasen and Grove, 1986; Snowdon et al., 1996; DeLisi, 2001; McKenna and Oh, 2005).

One important approach to the study of the neural basis of cognition relies heavily on mathematical models of neural networks. These models try to reveal the essential workings of the almost inextricably complex real neural networks of the brain. One of the main goals of neural modeling is to translate some basic empirical facts into transparent mathematical devices that could be used to explicitly establish some properties of network topologies and neuronal connections, including synaptic transmission (McCulloch and Pitts, 1943; Anderson, 1995; Arbib, 1995; Elman et al., 1997).

The language faculty includes several functional modules and so it is natural to try and devise explicit theoretical accounts that connect all the modules and explain how they function and how they breakdown. We suggest that adoption of mathematical and computational approaches may nurture a more explicit and unified account of language by demanding explicitly stated hypotheses. Such an approach does, of course, run the risk of naively ignoring important details, and in light of current knowledge this approach will be unable to do justice to the richness of current verbal or symbolic characterizations of the relevant phenomena. However, we believe that these limitations are at least partially outweighed by the heuristic value of a computationally explicit (and hence testable and readily falsifiable) model. Indeed, we suggest that the diversity of current models points to the need for a more constrained form of theorizing.

A core assumption underlying the present work is that many of the properties of language production and comprehension can be understood in terms of the dynamics of modular networks of neural associative memories. In our computational modeling that we present below each module has a core matrix associative memory. This representation is ample enough to accept connectionist approaches à la Anderson–Kohonen or McClelland–Rumelhart (Anderson, 1972; Kohonen, 1977; McClelland et al., 1986; Rumelhart et al., 1986a) where learning is essential, and also some eventual *a priori* neural circuits created by evolution and developed during brain embryogenesis.

The aim of this paper is to present this modular network of associative memories as a convenient theoretical tool to unify diverse aspects of cognitive neuroscience of language production, comprehension and distortion, and exemplify its potential with the analysis of some expected consequences of functional disconnection and its explicative power of empirical observations of language disorganization in some patients with schizophrenia.

Therefore, as a contribution toward a neurocomputational ‘theory’ of language we connect three theoretical and

empirical domains currently used to analyze language, namely (1) neural models of language processing and production, (2) statistical methods that are used to create functional brain images, and (3) corpus linguistic tools such as Latent Semantic Analysis (henceforth LSA) that are used to discover the topic organization of language. To this end we demonstrate how these neurocomputational models intertwine with LSA and the mathematical basis of functional neuroimaging. Subsequently, we apply these methods to some aspects of discourse disorganization.

We have recently presented an empirical analysis of discourse productions in patients with schizophrenia, where we introduce an entropy measure in order to quantify discourse disorganization (Cabana et al., 2011a). Here we show that this disorganization can be captured in our models of language production. We use our three-part approach (stated above) in order to evaluate the consequences of functional disconnection at the level of discourse disorganization and neuroimaging, solving the apparent paradox that disconnection at the synapse (and neuroimaging level) can nevertheless produce a certain level of semantic hyper-connection in the discourse of some patients with schizophrenia.

The organization of this paper is as follows: in Section 2 we present a graphical representation of linguistic productions that enables the measurement of the degree of order in discourse. This representation of empirical data was inspired by results obtained from neural models of associative memories, and motivates the exploration of the usage of neurocomputational models in the investigation of the nature of language anomalies in psychosis. In Sections 3 and 4 we review previous theoretical work that provides the foundation for the development of our neural models, and Sections 5–7 document the results from our lesioned models in terms of the deterioration of discourse production. In Section 3 we review the fundamentals of associative memory models and present a multiplicative model that we have been developing for several years, and that is at the core of all the models we employ in this paper. Then, in Section 4 we describe a multi-modular neural network that reproduces the thematic packing of lexical strings in a way that is closely related to LSA. In Section 6 we present a new context-dependent neural model that performs goal-oriented linguistic behavior. In Section 5 we link matrix associative memory models with the mathematics that underlies functional neuroimaging techniques and thus show the “functional brain imaging” that would be generated from the model described in Section 6. Finally, in Section 7 we use the numerical results emerging from this model to explore a possible explanation for the disconnection-hyperconnectivity paradox that is reported in some patients with schizophrenia.

2. Order and disorder in the topology of language

In addition to being an important human trait on its own, language is an observable that can point to underlying pathologies or disorders. The fact that language shows regularities at many levels allows for empirical quantitative descriptions to be devised that capture an operational

measure of putative pathological processes. A valuable objective is to find measures that are related to the underlying mechanisms involved in language production, and thus whose change we can model.

We have recently described the use of graphical representations for the trajectories of discourse (Cabana et al., 2011a) comparing the linguistic productions from patients with schizophrenia to healthy comparison participants. In the present paper, we are interested in establishing connections between the graphical representations and neural models. We first present an application of our method as an illustrative example. To do so, we selected a brief passage from the personal diaries of the famous Russian ballet dancer Vaslav Nijinsky (1890–1950) that was written during a period of illness (1919) and that is particularly lacking in coherence (taken from the unexpurgated edition, published in 1999). In the most comprehensive modern examination of Nijinsky's medical notes and life in general, Joseph H. Stephens, M.D. and Peter Ostwald, M.D. conclude: "Thus, our final diagnosis according to DSM¹ III of the tragic genius Vaslav Nijinsky must be Schizoaffective Disorder in a Narcissistic Personality" (Ostwald, 1991; p. 350). Naturally there are many problems and risks with retrospective psychiatric diagnosis, but we nonetheless present this case information simply to show that the writing samples from Nijinsky's diary were generated by a person who most certainly struggled with a form of psychosis. As a contrasting sample, we used the first paragraphs of "A Study in Scarlet" about the adventures of the fictional detective Sherlock Holmes written by Arthur Conan Doyle in 1887 that features a first-person introduction of Dr. John Watson's experiences in the British army (Doyle, 2005). The clearly written and narrative nature of the text presupposes few complex metaphors or other literary devices that could complicate analysis and hence obscure comparison. Both passages are displayed in Table 1 (see Fig. 1).

Both samples were subjectively evaluated in order to identify a small set of topics that were present in the text. Here, we illustrate a graphical representation of the trajectory of the discourse where the visits to different topics are used to quantitatively evaluate the degree of order of discourse. In Fig. 1, the text is represented as a line that traverses over the top of relevant nodes of an underlying semantic graph, and each topic is represented as a distinct layer on the graph. A text with a more disordered trajectory would show a greater degree of oscillation and change between topic layers.

As mentioned earlier, we have recently developed a quantitative measure to index the degree of disorder, namely "topic entropy" (see Cabana et al., 2011a). Specifically, this entropy results in higher values for samples in which one or more topics are "visited several times", and in lower values when each topic is "visited" only once. When we apply this measure to the literary examples in Fig. 1, crucially we obtain a higher topic entropy for the Nijinsky sample ($S = 1.75$) than for the Doyle sample ($S = 0$). This difference is what would be expected given that topic entropy measures the disorganization of discourse, and we have previously shown that this metric correlates systematically with the ratings of clinical

levels of thought disorder in patients with schizophrenia (Cabana et al., 2011a).

The graphical representations thus far presented, where the semantic space spanned by the discourse is projected in thematic subspaces, has been inspired by a class of context-dependent associative memory models in which the contexts allow the dissection of the conceptual structure of the memory (Pomi and Mizraji, 2004). In the next section we introduce the basic ideas underlying these models in terms of their potential usefulness to explore some aspects of the pathophysiological hypotheses in language disorganization.

3. Modeling cognitive neural processing with context-dependent matrix memories

The classical models of associative memories (Anderson, 1972; Cooper, 1973; Kohonen, 1977) are based on matrix algebra, and large dimensional vectors are used as a natural mathematical representation of neural data. A landmark in neural modeling was heralded by the emergence of the "Parallel Distributed Processing (PDP) Group" (Rumelhart et al., 1986b; McClelland et al., 1986), as well as the pioneering research previously conducted by Rumelhart and McClelland which developed the learning and representation abilities of original neural models. This work established the very foundation for subsequent investigations aimed at exploring deep psychological questions using neural models, specifically language and development (see Elman et al., 1997). Indeed, language processing has motivated some extremely influential neural models such as the one developed by Elman (1990), a multimodular neural model consisting of a perceptual neural device, a working memory and an associative memory.

It is interesting to emphasize that one of the main advantages provided by the kind of neural models we are employing in this work is the universality of vector coding of neural activity. If we consider the receptive aspects of language, we note that the phonetic reception of speech, the visual perception of written language or the tactile detection of texts coded in Braille involve different physical signals (acoustic waves, photons and mechanical pressures). In these three cases, each kind of sensory receptor transduces its associated signal into a bioelectric neuronal signal that neurons carry to a variety of neural processing relays. Finally, in each case, it is highly plausible that the neural modules responsible for decoding language receive large sets of electrochemical and neurochemical activities that are, by definition, neural vectors independent of the original sensory modality.

Similarly, in the case of language production, the large sets of neuronal signals that emerge during the cognitive and linguistic processes involved, also define neural vectors. The ensemble of neural vectors produced in this way, will after a number of preprocessing stages command the motor processes that make language communicable and observable. For these reasons, models involving large dimensional vectors as inputs and outputs are potentially apt to represent the neuronal dynamics underlying linguistic processing and production. The models of the pathophysiology of schizophrenia by Hoffman and McGlashan (1997) and Valle-Lisboa et al (2005)

¹ Diagnostic and Statistical Manual of Mental Disorders

Table 1 – The two texts used for the illustration of the notion of the underlying thematic structure of language production. The first text is extracted from “A study in scarlet” (Doyle, 2005) and the other from the diaries of Nijinsky (1999).

“A Study in Scarlet”, Chapter 1, A. Conan Doyle

In the year 1878 I took my degree of Doctor of Medicine of the University of London, and proceeded to Netley to go through the course prescribed for surgeons in the army.

Having completed my studies there, I was duly attached to the Fifth Northumberland Fusiliers as Assistant Surgeon. The regiment was stationed in India at the time, and before I could join it, the second Afghan war had broken out. On landing at Bombay, I learned that my corps had advanced through the passes, and was already deep in the enemy's country. I followed, however, with many other officers who were in the same situation as myself, and succeeded in reaching Candahar in safety, where I found my regiment, and at once entered upon my new duties. The campaign brought honors and promotion to many, but for me it had nothing but misfortune and disaster. I was removed from my brigade and attached to the Berkshires, with whom I served at the fatal battle of Maiwand. There I was struck on the shoulder by a Jezail bullet, which shattered the bone and grazed the subclavian artery. I should have fallen into the hands of the murderous Ghazis had it not been for the devotion and courage shown by Murray, my orderly, who threw me across a pack-horse, and succeeded in bringing me safely to the British lines. Worn with pain, and weak from the prolonged hardships which I had undergone, I was removed, with a great train of wounded sufferers, to the base hospital at Peshawar. Here I rallied, and had already improved so far as to be able to walk about the wards, and even to bask a little upon the verandah, when I was struck down by enteric fever, that curse of our Indian possessions. For months my life was despaired of, and when at last I came to myself and became convalescent, I was so weak and emaciated that a medical board determined that not a day should be lost in sending me back to England. I was dispatched, accordingly, in the troopship “Orontes,” and landed a month later on Portsmouth jetty, with my health irretrievably ruined, but with permission from a paternal government to spend the next 9 months in attempting to improve it.

Nijinsky's Book I, On Life, pp 3–4 (Unexpurgated Edition).

I have had a good lunch, for I ate two soft-boiled eggs and fried potatoes and beans. I like beans, only they are dry. I do not like dry beans, because there is no life in them. Switzerland is sick because it is full of mountains. In Switzerland people are dry because there is no life in them. I have a dry maid because she does not feel. She thinks a lot because she has been dried out in another job that she had for a long time. I do not like Zurich, because it is a dry town. It has a lot of factories and many business people. I do not like dry people, and therefore I do not like business people.

The maid was serving lunch to my wife, to my first cousin (this, if I am not mistaken, is how someone related to me by being my wife's sister is called), and to Kyra, together with the Red Cross nurse. She wears crosses, but she does not realize their significance. A cross is something that Christ bore. Christ bore a large cross, but the nurse wears a small cross on a little ribbon that is attached to her headdress, and the headdress has been moved back so as to show the hair. Red Cross nurses think that it is prettier this way and have therefore abandoned the practice that doctors wanted to instill in them. The nurses do not obey doctors, because they do not understand the instructions they have to carry out. The nurse does not understand the purpose she is here for, because when the little one was eating, she wanted to tear her away from her food, thinking that the little one wanted dessert. I told her that “she would get dessert when she had eaten what was on the plate”. The little one was not offended, because she knew I loved her, but the nurse felt otherwise. She thought I was correcting her. She is not getting any better, because she likes eating meat. I have said many times that it is bad to eat meat. They don't understand me. They think that meat is an essential thing. They want a lot of meat. After eating lunch they laugh. I am heavy and stale after eating, because I feel my stomach. They do not feel their stomachs, but feel blood playing up. They get excited after eating. Children also get excited. They are put to bed because people think they are weak creatures. Children are strong and do not need help. I cannot write, my wife disturbs me. She is always thinking about things I have to do.

illustrate a scenario with phonetic inputs and conceptual outputs. However, it should be emphasized that we employ no specific phonetic particularities and we assume that the output of the phonological module is a vector representing a word. This does not imply that the phonological processes cannot be the locus of or the basis for language associated pathologies. Indeed, a strong case has been made for the role of phonological aspects of working memory in language processing (see [Baddeley, 2007](#) for a review), and consequently in various language disorders ([Gathercole and Baddeley, 1990](#)).

Matrix models of distributed associative memories are, naturally, only approximations of real memories and thus they have obvious limitations. Nevertheless, these models are capable of capturing important aspects of collective neuronal behaviors, especially considering neural networks composed of a very large number of neurons. In his early work on the biological plausibility of these matrix models, [Cooper \(1973\)](#) analyzed the manner in which similar patterns corresponding to the same percept could be identified with a prototype. This observation arises because similar patterns processed during learning generate a type of statistical average, and this average is by definition the prototype of the experienced pattern. For instance, if we are introduced to a new friend (we

use the letter k to indicate this particular friend) and we observe her from different angles, each visual experience is coded by the neural system as a set of neural vectors $f_1(k)$, $f_2(k)$, $f_3(k)$, $f_4(k)$, ..., $f_n(k)$, corresponding to the different views of the face of this friend (different perceptual angles, for example). In this case, what the memory stores is an average pattern $f(k)$, a large n -dimensional vector whose dimension n depends on anatomical connectivity. A new pattern similar to $f(k)$ produced by an unknown photo from our friend k , say $f^*(k)$, is finally projected onto $f(k)$, and due to the large dimensionality, it is in fact confounded with $f(k)$ (the small differences are made null) and this pattern is identified as the face of our friend. This is the basic reason that enables these matrices to act as statistical identifiers. As was demonstrated by [Kohonen \(1972\)](#) and [Anderson \(1972\)](#) large dimensionalities enable the sharing of the same matrix support with a number of other different patterns (e.g., the faces of different friends k' , k'' , etc.) We illustrate this central property of associative matrix models in [Fig. 2](#). Note that these averages behave in the framework of the model in a similar manner to prototypical concepts used in cognition. In this way the emergence of concepts, a high level cognitive brain ability, can be traced back to the imperfection of memory and its propensity to

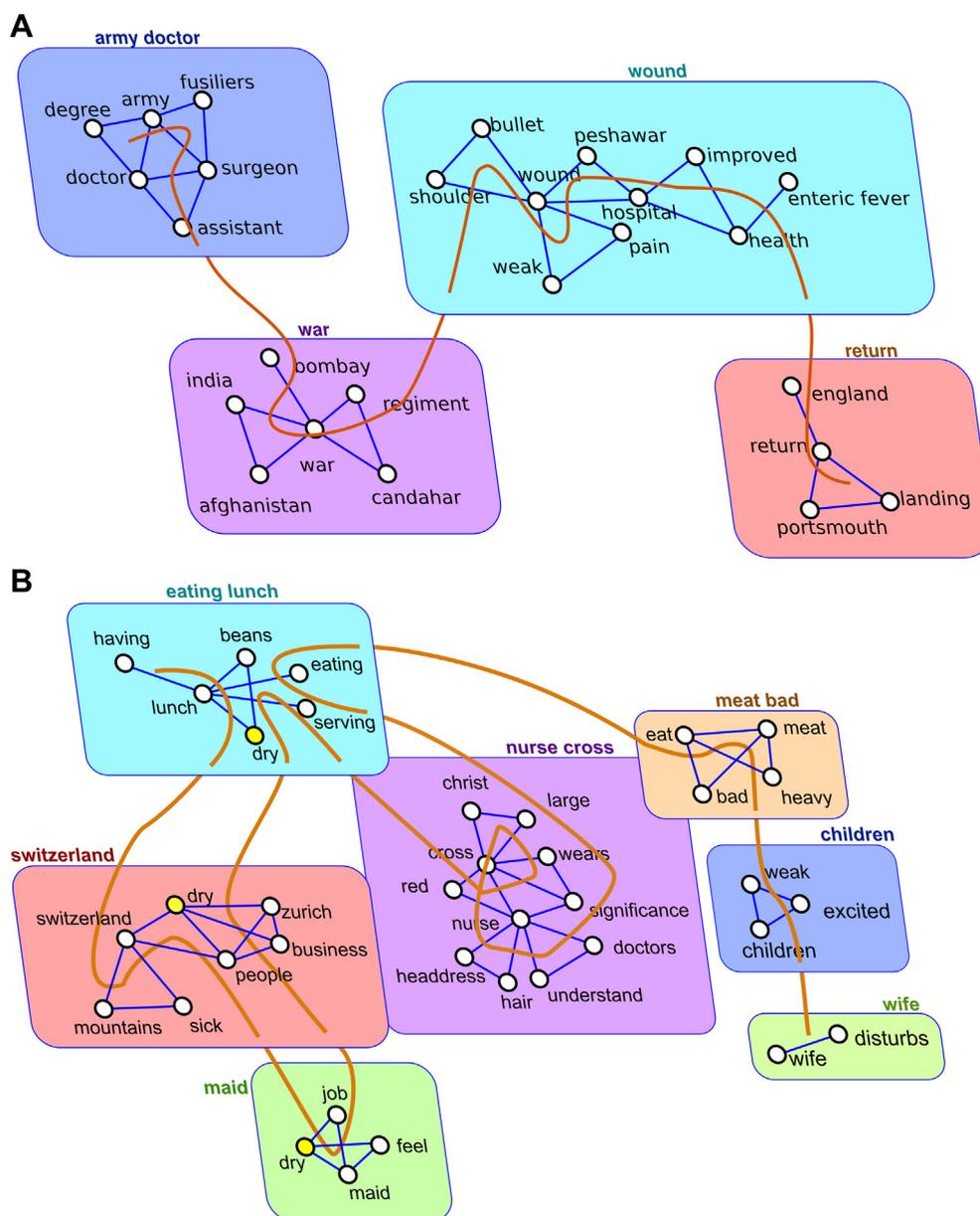


Fig. 1 – Topic graphs obtained from manual label assignment for (A) Dr. Watson’s “speech” and (B) Vaslav Nijinsky’s diary. Discourse trajectories are shown as lines over the graphs. From visual inspection alone (and supported by the corresponding entropies, see text) it can be seen that the latter sample is more disordered with a more recurrent trajectory. The yellow “dry” node in (B) indicates perseveration of a concept over different topics.

confound similar patterns (for an analysis of this issue, see Cooper, 1973, and also Levi-Montalcini, 1989). An interesting consequence of the capacity to project onto prototypes is the ability of these matrix memories to correctly identify patterns that have deteriorated, a remarkable phenomenon beautifully illustrated by the classical numerical experiments of Kohonen and coworkers (Kohonen, 1977; Kohonen et al., 1977).

It is noteworthy that matrix memories can retrieve patterns according to the frequency of presentation of the patterns during the learning process, something which can be interpreted as a statistical procedure (Pomi and Mizraji, 1999; Mizraji, 2008; see also Oja, 1982). In its simplest form,

each association is weighted by its frequency which leads directly to the Singular Value Decomposition (henceforth SVD) of the matrix. SVD is a technique that decomposes a matrix as a sum of matrices in order of decreasing importance. In matrix memories each term of the SVD is a particular associative memory and the highest weighted terms retain the most important associations. In this way the effect of learning can be seen as an enhancement of prototypical associations and a reduction of the importance of infrequent or underrepresented patterns, effectively performing a kind of dimensionality reduction (see below, and Mizraji, 2008).

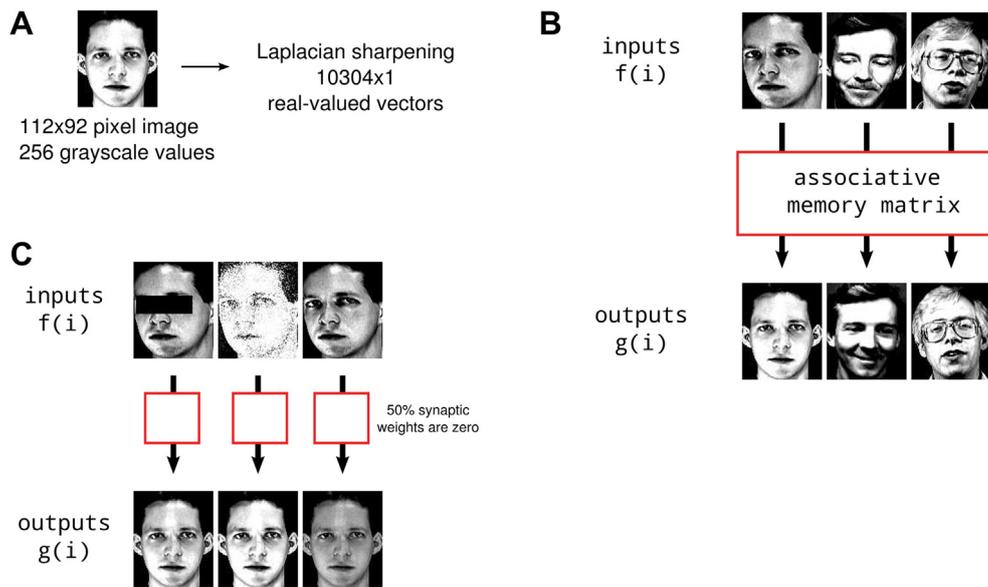


Fig. 2 – Example of an associative memory matrix whose input and output vectors are images of human faces [the images we used in these simulations were adapted from the set of [Samaria and Harter \(1994\)](#)]. (A) For each grayscale image, the edges were enhanced and the pixel values stored in a single column vector. (B) An associative memory matrix was built with two faces of the same individual used as inputs and outputs. (C) The resulting memory was tested for its ability to restore the corresponding output given an incomplete input (a noisy input) and the destruction of half the synaptic weights stored in the matrix.

The potential of the matrix memory models can be largely extended by employing multiplicative vector contexts. Multiplicative interactions have been postulated in different cognitive contexts ([Humphreys et al., 1989](#); [Mizraji, 1989](#); [Smolensky, 1990](#)). In our models ([Pomi and Mizraji, 2004](#); [Mizraji et al., 2009](#)), the context vectors allow the storage in memories of a large variety of potentially adaptive behaviors. This kind of contextualization requires performing some form of multiplication of signals at the synaptic level, or through

the existence of coincidence detectors, or due to the display of AND functions at the level of small neural networks capable of acting as units in a larger neural model (for further detail on these multiplicative performances, see [Koch and Poggio, 1992](#); [Peña and Konishi, 2001](#)).

An alternative to multiplicative contexts is to assume that contexts are part of a network that interacts additively with the input. This is precisely the approach adopted in classical multilayer perceptron models trained by backpropagation

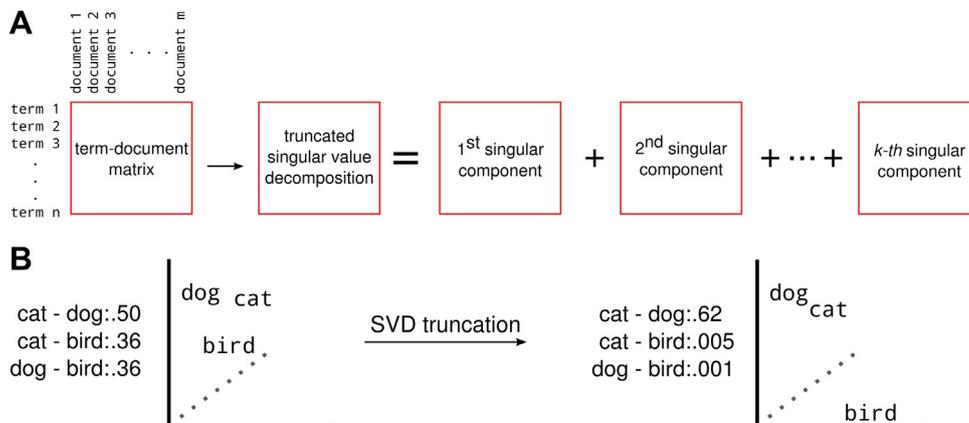


Fig. 3 – Schematic description of LSA. (A) A term-by-document matrix (TD-matrix) is constructed, where each element is a function of the frequency of a word in a document from the corpus. Then, SVD is applied to the matrix, but only k singular vectors and values are retained (with k heuristically selected in the order of several hundreds), resulting in a truncated TD-matrix. (B) Pairwise word (cosine) similarities for “cat”, “dog” and “bird” before versus after SVD truncation of the TD-matrix. This procedure enhances the similarity between “cat” and “dog” while dramatically reducing the similarity between the other two. A depiction of the vector positions in the semantic space enabled by LSA is also shown.

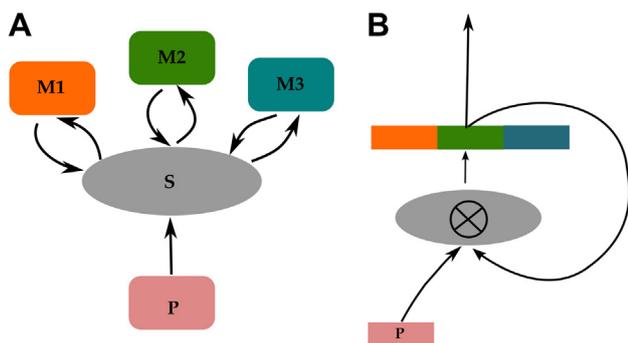


Fig. 4 – The neural architecture underlying the topic detection model. (A) Neural processing of language requires phonetic and phonological processing achieved by module (P), a Sketchpad where integration of information is produced (S), and several memory modules (corresponding to different modalities or association cortices) that are queried for an interpretation of the representation present in the sketchpad. (B) Our version of this model connects LSA and neural models by collapsing the model in (A). The module (P) produces phonological representations. The integration is achieved through the Kronecker product (symbolized by “ \otimes ”) and it is used to query memory systems that produce a vector of activities conceived as the interpretation. Therefore, each pair (word, interpretation) produces a new interpretation which is the “gist” of the linguistic input. It is in this sense that the model detects the “topic” of a processed discourse. In the same way that a string of words is mapped to a vector that represents the aggregated meaning of the words in LSA, here the model gives the composite meaning as a topic vector.

(Rumelhart et al., 1986a, 1986b). This approach is of neuro-computational interest despite the fact that the biological basis of backpropagation training remains uncertain, although there are recent proposals that aim to solve this lack of biological realism while preserving the desirable computational properties (Grüning, 2007). Instead, the multiplicative networks, that we adopt here in the simulations displayed in Figs. 5–8, can be trained with simpler Hebbian-like learning procedures. A caveat is that potential biophysical mechanisms that can support multiplication have not yet been definitively established. This limitation notwithstanding, a great advantage of a neurocomputational approach based on multiplicative contexts is that it provides a clear theoretical foundation for exploring neural networks and allows a straightforward vector-matrix representation of some of the networks; this very fact presents a unique opportunity to exploit powerful mathematical instruments to evaluate the potential of the modeled neural systems (Pomi and Mizraji, 2004; Mizraji et al., 2009).

4. A model for the thematic packing of lexical strings and its relation to LSA

A second approach that can bridge the gap between neurobiological and cognitive levels is research in the domain of machine learning and information retrieval. These fields have created a set of methodologies to automatically retrieve information, extract topics and organize information derived from stored textual documents. At least one of these methods – LSA (Deerwester et al., 1990) – shows similarities with procedures plausibly employed by the human brain to acquire and decode linguistic productions (Mizraji, 2008; Pulvermüller, 2012). LSA is interesting since although its original application was as a method for information retrieval (Deerwester et al., 1990), early on it was proposed as an influential psychological model of language acquisition and analysis (Landauer and Dumais, 1997; Landauer et al., 1998; Kintsch, 2001; Dumais, 2003; Landauer et al., 2007). LSA exploits well-understood and powerful mathematical methods based on matrix and vector algebra particularly apt for computer implementations.

In LSA documents are coded as large dimensional (column) vectors, whose components correspond to some defined function of the frequency in each document of particular words in the vocabulary. The whole set of documents defines a matrix, the “term-by-document matrix” which is processed using SVD (Berry and Browne, 2005). This method captures indirect links between different documents (tens of thousands) and thus reveals “latent” semantic relationships between them. The resulting matrices can be huge, but easily available computational power makes it possible to compute a term-by-document matrix of hundreds of millions of components using a simple desktop computer.

We have already mentioned the implicit connection between matrix memories and SVD, but in LSA the SVD is explicitly employed. The processing of the matrix through SVD leads to a set of non-negative numbers of decreasing magnitude namely the “singular values”, and a set of associated vectors called “singular vectors”. The final result of the SVD is that the original term-by-document matrix is

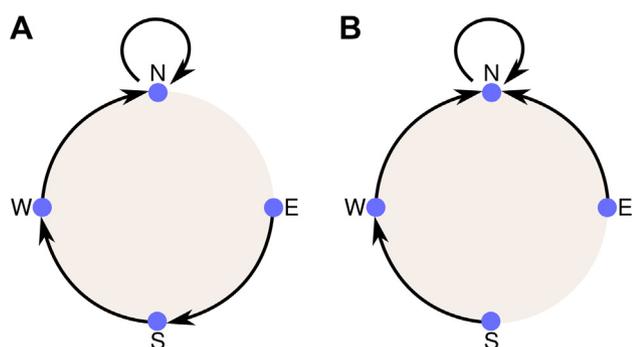


Fig. 5 – Graphical representation of the associations stored in the simple model. Two examples of the pathway to North. In both cases the associations are trained between neighboring neural (cognitive) states but with a projected overall design such that the system can achieve the explicitly stated goal. The global design could be to perform only clockwise steps (A) or to reach the goal in only one step if the initial point is adjacent to it (B). The successive states of the system will be blind local steps but they are globally organized as actions aimed toward reaching the overall goal.

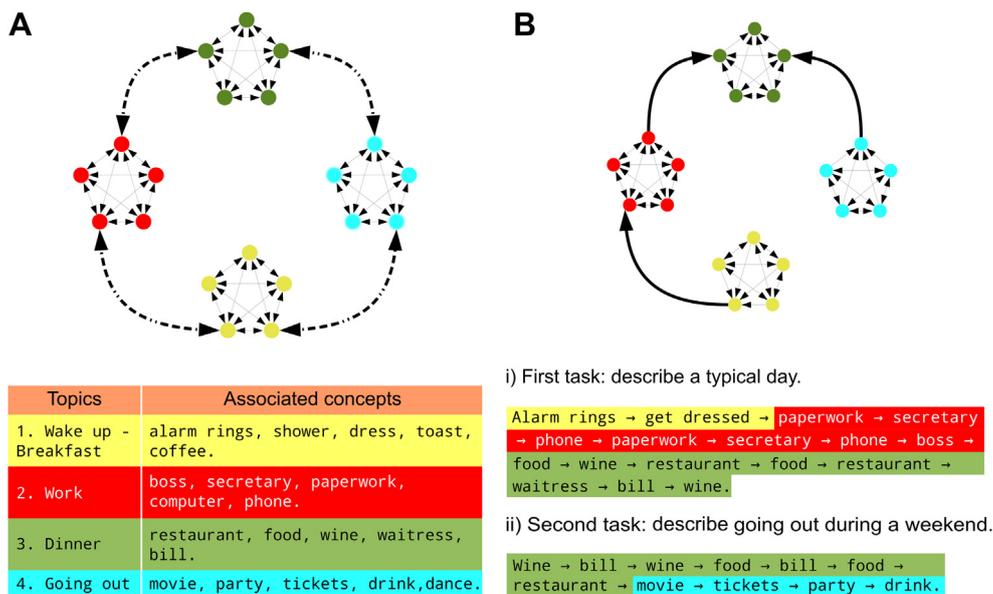


Fig. 6 – A model of language production as an ordered trajectory through topics. (A) The installed graph of concepts. Every color represents a topic, where each node refers to a concept. Full lines between nodes are context specific links; dashed lines represent connections between concepts in different topics. The table under the graph describes every topic and every associated concept. (B) Output of the model when asked to “describe a typical day, starting in the morning” (i) or (ii) “going out during a weekend”. The graph represents the first situation. Below the graph some example trajectories are depicted. Note that there is some repetition of the concepts in the target topic given that we did not include a stopping criterion. The imposition of the topic selects a set of links between concepts in such a way that from every concept stored there is at least a route to the target topic (see Appendix for the mathematical details). Whenever there are multiple routes the system decides probabilistically which route to take.

expressed as a sum of matrices that result from the product of singular vectors, with each one of these new matrices weighted by the corresponding singular values (Berry and Browne, 2005). Usually (but not always) raw word frequencies are not used in LSA, but are substituted by one particular function. The functions weight differently each term according to its distribution in a corpus of texts (see Salton and Lesk, 1965; Dumais, 1991; Landauer et al., 2007). For instance highly frequent words (which are not informative about the topic of the document) are weakly weighted and some other words (e.g., articles or prepositions) are discarded. Words that appear only in a single document are also discarded. This defines a meaningful vocabulary, and the dimension of the document vector is the length of the meaningful vocabulary used in the full set of documents. As a consequence, the vector dimension is usually on the order of thousands (Landauer et al., 1998). The central idea of LSA is the retention of hundreds (out of thousands) of those terms of the decomposition that are associated with the largest singular values (see Fig. 3 for an illustration of this idea).

It was empirically discovered that the retained singular vectors can usually be regarded as “topic markers” and that the singular vectors involved act as symbolic conceptual vectors that produce a form of averaging of the real documents involved in a topic (Hofmann, 1999; Papadimitriou et al., 2000; Valle-Lisboa and Mizraji, 2007).

The relevance of LSA to language processing models has been discussed previously (e.g., Landauer et al., 1998; Foltz

et al., 1998). Additionally, it has been shown to be useful in understanding some of the unique features of language in Broca’s aphasia (Roll et al., 2011), and also in accounting for word frequency differences in semantic aphasia and semantic dementia (Hoffman et al., 2011). Moreover, it has been demonstrated that it is possible to evaluate patients with schizophrenia based on open-ended verbalizations and using LSA automatically derived language scores to accurately distinguish patients from controls, patients from other patients, and also from their family members (Elvevåg et al., 2007, 2010). In this sense, besides being successful for the description of putatively “normal” language use and acquisition (Landauer and Dumais, 1997), LSA also provides a framework with which to examine its breakdown.

Inspired by LSA and its relationships with neural models, we present below a novel neural model that can make these connections explicit. As our main concern is the exploration of possible links between memories, neural imaging and LSA, we assume that the input of our model is a sequence of vectors, each consisting of a particular pattern of activity of a neural population that represents each word in the input. Each of these internally represented words operates as a query to the memory systems of the “brain” in such a way that together with the current understanding of the previous linguistic input, it elicits new conceptual understanding (see Fig. 4 for a general outline of the model).

The various memory systems that are queried by the internally represented word can be the modality specific

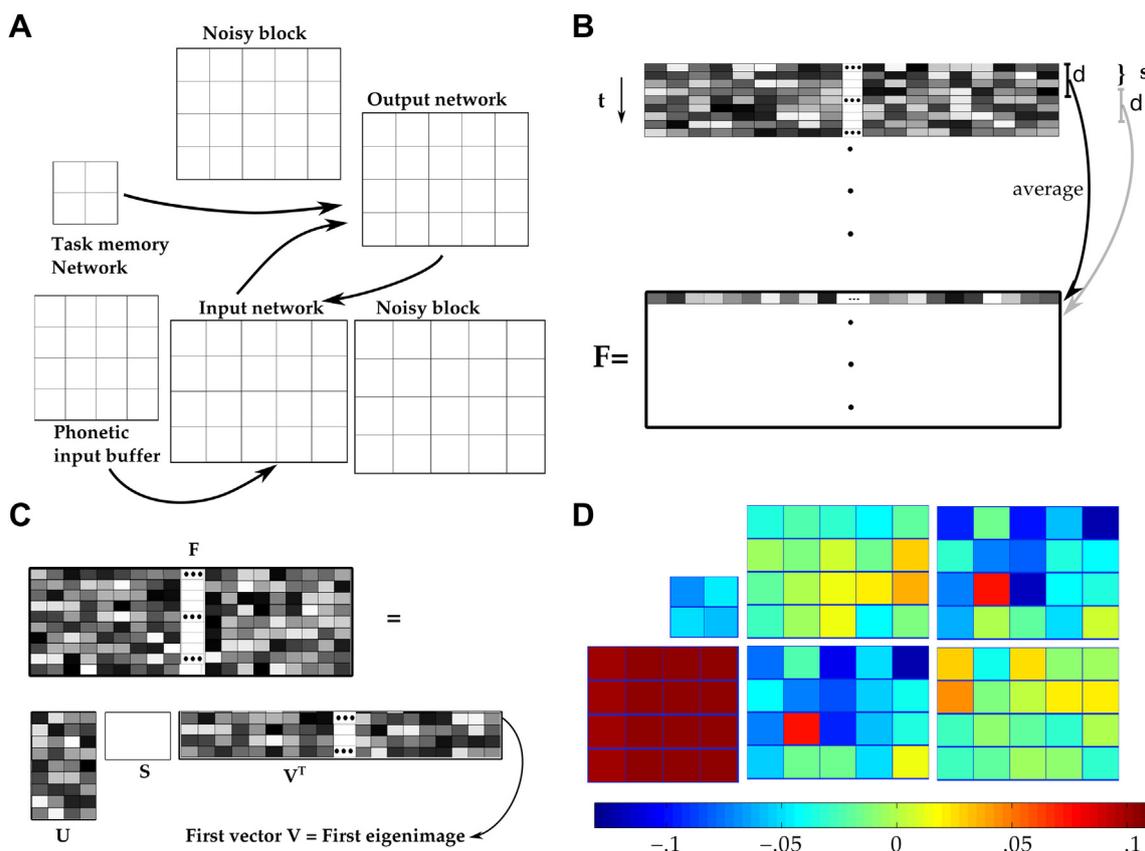


Fig. 7 – (A) Structure of the neural blocks involved in the simulations. Each block corresponds to a vector in the model. The noisy blocks vectors do not participate in the computations and each of their cells is assumed to have Gaussian noise. The task memory buffer is the vector that codes for the goal or target of the discourse. The phonetic input buffer has a special coding of the input, assumed to be different to the input network which represents the activity that goes into the computational module. Finally the output network is the output activity of the computational module. (B) Time averaging and matrix construction. Each unit in the model – including unrelated blocks – has an associated activity that varies with time. In order to have a more realistic time resolution, we use a moving window of width d that moves s time steps, in order to average the activity of each unit. If the whole set of tasks takes T units of time we end up with T/s time epochs. In each epoch the average activity of each unit is placed in a row vector as shown. (C) SVD of the matrix that results from B yields a set of V and U vectors. The first V vector is the (first) eigenimage (D) projecting the eigenimages obtained from the model of Fig. 6. The models were first subjected to the task of going from topics $3 \rightarrow 1$, then paused (p), followed by $1 \rightarrow 3$, p , $1 \rightarrow 3$, p , $3 \rightarrow 1$. The images are the mean eigenvectors obtained over twenty different simulation sets. The colorbar in the bottom of Panel (B) shows the deviation of activity levels from baseline; in red the highest activity increase and in blue the biggest decrease.

cortex or even category specific areas (Pulvermüller, 2010). The data retrieved from this query is a vector of activities representing the various pieces of information that each word in the specific context elicits in each memory system. We can formally represent this network by a distributed memory whose output is an interpretation of the context and which depends both on the previous interpretation and the current input. This module is an example of a multiplicative context-dependent memory as described above that associates each word with a subset of interpretations. The strength of the association reflects the likelihood of the word given the interpretation. In the absence of input, the interpretation is a combination of possible interpretations weighted by the prior probability. Therefore, it can be shown that the output of this context-dependent module at each time step is a combination of interpretations weighted by the *a posteriori* probability. In

this simple and approximate conception the interpreter acts as a naive Bayesian classifier (Valle-Lisboa, 2007).

Richer diagnosis capabilities can be obtained by subtle modifications such as having a short-term memory that keeps track of several inputs (see Pomi and Olivera, 2006; for an application to medical diagnosis) leading to enhancements of recognition capabilities in artificial models aimed at Natural Language Processing tasks (Valle-Lisboa, 2007; Cabana and Valle-Lisboa, in preparation). What is of critical relevance here is the fact that a linguistic input consisting of a sequence of words is transformed into a sequence of internally represented words and associated concepts, which can thus be processed further. But what is unique about this inner conceptual representation, as opposed to the bare (original) input? The importance of the process depicted in Fig. 4 lies in the fact that in order to understand an utterance, individuals first need to generalize

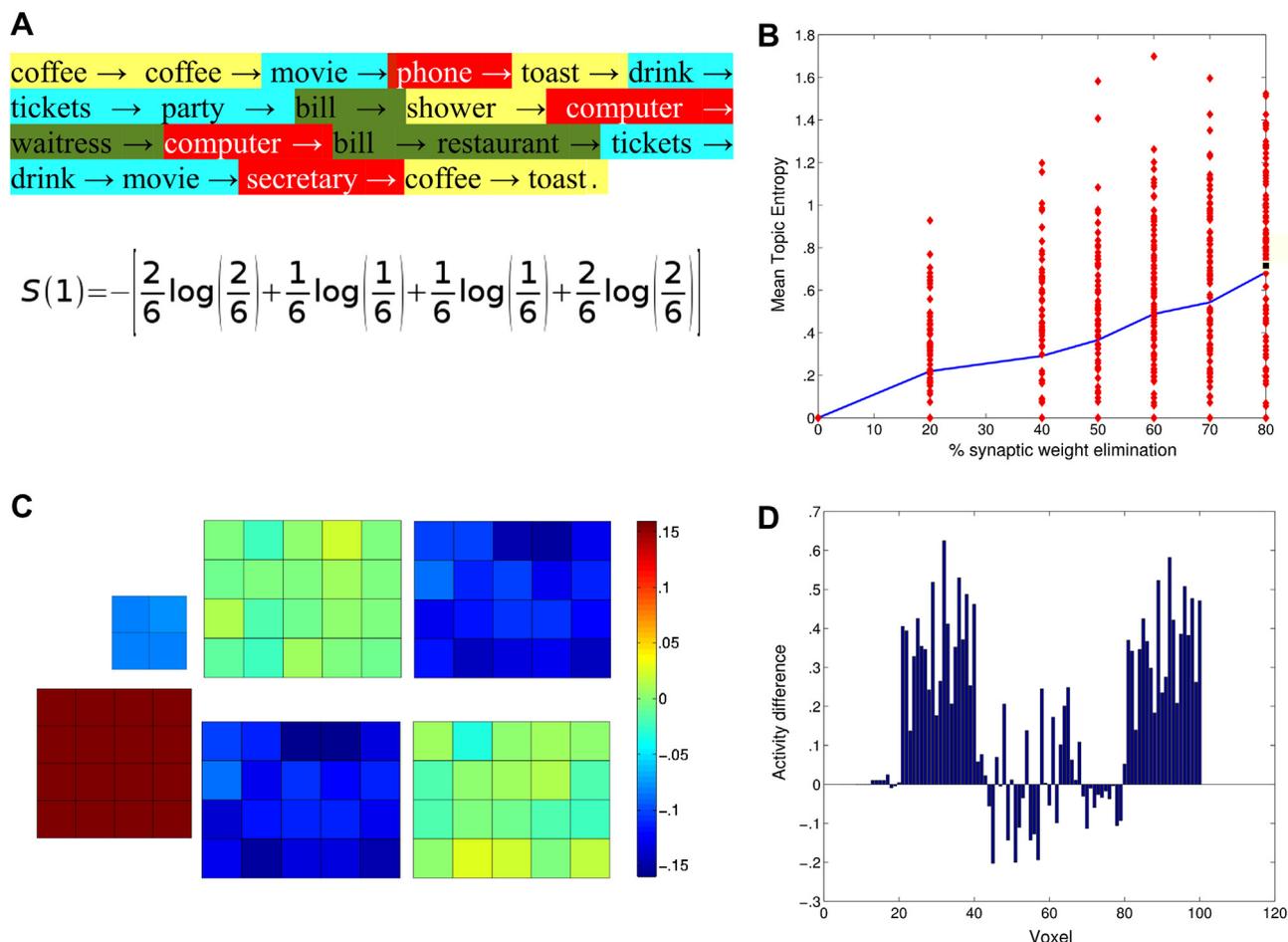


Fig. 8 – The effect of synaptic weight elimination in the workings of the simplified neural model of Fig. 6. The performance is evaluated on one of the tasks as in Fig. 7. (A) Example of a trajectory (first task) of an “affected” individual (simulation). Notice the jump between different topics. Under the discourse trajectory we show the calculation that leads to the entropy of topic 1, $S(1)$. As there are six words belonging to topic 1 every term has a denominator equal to 6. Each term adds the contribution of each segment of topic 1; the first segment has two words, the second and third have one word, and the last segment belonging to topic 1 has two words. Thus the topic 1 entropy is 1.33. The mean entropy for the discourse is calculated by averaging the entropy over the four topics, and this mean entropy is 1.19. (B) Mean entropy as a function of synaptic weight elimination. For each level of destruction we ran 100 simulations and calculated the mean entropy of the output (red diamonds). The blue line shows the grand average of entropy at each level of synaptic elimination. (C) The average eigenimage of the simulated individuals with 80% of the network’s weights pruned. (D) “Voxel” to “voxel” difference in activity between the average eigenimage of the non-pruned networks and the average eigenimage of the pruned networks. The first block was used as normalization and then the images were subtracted. Notice the predominance of the non-pruned network image.

from what they encountered to novel situations. In our extremely simplified model this is accomplished by the sequence of interpretation vectors that produce a form of contextual interpretation of each term. In this sense, our model is closely related to LSA (Mizraji, 2008), in which the domain or topic that a linguistic expression refers to is the final output of the interpreter module after the sequence of words has been presented. In LSA a string of words is projected onto a semantic space and the resulting semantic vector can be said to represent the interpretation of that string. Other corpus based methods have the same objectives, despite different technical implementations (Hofmann and Puzicha, 1999; Griffiths and Steyvers, 2002). Our model’s interpretations are what in other models can

be termed “topics” (Griffiths and Steyvers, 2002). Naturally this is only a small part of the relevant processes, but in the next section we demonstrate that motor activities related to language can also be modeled in our neural networks and that the destruction of part of the network results in a confusion of discourse targets that can be used to model certain aspects of psychopathology as expressed via language.

5. Functional imaging

In recent years substantial advances in technologies designed to capture and process brain images have greatly expanded

available classical neuroanatomical data concerning neural connectivity. Neuroimages obtained using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) open up new avenues to investigate the relationship between the connectivity of different neural networks and the functional behaviors reliant upon these networks (Sporns, 2010). Within this framework, three different empirical approaches to connectivity can be considered: (1) structural or anatomical connectivity, defined by the specific physical links between neurons; (2) functional connectivity, that explores the statistical correlations of different networks during behavioral tasks performed during a defined time interval; and (3) effective connectivity, that goes beyond the functional correlations and looks for the network of causal dependencies underlying a given behavior (Friston, 1995, 2011; Sporns, 2010). Recently the resting state activity (namely activity not due to any particular experimental task) has been shown to be organized according to similar functional networks as those revealed by conventional activation paradigms. In brief, there is a dramatic explosion of studies designed to understand those networks in terms of neural models (Bienenstock and Lehmann, 1998; Rubinov and Sporns, 2010; Tagliazucchi et al., 2011).

Functional connectivity is a method that enables the exploration of how neural activities correlate during particular tasks and has been applied to a variety of lexical and linguistic data (Xiang et al., 2010). During the execution of a defined verbal production, a sequence of “functional” brain images is captured. These images are represented mathematically as a very large vector, each component of this vector being a voxel spatially located in the brain at a specific point and time. The “intensity” of a voxel is a measure of the level of the neural activity in the corresponding spatial location. Thus, the resulting data from a verbal task would generate a sequence of vectors that are subsequently arranged to form a rectangular matrix (Friston, 1995). This matrix is processed using SVD and the first few singular vectors are retained. Each singular vector is thus an image, usually called an “eigenimage”. The eigenimage is one of the singular vectors of this SVD associated with the larger singular values. Given what SVD does, this is effectively retaining most of the variability in the data.

Here again, as is the case in LSA and in matrix memories, SVD applied to the appropriate matrix, produces vectors that code for a strategic synthesis of the data under investigation. Obviously, in each one of the three different situations, the nature of the matrices is unique. In LSA the matrices store abstract versions of the documents, in the neural memories, the matrices store associative information in their synaptic weights, and in the case of neuroimages the matrices are constructed from time sequences of large voxel vectors.

An appealing possibility is the creation of a model system that allows the viewing of the same kind of linguistic task using the three approaches in combination. However, at this point it is not possible to completely reach this objective with real data, primarily because there is as of yet no universally acceptable neural model for linguistic decoding and production and also because the spatial resolution of imaging does not reach the level of individual neurons. Thus, in order to illustrate how this connection potentially could be achieved we present a miniature system. This model is a framework

with which to explore aspects of language production and additionally to model its breakdown due to anatomical and functional disconnections as may be the case in psychopathology. This is the issue that we address in the next two sections.

6. Context-dependent associative models for goal-oriented linguistic behaviors

The organization of goal-directed sequences of behavior is a basic operation of neural systems. This ranges from the simplest animals and motor actions to the most complex abstract human behavior such as navigation, language production or the searching for solutions in the presence of constraints (which naturally is the foundation of scientific inquiry).

Context-dependent matrix memories enable a neural system to store (eventually in the same neural substrate) different sequences (also rhythmic sequences) that can be accessed depending upon the neural context (Mizraji, 1989). In the model presented above (Fig. 4) the context was the topic or interpretation arrived at by the model. Here we illustrate the generation of sequences using a geographical metaphor: a context-dependent memory trained in such a way that starting from any one of the cardinal points it is able to reach any other point. This can be thought of as a minimalistic system displaying a goal-directed behavior, an essential feature of any model of language production. In our simple example presented previously (Mizraji et al., 1994), given a cardinal destination (a neural activity acting as context), a particular memory was trained to assign a step clockwise or counterclockwise in the compass rose, starting from each possible cardinal point (see Fig. 5).

Many aspects of language production are naturally goal-directed behaviors. To describe a daily routine such as the act of dressing, or how to reach some address in a city, or simply to tell a story, requires the ability to organize a trajectory in discourse so as to meet the objective. In those cases, the discourse develops a trajectory of associations within a theme or a topic and smoothly crosses to a neighboring topic to continue its associative “navigation”. A thematic ambient can be represented by a vector context in the model. Within each topic, different chains of associations can be produced depending on another contextual signal, namely the goal to be reached. This context marks the transition between the different themes or topics of the discourse and the sequence of associations within each theme.

Consider for instance the following extension of the geographic model in which patients' flow of speech is evaluated by simple sequential “script-like” questions, such as describing daily routines. The speaker aims to communicate the routine starting with their morning activities, then going to work, and so on. A simple neural model of these activities is devised below using a single context-dependent memory module that is fed by a neural activity representing the target endpoint in the discourse (e.g., having dinner) and a starting state (e.g., waking up). During the operation of the model, the memory module outputs the words concerning any intermediate themes (for instance going to work).

The behavior of the model is illustrated in Fig. 6, with every topic imposing associative links between concepts, even between terms that are not directly related to the topic. The set of concepts associated with the topic that is the target of the discourse acts as an “attractor”, in such a way that every concept trajectory ends in the intended domains. The links imposed on concepts associated with other topics are unspecified in the sense of being context-independent (in Fig. 6 we describe the workings of the model when it is asked to describe the activities of a typical day starting in the morning and then progressing through work and up until dinner). When the input target is noise (assumed to be a combination of all topics) every concept within each topic is associated to each of the concepts belonging to the same topic and there are special links between all topics (see the dashed lines in the Fig. 6A) that connect some concepts in one topic with nodes in other topics. Notice that the target imposes that every thematic network organizes in such a way that trajectories will go in the direction of the relevant topic. During each step, the output activity of the module enters a probabilistic decision process which selects only one concept; the probability of selecting each concept being proportional to the weight of each concept in the output.

7. Using neurocomputational models to solve the disconnection–hyper-connection paradox

A very elegant and influential model of the pathophysiology of schizophrenia states that part of its neural substrate is the reduction of synapses, possibly as an overshoot of the normal pruning process and importantly this has consequences at the behavioral-symptom level, specifically concerning hallucinations (Hoffman et al., 1995; Hoffman and McGlashan, 1997). The authors proposed a standard connectionist model to show that when this overpruning was applied to a language processing module spontaneous activity resulted, something that they linked to the resulting “hallucinated voices”. This is because the intact model forms linguistic expectations in the sense that the recognition of words depends on the previous linguistic context and when it is pruned it generalizes these expectations to spontaneous neural activity, in a sense confusing noise with input. Essentially the same results are obtained when using the multiplicative context models that constitute the building blocks of our models (Valle-Lisboa et al., 2005).

In line with this previous work, here we analyze the effect of synaptic pruning in discourse production by applying a disconnection regime to the neural model depicted in Fig. 6. Synaptic pruning involves turning to zero a certain percentage of randomly selected weights, a manipulation that can be conceived of as the result of anatomical disconnection but that could also result from neurochemical alterations. We also obtain the imaging counterpart of the neural model. In order to do so, we devise the anatomical model system as shown in Fig. 7, where each square represents a specific “brain” region, and within the squares, each cell represents units in the model. To calculate the unit’s activity, the model is run to execute a specific set of tasks that involve many time steps.

The activity of each neuron is recorded and averaged over time during the execution of each task, in this case using a sliding window. Averaging over sets of units is also possible but it is not shown, basically because it adds complexity but no additional insights. The resulting succession of vectors of averaged activity is used to build a matrix [which is processed as detailed in Friston (2011)]. After obtaining the first eigenvector of this matrix (the ‘eigenimage’ in Friston nomenclature) each unit activity is color-coded and placed in the corresponding cell in the anatomical model. A scheme of the process is shown in Fig. 7.

The results of the pruning simulations are illustrated in Fig. 8. To quantify language disorder we use the “topic entropy measure” (introduced in Cabana et al., 2011a), which essentially measures the tendency of words belonging to the same topics to be produced separately. Elimination of 80% of the synaptic weights resulted in the majority of the simulation sets producing trajectories that were either disordered or deviated from their target when the models were asked to reconstruct “what they did during the day” (notice the dispersion of entropies in Fig. 8 panel B). In Fig. 8 panels C and D we present the average eigenimage of a pruned model with 80% of its connections set to zero, and in Fig. 8 panel D the difference between the images of an intact network and a pruned network.

Although seemingly large, elimination of 80% of the model’s synapses is not unrealistic, since this percentage is calculated by only taking into account the context-dependent memory module, and not other parts of the system which despite being part of the eigenimage are modeled as noise. Moreover, as argued in Valle-Lisboa et al. (2005; see also Mizraji et al., 1994) multiplicative models are idealized in the sense that they assume that the full Kronecker product is implemented. We suggest that this only approximates reality and we have shown previously that removal of a certain number of synapses does not significantly affect the model’s behavior (Mizraji et al., 1994). However, it is important to emphasize that, as is the case in the physical sciences and in some engineering applications, a rigorous application of the quantitative conclusions emerging from neurocomputational models to real neural systems requires scaling. These scaling laws determine the manner in which many properties scale with size, and are very specific to each system. In the case of neuronal models, these laws have only just begun to be explored and likely will become crucial topics in the near future (see Bassett et al., 2010).

The deterioration of neural connectivity can be due to physical disconnection or to modulatory disturbances. We have shown previously that in language processing networks the effect of synaptic pruning can be mimicked by changing each unit’s threshold and that these changes are reversible (Valle-Lisboa et al., 2005). This deterioration in synaptic connectivity does not imply a disconnection of the conceptual semantic network; the way information is coded could make synaptic disconnection into conceptual over-linkage, leading to confusion, or strange associations. Crucially, associative memory models can clarify this issue and resolve the paradox. Distinguishable concepts in the model are coded by orthogonal or quasi-orthogonal vectors. The effect of the physical or functional disconnection of synaptic elements provokes the

loss of the mathematical orthogonality among vectors coding some concepts, creating confusion in the associations and resulting in the linking of concepts that were not previously associated. This linkage does not depend on the semantic similarity but on the arbitrary neural coding of the concepts, which is not pre-determined.

The presence of rare associations induces “jumps” in the thematic sequence of the discourse. The possibility that the synaptic pruning also affects the separateness of the concept or theme acting as a target (these are also coded by vectors that could have lost their orthogonality) can create a loss of polarity in the trajectory, and even result in the random “wandering” through the semantic landscape. Thus, in the model, the synaptic disconnection is evident at the same time as a disconnection in the neuroimages, and as an increase in semantic connectivity leading to conceptual confusion (for a full explanation of the differences between anatomical networks and semantic networks see [Pomi and Mizraji, 2004](#)).

8. Discussion

The relationship between language and the brain has traditionally been studied by a combination of computational and linguistic tools as well as by functional neuroimaging. The addition of neural network models has contributed to new perspectives but also introduced new challenges. We suggest that neural models are the key to connect these approaches and provide a unified theory of normal brain functioning and the ways in which language can be disrupted. The present work is a step in this direction, but much remains to be done. The neural models based on multiplicative contextualization that we use in this work considerably simplify the implementation of modulatory effects of context in associations. Instead of using multiple hidden layers and being trainable by a powerful but biologically hard to sustain algorithm (as it is the case of backpropagation), our multiplicative models can be based on coincidence detection at the synaptic level, and learning can be achieved using simple Hebbian processes. In addition, the mathematical structure implied by multiplicative contextualization via tensor products gives us a powerful symbolic representation that allows deep theoretical insights, capable for instance of guiding the design of computer experiments. This formalism reveals links between the statistics of discourse organization displayed by LSA and the mathematics that construct the functional neuroimages. This fact gives rise to a challenging perspective: the possibility of creating a mutual feedback process that improves our understanding in the three domains. In this way, we can imagine that a deep understanding of linguistic structures in close correspondence with refined functional neuroimages, can help to improve the neural models that describe the underlying (and usually non observable) neuronal dynamics.

The model presented in Sections 5–7 can be used to understand the processes underlying the disorganization of discourse as determined by the entropy we presented in Section 2. To obtain quantitative agreement, we should employ a detailed model of language production, for instance one including a larger vocabulary than what we have used here by way of illustration. As we have discussed elsewhere ([Cabana](#)

[et al., 2011a](#)) we also need to obtain entropy measures from more speech samples from patients and control participants, something that will contribute to further improvements in the methods used to quantify disorder. Toward this end, we are developing a combination of machine learning techniques and LSA in order to obtain automatically the graphical representations and the associated analysis ([Cabana et al., 2011b](#)).

Some further issues merit comment: First, the type of models we apply here are quite basic. In recent years it has become possible to perform large scale simulations of a huge number of realistic, or at least rich, spiking neurons ([Izhikevich, 2006](#)). Although this line of work is increasingly important and likely will figure prominently in future attempts to understand the relationship between brain associative networks and emergent activities such as language, the number of parameters needed to be set can be extremely large and difficult to determine. Potentially this can lead to a combinatorial explosion of possible parameter values, something that is highly impractical and leaves several interpretations open. More importantly, even if the correct parameter set can be found and the model performs well, the basis of its functioning can be hard to discern. In contrast, the model type we have employed in this paper aims to balance realism and interpretability.

Second, the models we have presented above aim to capture the assumed basis of mental functioning, which is that mental activities are the result of the concurrent operation of a large number of interconnected units. Central to this approach, the nature of the local computations together with the wiring diagram and the plasticity rules determine mental activities. Thus, we have sacrificed realism because we have employed context-dependent matrix memories as the core computational devices within each module. Obviously, the level of required detail in any model is determined by the problem that needs to be addressed. In particular, language phenomena belong to a level of organization that requires highly collective neural behaviors to be implemented: language manages abstract symbolic syntactic or logical structures, organizes the received and produced conceptual data and interfaces with many other cognitive capacities.

These limitations notwithstanding, we propose that the models and methods based on matrix algebra provide us with an appropriate level of description that can be used as the foundation for more detailed models (see for example [Cooper, 2000](#); [Lee et al., 2006](#)), or at the very least as building blocks of complex multimodular systems ([Mizraji et al., 2009](#); [Mizraji and Lin, 2011](#)).

The ubiquity of matrix formalisms evident in several of the neural models of language, in the statistical procedures developed to capture the properties of functional brain images and in procedures such as LSA that discover topic organization in language corpora, is not a coincidence. There are two core properties that these approaches have in common that require the use of matrix techniques. One obvious property is the large number of interacting parts, be it voxels, terms or neurons. The other, more subtle, property is the inclusion, in all the aforementioned approaches, steps of dimensionality reduction as a fundamental procedure. Indeed, Friston eigenimages, LSA topics and associative memory prototypes, all emerge from procedures such as SVDs and retention of only

part of the latent structure that the decomposition produces. In image processing SVD is a standard procedure to eliminate noise. In LSA dimensionality reduction discovers underlying hidden links between terms that are not superficially related. The rules of synaptic change in neural networks can produce a similar dimensionality reduction that ends up storing prototypes of the experienced activities.

These similarities should not be taken to imply that the connection between the approaches is straightforward. On the contrary, as we have demonstrated here, together with the many similar properties there are many important differences. For instance, each of the voxels that fMRI or PET register involve the activity of tens of thousands of neurons that within models are usually (but not always, see Poirazi et al., 2003) represented by single units. Language production relies on hidden structures and variables (in addition to word pronunciation) and so a neural instantiation of latent structures is expected, but that does not necessarily imply that the eigenimages, or our vector topics, coincide with latent variables. However, we do think that the similarity of the approaches and tools calls for a principled integration of the different levels involved. In fact, neural models such as the one we described in Section 4 (Mizraji et al., 2009) can produce a type of neural LSA and in this way establish contact with the underlying mathematics of classical LSA (and the results of Landauer and Dumais, 1997).

We have explored here the other part of the connection, namely the relationship between neural models and functional neuroimaging (see Sporns, 2010). Each latent variable in our model implies relationships between concepts but there is not a simple one-to-one mapping between eigenimages and either concepts or latent variables. Naturally there are many other possible ways of implementing an imaging output from the model. In our case we opted to accumulate the units of neural activity as a measure of voxel activity but we did not include any spatial coarsening. In this sense our data is more detailed than what standardly would be obtained experimentally. Nevertheless, there are clear differences between the images obtained in a damaged model compared to the intact model. In particular the difference between the average eigenimage of the intact models and that of the damaged models shows a clear decrease in correlation between areas in the lesioned models.

The relevance of our experiments that damage the synapses of the model is the relationship to phenomena that previously have been postulated as the basic pathophysiological mechanisms underlying some symptoms in psychosis. As mentioned, a well known example is the neural model of Hoffman and McGlashan that aims to connect the neurodevelopmental hypothesis of schizophrenia with the subsequent experience of hallucinated voices (Hoffman and McGlashan, 1997). Following on from this pioneering work, we have previously shown that context-dependent matrix associative memories, adopting sigma-pi neurons and multiplicative contexts, can also be used to investigate the same problem with very similar results (Valle-Lisboa et al., 2005).

Here we adopt a simple geographical model based on multiplicative contexts, to explore aspects of language production with the view that this production is a target-oriented neural activity. The deterioration of the connectivity of parts of the neural modules results in a disorganization of the

network trajectories. During the putatively “normal” operation of the model the sequence of outputs ends within the desired target of the discourse. This models the task where patients with schizophrenia or healthy participants are asked for example to “describe a typical day”. Interestingly, disconnecting the synapses in our model leads to a confusion of targets and in some cases to random “jumps” between topics, as if the model has lost its target-oriented capabilities. A paradoxical feature of these simulations is the fact that increasing the level of disconnection produces an increase in statistical connectivity of regions in the neural model. Moreover, targets that were originally far apart (due to vector orthogonality) become closer after damage (due to the loss of orthogonality) and this leads to target confusion.

The results obtained from our model suggest new explorations at the level of language production, and in particular in some cases of schizophrenia. For instance, can the functional disconnection replicate the paradoxical “shrinking” (hyperconnectivity) at the level of semantic space postulated in some cases of psychosis? Many studies have been conducted to explore natural semantic spaces (Jones et al., 2006; Griffiths et al., 2007), and thus we have available many methods to test our hypotheses. Indeed, our explorations are related to several recent computational attempts to understand how semantic knowledge is stored and represented in the brain, how it is learned through development, and how this knowledge is affected and degraded by acquired injury and illness (for a review, see Rogers and McClelland, 2004). A core premise in this neurocomputational work is that explicit theorizing helps clarify ambiguity and it motivates very specific hypotheses. Additionally there is the promise of being able to resolve clinical paradoxes and thereby provide unified accounts of language and its breakdown.

Acknowledgments

AP, EM and JCVL acknowledge the partial financial support by PEDECIBA and CSIC-UdelaR. AC was supported by PEDECIBA. BE was supported by the Northern Norwegian Regional Health Authority (Helse Nord RHF).

Appendix

Context-dependent associative matrix memories (CDAMM) can be used to implement two-argument vector-valued functions.

In general a CDAMM has the following matrix structure:

$$M = \text{Output}_1(\text{Input}_1 \otimes \text{Context}_1)^T + \text{Output}_2(\text{Input}_1 \otimes \text{Context}_2)^T \\ + \text{Output}_3(\text{Input}_2 \otimes \text{Context}_1)^T + \dots$$

where $\text{Output}_1, \text{Output}_2, \dots$, are different output vectors, $\text{Input}_1, \text{Input}_2, \dots$ are different input vectors, $\text{Context}_1, \text{Context}_2, \dots$ are different contexts and \otimes represents the Kronecker product between vectors. When both an input, say Input_2 , and a context, say, Context_1 are presented to the memory, the output will be Output_3 .

The core of the model depicted in Fig. 6 is a context-dependent matrix memory consisting of four contexts (one for

each topic) and five concepts in each topic that are associated both as inputs and as outputs. All concepts of the same topic are associated with each other regardless of context in the sense that the matrix has a term such that for every context, every concept in the input elicits a concept of the same topic. If the sum of all topic vectors is *All topics*, then the matrix

$$M_1 = \sum \text{Concepts}_{\text{topic } 1} \left(\sum \text{Input}_{\text{topic } 1} \otimes \text{All topics} \right)^T \\ + \sum \text{Concepts}_{\text{topic } 2} \left(\sum \text{Input}_{\text{topic } 2} \otimes \text{All topics} \right)^T \\ + \sum \text{Concepts}_{\text{topic } 3} \left(\sum \text{Input}_{\text{topic } 3} \otimes \text{All topics} \right)^T \\ + \sum \text{Concepts}_{\text{topic } 4} \left(\sum \text{Input}_{\text{topic } 4} \otimes \text{All topics} \right)^T$$

incorporates this part of the model.

Associations that connect topics are also stored in such a way that when a given context is active it is always possible to travel from one concept in another topic to concepts corresponding to the active topic. This is achieved by terms in the equation associating a concept in one topic to a concept in another topic, in such a way that the concepts associated to the target topic can be reached by a chain of associations.

Thus the term:

$$\text{Concept}_{2\text{topic}1} (\text{Concept}_{1\text{topic}2} \otimes \text{topic}_1)^T \\ + \text{Concept}_{3\text{topic}2} (\text{Concept}_{1\text{topic}3} \otimes \text{topic}_1)^T$$

where $\text{Concept}_{i \text{ topic } k}$ denotes a concept belonging to topic k , together with the previous matrix M_1 ensures that when the topic 1 is active starting from any concept in topic 3 there is a trajectory that goes from concept 1 in topic 3, to concept 3 in topic 2. When this reaches concept 1 in topic 2 the trajectory can jump to topic 1.

The construction of the model implies that from every concept several outputs can be reached. The particular path followed is based on a probabilistic decision using the same probability for each available association.

REFERENCES

- Anderson JA. *An Introduction to Neural Networks*. Cambridge, MA: MIT Press, 1995.
- Anderson JA. A simple neural network generating an interactive memory. *Mathematical Biosciences*, 14: 197–220, 1972.
- Andreasen NC and Grove WM. Thought, language and communication in schizophrenia: Diagnosis and prognosis. *Schizophrenia Bulletin*, 12: 348–359, 1986.
- Arbib MA (Ed), *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press, 1995.
- Baddeley A. Working Memory, Thought, and Action. In *Oxford Psychology Series*. New York, NY: Oxford University Press, 2007.
- Bassett DS, Greenfield DL, Meyer-Lindenberg A, Weinberger D, Moore SW, and Bullmore ET. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Computational Biology*, 6: e1000748, 2010.
- Berry MW and Browne M. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. Philadelphia: SIAM, 2005.
- Bienenstock E and Lehmann D. Regulated criticality in the brain? *Advances in Complex Systems*, 1: 361–384, 1998.
- Cabana Á, Valle-Lisboa JC, Elvevåg B, and Mizraji E. Detecting order–disorder transitions in discourse: Implications for schizophrenia. *Schizophrenia Research*, 131: 157–164, 2011a.
- Cabana Á, Valle-Lisboa JC, Elvevåg B, and Mizraji E. *Using Machine Learning Techniques to Study Discourse Alterations in Patients with Schizophrenia*. São Paulo, Brazil: Schizophrenia International Research Society (SIRS) South America Meeting. Abstract published in *Revista de Psiquiatria Clínica*, <http://www.hcnet.usp.br/ipq/revista/vol38/s1/index.html>; 2011b (downloaded December 18, 2012).
- Cooper LN. A possible organization of animal memory and learning. In *Proceedings of the Nobel Symposium on Collective Properties of Physical Systems*. New York: Academic Press, 1973.
- Cooper LN. Memories and memory: A physicist's approach to the brain. *International Journal of Modern Physics A*, 15: 4069–4082, 2000.
- Deerwester S, Dumais S, Furnas G, Landauer T, and Harshman R. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41: 391–407, 1990.
- Doyle AC. A study in scarlet. In *The Complete Sherlock Holmes by Sir Arthur Conan Doyle (Collector's Library Editions)*. London: CRW Publishing Limited, 2005 [First published by London: Ward Lock & Co, 1887].
- Dumais S. Data-driven approaches to information access. *Cognitive Science*, 27(3): 491–524, 2003.
- Dumais S. Improving the retrieval of information from external sources. *Behavior Research Methods: Instruments and Computers*, 23: 229–236, 1991.
- DeLisi LE. Speech disorder in schizophrenia: Review of the literature and exploration of its relation to the uniquely human capacity for language. *Schizophrenia Bulletin*, 27(3): 481–496, 2001.
- Elman JL, Bates EA, Johnson MH, Karmiloff-Smith A, Parisi D, and Plunkett K. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: The MIT Press, 1997.
- Elman J. Finding structure in time. *Cognitive Science*, 14: 179–211, 1990.
- Elvevåg B, Foltz PW, Rosenstein M, and DeLisi LE. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics*, 23: 270–284, 2010.
- Elvevåg B, Foltz PW, Weinberger DR, and Goldberg TE. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1–3): 304–316, 2007.
- Foltz PW, Kintsch W, and Landauer TK. The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25: 285–307, 1998.
- Friston KJ. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2: 56–78, 1995.
- Friston KJ. Functional and effective connectivity: A review. *Brain Connectivity*, 1(1): 13–36, 2011.
- Gathercole SE and Baddeley AD. Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory & Language*, 22: 103–127, 1990.
- Griffiths TL, Steyvers M, and Tenenbaum JB. Topics in semantic representation. *Psychological Review*, 114(2): 211–244, 2007.
- Griffiths TL and Steyvers M. *A Probabilistic Approach to Semantic Representation*. Proceedings of the 24th Annual Meeting of the Cognitive Science Society.
- Grünig A. Elman backpropagation as reinforcement for simple recurrent networks. *Neural Computation*, 19(11): 3108–3131, 2007.
- Hoffman P, Rogers T, and Lambon Ralph MA. Semantic diversity accounts for the “missing” word frequency effect in stroke aphasia: Insights using a novel method to quantify contextual variability in meaning. *Journal of Cognitive Neuroscience*, 23(9): 2432–2446, 2011.
- Hoffman RE and McGlashan TH. Synaptic elimination, neurodevelopment, and the mechanism of hallucinated

- "Voices" in Schizophrenia. *American Journal of Psychiatry*, 154: 1683–1689, 1997.
- Hoffman RE, Rapaport J, Ameli R, McGlashan TH, Harcherik D, and Servan-Schreiber D. A neural network simulation of hallucinated "voices" and associated speech perception impairments in schizophrenia patients. *Journal of Cognitive Neuroscience*, 7: 479–497, 1995.
- Hofmann T. *Probabilistic Latent Semantic Analysis*. Proceedings of Uncertainty in Artificial Intelligence, UAI'99.
- Hofmann T and Puzicha J. Latent class models for collaborative filtering. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*: 688–693.
- Humphreys MS, Bain LD, and Pike R. Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, 96: 208–233, 1989.
- Izhikevich E. *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Cambridge, MA: The MIT Press, 2006.
- Jones M, Kintsch W, and Mewhort D. High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55(4): 534–552, 2006.
- Kintsch W. Predication. *Cognitive Science*, 25: 173–202, 2001.
- Koch C and Poggio T. *Single Neuron Computation*. Cambridge, MA: Academic Press, 1992.
- Kohonen T. Correlation matrix memories. *IEEE Transactions on Computers*, C-21: 353–359, 1972.
- Kohonen T. *Associative Memory: A System-Theoretical Approach*. New York: Springer-Verlag, 1977.
- Kohonen T, Lehtio P, Rovamo J, Hyvärinen J, Bry K, and Vainio L. A principle of neural associative memory. *Neuroscience*, 2(6): 1065–1076, 1977.
- Landauer TK, Foltz P, and Laham D. An introduction to latent semantic analysis. *Discourse Processes*, 25: 259–284, 1998.
- Landauer T and Dumais S. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104: 211–240, 1997.
- Landauer TK, McNamara DS, Dennis S, and Kintsch W (Eds), *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, 2007.
- Lee L, Friston K, and Horwitz B. Large-scale neural models and dynamic causal modelling. *NeuroImage*, 30(4): 1243–1254, 2006.
- Levi-Montalcini R. *Praise of Imperfection: My Life and Work*. New York: Basic Books, 1989.
- McClelland JL, Rumelhart DE, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. In *Psychological and Biological Models*. Cambridge, MA: MIT Press, 1986.
- McCulloch WS and Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5: 115–133, 1943.
- McKenna P and Oh T. *Schizophrenic Speech: Making Sense of Bathrooms and Ponds that Fall in Doorways*. Cambridge, UK: Cambridge University Press, 2005.
- Mizraji E, Pomi A, and Alvarez F. Multiplicative contexts in associative memories. *BioSystems*, 32: 145–161, 1994.
- Mizraji E. Neural memories and search engines. *International Journal of General Systems*, 37(6): 715–732, 2008.
- Mizraji E. Context-dependent associations in linear distributed memories. *Bulletin of Mathematical Biology*, 51: 195–205, 1989.
- Mizraji E and Lin J. Logic in a dynamic brain. *Bulletin of Mathematical Biology*, 73: 373–397, 2011.
- Mizraji E, Pomi A, and Valle-Lisboa JC. Dynamic searching in the brain. *Cognitive Neurodynamics*, 3: 401–414, 2009.
- Nijinsky V. In Acocella JR (Ed), *The Diary of Vaslav Nijinsky (Unexpurgated)*. New York: Farrar, Straus, and Giroux, 1999.
- Oja E. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15: 267–273, 1982.
- Ostwald P. *Vaslav Nijinsky: A Leap into Madness*. London: Robson Books Ltd, 1991.
- Papadimitriou C, Raghavan P, Tamaki H, and Vempala S. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2): 217–235, 2000.
- Peña JL and Konishi M. Auditory spatial receptive fields created by multiplication. *Science*, 292(5515): 249–252, 2001.
- Poirazi P, Brannon T, and Mel BW. Pyramidal neuron as two-layer neural network. *Neuron*, 37(6): 989–999, 2003.
- Pomi A and Mizraji E. Memories in context. *Biosystems*, 50: 173–188, 1999.
- Pomi A and Mizraji E. Semantic graphs and associative memories. *Physical Review E*, 70(6): 066136, 2004.
- Pomi A and Olivera F. Context-sensitive autoassociative memories as expert systems in medical diagnosis. *BMC Medical Informatics and Decision Making*, 6: 39, 2006.
- Pulvermüller F. Brain-language research: Where is the progress? *Biolinguistics*, 4(2–3): 255–288, 2010.
- Pulvermüller F. Meaning and the brain: The neurosemantics of referential, interactive, and combinatorial knowledge. *Journal of Neurolinguistics*, 25: 423–459, 2012.
- Rogers TT and McClelland JL. *Semantic Cognition. A Parallel Distributed Processing Approach*. Cambridge, MA: The MIT Press, 2004.
- Roll M, Mårtensson F, Sikström S, Apt P, Arnlind-Bååth R, and Horne M. Atypical associations to abstract words in Broca's aphasia. *Cortex*, 48(8): 1068–1072, 2011.
- Rubinov M and Sporns O. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52: 1059–1069, 2010.
- Rumelhart DE, Hinton GE, and Williams RJ. Learning representations by back-propagating errors. *Nature*, 323(6088): 533–536, 1986a.
- Rumelhart DE, McClelland JL, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. In *Foundations*. Cambridge, MA: MIT Press, 1986b.
- Salton G and Lesk ME. The SMART automatic document retrieval systems – An illustration. *Communications of the ACM*, 8: 391–398, 1965.
- Samaria F and Harter A. *Parameterisation of a Stochastic Model for Human Face Identification*. Sarasota (Florida): 2nd IEEE Workshop on Applications of Computer Vision, 1994.
- Smolensky P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46: 159–216, 1990.
- Snowdon DA, Kemper SJ, Mortimer JA, Greiner LH, Wekstein DR, and Markesbery WR. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. *The Journal of the American Medical Association*, 275(7): 528–532, 1996.
- Sporns O. *Networks of the Brain*. Cambridge: The MIT Press, 2010.
- Tagliazucchi E, Balenzuela P, Fraiman D, Montoya P, and Chialvo DR. Spontaneous BOLD event triggered averages for estimating functional connectivity at resting state. *Neuroscience Letters*, 488(2): 158–163, 2011.
- Valle-Lisboa JC. *Las redes neuronales y el procesamiento del lenguaje natural*. PhD Thesis. Montevideo, Uruguay: PEDECIBA-Universidad de la República, 2007.
- Valle-Lisboa JC and Mizraji E. The uncovering of hidden structures by latent semantic analysis. *Information Sciences*, 177: 4122–4147, 2007.
- Valle-Lisboa JC, Reali F, Anastasia H, and Mizraji E. Elman topology with sigma-pi units: An application to the modeling of verbal hallucinations in schizophrenia. *Neural Networks*, 18: 863–877, 2005.
- Xiang H, Fonteijn HM, Norris DG, and Hagoort P. Topographical functional connectivity pattern in the perisylvian language networks. *Cerebral Cortex*, 20: 549–560, 2010.

Parte III
Apéndices

Apéndice A

Redes neuronales

A.1 Derivación de la regla delta

Siguiendo a [Bechtel y Abrahamsen \(2002\)](#); [Bose y Liang \(1996\)](#), repasaremos la obtención de la expresión para el cambio de los pesos sinápticos según el algoritmo de gradiente descendiente ideado por [Widrow y Hoff \(1960\)](#), también conocido como *regla delta*.

En los algoritmos de gradiente descendiente, el objetivo del entrenamiento es minimizar el error en la salida. Una expresión estándar para este tipo de errores es el error cuadrático medio:

$$E(W) = \frac{1}{2N} \sum_k (y^k - g^k)^2 \quad (\text{A.1})$$

donde y^k es la salida deseada de la red, y g^k son los valores de actividad de salida reales.

Siguiendo la estrategia de gradiente descendiente, necesitamos conocer el cambio en el error para todos los pesos sinápticos:

$$\nabla E = \frac{\partial E}{\partial W}$$

Muchas veces en lugar de esta expresión, se trabaja con el error cuadrático para cada ítem k del conjunto de entrenamiento:

$$e^k = \frac{1}{2} (y^k - g^k)^2 \quad (\text{A.2})$$

y su gradiente:

$$\nabla e^k = \frac{\partial e_k}{\partial W} = \left\{ \frac{\partial e_{j'}^k}{\partial w_{ij}} \right\}_{j', i \times j} \quad (\text{A.3})$$

donde $e_{j'}^k$ es la componente j' del vector error (e^k). Las derivadas parciales $\frac{\partial e_{j'}^k}{\partial w_{ij}}$ sólo tienen componente no nula cuando el índice $j = j'$, lo que nos permite

escribir:

$$\nabla e^k = \left\{ \frac{\partial e_j^k}{\partial w_{ij}} \right\}_{i,j} \quad (\text{A.4})$$

Aplicando la regla de la cadena puede obtenerse el valor de este gradiente:

$$\frac{\partial e_j^k}{\partial w_{ij}} = \frac{\partial e_j^k}{\partial g_j^k} \frac{\partial g_j^k}{\partial w_{ij}} \quad (\text{A.5})$$

La derivada parcial del error con respecto a la j -ésima actividad de salida g_j^k es:

$$\frac{\partial e_j^k}{\partial g_j^k} = -(y_j^k - g_j^k) \quad (\text{A.6})$$

Nótese que las siguientes derivadas se anulan:

$$\frac{\partial e_{j'}^k}{\partial g_j^k} = 0 \quad \text{si} \quad j' \neq j$$

Como las neuronas tienen activación lineal, es decir:

$$g_j^k = \sum_i w_{ij} f_i^k$$

entonces se tiene que:

$$\frac{\partial g_j^k}{\partial w_{ij}} = f_i^k \quad (\text{A.7})$$

Sustituyendo A.7 y A.6 en la expresión A.5, obtenemos:

$$\frac{\partial e_j^k}{\partial w_{ij}} = (y_j^k - g_j^k) f_i^k \quad (\text{A.8})$$

Entonces la matriz gradiente del error está dada por el producto externo:

$$\nabla e^k = (y^k - g^k) f^{kT} \quad (\text{A.9})$$

Según el método de gradiente descendiente, el cambio en la matriz de pesos W es:

$$\Delta W^k = -\eta \nabla e^k = -\eta (y^k - g^k) f^{kT} \quad (\text{A.10})$$

Apéndice B

Entropías temáticas

B.1 Máximo de entropía temática

B.1.1 Definición

La entropía de un tema α puede definirse como:

$$S(\alpha) = - \sum_{\alpha_i}^{N_\alpha} p(\alpha_i) \log p(\alpha_i) \quad (\text{B.1})$$

donde α_i son los bloques contiguos de texto pertenecientes al tema α , $p(\alpha_i)$ es la fracción del tamaño de cada tema que tienen los bloques, y N_α es el número de bloques en que está segmentado el tema.

La entropía de un texto S es la suma de las entropías de cada tema:

$$S(\mathcal{D}) = - \sum_{\alpha} S(\alpha). \quad (\text{B.2})$$

B.1.2 Caso 2 temas

En el caso de que sólo hayan dos temas, α y β , supongamos que uno está constituido por más palabras que el otro:

$$n_\alpha > n_\beta$$

Una primera aproximación a la partición que maximiza la entropía temática, puede ser particionar totalmente los segmentos del tema más pequeño por el más grande:



En tal caso, la entropía de β será la máxima posible:

$$S^{max}(\beta) = \log n_\beta. \tag{B.3}$$

Sin embargo, la entropía de α no es máxima, ya que está lejos de la equipartición.

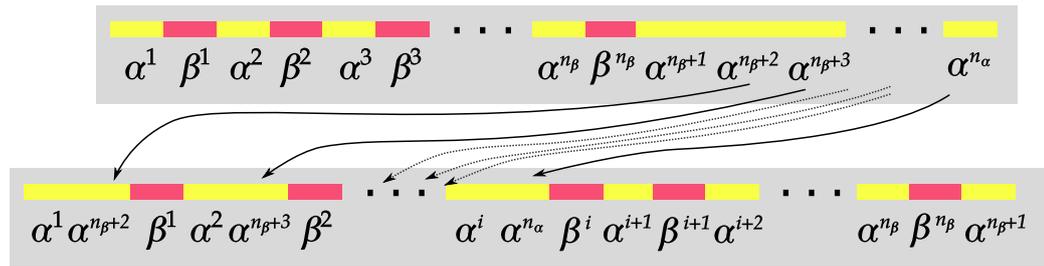
Una aproximación más refinada sería intercalar duplas de α entre los bloques de β , a modo de reducir el tamaño del bloque de α que quedó al final. En particular, si asumimos a modo ilustrativo que se cumple que:

$$n_\alpha < 2(n_\beta + 1)$$

podemos hacer esto de modo de acabar con todo el bloque final de α , formando secuencias $\alpha\alpha\beta$ de la siguiente manera:

$$\alpha\alpha\beta\alpha\alpha\beta \dots \alpha\alpha\beta\alpha\beta\alpha\beta\alpha$$

o más en detalle ...



por ejemplo.

De esta forma, la entropía del tema α ahora es máxima, ya que está fragmentada en el mayor número posible de fragmentos ($n_\beta + 1$). El número total de bloques (s) es igual a la suma de las duplas (r bloques de largo 2, $\alpha\alpha$) y las palabras aisladas (u bloques de largo 1):

$$s = n_\beta + 1 = r + u$$

y por tanto

$$u = s - r$$

Como las duplas se generan con el remanente de palabras α que exceden los posibles bloques unitarios $n_\beta + 1$, el número de duplas puede calcularse como:

$$r = \text{resto}\{n_\alpha / (n_\beta + 1)\}$$

Para calcular el valor de entropía máxima, se pueden sumar las contribuciones de las duplas $\alpha\alpha$ ($S_r(\alpha)$) y las palabras en solitario ($S_u(\alpha)$):

$$S(\alpha) = S_r(\alpha) + S_u(\alpha) \quad (\text{B.4})$$

$$\begin{aligned} &= -\sum^r \frac{2}{n_\alpha} \log \frac{2}{n_\alpha} - \sum^u \frac{1}{n_\alpha} \log \frac{1}{n_\alpha} \\ &= \frac{2r}{n_\alpha} \log \frac{n_\alpha}{2} + \frac{s-r}{n_\alpha} \log n_\alpha \\ &= \frac{(s+r)}{n_\alpha} S^{max}(\alpha) - \frac{2r}{n_\alpha} \log 2 \end{aligned} \quad (\text{B.5})$$

B.1.3 Caso general

Si se consideran más de 2 temas (N_T temas), el de mayor tamaño podrá “particionar” perfectamente a todos los demás en bloques de largo 1, que tendrán entropía máxima según la ecuación B.3. Para el tema de mayor tamaño (llamémosle α), es necesario considerar las siguientes cantidades. Por un lado, el número máximo de bloques posibles generados por los temas menores:

$$s = 1 + \sum_{\gamma \neq \alpha}^{N_T} n_\gamma \quad (\text{B.6})$$

Si el tamaño del tema mayor es menor o igual a la suma de todos los otros temas ($s \geq n_\alpha$), entonces éste puede ser equiparticionado totalmente por todos los demás temas, alcanzando así su entropía máxima (ecuación B.3).

Si por el contrario, tenemos la situación opuesta, es decir que:

$$s < n_\alpha$$

el tema no puede ser equiparticionado totalmente por los demás, y debemos considerar la emergencia de “bloques” de temas de largo mayor a 1. Entonces, tenemos que el tamaño máximo de bloque al que se llega en caso de una partición del tema por parte de los otros es:

$$c = \left\lceil \frac{n_\alpha}{s} \right\rceil$$

y por último, el número de estos bloques de tamaño máximo:

$$r = \text{resto}\{n_\alpha/s\}$$

El resto de los bloques ($u = s - r$) tendrá un tamaño de $c - 1$, de manera similar a lo que se ilustra para el caso de dos temas, donde hay r bloques de dos palabras y $u = s - r$ bloques de una palabra.

Es importante notar, que si este resto es igual a cero, quiere decir que el largo del tema (n_α) es múltiplo entero del número de particiones (s), y por

lo tanto, *todos* los bloques son de tamaño máximo c . Por tanto, en ese caso, debe elegirse $r = s$, y quedará entonces $u = 0$, es decir, ningún bloque de largo $c - 1$. Una expresión más general para r es, por tanto, la siguiente:

$$r = \begin{cases} \text{resto}\{n_\alpha/s\} & \text{si } \text{resto}\{n_\alpha/s\} > 0 \\ s & \text{si } \text{resto}\{n_\alpha/s\} = 0 \end{cases} \quad (\text{B.7})$$

Puede calcularse entonces la expresión general para la entropía máxima puede encontrarse tomando en cuenta los aportes de los dos tipos de bloques, los de tamaño $c - 1$ ($u = s - r$ bloques), y los de tamaño c (r bloques):

$$\begin{aligned} S(\alpha) &= S_r(\alpha) + S_u(\alpha) & (\text{B.8}) \\ &= -\sum^r \frac{c}{n_\alpha} \log \frac{c}{n_\alpha} - \sum^u \frac{c-1}{n_\alpha} \log \frac{c-1}{n_\alpha} \\ &= \frac{rc}{n_\alpha} \log \frac{n_\alpha}{c} + \frac{(s-r)(c-1)}{n_\alpha} \log \frac{n_\alpha}{c-1} \\ &= (a+b) \log n_\alpha - a \log c - b \log(c-1) \\ &= (a+b) S^{\max}(\alpha) - a \log c - b \log(c-1) & (\text{B.9}) \end{aligned}$$

donde

$$a = \frac{rc}{n_\alpha}$$

y

$$b = \frac{(s-r)(c-1)}{n_\alpha}$$

Consideremos entonces el caso en que $s = n_\alpha - 1$, en el que la entropía debería ser muy cercana a la máxima, ya que es posible una equipartición casi total del tema, a excepción de un bloque que quedará de largo 2. En tal caso, tenemos que por definición $c = 2, r = 1$ y $a = \frac{2}{n_\alpha}$ y $b = \frac{s-1}{n_\alpha}$, y tenemos que $a + b = \frac{s+1}{n_\alpha} = 1$, y por tanto:

$$S(\alpha) = S^{\max}(\alpha) - \frac{2}{n_\alpha} \log 2$$

y puede apreciarse claramente en esta expresión que la entropía es casi igual a la máxima, a diferencia de una “penalización” por tener 1 bloque de largo 2 palabras.

En la figura B.1 se ilustra un ejemplo en el que se aplicó la fórmula obtenida para el máximo de entropía de tema (ecuación B.9)). Se construyó un discurso artificial de 6 temas, donde un tema es de mayor tamaño que todos los otros sumados. Para este ejemplo, se calculó la entropía máxima posible, y el resultado es $S(\mathcal{D}) = 9,5684$. En la figura B.1 A se muestran dos ordenamientos posibles de los temas: la primera, una de mínima entropía, en donde todos los temas están presentes en un bloque contiguo. La segunda

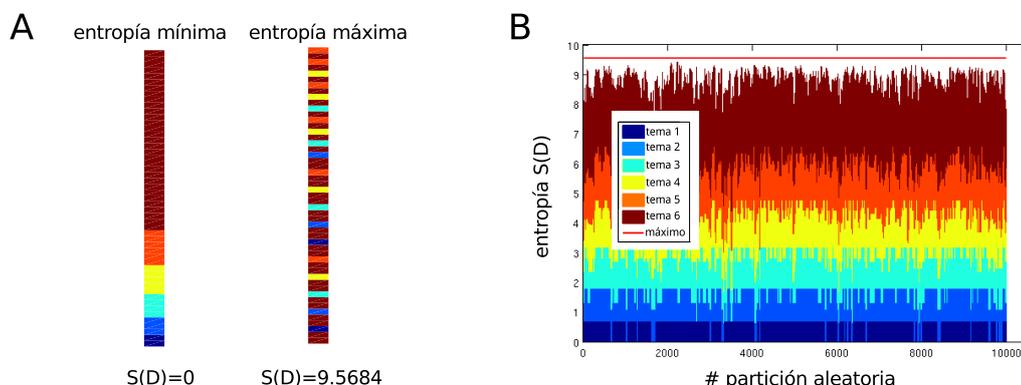


Figura B.1: Ejemplo de discurso artificial de seis temas. A. Particiones de mínima y máxima entropía de tema. B. Entropías de tema para 10000 particiones aleatorias.

es una partición de máxima entropía, construida mediante el procedimiento descrito anteriormente. Al calcular la entropía de esta partición utilizando la definición de entropía de tema (ecuación 7.3), el valor obtenido coincide con el máximo derivado en esta sección, según la ecuación B.9. En la figura 7.3 B se muestran las entropías por tema (el acumulado es la suma) correspondiente a 10000 particiones aleatorias del mismo discurso, realizadas mediante el procedimiento ilustrado en la figura 7.2 A. Los valores alcanzados por la entropía de tema nunca superan el valor de entropía máxima.

B.2 Máximo de entropía de transición

A diferencia del caso de la entropía de tema, esta medida no toma en cuenta la “longitud” de los bloques temáticos, sino la sucesión con que ocurren unos después de otros. Por tanto, la situación de la entropía de transición es más sencilla que en el caso anterior. Si existen sólo dos tópicos, la entropía de transición es siempre igual a cero. En caso de haber dos o más tópicos, casi siempre puede alcanzarse el máximo de entropía de transición en situaciones donde la entropía de tema dista mucho de ser máxima.

B.2.1 Definición

La entropía de transición de un tema α , $T(\alpha)$ se calcula mediante la expresión:

$$T(\alpha) = - \sum_{\tau \neq \alpha} p_{\alpha}(\tau) \log p_{\alpha}(\tau) \quad (\text{B.10})$$

donde $p_\alpha(\tau)$ es la probabilidad de que después del tópico *alpha* comience el tópico τ . La entropía del discurso $T(D)$ es la suma de todas las entropías de transición de los temas:

$$T(D) = \sum_{\alpha} T(\alpha) \quad (\text{B.11})$$

B.2.2 Caso 3 tópicos

En el caso particular de 3 tópicos, α , β y γ , basta con que uno de los tres tópicos esté particionado 3 veces (por ejemplo α), y los otros 2 veces, para que exista un arreglo tal que maximice las entropías de transición. Un arreglo podría ser en este ejemplo:

$$\alpha_1\beta_1\alpha_2\gamma_1\beta_2\gamma_2\alpha_3. \quad (\text{B.12})$$

Puede verificarse que en este caso que todas las entropías son máximas, ya que todas las transiciones posibles ocurren el mismo número de veces, resultando en una entropía máxima igual a:

$$\begin{aligned} T^{max}(\alpha) &= -p_\alpha(\beta) \log p_\alpha(\beta) - p_\alpha(\gamma) \log p_\alpha(\gamma) \\ &= \frac{1}{2} \log 2 + \frac{1}{2} \log 2 \\ &= \log 2 \end{aligned} \quad (\text{B.13})$$

y este valor es igual para todos los tópicos.

B.2.3 Caso general

En el caso en que hayan N_T tópicos, el mínimo de particiones necesarias para que todos los tópicos alcancen la máxima entropía posible, es de un tópico con N_T particiones, y el resto con $N_T - 1$ particiones. Cuando todas las transiciones están presentes en igual cantidad, las probabilidades de transición se vuelven $p_\alpha(\tau) = \frac{1}{N_T - 1}$ y entonces la entropía máxima se vuelve:

$$\begin{aligned} T^{max}(\alpha) &= \sum_{\tau \neq \alpha} \frac{1}{N_T - 1} \log(N_T - 1) \\ &= \log(N_T - 1) \end{aligned} \quad (\text{B.14})$$

Sin embargo, puede suceder que el tema conste de menos palabras que el número total de temas:

$$n_\alpha < N_T - 1$$

y en ese caso, sólo existen como máximo n_α transiciones a temas distintos, con probabilidades $p_\alpha(\tau) = \frac{1}{n_\alpha}$. El máximo de entropía de transición se vuelve entonces:

$$T^{max}(\alpha) = \log(n_\alpha),$$

Tomando en cuenta esta posibilidad, la expresión general queda:

$$T^{max}(\alpha) = \log \min(N_T - 1, n_\alpha). \quad (\text{B.15})$$

Apéndice C

Espacios semánticos

C.1 Obtención de un espacio semántico mediante Análisis Semántico Latente

En esta sección se describen brevemente los pasos realizados para la obtención de un espacio semántico a partir de una colección de documentos de la enciclopedia en línea *Wikipedia* en inglés, utilizando Análisis Semántico Latente. El espacio utilizado en este trabajo es el mismo obtenido y utilizado en mi tesis de maestría (Cabana, 2009): aquí reproducimos aspectos técnicos del trabajo que se encuentran en mayor detalle desplegados en la tesis.

Se utilizó el volcado (*dump*) de la enciclopedia que contiene más de 7 millones de documentos, de los cuales se muestreó un 5% aproximadamente. Después de limpiar marcas de formato (XML y Wiki), se eliminaron las palabras funcionales o “stop” en base a una lista comúnmente utilizada en la literatura (Salton y Lesk, 1965). Luego, sólo se consideraron los documentos que tuvieran más de 50 palabras de largo, y de los mismos se retuvieron sólo las primeras 200 palabras. Además, sólo se retuvieron las palabras que ocurrían en al menos dos documentos, y al menos 8 veces en todo el *corpus*. Luego de aplicar estos criterios tres veces seguidas (la remoción de palabras provoca remoción de documentos, y viceversa), se llegó a la cuenta final de 56.108 palabras (*types*) en 53.956 documentos, con un total de 4.616.285 ocurrencias (*tokens*).

A partir de estos documentos se construyó una matriz palabra–documento A , en donde cada entrada conformaba la expresión:

$$A_{ij} = \log(1 + t_{ij})(1 - H(i)) \quad (\text{C.1})$$

donde t_{ij} es la frecuencia de la palabra i en el documento j , y $H(i)$ es la entropía normalizada de la palabra i :

$$H(i) = -\frac{1}{\log_2(n)} \sum_j p_i(j) \log_2(p_i(j)), \quad (\text{C.2})$$

con $p_i(j)$ la probabilidad de que la palabra i esté en el documento j :

$$p_i(j) = \frac{t_{ij}}{\sum_j t_{ij}} \quad (\text{C.3})$$

Se realizó la descomposición en valores singulares de la matriz A , obteniéndose los primeros 500 vectores y valores singulares (véase la sección 3.2.1). Para determinar el número de dimensiones k a utilizar se evaluó el desempeño del espacio en el test de sinónimos TOEFL a diferentes valores de k , y se obtuvo un máximo de 35/55 respuestas correctas para $k = 390$ (63.6 %, un valor similar al reportado para implementaciones de referencia (Lan-dauer y Dumais, 1997)). Por tanto, el espacio se generó con las primeras 390 dimensiones de descomposición singular.

El espacio obtenido se comparó con la implementación de la Universidad de Colorado (LSA@Boulder) (Dennis, 2007), que está basada en el *corpus* TASA de textos de interés general y posee 300 dimensiones. Esta implementación ha sido ampliamente utilizada en trabajos científicos (por ejemplo), y es considerada de referencia (ver por ejemplo: (Dumais, 2003; Elvevåg *et al.*, 2007; Hare *et al.*, 2009)). La comparación mostró que ambos espacios asignaban distancias muy similares entre sí, siendo la media de las diferencias muy cercana a cero, y el desvío estándar alrededor de 0.1 (Cabana, 2009). Puede concluirse entonces que el espacio obtenido funciona en forma similar a una implementación de referencia.

C.1.1 *Corpus Wikipedia en español*

Para obtener el *corpus Wikipedia* en español utilizado en la sección 6.3 se utilizó un procedimiento similar al descrito en la sección anterior. Las diferencias se señalan a continuación.

En primer lugar, no se removieron las palabras “stop”, ni los signos de puntuación. De cada documento, se retuvieron la oraciones que no contenían dígitos ni símbolos extraños. Esto resultó en un total de 39709 oraciones pertenecientes a 20217 documentos en las que ocurren 20287 palabras un total de 374361 veces.

Al procesar este *corpus* para obtener un espacio mediante LSA, el número total de palabras se redujo a 16317. Esto se debe en parte debido a la remoción de las palabras “stop”, y al hecho de que se eliminaron todas las palabras que aparecieron menos de 8 veces en total y en menos de 4 documentos.

C.2 Búsquedas en un espacio semántico

El Análisis Semántico Latente (LSA) prescribe una manera para obtener un espacio semántico a partir de una matriz de ocurrencia de palabras y documentos A . En particular, implica el cálculo de la descomposición en valores

singulares de esta matriz, y truncar la aproximación en número de valores k menor al rango de la matriz, obteniéndose una matriz aproximada A_k :

$$A_k = \sum_k u_k \sigma_k v_k^T = U_k \sigma_k v_k^T \quad (\text{C.4})$$

La matriz A (y su aproximada A_k) corresponde una transformación lineal entre un espacio de palabras y un espacio de documentos. En particular, la imagen de A_k es igual a la imagen de U_k , mientras que la imagen de V_k es la imagen de A_k^T . Si se quiere evaluar la proyección de una combinación de palabras sobre el espacio de documentos (la imagen de A_k), la matriz $U_k U_k^T$ es la matriz que da las coordenadas de la proyección.

Por tanto, una forma posible de comparar dos “pseudo-documentos” (oraciones o palabras, por ejemplo), es construir dos vectores de búsqueda, q_1 y q_2 , pertenecientes al espacio de palabras (una coordenada por palabra). Las componentes de estos vectores son no nulas sólo para las palabras presentes en el fragmento que se quiera representar. El coseno del ángulo $\phi_{1,2}$ entre las proyecciones de q_1 y q_2 , \tilde{q}_1 y \tilde{q}_2 , puede calcularse como:

$$\begin{aligned} \cos(\phi_{1,2}) &= \frac{q_1^T U_k U_k^T U_k U_k^T q_2}{\|q_1^T U_k U_k^T\| \|U_k U_k^T q_2\|} \\ &= \frac{q_1^T U_k U_k^T q_2}{\|\tilde{q}_1\| \|\tilde{q}_2\|} \end{aligned} \quad (\text{C.5})$$

C.3 Generación de discursos artificiales

Para generar discursos artificiales usando un espacio semántico, se diseñó el siguiente procedimiento. Primero se definen dos palabras, que van a officiar de inicio y objetivo del discurso. El procedimiento consiste en calcular para la palabra de inicio, las 6 palabras más cercanas en el espacio, y se generan 6 discursos candidatos agregando estas palabras a la palabra inicial. Luego se calculan las distancias de los discursos generados con la palabra objetivo, y se elige el más cercano. La palabra que originó el discurso “ganador”, es utilizada como semilla junto a la que le precede (la inicial en este caso), y se vuelve a repetir el procedimiento.

En cada iteración, se elige el discurso más cercano a la palabra objetivo, y se utilizan como nueva semilla para obtener palabras cercanas las tres últimas palabras del discurso generado hasta el momento. El procedimiento se detiene una vez que se genera un discurso que está a una distancia menor a un umbral predefinido de la palabra objetivo. En el caso del discurso generado en la figura 7.7, el umbral fue de 0.18, y las palabras inicial y final fueron “science” y “team”. Puede apreciarse en esa figura que el discurso generado aprovecha la asociación entre “science” y “fiction”, la de “fiction” con “story” y “play”, y la relación de “play” con palabras relacionadas al deporte.

El siguiente pseudocódigo ilustra algunos detalles del procedimiento aquí descrito:

```
· origen ←  
· destino ←  
· discurso ← origen  
· while dist(discurso,destino)<umbral  
  · semilla ← últimas 3 palabras de discurso  
  · vecinos ← 6 palabras más cercanas a semilla  
  · foreach vecino  
    · discursoCandidato ← discurso+vecino  
    · distancia ← dist(discursoCandiato,destino)  
    · discurso ← discurso+vecino[min distancia]  
· discurso ← discurso + destino
```

Tabla C.1: Pseudocódigo del algoritmo para la generación de discursos artificiales.

Bibliografía

- Aggarwal, Charu C y Zhai, ChengXiang (2012). «A survey of text classification algorithms». En: *Mining text data*, pp. 163–222. Springer.
- Alexander, Michael P (1997). «Aphasia: clinical and anatomic aspects». *Behavioral neurology and neuropsychology*, pp. 133–149.
- Amari, S.-I. (1972). «Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements». *IEEE Transactions on Computers*, **C-21**, pp. 1197–1206.
- American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders*. 3ª edición.
- Anderson, J.A. (1972). «A Simple Neural Network Generating an Interactive Memory». *Mathematical Biosciences*, **14**, pp. 197–220.
- Andreasen, Nancy C (1979). «Thought, language, and communication disorders: I. Clinical assessment, definition of terms, and evaluation of their reliability». *Archives of general Psychiatry*, **36(12)**, pp. 1315–1321.
- Andreasen, Nancy C (1986). «Scale for the assessment of thought, language, and communication (TLC).» *Schizophrenia Bulletin*, **12(3)**, p. 473.
- Andreasen, Nancy C. (2011). *Concept of schizophrenia: past, present, and future*. pp. 3–8. En: [Weinberger y Harrison \(2011\)](#), 3ª edición.
- Andreasen, Nancy C y Grove, William M (1986). «Thought, language, and communication in schizophrenia: Diagnosis and prognosis.» *Schizophrenia Bulletin*, **12(3)**, p. 348.
- Apté, Chidanand; Damerau, Fred y Weiss, Sholom M. (1994). «Automated learning of decision rules for text categorization». *ACM Trans. Inf. Syst.*, **12**, pp. 233–251. ISSN 1046-8188. doi: 10.1145/183422.183423.
- Arango, Celso y Carpenter, William T. (2011). *The schizophrenia construct: symptomatic presentation*. pp. 9–23. En: [Weinberger y Harrison \(2011\)](#), 3ª edición.

- Baronchelli, Andrea; Ferrer i Cancho, Ramon; Pastor-Satorras, Romualdo; Chater, Nick y Christiansen, Morten H (2013). «Networks in cognitive science». *Trends in cognitive sciences*, **17(7)**, pp. 348–360.
- Barsalou, Lawrence W; Santos, Ava; Simmons, W Kyle y Wilson, Christine D (2008). «Language and simulation in conceptual processing». *Symbols, embodiment, and meaning*, pp. 245–283.
- Bechtel, W. y Abrahamsen, A. (2002). *Connectionism and the Mind*. Blackwell Publishing, 2ª edición.
- Blei, David M (2012). «Probabilistic topic models». *Communications of the ACM*, **55(4)**, pp. 77–84.
- Blei, D.M.; Ng, A.Y. y Jordan, M.I. (2003). «Latent Dirichlet Allocation». *Journal of Machine Learning Research*, **3**, pp. 993–1022.
- Blouw, Peter y Eliasmith, Chris (2013). «A Neurally Plausible Encoding of Word Order Information into a Semantic Vector Space». En: *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, .
- Borensztajn, Gideon; Zuidema, Willem y Bechtel, William (2014). *7 Systematicity and the Need for Encapsulated Representations*. p. 165. MIT Press.
- Bose, N.K. y Liang, P. (1996). *Neural Network Fundamentals with Graphs, Algorithms and Applications*. McGraw-Hill.
- Botvinick, Matthew y Plaut, David C (2004). «Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action.» *Psychological review*, **111(2)**, p. 395.
- Cabana, Á. (2009). *Representación de la estructura del lenguaje escrito mediante Grafos y Espacios semánticos*. Tesina o Proyecto, PEDECIBA, Universidad de la República.
- Cabana, Álvaro; Valle-Lisboa, Juan C; Elvevåg, Brita y Mizraji, Eduardo (2011). «Detecting order–disorder transitions in discourse: Implications for schizophrenia». *Schizophrenia research*, **131(1)**, pp. 157–164.
- Cameron, Norman (1938). «Reasoning, regression and communication in schizophrenics.» *Psychological Monographs*, **50(1)**, p. i.
- Cannon, Tyrone D. (2009). *Prediction of Psychosis Through the Prodromal Syndrome*. pp. 251–266. En: [Gattaz y Busatto \(2009\)](#).
- Choi, Freddy YY (2000). «Advances in domain independent linear text segmentation». En: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pp. 26–33. Association for Computational Linguistics.

- Chomsky, Noam (1993). *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.
- Christiansen, Morten y Chater, Nick (1994). «Generalization and connectionist language learning». *Mind and Language*, **9**, pp. 273–287.
- Cohen, Jonathan D y Servan-Schreiber, David (1992). «Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia.» *Psychological review*, **99(1)**, p. 45.
- Coltheart, Max (2004). «Brain imaging, connectionism, and cognitive neuropsychology». *Cognitive neuropsychology*, **21(1)**, pp. 21–25.
- Coltheart, Max (2006). «Dual route and connectionist models of reading: An overview». *London Review of Education*, **4(1)**, pp. 5–17.
- Coltheart, Max; Rastle, Kathleen; Perry, Conrad; Langdon, Robyn y Ziegler, Johannes (2001). «DRC: a dual route cascaded model of visual word recognition and reading aloud.» *Psychological review*, **108(1)**, p. 204.
- Cuetos, Fernando; Glez-Nosti, Maria; Barbón, Analía y Brysbaert, Marc (2011). «SUBTLEX-ESP: Spanish word frequencies based on film subtitles.» *Psicologica: International Journal of Methodology and Experimental Psychology*, **32(2)**, pp. 133–143.
- De Deyne, Simon; Navarro, Daniel J y Storms, Gert (2013). «Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations». *Behavior research methods*, **45(2)**, pp. 480–498.
- de Saussure, F. (1990). *Curso de lingüística general*.
- Deerwester, S.C.; Dumais, S.T.; Landauer, T.K.; Furnas, G.W. y Harshman, R.A. (1990). «Indexing by Latent Semantic Analysis». *Journal of the American Society of Information Science*, **41(6)**, pp. 391–407.
- DeLisi, Lynn E (2001). «Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language.» *Schizophrenia Bulletin*, **27(3)**, p. 481.
- Dennis, S. (2007). *How to use the LSA website*. pp. 57–70. En: [Landauer et al. \(2007\)](#).
- Dennis, S. y Harrington, M. (2001). «The Syntagmatic Paradigmatic Model: An distributed instance-based model of sentence processing». En: *The Second Workshop on Natural Language Processing and Neural Networks*, .

- Dominey, P (2003). «Neurological basis of language and sequential cognition: Evidence from simulation, aphasia, and ERP studies». *Brain and Language*, **86(2)**, pp. 207–225. ISSN 0093934X. doi: 10.1016/S0093-934X(02)00529-1.
- Doxas, I.; Dennis, S. y Oliver, W. (2007). «The dimensionality of language.» En: *Proceedings of the 29th Conference of the Cognitive Science Society*, .
- Dronkers, N.; Pinker, S. y Damasio, A.R. (2000). *Language and the aphasias*. pp. 1170–1188. McGraw-Hill Medical.
- Dumais, S. (2003). «Data-driven approaches to information access». *Cognitive Science*, **27(3)**, pp. 491–524.
- Eliasmith, Chris y Anderson, Charles H (2004). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT Press.
- Elman, Jeffrey L. (1990). «Finding Structure in Time». *Cognitive Science*, **14(2)**, pp. 179–211.
- Elman, Jeffrey L (1993). «Learning and development in neural networks: the importance of starting small.» *Cognition*, **48(1)**, pp. 71–99. ISSN 0010-0277.
- Elman, Jeffrey L (2009). «On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon.» *Cognitive science*, **33(4)**, pp. 547–582. ISSN 1551-6709. doi: 10.1111/j.1551-6709.2009.01023.x.
- Elman, J.L. (1991). «Distributed Representations, Simple Recurrent Networks, and Grammatical Structure». *Machine Learning*, **7**, pp. 195–225.
- Elman, J.L.; Bates, E.A.; Johnson, M.H.; Karmiloff-Smith, A.; Parisi, D. y Plunkett, K. (1997). *Rethinking Innateness: A Connectionist Perspective on Development (Neural Networks and Connectionist Modeling)*. The MIT Press.
- Elvevåg, B.; Foltz, P.W.; Rosenstein, M. y Delisi, L.E. (2010). «An automated method to analyze language use in patients with schizophrenia and their first-degree relatives». *Journal of Neurolinguistics*, **23(3)**, pp. 270–284. doi: 10.1016/j.jneuroling.2009.05.002.
- Elvevåg, B.; Foltz, P.W.; Weinberger, D.R. y Goldberg, T.E. (2007). «Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia». *Schizophrenia Research*, **93(1-3)**, pp. 304–316.
- Ferrer i Cancho, Ramón y Solé, Ricard V (2003). «Least effort and the origins of scaling in human language». *Proceedings of the National Academy of Sciences*, **100(3)**, pp. 788–791.
- Fodor, Jerry A y Pylyshyn, Zenon W (1988). «Connectionism and cognitive architecture: A critical analysis». *Cognition*, **28(1)**, pp. 3–71.

- Foltz, F.W. (2007). *Discourse coherence and LSA*. pp. 167–184. En: [Landauer et al. \(2007\)](#).
- Foltz, P.W.; Kintsch, W. y Landauer, T.K. (1998). «The measurement of textual coherence with Latent Semantic Analysis». *Discourse Processes*, **25(2 & 3)**, pp. 285–307.
- Frith, C.D. (1995). *La esquizofrenia. Un Enfoque psicológico*. Ariel Psicología.
- Gattaz, Wagner F. y Busatto, Geraldo (Eds.) (2009). *Advances in Schizophrenia Research*. Springer.
- Gradin, Victoria B y Pomi, Andrés (2008). «The role of hippocampal atrophy in depression: a neurocomputational approach». *Journal of biological physics*, **34(1-2)**, pp. 107–120.
- Griffiths, Thomas L. y Steyvers, Mark (2011). «A probabilistic approach to semantic representation». En: *Proceedings of the 33rd annual conference of the cognitive science society*, pp. 3444–3449.
- Griffiths, Thomas L; Steyvers, Mark y Tenenbaum, Joshua B (2007). «Topics in semantic representation.» *Psychological review*, **114(2)**, p. 211.
- Hare, Mary; Elman, Jeffrey L; Tabaczynski, Tracy y McRae, Ken (2009). «The wind chilled the spectators, but the wine just chilled: Sense, structure, and sentence comprehension.» *Cognitive science*, **33(4)**, pp. 610–628. ISSN 1551-6709. doi: 10.1111/j.1551-6709.2009.01027.x.
- Harm, Michael W y Seidenberg, Mark S (1999). «Phonology, reading acquisition, and dyslexia: insights from connectionist models.» *Psychological review*, **106(3)**, p. 491.
- Harm, Michael W y Seidenberg, Mark S (2004). «Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes.» *Psychological review*, **111(3)**, p. 662.
- Harris, Z.S. (1968). *Mathematical Structures of Language*. Wiley, New York, NY, USA.
- Harrow, Martin y Quinlan, Donald M (1985). *Disordered thinking and schizophrenic psychopathology*. Gardner press.
- Hayes, Philip J. y Weinstein, Steven P. (1991). «CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories». En: *Proceedings of the The Second Conference on Innovative Applications of Artificial Intelligence, IAAI '90*, pp. 49–64. ISBN 0-262-68068-8.
- Hebb, Donald Olding (1949). *The organization of behavior: A neuropsychological theory*. Psychology Press.

- Hersh, W.; Buckley, C.; Leone, T. J. y Hickam, D. (1994). «OHSUMED: an interactive retrieval evaluation and new large test collection for research.» En: *Proceedings of the 17th Annual ACM SIGIR Conference.*, pp. 129–2001.
- Hoffman, Ralph E y Dobscha, Steven K (1989). «Cortical pruning and the development of schizophrenia: a computer model.» *Schizophrenia bulletin*, **15(3)**, p. 477.
- Hoffman, Ralph E y McGlashan, Thomas H (1997). «Synaptic elimination, neurodevelopment, and the mechanism of hallucinated “voices” in schizophrenia.» *American Journal of Psychiatry*, **154(12)**, pp. 1683–1689.
- Hoffman, Ralph E; Rapaport, Jill; Ameli, Rezvan; McGlashan, Thomas H; Harcherik, Diane y Servan-Schreiber, David (1995). «A neural network simulation of hallucinated “voices” and associated speech perception impairments in schizophrenic patients.» *Journal of cognitive neuroscience*, **7(4)**, pp. 479–496.
- Hoffman, R.E.; Grasemann, U.; Gueorguieva, R.; Quinlan, D.; Lane, D. y Mikkulainen, R. (2011). «Using computational patients to evaluate illness mechanisms in schizophrenia.» *Biological Psychiatry*, **69(10)**, pp. 997–1005.
- Hoffman, R.E. y McGlashan, T.H. (2001). «Neural network models of schizophrenia.» *Neuroscientist*, **7(5)**, pp. 441–454.
- Hofmann, T. (1999). «Probabilistic Latent Semantic Analysis.» En: *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, .
- Hollis, Chris y Rapoport, Judith (2011). *Child and adolescent schizophrenia*. pp. 24–46. En: [Weinberger y Harrison \(2011\)](#), 3ª edición.
- Ingram, John CL (2007). *Neurolinguistics: An introduction to spoken language processing and its disorders*. Cambridge University Press.
- Jackendoff, Ray (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.
- Jackendoff, Ray (2009). «Natural Logic of Morals and of Laws, The.» *Brook. L. Rev.*, **75**, p. 383.
- Janicak, Philip G.; Marder, Stephen R.; Tandon, Rajiv y Goldman, Morris (2014a). *Chapter 1. Introduction*. pp. 3–10. En: [Janicak et al. \(2014b\)](#).
- Janicak, Philip G.; Marder, Stephen R.; Tandon, Rajiv y Goldman, Morris (Eds.) (2014b). *Schizophrenia. Recent advances in diagnosis and treatment*. Springer.

- Joachims, Thorsten (1998). «Text categorization with support vector machines: Learning with many relevant features». En: *ECML Conference*, pp. 2–7.
- Jones, M.; Kintsch, W. y Mewhort, D. (2006). «High-dimensional semantic space accounts of priming». *Journal of Memory and Language*, **55(4)**, pp. 534–552.
- Jones, Michael N; Gruenenfelder, Thomas M y Recchia, Gabriel (2011). «In defense of spatial models of lexical semantics». En: *Proceedings of the 33rd annual conference of the cognitive science society*, pp. 3444–3449.
- Jones, M.N. y Mewhort, D.J. K. (2007). «Representing word meaning and order information in a composite holographic lexicon.» *Psychological Review*, **114(1)**, pp. 1–37.
- Jordan, M.I. (1986). «Serial order: a parallel distributed processing».
- Kanerva, P. (2009). «Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors». *Cognitive Computation*, **1**, pp. 139–159.
- Kintsch, W. (1988). «The role of knowledge in discourse comprehension: a construction-integration model.» *Psychol Rev*, **95(2)**, pp. 163–182.
- Kintsch, W. (2002). *On the notions of theme and topic in psychological models of text comprehension*. pp. 157–170. John Benjamins Publishing Company, 2ª edición.
- Kiss, G.R.; Armstrong, G.; Milroy, R. y Piper, J. (1973). *An associative thesaurus of English and its computer analysis*. Edinburgh University Press.
- Kohonen, T. (1972). «Correlation Matrix Memories». *IEEE Transactions on Computers C*, **21**, pp. 353–359.
- Kohonen, Teuvo (1982). «Self-organized formation of topologically correct feature maps». *Biological cybernetics*, **43(1)**, pp. 59–69.
- Kohonen, Teuvo (1988). «The 'neural' phonetic typewriter». *Computer*, **21(3)**, pp. 11–22.
- Kuperberg, Gina R (2007). «Neural mechanisms of language comprehension: challenges to syntax.» *Brain research*, **1146**, pp. 23–49. ISSN 0006-8993. doi: 10.1016/j.brainres.2006.12.063.
- Kuperberg, Gina R (2010a). «Language in schizophrenia part 1: An introduction». *Language and linguistics compass*, **4(8)**, pp. 576–589.

- Kuperberg, Gina R (2010b). «Language in schizophrenia Part 2: What can psycholinguistics bring to the study of schizophrenia... and vice versa?» *Language and linguistics compass*, **4(8)**, pp. 590–604.
- Kutas, Marta y Hillyard, Steven A (1980). «Reading senseless sentences: Brain potentials reflect semantic incongruity». *Science*, **207(4427)**, pp. 203–205.
- Laisney, Mickaël; Giffard, Bénédicte; Belliard, Serge; De La Sayette, Vincent; Desgranges, Béatrice y Eustache, Francis (2011). «When the zebra loses its stripes: Semantic priming in early Alzheimer’s disease and semantic dementia». *Cortex*, **47(1)**, pp. 35–46.
- Lakoff, George (1990). *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge Univ Press.
- Landauer, T.K. y Dumais, S.T. (1997). «A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge». *Psychological Review*, **104(2)**, pp. 211–240.
- Landauer, T.K.; Mcnamara, D.S.; Dennis, S. y Kintsch, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum.
- Lashley, Karl S (1950). «In search of the engram». En: *Symposia of the society for experimental biology*, volumen 4, p. 30.
- Lawrie, Stephen M. y Pantelis, Christos (2011). *Structural brain imaging in schizophrenia and related populations*. pp. 334–352. En: [Weinberger y Harrison \(2011\)](#), 3ª edición.
- Lewis, David D y Ringuette, Marc (1994). «A comparison of two learning algorithms for text categorization». En: *Third annual symposium on document analysis and information retrieval*, volumen 33, pp. 81–93.
- Lewis, David D; Yang, Yiming; Rose, Tony G y Li, Fan (2004). «RCV1: A new benchmark collection for text categorization research». *The Journal of Machine Learning Research*, **5**, pp. 361–397.
- Lewis, John D y Elman, Jeffrey (2001). «Learnability and the statistical structure of language: Poverty of stimulus arguments revisited». En: *Proceedings of the 26th annual Boston University conference on language development*, volumen 1, pp. 359–370.
- Li, Fan y Yang, Yiming (2003). «A loss function analysis for classification methods in text categorization». En: *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pp. 472–479. Washington DC..
- Liddle, Peter F (1987). «The symptoms of chronic schizophrenia. A re-examination of the positive-negative dichotomy.» *The British Journal of Psychiatry*, **151(2)**, pp. 145–151.

- Lund, K.; Burgess, C. y Atchley, R.A. (1995). «Semantic and Associative Priming in High-Dimensional Semantic Space». En: J.D. Moore y J.F. Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pp. 660–665. Lawrence Elbaum.
- Mamah, Daniel y Barch, Deanna M. (2011). *Diagnosis and classification of the schizophrenia spectrum disorder*. pp. 45–83. Volumen 1 de [Ritsner \(2011\)](#).
- Manning, C D y Schütze, H (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.
- Marcus, Mitchell; Kim, Grace; Marcinkiewicz, Mary Ann; MacIntyre, Robert; Bies, Ann; Ferguson, Mark; Katz, Karen y Schasberger, Britta (1994). «The Penn Treebank: annotating predicate argument structure». En: *Proceedings of the workshop on Human Language Technology*, pp. 114–119. Association for Computational Linguistics.
- McClelland, J.L. y Rumelhart, D.E (Eds.) (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*. MIT Press.
- McCulloch, W.S. y Pitts, W. (1943). «A logical calculus of the ideas immanent in nervous activity». *Bulletin of Mathematical Biophysics*, **5**, pp. 115–133.
- McKenna, Peter J y Oh, Tomasina M (2005). *Schizophrenic speech: making sense of bathroofs and ponds that fall in doorways*. Cambridge University Press.
- Menezes, Paulo R. (2009). *Incidence and Outcome of Schizophrenia Across the Globe*. pp. 3–17. En: [Gattaz y Busatto \(2009\)](#).
- Merck, Catherine; Jonin, Pierre-Yves; Laisney, Mickaël; Vichard, Hélène y Belliard, Serge (2014). «When the zebra loses its stripes but is still in the savannah: Results from a semantic priming paradigm in semantic dementia». *Neuropsychologia*, **53**, pp. 221–232.
- Meyer, Carl D (2000). *Matrix analysis and applied linear algebra*. Siam.
- Miller, George A; Chodorow, Martin; Landes, Shari; Leacock, Claudia y Thomas, Robert G (1994). «Using a semantic concordance for sense identification». En: *Proceedings of the workshop on Human Language Technology*, pp. 240–243. Association for Computational Linguistics.
- Minsky, M. y Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press.
- Mizraji, E. (1989). «Context-dependent associations in linear distributed memories». *Bulletin Mathematical Biology*, **51**, pp. 195–205.
- Mizraji, E. (1992). «Vector logics: the matrix-vector representation of logical calculus». *Fuzzy Sets and Systems*, **50**, pp. 179–185.

- Mizraji, E.; Pomi, A. y Alvarez, F. (1994). «Multiplicative contexts in associative memories». *Biosystems*, **32**, pp. 145–161.
- Mizraji, E. y Valle-Lisboa, J.C. (2007). «Schizophrenic speech as a disordered trajectory in a collapsed cognitive "Small-World"». *Medical Hypotheses*, **68(2)**, pp. 347–352.
- Mizraji, Eduardo; Pomi, Andrés y Valle-Lisboa, Juan C. (2009). «Dynamic searching in the brain.» *Cognitive neurodynamics*, **3(4)**, pp. 401–414–414. ISSN 1871-4099. doi: 10.1007/s11571-009-9084-2.
- Montemurro, M.A. y Zanette, D.H. (2001). «Entropic analysis of the role of words in literary texts». *Arxiv preprint cond-mat/0109218*.
- Mota, Natália B; Furtado, Raimundo; Maia, Pedro PC; Copelli, Mauro y Ribeiro, Sidarta (2014). «Graph analysis of dream reports is especially informative about psychosis». *Scientific reports*, **4**.
- Mota, Natalia B; Vasconcelos, Nivaldo AP; Lemos, Nathalia; Pieretti, Ana C; Kinouchi, Osame; Cecchi, Guillermo A; Copelli, Mauro y Ribeiro, Sidarta (2012). «Speech graphs provide a quantitative measure of thought disorder in psychosis». *PloS one*, **7(4)**, p. e34928.
- Murray, Christopher J; Lopez, Alan D *et al.* (1996). *Global burden of disease*. volumen 1. Harvard University Press Boston.
- Nakamura, Kimihiro; Kuo, Wen-Jui; Pegado, Felipe; Cohen, Laurent; Tzeng, Ovid JL y Dehaene, Stanislas (2012). «Universal brain systems for recognizing word shapes and handwriting gestures during reading». *Proceedings of the National Academy of Sciences*, **109(50)**, pp. 20762–20767.
- Nass, Menasche M y Cooper, Leon N (1975). «A theory for the development of feature detecting cells in visual cortex». *Biological cybernetics*, **19(1)**, pp. 1–18.
- Nelson, D.L.; McEvoy, C.L. y Schreiber, T.A. (1998). «The University of South Florida word association, rhyme, and word fragment norms».
- Osgood, C. (1952). «The nature and measurement of meaning». *Psychological Bulletin*, **49(3)**, pp. 197–237.
- Papadimitriou, C.; Raghavan, P.; Tamaki, H. y Vempala, S. (2000). «Latent Semantic Indexing: A Probabilistic Analysis». *Journal of Computer and System Sciences*, **61(2)**, pp. 217–235.
- Pennebaker, James W; Mehl, Matthias R y Niederhoffer, Kate G (2003). «Psychological aspects of natural language. use: our words, our selves.» *Annual review of psychology*, **54**, pp. 547–77. ISSN 0066-4308. doi: 10.1146/annurev.psych.54.101601.145041.

- Peterson, Robin L y Pennington, Bruce F (2012). «Developmental dyslexia.» *Lancet*, **379(9830)**, pp. 1997–2007. ISSN 1474-547X. doi: 10.1016/S0140-6736(12)60198-6.
- Phipson, Belinda y Smyth, Gordon K (2010). «Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn». *Statistical applications in genetics and molecular biology*, **9(1)**.
- Pinker, Steven (1999). *Words and rules: The ingredients of language*. Basic Books.
- Pinker, Steven (2007). *The stuff of thought: Language as a window into human nature*. Penguin.
- Pinker, Steven y Prince, Alan (1988). «On language and connectionism: Analysis of a parallel distributed processing model of language acquisition». *Cognition*, **28(1)**, pp. 73–193.
- Pollard, Carl (1994). *Head-driven phrase structure grammar*. University of Chicago Press.
- Pulvermüller, F. (2002). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge University Press.
- Qi, Zhen; Miller, Gary W. y Voit, Eberhard O. (2011). *Mathematical Models in Schizophrenia*. pp. 305–325. Volumen 1 de [Ritsner \(2011\)](#).
- Real Academia Española (2014). «Diccionario de la real academia española».
- Reali, F. (2002). *Interacciones multiplicativas en modelos de redes neuronales: Algunas aplicaciones en redes de procesamiento del lenguaje*. Tesina o Proyecto, PEDECIBA, Universidad de la República.
- Reali, Florencia y Christiansen, Morten H (2005). «Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence». *Cognitive Science*, **29(6)**, pp. 1007–1028.
- Recchia, Gabriel y Jones, Michael N (2009). «More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis.» *Behavior research methods*, **41(3)**, pp. 647–56. ISSN 1554-351X. doi: 10.3758/BRM.41.3.647.
- Riordan, Brian y Jones, Michael N (2011). «Redundancy in Perceptual and Linguistic Experience: Comparing Feature-Based and Distributional Models of Semantic Representation». *Topics in Cognitive Science*, **3(2)**, pp. 303–345.
- Ritsner, Michael S. (Ed.) (2011). *Handbook of Schizophrenia Spectrum Disorders*. volumen 1. Springer.

- Ritsner, Michael S. y Gottesman, Irving I. (2011). *The Schizophrenia Construct After 100 Years of Challenges*. pp. 45–83. Volumen 1 de [Ritsner \(2011\)](#).
- Ritter, Helge y Kohonen, Teuvo (1989). «Self-organizing semantic maps». *Biological cybernetics*, **61(4)**, pp. 241–254.
- Rogers, Timothy T y McClelland, James L (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.
- Rosenblatt, Frank (1958). «The perceptron: a probabilistic model for information storage and organization in the brain.» *Psychological review*, **65(6)**, p. 386.
- Rumelhart, D.E.; Hinton, G.E. y Williams, R.J. (1986). *Learning Internal Representations By Error Propagation*. En: [McClelland y Rumelhart \(1986\)](#).
- Rumelhart, D.E. y McClelland, J.L. (1986). *On learning the past tenses of English verbs*. pp. 216–271. MIT Press.
- Russell, Stuart y Norvig, Peter (1995). *A modern approach*. Prentice Hall, New Jersey.
- Sahlgren, M. (2006). *The word-space model*. Tesis doctoral, Stockholm University, Stockholm, Sweden.
- Salton, G. y Lesk, M.E. (1965). «The SMART automatic document retrieval systems—an illustration». *Communications of the ACM*, **8(6)**, pp. 391–398.
- Samaria, F. y Harter, A. (1994). «Parameterisation of a Stochastic Model for Human Face Identification.» En: *2nd Workshop on Applications of Computer Vision*, .
- Sebastián-Gallés, Núria (2000). *LEXESP: Léxico informatizado del español*. Edicions Universitat Barcelona.
- Seidenberg, Mark S y McClelland, James L (1989). «A distributed, developmental model of word recognition and naming.» *Psychological review*, **96(4)**, p. 523.
- Shankar, K.H.; Jagadisan, U.K.K. y Howard, M.C. (2009). «Sequential learning using temporal context». *Journal of Mathematical Psychology*. doi: 10.1016/j.jmp.2009.07.005. En prensa.
- Shannon, Claude E (1949). «Communication Theory of Secrecy Systems*». *Bell system technical journal*, **28(4)**, pp. 656–715.
- Siegle, Greg J (1999). «A neural network model of attention biases in depression». *Progress in Brain Research*, **121**, pp. 407–432.

- Siekmeier, Peter J y Hoffman, Ralph E (2002). «Enhanced semantic priming in schizophrenia: a computer model based on excessive pruning of local connections in association cortex». *The British Journal of Psychiatry*, **180(4)**, pp. 345–350.
- Spitzer, Manfred (1999). *The mind within the net: Models of learning, thinking, and acting*. MIT Press.
- Stein, Dan J y Ludik, Jacques (1998a). *Neural networks and psychopathology: an introduction*. En: [Stein y Ludik \(1998b\)](#).
- Stein, Dan J y Ludik, Jacques (Eds.) (1998b). *Neural Networks and Psychopathology: Connectionist Models in Practice and Research*. Cambridge University Press.
- Stemmer, Brigitte y Whitaker, Harry A (2008). *Handbook of the Neuroscience of Language*. Academic Press.
- Steyvers, M. y Griffiths, T. (2007). *Probabilistic Topic Models*. pp. 427–448. En: [Landauer et al. \(2007\)](#).
- Steyvers, M.; Shiffrin, R.M. y Nelson, D.L. (2004). *Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory*. American Psychological Association.
- Steyvers, M. y Tenenbaum, J.B. (2005). «The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth». *Cognitive Science*, **29(1)**, pp. 41–78.
- Tandon, Rajiv y Bruijnzeel, Dawn (2014). *The Evolving Nosology of Schizophrenia: Relevance for Treatment*. pp. 13–23. En: [Janicak et al. \(2014b\)](#).
- Tankus, Ariel; Fried, Itzhak y Shoham, Shy (2012). «Structured neuronal encoding and decoding of human speech features». *Nature communications*, **3**, p. 1015.
- Tversky, Amos (1977). «Features of similarity». *Psychological review*, **84(4)**, pp. 327–352.
- Tversky, Amos y Gati, Itamar (1982). «Similarity, separability, and the triangle inequality.» *Psychological review*, **89(2)**, p. 123.
- Valle-Lisboa, J.C. (2007). *Las redes neuronales y el procesamiento del lenguaje natural*. Tesis doctoral, PEDECIBA, Universidad de la República.
- Valle-Lisboa, J.C. y Mizraji, E. (2007). «The uncovering of hidden structures by Latent Semantic Analysis». *Information Sciences*, **177(19)**, pp. 4122–4147.

- Valle-Lisboa, J.C.; Reali, F.; Anastasia, H. y Mizraji, E. (2005). «Elman topology with sigma-pi units: an application to the modeling of verbal hallucinations in schizophrenia». *Neural Networks*, **18(7)**, pp. 863–877.
- Valle-Lisboa, Juan C; Pomi, Andrés; Cabana, Álvaro; Elvevåg, Brita y Mizraji, Eduardo (2014). «A modular approach to language production: Models and facts». *Cortex*, **55**, pp. 61–76.
- van Dijk, Teun A (1993). «Texto y contexto: Semántica y pragmática del discurso».
- Vistoli, Damien; Passerieux, Christine; Houze, Bérengère; Hardy-Baylé, Marie-Christine y Brunet-Gouet, Eric (2011). «Neural basis of semantic priming in schizophrenia during a lexical decision task: A magnetoencephalography study». *Schizophrenia research*, **130(1)**, pp. 114–122.
- Weinberger, Daniel R. y Harrison, Paul J. (Eds.) (2011). *Schizophrenia*. Wiley-Blackwell, 3ª edición.
- Wendykier, Piotr y Nagy, James G (2010). «Parallel colt: a high-performance Java library for scientific computing and image processing». *ACM Transactions on Mathematical Software (TOMS)*, **37(3)**, p. 31.
- Whitehorn, John C y Zipf, George Kingsley (1943). «Schizophrenic language». *Archives of Neurology & Psychiatry*, **49(6)**, pp. 831–851.
- Widrow, B. y Hoff, M. (1960). «Adaptive switching circuits». En: *1960 IRE WESCON Convention Record*, pp. 96–104. IRE, New York.
- Wing, John Kenneth; Cooper, John Edward y Sartorius, Norman (1974). *Measurement and classification of psychiatric symptoms: an instruction manual for the PSE and Catego Program*. Cambridge University Press.
- Yang, Yiming (1996). «An evaluation of statistical approaches to MEDLINE indexing.» En: *Proceedings of the AMIA Annual Fall Symposium*, pp. 358–362.
- Yang, Yiming (1999). «An evaluation of statistical approaches to text categorization». *Information retrieval*, **90**, p. 69–90.
- Yoon, Yongwook; Lee, C y Lee, GG (2006). «An effective procedure for constructing a hierarchical text classification system». *Journal of the American Society for Information Science and Technology*, **57(3)**, pp. 431–442. doi: 10.1002/asi.
- Ziegler, Johannes C y Goswami, Usha (2005). «Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory.» *Psychological bulletin*, **131(1)**, p. 3.
- Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley.