UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIDAD IZTAPALAPA

DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

POSGRADO EN CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

“Sistema de clasificación de deterioro cognitivo mediante el análisis de redes léxico-semánticas en adultos mayores de habla hispana (SIDERLEX)”

TESIS PARA OBTENER EL GRADO ACADÉMICO DE MAESTRA EN CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

PRESENTA:

LIC. MARÍA ANGÉLICA GUTIÉRREZ LOZANO

MATRICULA: 2222800233

CORREO: gutangie04@gmail.com

ASESOR: DR. BENJAMIN MORENO MONTIEL

JURADO:

PRESIDENTA: DRA. GEMMA BEL ENGUIX

SECRETARIO: DR. BENJAMIN MORENO MONTIEL

VOCAL: DR. RENÉ MACKINNEY ROMERO

IZTAPALAPA, CIUDAD DE MÉXICO, 7 DE AGOSTO DE 2025

Resumen

Dado el aumento progresivo de los casos de Alzheimer a nivel mundial, esta enfermedad se ha convertido en una de las principales preocupaciones de salud pública, impulsando la búsqueda de herramientas más precisas para su detección temprana.

El Alzheimer es una enfermedad neurodegenerativa progresiva que se manifiesta a través del deterioro cognitivo y trastornos de la conducta. Entre los primeros signos de la enfermedad, se ha identificado la alteración del proceso lingüístico, la cual puede ser una señal temprana de afectación en las funciones cognitivas del paciente. Este cambio indica que el análisis del lenguaje podría facilitar un diagnóstico en etapas iniciales, permitiendo intervenciones más oportunas en el tratamiento.

El objetivo de este estudio es analizar redes léxico-semánticas construidas a partir de conversaciones de pacientes con la enfermedad de Alzheimer (AD, por sus siglas en inglés *Alzheimer's disease*) y sujetos sanos (HC, por sus siglas en inglés *Healthy Controls*), con el propósito de identificar patrones estructurales que permitan diferenciar ambos grupos y agrupar a los pacientes según su nivel de deterioro cognitivo. Para ello, se examinaron las redes mediante métricas de teoría de redes complejas y se aplicaron técnicas de agrupamiento con el algoritmo K-medias, lo que permitió caracterizar la conectividad y cohesión de las redes semánticas en función del avance de la enfermedad.

Como resultado, se observaron diferencias significativas en la estructura y cohesión de las redes entre los sujetos HC y pacientes con AD. La segmentación alcanzó una exactitud y precisión del 100 % en la categorización de ambos grupos, lo que respalda el uso de métricas de redes complejas como

indicadores fiables del deterioro semántico.

En contraste con la clasificación binaria entre sujetos HC y pacientes con AD, la segmentación por niveles de deterioro cognitivo según la Escala de Deterioro Global (*Global Deterioration Scale*, GDS) —niveles 4, 5 y 6— evidenció ciertas limitaciones. Aunque el modelo alcanzó una exactitud global del 85.19 %, la precisión varió entre los diferentes niveles de afección, con un 91.67 % en GDS 4, 75 % en GDS 5 y 85.71 % en GDS 6. Estas diferencias sugieren que la distinción entre algunos grados de deterioro no es completamente precisa, posiblemente atribuible a inconsistencias en la asignación inicial de los niveles GDS.

En conclusión, los resultados obtenidos a lo largo de esta investigación permiten establecer que el análisis de redes léxico-semánticas, mediante métricas de teoría de redes complejas y técnicas de agrupamiento, constituye un enfoque eficaz para caracterizar el deterioro semántico y diferenciar a los sujetos HC de los pacientes con AD, lo que abre la posibilidad de implementar esta metodología en procesos de detección temprana de la enfermedad.

Agradecimientos

Agradezco a la Universidad Autónoma Metropolitana, Unidad Iztapalapa, y al Posgrado en Ciencias y Tecnologías de la Información por brindarme la oportunidad de formarme académicamente y crecer en el ámbito profesional. Asimismo, quiero expresar mi agradecimiento al CONAHCyT por el apoyo económico recibido durante este proceso.

Agradezco sinceramente al Dr. Benjamin Moreno Montiel, por su guía constante, sus valiosas observaciones y el compromiso demostrado a lo largo de este trabajo. Su experiencia y orientación fueron fundamentales para el desarrollo de esta investigación.

Agradezco al Dr. René MacKinney Romero de la UAM-I y a la Dra. Gemma Bel Enguix de la UNAM por haber aceptado participar como sinodales en esta tesis. Sus valiosos comentarios y sugerencias contribuyeron de manera significativa a enriquecer este trabajo.

Contenido

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La demencia describe, a nivel cognitivo, una disminución general de la memoria y otras habilidades cognitivas lo suficientemente grave como para reducir la capacidad de una persona para realizar las actividades cotidianas. Se caracteriza por el deterioro progresivo y persistente de la función cognitiva. Los pacientes afectados suelen tener pérdida de memoria y una falta parcial o significativa de comprensión de sus carencias [14].

La demencia es causada por un conjunto de trastornos o afecciones a nivel cerebral, usualmente de naturaleza crónica o progresiva, que provoca una alteración de múltiples funciones corticales superiores, incluyendo la memoria, el pensamiento, la orientación, la comprensión, el lenguaje, la capacidad de aprender y de realizar cálculos, y la toma de decisiones. En la actualidad afecta a más de 55 millones de personas adultas mayores en el mundo según la Organización Mundial de la Salud (OMS) y cada año se presentan alrededor de 10 millones de casos nuevos y se prevé que el número de pacientes que padecen demencia superará los 150 millones en 2050 [1].

La demencia puede ser debida a diferentes tipos de enfermedades como: el Alzheimer, la demencia vascular, la afasia, cuerpos de Lewy, entre otras. Sin embargo, la enfermedad de Alzheimer es la forma más predominante de demencia, ya que, representa del 60 al 70 por ciento de los casos de demencia. En cuanto a las estadísticas de mortalidad, la OMS reportó que en 2019 la enfermedad de Alzheimer y otras demencias se posicionaron como una de las principales causas de muerte en el mundo, y que 1.8 millones de

personas murieron debido a esta causa [2].

El Alzheimer es una enfermedad neurodegenerativa que provoca el deterioro progresivo de funciones cognitivas superiores como la memoria, el cálculo, la praxia ideomotriz, el pensamiento abstracto, la capacidad de razonamiento o el lenguaje [26]. Las afecciones evolucionan, de tal forma, que el paciente en la fase final es totalmente dependiente.

Aunque el Alzheimer no tiene cura, la posibilidad de un diagnóstico temprano permitiría un mejor tratamiento de la enfermedad y por ende, ayudaría a elevar la calidad de vida tanto de los pacientes como de sus familiares. Por lo cual, surge la posibilidad de realizar investigaciones sobre la detección del Alzheimer utilizando técnicas de inteligencia artificial como el procesamiento de lenguaje natural (PLN) y el aprendizaje automático para analizar y evaluar el deterioro cognitivo de los pacientes que padecen Alzheimer con el objetivo de diagnosticar la enfermedad en etapa precoz.

1.1. Problemática

El diagnóstico de la enfermedad de Alzheimer (AD, *Alzheimer's Disease*) en etapas tempranas se torna complejo debido a que el proceso de desarrollo de la enfermedad dura décadas. Las personas afectadas sufren *déficits* cognitivos identificables sin disfunción sustancial en la etapa prodrómica de la AD. El diagnóstico habitual de la AD se ha basado en gran medida en evaluaciones cognitivo-mentales posteriores a los síntomas, en las imágenes del cerebro y otras herramientas de diagnóstico que desempeñan un papel más complementario [37].

En el área de neuropsicología, los especialistas en psicología y psiquiatría emplean algunas pruebas para el diagnóstico de la enfermedad de Alzheimer como:

- Mini Mental State Examination (MMSE)
- Montreal Cognitive Assessment (MoCA)
- Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog)

Sin embargo, este tipo de pruebas suelen ser generales, ya que, algunas no

hacen uso de criterios relacionados con el procesamiento linguistico o comunicativo no discriminan entre las diferentes enfermedades que provocan la demencia y su valor clínico puede ser limitado debido a ciertas características de los pacientes que influyen en los resultados como el nivel cultural y educativo. Asimismo, muestran una carencia de sensibilidad en etapas tempranas de la AD [10, 34].

Por otro lado, especialistas en neurología emplean estudios de imágenes cerebrales, siendo las más comunes la imagen por resonancia magnética (IRM) y la tomografía computarizada (TC) que son usadas para detectar alteraciones cerebrales asociadas con el deterioro cognitivo. No obstante, en las fases iniciales de la enfermedad los depósitos de las proteínas beta-amiloide o tau son demasiado pequeños para ser detectados. Además, estos recursos no suelen ser usados en la clínica habitual, ya que su análisis debe estar altamente mediado por personal médico especializado. Asimismo, la interpretación de las imágenes de resonancia magnética puede ser subjetiva [37].

Investigaciones recientes emplean técnicas de Inteligencia Artificial (IA), especialmente con algoritmos de aprendizaje automático y aprendizaje profundo para un diagnóstico más preciso en etapas tempranas. Estas investigaciones emplean dichos algoritmos para analizar las imágenes de resonancia magnética cerebral [37] y también para el reconocimiento del deterioro del proceso lingüístico de las personas afectadas por la AD [3].

Especifícamente, para este trabajo de investigación el proceso lingüistico es relevante, puesto que algunos estudios revelan que, las alteraciones en las características lingüisticas del habla, probablemente sean uno de los primeros signos de la aparición de la AD [27]. Con base en esto, el análisis del lenguaje mediante la IA posibilitaría un diagnóstico en etapas tempranas de la AD.

Aunque se han obtenido avances significativos en la investigación de la detección en etapas tempranas de la AD aplicando la IA, aún existen varios desafíos que hacen complejo este tipo de estudios. Algunos de los problemas que se presentan es la disponibilidad limitada de bases de datos abiertas y datos reales sobre pacientes que padecen Alzheimer, así como la cantidad de información contenida en las bases de datos.

En este trabajo de investigación se propone el modelo SIDERLEX (Sistema de Clasificación de Deterioro Cognitivo Mediante el Análisis de Redes Léxico-Semánticas en Adultos Mayores de Habla Hispana), el cual emplea técnicas de procesamiento de lenguaje natural (PLN) y algoritmos de aprendizaje automático para el análisis de la alteración del proceso lingüístico provocada por un deterioro cognitivo generalizado que afecta la memoria y funciones ejecutivas como el razonamiento y la capacidad de planificación.

Este modelo contempla ciertas complejidades inherentes a su desarrollo, como la disponibilidad de una base de datos acorde a las necesidades de la investigación, especialmente en lo relativo a la cantidad y calidad de información requerida para la construcción de un corpus. Asimismo, considera los desafíos relacionados con la creación de redes léxico-semánticas a partir del corpus y la comparación de los resultados con los obtenidos de las pruebas neuropsicológicas tradicionales.

1.2. Hipótesis

- Es factible que a través de patrones de redes léxico-semánticas se pueda representar el estado léxico-semántico del ser humano y estimar el deterioro cognitivo en las personas adultas mayores.
- Mediante el análisis de la representación vectorial de palabras y la similitud y representación de esta, es posible crear redes léxico-semánticas.
- Algunas características de las redes complejas obtenidas a partir de las redes léxico-semánticas permiten categorizar el deterioro cognitivo en personas adultas mayores.
- Mediante un modelo del aprendizaje maquinal y minería de datos es factible estimar el factor de deterioro cognitivo en adultos mayores.

1.3. Objetivos

1.3.1. Objetivo general

Implementar el Sistema de Clasificación de Deterioro Cognitivo Mediante el Análisis de Redes Léxico-Semánticas en Adultos Mayores de Habla Hispana

(SIDERLEX).

1.3.2. Objetivo específicos

- Revisión del estado del arte sobre métodos de evaluación estructural de redes léxico-semánticas, a través de métricas propias de la teoría de redes complejas.
- Incorporar la noción de representación vectorial de palabras para crear matrices de coocurrencia, con el propósito de automatizar la creación de redes léxico-semánticas.
- Establecer métricas específicas de redes complejas para la evaluación de la estructura y propiedades de las redes léxico-semánticas.
- Establecer un esquema de clasificación y factores sobre las características de las redes complejas que permitan estimar deterioro cognitivo.
- Realizar pruebas para validar el sistema propuesto, con la finalidad de obtener sus tasas de aprendizaje.

1.4. Metodología

La metodología desarrollada para la implementación del sistema de clasificación de deterioro cognitivo mediante el análisis de redes léxico-semánticas en adultos mayores de habla hispana (SIDERLEX) consiste en las siguientes fases:

1. La primera fase se enfoca a la recolección de datos sobre pacientes con la enfermedad de Alzheimer. La información se obtiene de una base de datos creada por el doctor José Luis Pérez Mantero, en Valencia, España. Esta base de datos está conformada por 27 conversaciones en idioma español de pacientes que padecen Alzheimer. Dichas conversaciones se llevaron a cabo en las casas o centros de atención de los pacientes en compañía de familiares y/o amigos mediante audio grabaciones para su posterior transcripción.
2. En la segunda fase se inicia el preprocesamiento de las transcripciones de las conversaciones para la creación de la base de datos del corpus.

Esta etapa consiste en la limpieza y estructuración de los datos mediante ténicas de mineria de datos, con el objetivo de preparar adecuadamente el texto para su análisis léxico-semántico.

Para ello, se elimina la información general sobre las conversaciones y se excluyen todos los segmentos de la conversación que no pertenecen al paciente con AD. A continuación, se realiza el proceso de tokenización, que radica en dividir las oraciones en unidades lingüísticas más pequeñas llamadas *tokens* (unidades léxicas). En este estudio, cada *token* corresponde a una palabra individual dentro de la conversación. A partir de este punto, se utilizará el término *token* para referirse a estas unidades.

Posteriormente, se lleva a cabo la eliminación de las palabras irrelevantes y la lematización de las palabras resultado del proceso de tokenización. La eliminación de las palabras irrelevantes consiste en descartar términos que no aportan información significativa, es decir, que no contribuyen en gran medida a comprender el texto desde el punto de vista léxico como artículos, conjunciones, pronombres o preposiciones.

Por último, se realiza la lematización cuyo propósito es normalizar el texto. Este proceso consiste en reducir cada palabra a su forma base o lema. Por ejemplo, las formas verbales caminando, caminó y caminan se transforman en el lema caminar. De esta manera, se unifican las diferentes formas gramaticales que una palabra puede adoptar, facilitando un análisis más consistente del contenido lingüístico.

3. El objetivo de la fase tres es generar las redes léxico-semánticas. Para ello, primero se vectoriza el corpus de palabras obtenido en la segunda fase mediante un algoritmo específico. A continuación, se construyen las matrices de co-ocurrencia, que sirven como base para la creación de las redes léxico-semánticas.

4. En la fase cuatro de la investigación se hace un análisis detallado de las características más significativas de las redes léxico-semánticas, empleando métricas fundamentadas en las propiedades de los grafos. Este análisis tiene como finalidad facilitar la clasificación posterior de las redes.

5. En la etapa cinco, se adapta el algoritmo *K-medias* utilizado para el agrupamiento no supervisado de datos [19], con el fin de agrupar las conversaciones en categorías de acuerdo con sus características léxico-semánticas.

6. Finalmente, en la fase seis, se lleva a cabo la evaluación de la metodología empleada. Se realizan algunos análisis usando algunas métricas de evaluación para conocer la efectividad del modelo propuesto.

1.5. Justificación

En la actualidad, existen diferentes métodos de diagnóstico de la AD que involucran pruebas neuropsicológicas y estudios clínicos; sin embargo suelen ser procesos que toman períodos prolongados de tiempo, ya que, la enfermedad se caracteriza por un inicio gradual. Cuando los síntomas clínicos son identificables, la mayoría de los pacientes se encuentran en estadios moderados o avanzados de la AD, no existiendo tratamientos que detengan la progresión de la enfermedad. Además, estos métodos tienen ciertas limitaciones que afectan los resultados del diagnóstico, lo que implica que, en ocasiones no sean sensibles para hacer una detección temprana de la AD.

Estudios existentes sobre el diagnóstico en etapas tempranas de AD basados en el análsis del lenguaje en pacientes con AD y adultos mayores cognitivamente sanos, mediante el uso de técnicas de PLN y algoritmos de aprendizaje automático han permitido identificar patrones lingüisticos vinculados al deterioro del lenguaje en presencia de la AD. Con estos resultados, existe la posibilidad de detectar la enfermedad de forma oportuna.

El PLN permite realizar análisis puntuales del proceso lingüístico de las personas, permitiendo detectar patrones específicos en el lenguaje, como la repetición de palabras o frases, la pérdida de coherencia en la conversación y la disminución del vocabulario. Estos indicadores pueden ayudar a identificar patrones lingüísticos asociados con la AD.

En cuanto al aprendizaje automático permite crear modelos predictivos basados en datos. Estos modelos pueden clasificar en el contexto de la AD, a los pacientes en diferemtes categorías en función de las características lingüísticas extraídas mediante el PLN y las redes léxico-semánticas.

Por otra parte, las redes léxico-semánticas posibilitan la exploración de las relaciones entre palabras y su significado en un contexto específico. Mediante estas redes se pueden identificar cambios en las relaciones semánticas entre las palabras. Asimismo, se pueden identificar patrones específicos de palabras o conceptos que son más frecuentes en el habla de los pacientes con la EA, lo cual puede ser útil para el diagnóstico de la enfermedad. Por lo tanto, se han empleado modelos de redes léxico-semánticas que emulan representar la memoria semántica, lo cual permite, que a través del uso de PLN y algoritmos de aprendizaje automático, construir sistemas predictivos inteligentes para el diagnóstico de Alzheimer en etapas tempranas.

1.6. Organización del trabajo de tesis

Este documento refleja los resultados de una investigación con base tecnológica que pretende implementar el Sistema de Clasificación de Deterioro Cognitivo Mediante el Análisis de Redes Léxico-Semnticas en Adultos Mayores de Habla Hispana (SIDERLEX), y que tiene como objetivo la categorización de conversaciones de pacientes con AD y personas sanss (HC, por sus siglas en inglés *Healthy Controls*). El contenido está estructurado en seis capítulos cuya información explica el proceso de desarrollo del trabajo de investigación.

En el presente capítulo se abordo el contexto de lo que implica la AD, así como, la problemática que existe en el diagnóstico de la enfermedad en la actualidad. Asimismo, el planteamiento de la base de la investigación, por lo que incluye tanto el objetivo general como los objetivos específicos que se buscan cumplir, la metodología de la investigación y la importancia que tiene la misma.

El segundo capítulo engloba teorías y conceptos fundamentales que sirven como base para la implementación de SIDERLEX. Asimismo, contempla la explicación de áreas tecnológicas como la mineria de datos y técnicas de IA como PLN y el aprendizaje automático.

El tercer capítulo presenta la revisión de la literatura, la cual incorpora investigaciones relevantes vinculadas con el presente estudio y contribuye a plantear las bases para implementar el modelo propuesto.

El cuarto capítulo contiene la información correspondiente a la forma en cómo se desarrolla SIDERLEX. El proceso inicia con la recolección de las las conversaciones de pacientes con AD y el grupo HC, posteriormente se lleva a cabo el preprocesamiento y la construcción de las redes léxico-semánticas, donde se emplean matrices de coocurrencia para la crear dichas redes, seguido del análisis de las características relevantes de las redes léxico-semánticas con base a métricas basadas en la teoría de grafos. Después, se aplica el algoritmo K-medias para obtener la clasificación de las conversaciones y finalmente realizar la evaluación del modelo mediante métricas como exactitud, precisión, entre otras, para determinar la efectividad del modelo propuesto.

El quinto capítulo describe los procedimientos experimentales llevados a cabo en la investigación, así como los hallazgos obtenidos a partir del análisis de las redes léxico-semánticas de los pacientes con AD y los sujetos HC. Para ello, se emplearon métricas de teoría de grafos y técnicas de agrupamiento con K-medias, permitiendo evaluar las diferencias estructurales en la organización del lenguaje entre ambos grupos.

El sexto capítulo presenta una serie de conclusiones de acuerdo a lo planteado en la hipótesis formulada. Estas conclusiones están basadas en los resultados obtenidos en la implementación de SIDERLEX y los objetivos expuestos. Además de considerar posibles trabajos futuros de la implementación del modelo y recomendaciones para posteriores investigaciones.

Capítulo 2

Marco teórico

Este capítulo contempla los conceptos fundamentales que intervienen en el desarrollo del Sistema de Clasificación de Deterioro Cognitivo Mediante el Análisis de Redes Léxico-Semánticas en Adultos Mayores de Habla Hispana (SIDERLEX).

2.1. Deterioro cognitivo

En general la cognición se entiende como el funcionamiento intelectual que permite al ser humano a interactuar con el medio en el que se desenvuelve. En el proceso de envejecimiento, las personas sufren cambios en el funcionamiento cognitivo; esto implica que el cerebro presenta modificaciones a nivel morfológico, bioquímico, metabólico y circulatorio. Sin embargo, depende de la plasticidad del cerebro y de las redundancias de muchas funciones cerebrales que se presenten alteraciones cognitivas o que prevalezca un funcionamiento normal en el cerebro en las personas adultas mayores, derivados de estos cambios [7] [28].

Por tanto, el hecho que se desarrolle o no deterioro cognitivo asociado al envejecimiento dependerá de cómo se produzcan las adaptaciones compensatorias con la edad. El deterioro cognitivo es la pérdida de funciones cognitivas, específicamente en memoria, atención y velocidad de procesamiento de la información, que se produce en el envejecimiento normal. Aunque, las alteraciones que se presentan en el proceso de envejecimiento y las que im-

plican la existencia de una enfermedad no siempre se pueden observar con claridad; se conoce que el proceso de envejecimiento va acompañado de una serie de trastornos debido a las modificaciones que sufre el cerebro [28].

Algunas de estos trastornos consisten en la pérdida de la memoria especialmente para hechos recientes, permaneciendo conservada la remota y la inmediata. Además, existen alteraciones en la memoria episódica, no obstante, la memoria semántica se mantiene en buen estado, incluso puede mejorar con los años. Con respecto al lenguaje, principalmente en el aspecto léxico-semántico, y en el razonamiento verbal, así como en el vocabulario tampoco parecen verse alteradas, inclusive el vocabulario puede mejorar con la edad. En contraste, la fluencia verbal que involucra la velocidad, la atención y la producción motora suele disminuir [28]. En la sección 2.4 se realiza una explicación más extensa del aspecto léxico-semántico.

2.2. Memoria Semántica

El sistema semántico representa el conocimiento que las personas tienen de los objetos, de las relaciones entre los objetos y entre los eventos del mundo real [8]. De manera que, la memoria semántica es responsable de la adquisición, representación y procesamiento de la información conceptual. Esta memoria comprende una serie de funciones cognitivas como la capacidad de asignar significado a palabras y a oraciones, reconocer objetos, recordar información especifica de conceptos aprendidos previamente y obtener nueva información a partir de la experiencia real y perceptiva [30].

Asimismo, la memoria semántica está organizada por categorías, es decir, algunas de las propiedades de los conceptos representados pueden agruparse de una forma categórica que conjunta características y objetos, mediante procesos que pueden ayudar a organizar la inmensa cantidad de información sobre las experiencias significativas. Se piensa, que uno de los procesos está “basado en reglas” que, supone un análisis de un objeto de prueba para determinar las características necesarias y suficientes de un concepto. Un segundo proceso de categorización se basa en la “similitud” e implica una comparación de un objeto de prueba con un prototipo o con instancias recordadas de un concepto [17].

Por otro parte, cuando existe un deterioro cognitivo como la enfermedad de Alzheimer, se manifiestan deficiencias en la memoria semántica. El deterioro se presenta con problemas para nombrar objetos e imágenes, para definir objetos y por una mala comprensión del lenguaje oral y escrito [22].

2.3. Memoria de trabajo

Al igual, que la memoria semántica, la memoria de trabajo muestra alteraciones en presencia de la enfermedad de Alzheimer. La Memoria de Trabajo (MT) es un sistema de memoria activo responsable del almacenamiento temporal y procesamiento simultáneo de información necesaria para la realización de tareas cognitivas complejas [4]. Se ha demostrado que la capacidad de MT está fuertemente vinculada con una serie de tareas cognitivas de orden superior, como la comprensión lectora, el razonamiento lógico y la resolución de problemas, así como para la realización de actividades cotidianas como la participación en una conversación, leer una novela o usar una computadora [38].

Aunque, esta memoria es considerada una memoria de corto plazo, la memoria a corto plazo es un componente de la MT y se refiere al almacenamiento de información sin manipulación, mientras que la MT es responsable tanto del almacenamiento como la manipulación de la información y principalmente en la funcionalidad de la cognición compleja. Este concepto ha sido útil en muchas áreas de aplicación, incluidas las diferencias individuales en cognición, neuropsicología y neuroimagen [4].

El envejecimiento típico se relaciona con un desfase entre la producción y la comprensión lingüística, algo que ha sido detectado en la MT, ya que las asociaciones léxicas dependen de un criterio sintáctico y semántico. A diferencia de la memoria semántica, esta memoria al envejecer sufre una lentificación de las habilidades lingüísticas.

Por otro lado, Algunos estudios han encontrado que cuando existe un deterioro cognitivo leve los déficits de la MT son prominentes en pruebas complejas que exigen atención dividida, que aumentan la carga de memoria o que abarcan diferentes modalidades, como la MT verbal y visuoespacial [38]. Además, en otros estudios se han encontrado asociación entre los problemas

de la MT y el deterioro del lenguaje en pacientes que padecen la aparición temprana de la enfermedad de Alzheimer [32].

2.4. Redes léxico-semánticas

2.4.1. Definición

Las redes léxicas se definen como aquellos vínculos que se forman entre palabras, las cuales reflejan la organización de la memoria semántica, el procesamiento del lenguaje y el acceso léxico. Las redes léxicas son formas de emular la organización y el estado de la memoria semántica; son una herramienta útil para estudiar el lenguaje humano a partir de sus regularidades y la manera en que se desarrolla. Durante el envejecimiento, las redes léxicas se encuentran preservadas y resisten a la aparición de déficits cognitivos generales [6].

A lo largo del tiempo, se han propuesto diferentes modelos de redes léxico-semánticas, como redes jerárquicas, de ordenamiento semántico y basados en la teoría de grafos. El modelo de ordenamiento semántico propuesto por Collins y Loftus en 1975, está fundamentado en la teoría de la propagación de la activación, la cual consiste en que la activación de un nodo activa indirectamente a aquellos conectados a él de manera proporcional a su fuerza de conexión [16]. En este modelo los nodos se agrupan y conectan a partir de su cercanía semántica, como lo muestra la Figura 2.1. Por ejemplo, amanecer se encontraría muy cerca y unido a nube, pero no a así a vehículo. Sin embargo, es posible desplazarse de un nodo a otro a partir de conexiones indirectas, por ejemplo, se puede ir de amanecer hacia rojo y de rojo hacia naranja.

Este tipo de red permite la vinculación entre palabras que pertenecen a diferentes categorías gramaticales dando lugar a conexiones similares a las coocurrencias léxicas, así como a las relaciones sintácticas del habla. Asimismo, representa la magnitud de la asociación (conexión) entre dos nodos de una forma gráfica, donde la longitud del arco entre dos nodos representa la magnitud de la conexión. Así pues, entre más corta la distancia entre nodos, más fuerte la magnitud de la conexión entre ellas.

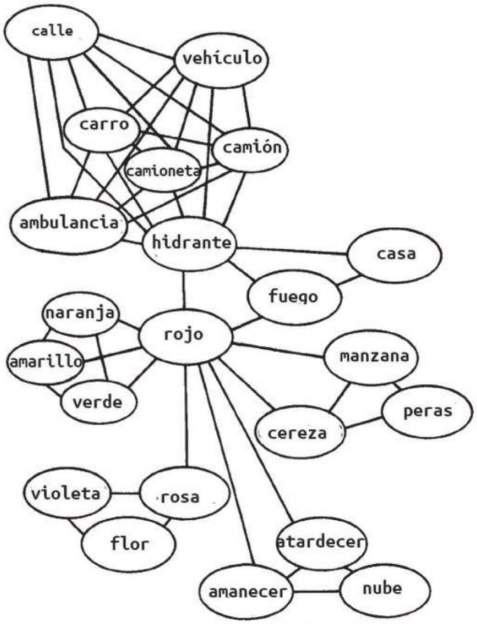


Figura 2.1: Ejemplo de red léxica de relaciones semánticas. [16]

Otro de los modelos es el basado en la teoría de grafos que nace del fenómeno del mundo pequeño (*small world*). En este modelo, cada elemento de la red se puede conectar indirectamente a cualquier otro elemento mediante una pequeña cantidad de nodos. Este tipo de redes, del mismo modo que el modelo de ordenamiento semántico, representan las palabras como nodos y la vinculación entre ellos por arcos. La unidad para medir el peso de la asociación o la dirección de la conexión depende de la implementación y sus fundamentos teóricos [16]. La Figura 2.2 es un ejemplo de una red léxica construida a partir de la coocurrencia de palabras en cuatro oraciones.

Alguna de las principales características de la teoría de grafos de tipo *small world* son:

- La estructura de las redes es semi-aleatoria y semi-ordenada.
- Los parámetros que rigen a las redes son el coeficiente de agrupamiento de los nodos (C), la distancia mínima entre nodos (L) y la distribución de los grados de conexión de los nodos ($P(k)$).
- El coeficiente de clusterización tiende a ser relativamente alto.
- L tiende a ser tan pequeña como la cantidad de nodos de una red lo permita.

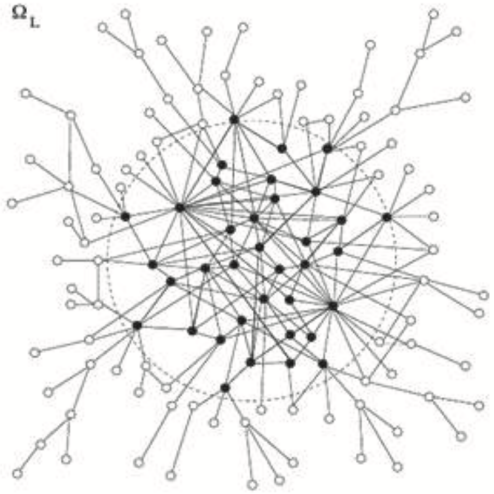


Figura 2.2: Ejemplo de red léxica basada en la teoría de grafos. [16]

Dado que estas redes están fundamentadas en la teoría de grafos, la información contenida en estas puede ser recuperadas por algoritmos computacionales.

Con base a lo anterior, se puede definir a una red léxico-semántica como un modelo computacional, lo cual es una manera de representar el conocimiento humano sobre el mundo basado en las relaciones que establecen los conceptos que lo conforman. En este modelo se representan los conceptos lexicalizados, es decir que las palabras están relacionadas entre sí mediante asociaciones semánticas y que es posible representar estas relaciones en forma de una red.

En una red léxico-semántica los nodos representan palabras o conceptos y las líneas o aristas representan diferentes relaciones semánticas como sinonimia (palabras con un mismo significado o semejante), antonimia (palabras con significado opuesto), hiperónimos (palabras que representan categorías más amplias, por ejemplo: deporte) e hipónimos (palabras que representan categorías más específicas, por ejemplo: fútbol).

Este tipo de redes se utilizan en la recuperación de la información, la traducción automática y la comprensión del lenguaje natural, asimismo, han sido usada en áreas de investigación de psicología y neurociencia para comprender cómo las personas almacenan y acceden al conocimiento semántico.

Para la construcción de redes léxico-semánticas existen distintos métodos basados en:

- Corpus.
- Tesauros.
- Conocimiento experto.
- Recursos lingüísticos existentes.

El método basado en corpus consiste en el análisis de conjuntos extensos de textos denominados corpus, con el propósito de identificar las relaciones entre las palabras. Para la construcción de la red léxico-semántica en este método, se pueden usar técnicas estadísticas o lingüísticas a partir de la frecuencia y la coocurrencia de las palabras.

2.5. Algoritmo de frecuencia de término - frecuencia inversa de documento (TF-IDF, *term frequency - inverse document frequency*)

TF-IDF es una técnica PLN para convertir palabras en vectores con cierta información semántica, asignando pesos a las palabras poco comunes. Este algoritmo es una combinación de dos métricas: frecuencia de término y frecuencia inversa de documento [29]. La definición formal de TF-IDF se puede representar como:

$$tfidf = tf \times idf$$

donde la frecuencia de término tf y la frecuencia inversa de documento idf representan las dos métricas.

La frecuencia de término en cualquier vector de documento se denota por el valor de frecuencia de ese término en un documento particular. Matemáticamente, se puede representar como:

$$tf(w, D) = f_{w_D}$$

donde f_{w_D} denota la frecuencia de la palabra w en el documento D, que se convierte en el término.

La frecuencia inversa de documento, denotada por idf, es el inverso de la frecuencia de documento para cada término. Se calcula dividiendo el número total de documentos en el corpus por la frecuencia de documento para cada término y luego aplicando una escala logarítmica al resultado, cuya representación matemática es:

$$idf(t) = 1 + log\frac{C}{1 + df(t)}$$

donde $idf(t)$ representa el idf para el término t, C representa el recuento del número total de documentos en el corpus, y $df(t)$ representa la frecuencia del número de documentos en los que el término t está presente.

2.6. Redes complejas

Una red es un grafo. Un grafo (G) está compuesto por dos conjuntos de información: un conjunto de vértices (vertex), $N = \{n_1, n_2, ..., n_m\}$ y un conjunto de aristas (edge), $L = \{l_1, l_2, ...l_\ell\}$ que conectan pares de vértices. Donde m representa el número de vértices y ℓ el número total de aristas [36]. La Figura 2.3 muestra una red pequeña compuesta de ocho vértices y diez aristas.

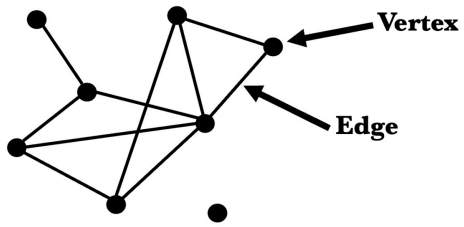


Figura 2.3: Ejemplo de una red pequeña. [24]

Muchos objetos de interés en las ciencias físicas, biológicas y sociales pueden concebirse como redes, lo que podría derivar en conocimientos nuevos y útiles [24].

Por otro lado, el análisis de las propiedades de las redes complejas —también denominadas métricas o descriptores topológicos— permite caracterizar su topología. A continuación, se definen algunas de estas propiedades:

- Grado (Degree). El grado de un vértice en una red es el número total

de sus conexiones, es decir, la cantidad de aristas que lo conectan con otros vértices de la red.

- Centralidad de Intermediación (Betweenness Centrality). La Centralidad de Intermediación se utiliza en el análisis de redes para medir la importancia o centralidad de un vértice. Esta medida se basa en cuántas veces un vértice actúa como intermediario en los caminos más cortos entre otros vértices en una red.

- Centralidad de Cercanía (Closeness Centrality). Esta propiedad se usa para evaluar la importancia o influencia de un vértice dentro de una red, midiendo cuán cercano se encuentra respecto al resto de los vértices de la red. Un vértice con alta centralidad de cercanía se caracteriza por tener distancias cortas a todos los demás vértices, lo que significa que puede acceder rápidamente a ellos o difundir información de manera eficiente a través de la red. Cuanto mayor es el valor de la centralidad de cercanía, más "central" o importante es el nodo en términos de su proximidad al resto de la red.

- Centralidad de Intermediación de Aristas (Edge Betweenness Centrality). Esta propiedad se emplea para evaluar la importancia de una arista dentro de una red. La centralidad de intermediación de una arista se define como el número de caminos más cortos entre pares de vértices que atraviesan dicha arista, lo que indica con qué frecuencia actúa como puente crítico en la conectividad global. Esta propiedad resulta particularmente útil para comprender el flujo de información y la estructura modular de la red.

- Densidad del Grafo (Graph Density). La densidad es una medida que indica el nivel de conectividad de la red en relación con el número máximo posible de enlaces. Esta propiedad proporciona información sobre la concentración de conexiones entre los vértices y resulta útil para analizar y comparar diferentes estructuras de redes.

- Diámetro del Grafo (Graph Diameter). El diámetro es la distancia máxima posible entre dos vértices, calculada siempre como la longitud del camino más corto que los conecta. Esta métrica indica la extensión máxima de la red y se obtiene identificando el valor más alto entre todas

las distancias geodésicas, como se observa en la Figura 2.4.

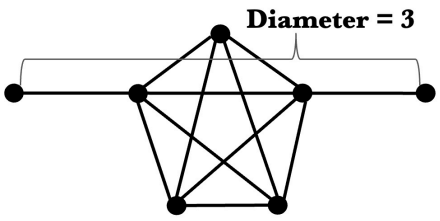


Figura 2.4: Ejemplo del diámetro de una red. [12]

Un diámetro largo significa que encontrar el camino más corto entre algunos pares de vértices puede ser complejo, debido a la existencia de múltiples rutas y bifurcaciones. En cambio, un diámetro pequeño,facilita alcanzar cualquier vértice en pocos saltos, normalmente uno o dos. Además, cuando el diámetro es grande, recorrer toda la red puede resultar inviable, especialmente si solo se dispone de información local sobre los nodos vecinos.

- Excentricidad (Eccentricity). La excentricidad de un vértice se define como la mayor distancia geodésica desde ese vértice a cualquier otro en la red. Esta métrica permite estimar la "distancia" relativa de un vértice respecto a los demás. A diferencia del diámetro, que es una medida global, la excentricidad es local, ya que se calcula individualmente para cada vértice.

- Coeficiente de Agrupamiento (Clustering Coefficient). El coeficiente de agrupamiento es una métrica utilizada para evaluar la tendencia de los vértices de una red a formar agrupamientos o grupos. Indica cuántos de los vecinos de un vértice están interconectados entre sí, proporcionando información sobre la cohesión local de la red. Los valores del coeficiente de agrupamiento varía entre cero y uno, siendo que el valor 1 significa que la red está densamente conectada y el valor cero significa una red dispersa.

- Transitividad (Transitivity). La transitividad en una red mide la tendencia a que los vecinos de un vértice estén también conectados entre sí. Conceptualmente, si el vértice u está conectado al vértice v y v al vértice w, la transitividad indica la probabilidad de que u también esté

conectado a w, formando un triángulo. En la Figura 2.5, se ilustra este concepto.

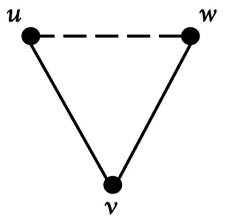


Figura 2.5: Ejemplo Transitividad. [24]

El camino uvw (aristas sólidas) se considera cerrado si la tercera arista que va directamente de u a w está presente (arista discontinua).

- Distribución de Grado (Degree Distribution). El grado de un vértice es una métrica local que describe únicamente sus conexiones directas, mientras que el grado medio ofrece una visión global de la conectividad de la red. Por su parte, la distribución de grados indica la proporción de vértices que presentan un grado específico, constituyendo una de las propiedades más relevantes para caracterizar la estructura de la red.

- Asortatividad (Assortativity). La asortatividad es una medida que indica la tendencia de los nodos en una red a conectarse con otros que comparten características similares. En su forma más común, se calcula mediante el coeficiente de correlación de Pearson aplicado a los grados de los vértices en los extremos de cada arista, determinando si existe correlación positiva (asortativa), negativa (disasortativa) o nula entre ellos.

Las métricas descritas anteriormente se fundamentan en definiciones del análisis de redes complejas, basadas en los trabajos de Newman [24] y Borge-Holthoefer y Arenas [9]. Estas herramientas resultan esenciales para caracterizar la estructura y propiedades de una red, lo que permite avanzar hacia la comprensión de cómo se organizan e interrelacionan distintos dominios en sistemas complejos.

Aunque, las funciones cognitivas pertenecen a ámbitos distintos, no existen aisladas unas de otras; el éxito en la realización de la mayoría de las tareas depende de la interdependencia de varios dominios cognitivos. La separación

e integración características de las capacidades cognitivas, por tanto, permite conceptualizar una red cognitiva en la que el rendimiento en cada tarea corresponde a un vértice y la interrelación o correlación entre rendimientos, a una arista.

2.7. Procesamiento de lenguaje natural (PLN)

El procesamiento de lenguaje natural (PLN) es una disciplina que combina informática, inteligencia artificial y lingüística (IA) con el propósito de estudiar y desarrollar métodos que permitan a las computadoras interpretar, comprender y procesar el lenguaje humano. Su finalidad principal es hacer posible el análisis y manejo eficiente de grandes volúmenes de datos en lenguaje natural mediante la aplicación de algoritmos y representaciones computacionales [13].

Dentro de este campo, el análisis del lenguaje ocupa un papel central, ya que permite comprender la estructura, el significado y la organización del texto. Para ello, el PLN integra enfoques provenientes de la lingüística, la estadística y el aprendizaje maquinal, lo que ha dado lugar a una amplia gama de técnicas. A continuación, se describen algunas de ellas:

- Resumen: Una técnica de PLN que genera una versión condensada de un texto, preservando su contenido esencial, lo cual contribuye a una mejor comprensión por parte de los lectores. Existen enfoques extractivos, que seleccionan fragmentos relevantes, y abstractivos, que reformulan el contenido mediante modelos de lenguaje.

- Extracción de palabras clave: Esta técnica se usa frecuentemente para identificar términos representativos del tema central del texto. Se utiliza en sistemas de recuperación de información, optimización para motores de búsqueda (SEO) y análisis empresarial.

- Tokenización: Procedimiento básico que divide un texto en unidades mínimas denominadas *tokens* (palabras, subpalabras o frases), constituyendo la base para tareas como el modelado de lenguaje, el análisis sintáctico y la representación semántica.

Estas técnicas no solo son esenciales para aplicaciones convencionales como

traductores automáticos o sistemas de búsqueda, sino que en el contexto de esta investigación resultan fundamentales para analizar patrones lingüísticos y construir representaciones semánticas en redes cognitivas, con el objetivo de detectar alteraciones del lenguaje asociadas a la enfermedad de Alzheimer (AD)

2.8. Aprendizaje maquinal

El aprendizaje maquinal es un subcampo de la IA, y se puede definir de forma general como el estudio sistemático de algoritmos y sistemas que mejoran su conocimiento o desempeño con experiencia [15]. Las garantías teóricas del aprendizaje de un algoritmo dependen de la complejidad de las clases de conceptos consideradas y del tamaño de la muestra de entrenamiento. Las técnicas de aprendizaje son métodos basados en datos que combinan conceptos fundamentales de la ciencia de la computación con ideas de estadística, probabilidad y análisis de datos [23]

Los métodos de aprendizaje maquinal utilizan patrones de entrenamiento para aprender o estimar la forma de un modelo clasificador. Dependiendo de la disponibilidad de los datos de entrenamiento y el resultado deseado de los algoritmos de aprendizaje, los algoritmos de aprendizaje maquinal se clasifican en aprendizaje supervisado y aprendizaje no supervisado.

En el aprendizaje supervisado, los pares de entrada y objetivo de salida son dados para entrenar una función, y un modelo de aprendizaje se entrena de tal forma que la salida de la función pueda predecirse con un costo mínimo. Los métodos de aprendizaje supervisado se agrupan en función de sus estructuras, como las redes neuronales artificiales (RNA), las máquinas de vectores de soporte (MVS) y los árboles de decisión.

Las RNA son modelos compuestos por capas de unidades (neuronas) que transforman secuencialmente las entradas mediante funciones de activación no lineares con el fin de aprender representaciones desde datos complejos. Las RNA pueden aproximar funciones arbitrarias siempre que tengan suficiente capacidad y datos para entrenamiento [23].

Por su parte, las SVM son técnicas utilizadas para clasificar tanto datos lineales como no lineales. Su funcionamiento se basa en aplicar una trans-

formación no lineal que proyecta los datos originales a un espacio de características de mayor dimensión, lo que permite encontrar una separación más efectiva entre clases [11].

El uso de los algoritmos de aprendizaje maquinal en investigaciones sobre el deterioro cognitivo producido por la enfermedad de Alzheimer, ha permitido desarrollar sistemas expertos que tienen la capacidad de emular la toma de decisiones de los médicos, ya se con base a la identificación de patrones inherentes a partir de datos de neuroimagenes o en la detección de anomalías en las narrativas del habla de los pacientes afectados por la enfermedad [3] [31].

2.9. Algoritmo de agrupamiento K-medias

El análisis de agrupamiento tiene como objetivo dividir un conjunto de observaciones en grupos ("*clusters*") de manera que las diferencias internas entre los elementos de un mismo agrupamiento sean menores en comparación con las diferencias entre elementos pertenecientes a agrupamientos distintos.

El método de agrupamiento K-medias es una técnica utilizada para identificar agrupamientos y centros de agrupamiento en un conjunto de datos no etiquetados. Para su aplicación, se define previamente el número deseado de centros, denotado como K, y el procedimiento K-medias ajusta de manera iterativa dichos centros con el objetivo de minimizar la varianza total dentro de cada agrupamiento, para ello ejecuta de manera alternada los siguientes pasos:

- Asignación: para cada K, se identifica el subconjunto de puntos de datos (su agrupamiento) que se encuentra más cercano a dicho centro que a cualquier otro.
- Actualización: se calcula la media de cada característica correspondiente a los puntos del agrupamiento, y el vector resultante se establece como el nuevo centro del mismo.

Estos dos pasos se repiten de manera iterativa hasta alcanzar la convergencia. Comúnmente, los centros iniciales corresponden a K puntos seleccionados aleatoriamente a partir del conjunto de datos [18]. Este procedimiento

se ilustra en la Figura 2.6.

Algorithm 1 *K*-means clustering algorithm

Input: K, number of clusters; D, a data set of N points
Output: A set of K clusters

1. Initialization.
2. **repeat**
3. **for** each point p in D **do**
4. find the nearest center and assign p to the corresponding cluster.
5. **end for**
6. update clusters by calculating new centers using mean of the members.
7. **until** stop-iteration criteria satisfied
8. **return** clustering result.

Figura 2.6: Algoritmo K-medias. [19].

La Figura 2.7 muestra un ejemplo de agrupamiento K-medias en un conjunto de puntos, con K = 2.

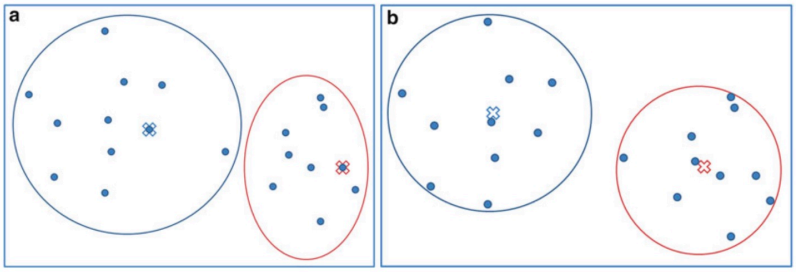


Figura 2.7: Ejemplo de agrupamiento K-medias (K = 2). El centro de cada grupo está marcado con una "x". (a) Inicialización. (b) Reasignación. [19]

En este caso, los agrupamientos se inicializan seleccionando aleatoriamente dos puntos como centros.

2.10. Métricas de evaluación

Para evaluar la calidad del agrupamiento generado por el algoritmo K-medias, resulta fundamental calcular métricas que permitan determinar en qué medida los clústeres obtenidos representan la estructura real de los datos. Las métricas de evaluación utilizadas para este análisis son:

- Exactitud: Mide el porcentaje de asignaciones correctas, es decir, cuántos sujetos fueron correctamente clasificados en el grupo correspon-

diente con respecto al total de casos evaluados. Esta métrica se calcula mediante la ecuación:

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precisión: Evalúa la proporción de sujetos correctamente clasificados dentro de cada clúster, midiendo qué tan confiable es la clasificación de un grupo específico y se calcula mediante la ecuación:

$$Precision = \frac{TP}{TP + FP}$$

En las ecuaciones TP, TN, FP y FN representan Verdadero Positivo , Verdadero Negativo, Falso Positivo y Falso Negativo, respectivamente.

Capítulo 3

Estado del arte

En este apartado se lleva a cabo una revisión de las investigaciones que se han hecho sobre los temas concernientes a este trabajo de investigación; cuyo contenido aborda los conceptos involucrados en el deterioro cognitivo, así como algunas técnicas de inteligencia artificial que se han utilizado para la detección de la enfermedad de Alzheimer en etapas tempranas.

3.1. Alteración del lenguaje asociado al alzhéimer

En el artículo "El déficit lingüístico en personas con demencia de tipo alzhéimer: breve estado de la cuestión" [27] se plantea como propósito crear una perspectiva global sobre el proceso de deterioro comunicativo en presencia de la demencia de tipo Alzheimer (DTA), enfatizando la importancia del análisis lingüístico en el diagnóstico de la enfermedad, ya que existe la posibilidad de encontrar cambios en las habilidades discursivas desde la etapa prodrómica de la demencia. Asimismo, el autor pretende tomar las principales aportaciones de las áreas de Neurología, Psicología o Lingüística que han coadyuvado a describir las alteraciones lingüísticas en personas con DTA.

En primera instancia, se señala que la demencia puede considerarse "un deterioro adquirido de varias funciones cognitivas, no necesariamente permanente ni progresivo, y de una importancia tal que impide al paciente el correcto desenvolvimiento de sus actividades sociolaborales". Las funciones

cognitivas comprenden fundamentalmente “la memoria, la gnosis, el lenguaje, el cálculo, la praxia ideomotriz, la habilidad visuoconstructiva o visuomotora, el pensamiento abstracto y la capacidad de razonamiento” [27].

Las características y diagnóstico de la enfermedad se enfocan en la memoria, en la habilidad visuomotora, y en la capacidad de razonamiento que afectan tanto a la capacidad comunicativa del hablante como a los resultados que se obtienen en los tests de habilidades lingüísticas.

Para realizar el diagnóstico de la DTA, existen test neuropsicológicos que evalúan el deterioro cognitivo y test específicos para analizar el lenguaje. Siendo que el test de las habilidades comunicativas ha demostrado ser relevante para la detección de la DTA, debido a que algunos estudios señalan que las alteraciones discursivas probablemente sean uno de los primeros signos de la enfermedad. Aparentemente, en la fase preclínica de la DTA, en el lenguaje de las personas se observan enunciados incompletos e incoherentes y surgen repeticiones de lo ya dicho.

De forma general, los estudios clínicos indican que los componentes semántico y pragmático del lenguaje sufren alteraciones en las DTA, esto se debe a que las habilidades semánticas y pragmáticas requieren procesos más complejos que dependen de otras funciones cognitivas deterioradas, como la memoria o la capacidad de planificación y razonamiento. Con base a estudios realizados en informantes que se situaban en distintos estadios de la enfermedad, se proponen tres fases en cuanto al deterioro comunicativo:

1. Etapa inicial. La duración de esta etapa suele ser entre dos y cinco años. Se presentan problemas de denominación y acceso a determinada palabra (anomia); el lenguaje se vuelve impreciso (ligera verbosidad o logorrea); existe una disminución en la iniciativa del habla y una dificultad creciente para detectar recursos humorísticos e irónicos.

2. Etapa moderada. Esta etapa puede durar aproximadamente entre tres y cinco años. Se agravan los déficits comunicativos existentes en la primera etapa, el individuo crea numerosos neologismos, abundan las parafasias semánticas (sustitución de un término por otro de significado similar o relacionado), hay presencia de circunloquios dando lugar a un discurso vacío de significado e incoherente y se hacen más evidentes los problemas de com-

prensión verbal.

3. Etapa avanzada. No existe una aproximación de la duración de esta etapa, ya que, es muy variable en el tiempo. Decrece el lenguaje espontáneo de la persona debido a una postura apática, la anomia está generalizada y surgen parafasias fonológicas; el discurso está repleto de ecolalias (repetición involuntaria del discurso ajeno) y glosomanías (desarrollo exclusivo de los temas de conversación preferidos). La sintaxis se ve afectada dando lugar a multitud de estereotipias (repetición involuntaria de cierta unidad lingüística, ya sea un sintagma, un neologismo, etc.). Asimismo, la comprensión se altera, lo cual implica que la persona no atiende a ordenes simples. En cuanto a la habilidades pragmáticas, la persona muestra poca iniciativa comunicativa derivando finalmente a un estado de mutismo.

Por otra parte, los componentes lingüísticos y las destrezas semióticas que pueden verse afectados por la demencia son el *déficit* fonético-fonológico, el déficit morfosintáctico, el déficit léxico-semántico, el déficit pragmático y el déficit en algunas destrezas semióticas, siendo afectados principalmente los tres últimos por la DTA.

La afectación del déficit léxico-semántico se ve reflejado en los primeros síntomas de tipo lingüístico, los cuales se relacionan principalmente con una disminución del vocabulario y una dificultad para recuperar la palabra buscada. Asimismo, aparecen los primeros problemas de denominación y de fluidez verbal y el uso de palabras relacionadas, pero no estrictamente sinónimas, en lugar de la palabra adecuada. El progreso del deterioro semántico finalmente puede hacer que las personas no responden a la petición de que nombren un objeto o una persona determinados.

El daño provocado en el déficit pragmático por la enfermedad DTA se ve reflejado en el aumento del uso de actos de habla directivos (peticiones y preguntas) y además se manifiesta una pequeña disminución de la iniciativa del habla por parte de las personas afectadas por la enfermedad. Por último, otra situación que revela el deterioro en este déficit es la falta de coherencia que muestran los textos producidos por las personas demenciadas debido a la alteración de la capacidad para organizarlos.

La afección en el déficit en algunas destrezas semióticas se presenta en la

última fase del deterioro lingüístico con aparición de la repetición de parte del discurso del interlocutor (ecolalia) y por la repetición de unidades o sintagmas producidos por el mismo individuo (perseverancia).

La conclusión de este artículo refiere que las investigaciones sobre el deterioro lingüístico en personas con DTA muestra la relevancia de la evaluación del leguaje en el momento de diagnosticar la enfermedad y de establecer diferentes etapas. Asimismo, se espera que, mediante los análisis de los componentes lingüísticos léxico-semántico y el pragmático, los cuales son los más deteriorados, se puedan crear protocolos de conversación con la finalidad de mejorar la calidad de vida de los pacientes y de sus familias.

3.2. Redes léxico-semánticas en el envejecimiento típico

Una propuesta para el análisis del funcionamiento de la memoria semántica en adultos mayores, a partir de la visión teórica de las redes léxicas, se plantea en el artículo “Estudio de la memoria semántica en adultos mayores desde la perspectiva de las redes léxicas” [16]. En este trabajo, los autores contemplan diferentes tipos de estructuras de red con el objetivo de comprender cómo se almacenan las palabras que una persona conoce, la manera en que estas se relacionan entre sí y el modo en que son recuperadas durante el proceso cognitivo. Asimismo, tratan la relación entre la estructura y la organización del léxico con otras habilidades cognitivas y la aplicación de las redes léxicas en la psicolingüística.

El proceso de envejecimiento en las personas con lleva una serie de cambios estructurales y funcionales a nivel físico, mental y psicosocial. Aunque, esto representa un deterioro en el funcionamiento cognitivo, las funciones cognitivas relacionadas con el lenguaje son las menos afectadas. El lenguaje está relacionado con la memoria semántica, la cual involucra el conocimiento de las palabras y los conceptos que hacen referencia a eventos u objetos, así como a las relaciones entre dichas palabras, conceptos y símbolos, es decir es la encargada de almacenar y recuperar la información lingüística a través del vocabulario de las personas [16].

Se considera que la memoria semántica está organizada en redes léxicas que

pueden ser analizadas mediante un método que posibilita estudiar la relación entre una palabra estímulo y su respuesta. Aunque el debilitamiento de la memoria y el daño cognitivo general se ha estudiado desde diversos enfoques, experimentalmente se suelen estudiar a partir de la comparación entre el desempeño de tareas lingüísticas y cognitivas en términos operacionales. Con base a dichos estudios se encontró que, a pesar del decaimiento de las habilidades relacionadas con la memoria y el procesamiento en la vejez, el vocabulario se encuentra en un mejor estado que el de los jóvenes.

Asimismo, existe evidencia que el tamaño del vocabulario se incrementa con respecto a la edad y el nivel educativo de una persona. Con base a lo anterior se plantea el entender cómo se almacenan las diferentes palabras que un individuo conoce, la forma en cómo éstas se interrelacionan entre sí, así como la manera de ser recuperadas por la persona. Los autores plantean que la solución teórica actual es proponer que las palabras (en fondo y forma) se encuentran almacenadas en redes, donde cada palabra equivale a un nodo, y la relación entre palabras -redes léxicas- así como la semántica está determinada por las uniones (arcos) entre uno o varios nodos (forma jerárquica) [16].

El primer tipo de red léxica que se contempla es la red jerárquica, donde el léxico se encuentra organizado en árboles jerárquicos. En esta red, las palabras sustantivas (el todo) fungen como nodos desde los cuales se subordinan otras palabras/nodos (al igual sustantivas) que corresponden al significado del todo. La Figura 3.1 representa un ejemplo de este tipo de red. En esta se puede observar que la palabra animal subordina a pájaro y pez y, al mismo tiempo, cada uno posee ramificaciones que las vinculan con sus respectivas características semánticas.

Este modelo propone que la fuerza de la conexión entre animal y pájaro es diferente en magnitud a la que existe entre animal y pez; lo que implica que es más fácil recuperar y asociar algunas palabras con relación a otras. No obstante, este modelo no considera las palabras que tienen más de un significado.

Otro modelo considerado por los autores es el de ordenamiento semántico, el cual se basa en la teoría de la propagación de la activación, donde la activación de un nodo activa indirectamente a aquellos conectados a él de

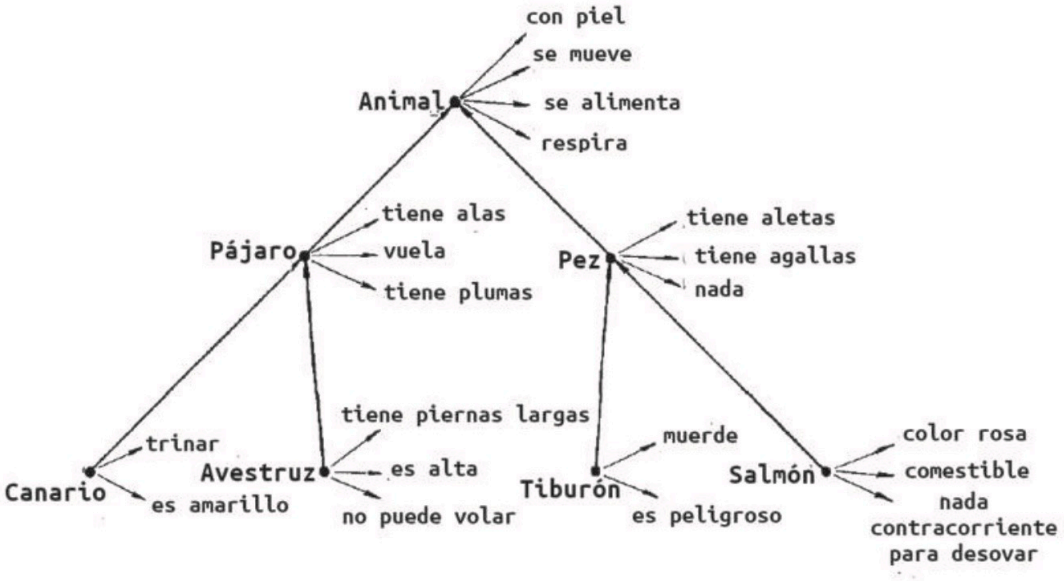


Figura 3.1: Ejemplo de red jerárquica. [16]

manera proporcional a su fuerza de conexión. En este modelo los nodos se agrupan y conectan a partir de su cercanía semántica,

Esta red léxica permite la conexión entre palabras de diferentes categorías gramaticales y posibilita conexiones similares a las coocurrencias léxicas, así como a las relaciones sintácticas del habla. Además, propone representar de una forma gráfica, la magnitud de la asociación (conexión) entre dos nodos, donde entre más grande la longitud del arco entre dos nodos, la magnitud de la conexión entre ellos será más débil. Un inconveniente es que algunos nodos se encuentran muy distanciados entre sí, por lo que haría falta propagar la activación por una cantidad variable de nodos, esto implicaría mayores recursos cognitivos y temporales cada que se acceda a la memoria semántica.

El último modelo planteado en este artículo está basado en la teoría de grafos. Este tipo de red está fundamentada en la estructura de *small world* en donde cada elemento de una red se puede conectar indirectamente a cualquier otro nodo mediante una cantidad mínima de nodos. En estas redes, como en los modelos anteriores, los nodos representan las palabras y los arcos la conexión entre ellos.

En este enfoque se plantea que la unidad para medir el peso de la asocia-

ción o la dirección de la conexión dependerá de la implementación y sus fundamentos teóricos.

Asimismo, se han realizado estudios de las relaciones léxicas en población adulta, lo cual posibilita tener una perspectiva del estado de la memoria semántica en el envejecimiento. Una de estas investigaciones contempló una muestra de 738 participantes angloparlantes de entre 18 y 87 años de edad con diferentes niveles de educación, dividiéndolos en dos grupos con respecto a la edad: el primer grupo se integro con personas de 18 a 21 años y el segundo con personas de 55 a 87 años. El estudio consistió en aplicarles una tarea de normas de asociación con 100 palabras estímulo para analizar las asociaciones de palabras en las diferentes etapas de la adultez.

Los resultados del estudio, mostró que los adultos jovenes tienden a dar menos respuestas distintas ante un estímulo en comparación con los adultos mayores. Esto puede deberse a que los adultos mayores poseen un vocabulario más amplio. No obstante, se afirmó que estos resultados se pudieron deber a que los adultos sostengan un mayor individualismo, a comparación de los jovenes que buscan alinearse a los rasgos de un grupo con el objetivo de pertenecer a dicho grupo. Con base a estas ideas, se llevó a cabo otro estudio, pero en este caso se dividió a los participantes en tres grupos de acuerdo a su edad: el primero de 18 a 33 años, el segundo de 34 a 49 años y el tercero de 50 a 87 años. Los resultados obtenidos demostraron que, conforme se incrementa la edad de los participantes, se aprecia un decremento en el uso de respuestas más usuales.

Otro de los estudios consistió en un análisis comparativo entre un grupo de adultos jóvenes de entre 18 a 33 años y otro grupo de adultos mayores de entre 62 a 87 años enfocado en hablantes nativos del inglés; sobre las relaciones de palabras que se generan en una tarea de asociación libre. El análisis consideró las normas de asociación para 113 palabras, entre ellas sustantivos, verbos, adjetivos y adverbios como palabras estímulo. La investigación obtuvo resultados similares al estudio descrito anteriormente, indicando que los adultos mayores generaron un mayor número de respuestas única, con esto se confirmó que el vocabulario es un buen predictor de la variabilidad de en las respuestas y no la edad en si misma.

Posteriormente, se volvió a realizar este estudio entre dos y tres meses des-

pués con una submuestra de los participantes; subrayando que las respuestas idiosincrásicas fueron repetidas solo en un 7 por ciento de las veces para los adultos jóvenes y 8 por ciento para los adultos mayores. Los resultados obtenidos demostraron que los adultos mayores no revelaron problemas de recuperación y con base a las consistencia de las respuestas expresadas en los dos momentos diferentes de la recolección de los datos se probó que las respuestas se generaron automáticamente desde la memoria semántica.

Otro estudio que se realizó con personas británicas residió en la comparación de normas de asociación de palabras entre un grupo de adultos jóvenes de 21 a 30 años y otro de adultos mayores de 66 a 81 años, los cuales respondieron a 90 palabras estímulo conformadas por sustantivos clasificados como concretos y no concretos. En los resultados se obtuvieron más respuestas de los adultos jóvenes para cada estímulo, así como más respuestas únicas en comparación con los adultos mayores. Esto llevó a la conclusión que los adultos jóvenes fueron menos consistentes en sus respuestas que los adultos mayores.

Los autores, con base en los resultados de los estudios descritos anteriormente, refieren que la memoria semántica se encuentra preservada en los adultos mayores y que se presentan pocos cambios en la organización de las palabras en la misma. Siendo estos estudios una evidencia de que las conexiones de la red semántica en los adultos mayores son resistentes a los déficits cognitivos generales.

Por otro lado, también se efectuaron algunos estudios que implican el vínculo entre las relaciones léxicas y otras habilidades cognitivas en el proceso de envejecimiento. Con respecto a este tema, se llevó a cabo una revisión de 91 estudios contenidos en 75 artículos de investigación, explicando los vínculos existentes entre las diferentes habilidades cognitivas y comparando dos tipos de modelos de procesamiento cognitivo con respecto a la edad. Basados en este análisis se halló la existencia de relaciones negativas entre la edad y el desempeño en tareas cognitivas, debido a la pérdida en la velocidad de procesamiento y la disminución en la capacidad de almacenamiento, los cuales son dos factores que están implicados en el deterioro de los aspectos que se relacionan con la edad.

Un estudio enfocado en la velocidad de procesamiento que consistió en dos

pruebas de velocidad perceptual con el fin de medir la velocidad de procesamiento a través del tiempo de reacción; indicó como resultado que las variables de velocidad (biológicas y conductuales) están altamente relacionadas con la edad. Basándose en el resultado, se pudo deducir que la memoria y la velocidad son factores que permiten conocer el estado del procesamiento cognitivo.

Con la finalidad de profundizar en la relación de la memoria y la velocidad de procesamiento con otras habilidades cognitivas en el proceso de envejecimiento, se hizo un estudio de comprensión del lenguaje con el objetivo de analizar los efectos de la edad. Esta investigación radicó en la evaluación de la capacidad de memoria de trabajo, así como de la velocidad de procesamiento en dos grupos de participantes, uno de adultos jóvenes y otro de adultos mayores. Obteniendo como resultado que, las medidas estándar de memoria de trabajo están relacionadas con medidas de comprensión del lenguaje global y que el tiempo de escuchar, lo cual se refiere a la diferencia entre la duración del estímulo y el tiempo de reacción, aumenta con la edad.

En otra investigación que contempla el estudio de la relación entre la memoria verbal y la velocidad de procesamiento en adultos con deterioro cognitivo leve mediante una tarea de asociación libre; arrojó como resultado que la capacidad de memoria verbal está asociada a la velocidad de procesamiento de las relaciones léxicas.

A partir de esto, se han propuesto metodologías que ofrecen una explicación sobre la memoria semántica y su estructura en forma de red (jerárquica) o mediante un ordenamiento por similitud semántica y posteriormente con el desarrollo de la teoría de grafos y su aplicación a la psicolingüística. Dichos estudios sobre las relaciones léxicas en adultos mayores han permitido entender como las personas organizan los conceptos, las palabras, los objetos y símbolos que conforman el mundo.

Sin embargo los autores consideran que aún se requieren más estudios sobre el almacenamiento, la estructura del léxico y sus relaciones. De igual forma, contemplan que los datos que se tienen sobre la evolución de estos elementos en el proceso de envejecimiento son aún menos concluyentes. Asimismo, mencionan que las tareas de asociación libre de palabras son una herramienta básica para explorar el estado de las redes léxicas.

Además, concluyen que la forma clásica de aproximación al estudio de las relaciones entre palabras ha mostrado información interesante sobre la organización de las redes léxicas. Aunque existe la posibilidad que, mediante el análisis de distintos parámetros de las redes small world se puedan revisar las diferencias más importantes entre la estructura léxica de los individuos en las distintas etapas de la vida.

3.3. Exploración de redes léxicas desde la infancia hasta la adultez

Las redes léxicas describen las relaciones que se establecen entre palabras, las cuales reflejan la organización del léxico mental, la memoria semántica y el procesamiento del lenguaje. Su estudio y formación habían sido exploradas mediante métodos empíricos y experimentales solo en la etapa adulta; no obstante, investigaciones recientes se han enfocado en la etapa infantil para comprender cómo emergen y se propagan, como se expone en “Métodos para explorar las redes léxicas desde la infancia hasta la etapa adulta” [6].

Este artículo contempla ciertos métodos que tienen la finalidad de estudiar las redes léxicas, cuyo desarrollo parten de diferentes enfoques teóricos. Estas metodologías son: nombramiento, decisión léxica, norma de asociación de palabras y priming, técnicas de tipo electrofisiológico mediante componentes como el N400, asociado a la incongruencia semántica. Asimismo, existen herramientas de grafos y coeficientes de correlación apoyadas en teorías de small worlds.

Las metodologías de nxombramiento, decisión léxica y asociación de palabras se basan en la presentación de palabras y/o imágenes como estímulo para el análisis de la relación que existe entre las palabras y/o imágenes; ya sea una relación semántica o asociativa.

En cambio, la metodología priming (facilitación) está basada en paradigmas de rastreo visual, lo cual permite reconocer a una palabra más rápido cuando esta es precedida por una palabra relacionada a cuando lo es por una no relacionada. La técnica componente N400 determina mediante una onda de actividad eléctrica cerebral la ausencia de relaciones léxicas o sintácticas entre palabras cuya mayor amplitud (negativa) se ha relacionado con la

incongruencia semántica.

Por otro lado, en la sintaxis, las gráficas para la representación de redes léxicas, se emplean para formalizar las relaciones de orden a partir de gráficas jerárquicas o árboles; en la morfología, las gráficas se han utilizado para representar relaciones entre conjugaciones, y en la semántica, es habitual representar las relaciones de significado a partir de gráficas no dirigidas o jerárquicas. Con las medidas de las gráficas se espera obtener información relevante sobre el comportamiento léxico de los hablantes ya que estas deben ser predictores de la accesibilidad y la similitud dentro del léxico mental.

Las principales conclusiones que señalan los autores es que la elección de cada método presentado en este artículo, depende de varios factores; como la pregunta de investigación planteada, del equipo técnico, y de los procesos asociados que se busca indagar. Además, de que cada método sobresale por sí solo según su grado de sistematicidad, su exactitud temporal, su ecología experimental o, bien, su indice de confiabilidad matemática. Además, resaltan que cada uno de los métodos pertenece a diferentes del conocimiento cómo la psicología, la lingüística, la neuropsicología, el área de las ingenierías, las matemáticas y la Inteligencia Artificial. Debido a esto, la contribución a las investigaciones de las redes léxicas ha sido enriquecida desde distintos disciplinas del conocimiento, lo cual coadyuva a que la investigación sea de más alto impacto.

3.4. Análisis lingüístico con procesamiento del lenguaje natural y aprendizaje maquinal para la detección del alzhéimer

Un diagnóstico preciso y en etapas tempranas de la enfermedad de alzhéimer (AD, por sus siglas en inglés *Alzheimer's disease*) puede permitir una intervención médica oportuna que contribuya a ralentizar su progresión mediante tratamientos adecuados. En la actualidad, diferentes investigaciones se han enfocado en el cambio del lenguaje como una señal temprana de que las funciones cognitivas de un paciente están afectadas, lo que potecialmente conduce a una detección temprana.

El deterioro del lenguaje en pacientes con AD se ha estado estudiando a través del análisis de patrones lingüísticos, mediante técnicas de procesamiento de lenguaje natural (PLN) y aprendizaje maquinal. No obstante, la mayoría de las investigaciones en esta área se han centrado en el idioma inglés, sin considerar lenguas locales como el nepalí, que poseen características lingüísticas particulares, como una morfología compleja y un sistema de escritura distinto (devánagari).

Estas particularidades requieren el desarrollo de recursos y enfoques específicos para su análisis. En respuesta a ello, el artículo "*Exploiting linguistic information from Nepali transcripts for early detection of Alzheimer's disease using natural language processing and machine learning techniques*" [3] propone la creación de un conjunto de datos en idioma nepalí para su posterior análisis mediante técnicas de PLN y aprendizaje maquinal, con el objetivo de facilitar la detección temprana de la AD.

El proceso de esta investigación comienza con la recopilación de los datos. Para ello, se obtuvieron transcripciones en idioma inglés del Pitt Corpus del DementiaBank [33] . En este estudio se utilizó las transcripciones basadas en la descripción de una imagen estímulo, títulada *Cookie Theft* (Robo de Galletas). Se pidió a cada participante que describiera la escena de la cocina como se puede observar en la Figura 3.2.

Figura 3.2: Imagen estímulo *Cookie Theft.* [3]

Se consideraron en total 296 transcripciones, de las cuales 168 corresponden a pacientes afectados por la AD y 98 personas normales de control (CN). La información caracteristica de los sujetos en estudio se muestra en la Tabla

3.1, así como la media de la edad, educación y puntos obtenidos en el Mini Mental State Examination (MMSE), junto con su correspondiente desviación estándar.

Después, las transcripciones se tradujeron al idioma nepalí de forma manual por hablantes nativos de nepalí que tenían al menos 13 años de educación formal en nepalí y fueron evaluadas y verificadas por un experto lingüístico independiente con la finalidad de conservar al máximo las características lingüísticas, la persistencia y la afirmación del tono general de los textos.

Atributos	**Control normal (CN)**	**Enfermedad de Alzheimer (EA)**
Número de participantes	98 sujetos CN	168 pacientes con EA
Género	31M/67F	55M/113F
Edad	64.7 (7.6)	71.2 (8.4)
Educación	14.0 (2.3)	12.2 (2.6)
MMSE	29.1 (1.1)	19.9(4.2)

Tabla 3.1: Información demográfica de los sujetos del DementiaBank. [3]

El siguiente paso, fue el preprocesamiento de las transcrpciones. En este caso, sólo se eliminaron todos los signos de puntuación y dejando las palabras "vacías" como: 'y', 'por lo tanto', etc. puesto que, los investigadores consideraron que este tipo de palabras preservan las características lingüísticas de los pacientes con AD.

Terminado el preprocesamiento, se prosiguió a la vectorización de las palabras. Este procedimiento permitió la extracción de las características de las mismas para su posterior clasificación. Para ello, se utilizaron dos métodos de vectorización: CountVectorizer y Term Frequency Inverse Document Frequency (TF-IDF) y dos técnicas de *word embedding*: Word2Vec y fastText. Para el modelo Word2Vec se utulizó el modelo creado por Rabindra Lamsal y la arquitectura de bolsa de palabras continua (CBOW, *continuous bag-of-words*). Este modelo previamente entrenado contiene vectores de 300 dimensiones para más de 0,5 millones de palabras y frases en nepalí.

Para los embeddings de fastText se usaron los vectores de palabras previamente entrenados en Common Crawl y Wikipedia. El modelo se entrenó al

igual que el modelo Word2Vec usando CBOW y con vectores de 300 dimensiones. Asimismo, se usaron n-gramas de caracteres de longitud 5, una ventana de tamaño 5 y 10 negativos.

Posteriormente, para la clasificación de las transcripciones de los sujetos CN y los pacientes con AD, se emplearon modelos de aprendizaje maquinal y aprendizaje profundo, con la finalidad de encontrar el modelo con mayor precisión, en cuanto a la clasificación. Los algoritmos de aprendizaje maquinal que fueron usados se listan a continuación:

- Decision Tree (DT)
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Naïve Bayes (NB)
- Random Forest (RF)
- AdaBoost
- XGBoost (XGB)

Con respecto a los modelos de aprendizaje profundo, se utilizaron tres algoritmos:

- Convolutional Neural Network (CNN)
- Bidirectional Long Short-Term Memory (BiLSTM)
- Combination of CNN and BiLSTM

Finalmente, para obtener el desempeño de los algoritmos usados en la investigación, los autores optaron el uso de la validación cruzada binaria como función de pérdida. Después de una validación cruzada estratificada de 10 veces, el rendimiento de los algoritmos propuestos se midió empleando cuatro métricas de evaluación:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{–}score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

donde TP, TN, FP y FN representan Verdadero Positivo, Verdadero Negativo, Falso Positivo y Falso Negativo respectivamente.

En los resultados obtenidos con los algoritmos de aprendizaje maquinal, el clasificador Naive Bayes logró el mejor rendimiento en ambos métodos de vectorización. Tanto CountVectorizer como TF-IDF condujeron al mismo rendimiento del modelo, alcanzando una puntuación F1 de 0.940.

En cuanto a los embeddings de palabras previamente entrenados, Word2Vec obtuvo el mejor rendimiento con el algoritmo XGBoost, logrando una puntuación F1 de 0.828. Por otra parte, con fastText previamente entrenado, el clasificador SVM superó a otros modelos con una puntuación F1 de 0.934.

Con base en los resultados anteriores, los autores infirieron que los métodos de vectorización obtuvieron un mejor rendimiento que los modelos basados en *word embeddings*, debido a que la similitud entre palabras no se representó adecuadamente, ya que el corpus era pequeño para entrenar dichas representaciones.

Por otro lado, en los resultados arrojados con los algoritmos de aprendizaje profundo; Word2Vec obtuvo una puntuación de F1 de 0,964 con Kim's Architecture. En cuanto, al modelo preentrenado fastText, con BiLSTM superó a otros modelos con una puntuación F1 de 0,887. Basándose en estos resultados, los algoritmos de aprendizaje profundo mostraron un desempeño ligeramente superior al utilizar *word embeddings* específicos del dominio, en comparación con aquellos previamente entrenados. Esto se debe a que dichas representaciones se forman a partir del propio corpus, lo que permite capturar con mayor precisión las particularidades léxicas y, en consecuencia, mejorar el rendimiento.

En cuanto a los resultados de los métodos de vectorización para los algoritmos de aprendizaje profundo, CountVectorizer superó a TF-IDF con la arquitectura CNN de Kim alcanzando una puntuación F1 de 0,897.

Además, también se aplicaron mecanismos de atención en los modelos de aprendizaje tanto para word embeddings como para los métodos de vectorización, siendo que CNN con Word2Vec obtuvó el mejor desempeño con una puntuación F1 de 0.968. Con base a los resultados se observó que, el mecanismo de atención dio mejores resultados, implicando que se le dio más peso a aquellas palabras que tenían más importancia en la oración. Asimismo, con este mecanismo, el desempeño de Word2Vec de dominio específico fue mejor que los word embeddings previamente entrenadas.

Con respecto a la aplicación de los mecanimos de atención en CountVectorizer, CNN alcanzó la puntuación F1 más alta con 0.766. Por lo tanto, se pudó inferir que los métodos de vectorización no tuvieron un buen desempeño como los word embeddings con este tipo de mecanismo.

Considerando los resultados de la investigación, se pudó observar que el modelo con el mejor desempeño fue la CNN basada en la atención con Word2Vec de dominio específico. Por lo que, los investigadores infieren que este modelo se puede utilizar para crear una aplicación intuitiva para los médicos, la cual, ayudaría a identificar la AD en sus primeras etapas, así como un canal con un mecanismo para la síntesis del habla con las metodologías proporcionadas para mejorar la detección de la enfermedad.

A modo de conclusión, los autores puntualizan que el modelo proporciona una metodología para la detección temprana de la AD utilizando un corpus traducido de forma manual en el idioma nepalí. Y con un modelo de este tipo se podría etiquetar, identificar y proporcionar un sistema de detección temprana de la enfermedad en idiomas de bajos recursos. Esto significaría predecir la presencia de la AD en un tiempo mucho menor reduciendo los costos del tratamiento y poder usarse durante una gran cantidad de ciclos para muchas personas.

De igual forma, subrayaron que el proyecto pudo alcanzar una precisión razonable inclusive sin recopilar datos específicamente en el idioma local.

Algunos de los trabajos futuros considerados por los autores es lograr que

los médicos verifiquen y validen el corpus traducido como una representación del uso real del idioma del paciente. Asimismo, buscar la posibilidad de combinar enfoques alternativos como el reconocimiento de imágenes de neuroanatomía basado en resonancias magnéticas, para el desarrollo de sistemas multimodales. Además, consideran que el trabajo se podría ampliar a otras formas de demencia, como el Parkinson entre los pacientes nepalíes.

3.5. Análisis del lenguaje mediante grafos en tareas de fluidez verbal

La fluidez verbal, de acuerdo con lo expuesto en el artículo "*Graph analysis of verbal fluency test discriminate between patients with alzheimer's disease, mild cognitive impairment and normal elderly controls*" [8], se define como la habilidad para generar una secuencia adecuada de palabras habladas en un período de tiempo determinado. Las pruebas de fluidez verbal están diseñadas para evaluar esta habilidad, mediante la producción de palabras que inician con una letra específica (fluidez verbal fonética) o que pertenecen a una categoría semántica (fluidez verbal semántica).

La fluidez verbal semántica, se utiliza comúnmente para examinar la memoria semántica y las habilidades lingüísticas en adultos mayores, ya que depende de un buen funcionamiento del lenguaje, de la memoria semántica y de las funciones ejecutivas. Este tipo de fluidez está relacionada con la capacidad para predecir el declive cognitivo y funcional, asi como la transición de un deterioro cognitivo leve (MCI) a la AD.

En este artículo los autores proponen un análisis gráfico de datos generados mediante la prueba de fluidez verbal semántica aplicada a adultos mayores cognitivamente sanos (NC), pacientes con deterioro cognitivo leve —en sus subtipos amnésico (aMCI) y dominio múltiple amnésico (a+mdMCI)—, así como a pacientes con AD. Las secuencias de palabras se representan como un grafo del discurso, en el que cada palabra corresponde a un nodo y los enlaces temporales entre palabras se representan mediante aristas dirigidas.

Una red de fluidez verbal semántica normal está representada por un gráfico dirigido con solo una ocurrencia por cada palabra. El análisis de las propiedades de la red ayuda a comprender la dinámica y la organización de

los procesos cognitivos y conductuales. En este artículo se llevo a cabo un análisis de las propiedades de las redes obtenidas mediante las pruebas de fluidez verbal semántica de individuos con MCI o AD.

Basándose en lo anterior, los autores plantean la siguiente hipótesis: el análisis de las propiedades de la red de fluidez verbal semántica puede ayudar a discriminar mejor entre adultos mayores con rendimiento cognitivo normal, deterioro cognitivo leve o enfermedad de Alzheimer. Este método previamente había demostrado ser útil para identificar pacientes con esquizofrenia y trastorno bipolar [8].

En la realización de este estudio se incluyeron cien adultos mayores, los cuales fueron evaluados en el Centro de Referência à Saúde do Idoso Jenny de Andrade Faria, Hospital Clínico de la Universidad Federal de Minas Gerais. Todos los participantes se sometieron a una evaluación clínica y neuropsicológica integral. El protocolo fue validado para la evaluación neuropsicológica de adultos mayores con bajo nivel educativo. Los sujetos se dividieron en cuatro grupos: rendimiento cognitivo normal (NC), n = 25; MCI amnésico de un solo dominio (aMCI), n = 25; MCI amnésico de dominio múltiple (a+mdMCI), n = 25 y AD, n = 25.

El diagnóstico se basó en los siguientes criterios: el diagnóstico de AD, el paciente debería presentar un deterioro cognitivo general y que empeorar, en dos o más dominios cognitivos, y un deterioro funcional en las actividades de la vida diaria. El diagnóstico de MCI, el adulto mayor deebría presentar un deterioro cognitivo en uno o más dominios cognitivos, pero preservando las actividades básicas e instrumentales de la vida diaria o presentar un deterioro mínimo. Con respecto a la división de subgrupos de MCI se consideró la clasificación de MCI amnésico (aMCI) para los participantes que solo presentaban deterioro de la memoria, y el MCI amnésico de dominio múltiple (a+mdMCI) para los participantes que presentaban deterioro de la memoria y otros dominios cognitivos, aunque cumpliendo con todos los criterios de MCI establecidos.

La prueba de fluidez verbal semántica consistió en la categoría de animales. Se les pidió a los participantes que produjeran el máximos de nombres de animales dentro de 60 segundos, asimismo, se les dieron instrucciones explícitas/implícitas para evitar repeticiones. Posteriormente, se registraron

todas las palabras, incluidas las repeticiones y los errores. El procedimiento de puntuación incluyó: el total de palabras producidas, el total de palabras correctas, el total de errores, el total de repeticiones y la fracción de repeticiones de acuerdo con el total de palabras producidas por cada participante. Las puntuaciones en esta tarea no se tomaron en cuenta en ell diagnóstico de cada participante.

En cuanto, al diseño del estudio, los autores incluyeron dos etapas de análisis, considerando tres (NC, MCI, AD) o cuatro grupos (NC, aMCI, a+mdMCI, AD), y aplicando el mismo análisis estadístico y medidas gráficas para comparar los tres o cuatro grupos. El grupo MCI comprendió tanto el grupo aMCI como el grupo a+mdMCI. La clasificación de grupos se implementó con un clasificador Naïve Bayes. La elección de atributos para el clasificador se basó en correlaciones significativas de los atributos con medidas clínicas establecidas de diagnóstico diferencial (estado cognitivo global y funcionalidad en la vida diaria).

La secuencia de palabras producida en la prueba de fluidez verbal semántica se representa como un gráfico de habla, utilizando el software SpeechGraphs. El programa representa, en este caso, la secuencia de palabras producidas por la prueba de fluidez verbal como un gráfico, donde cada palabra es un nodo y el vínculo temporal entre las palabras es una arista como se muestra en la Figura 3.3.

Posteriormente, se calcularon la cantidad de palabras (WC) y 13 atributos adicionales del grafo de habla (SGA) que comprenden atributos generales: total de nodos (N) y aristas (E); componentes conectados: el componente fuertemente conectado más grande (LSC); atributos de recurrencia: aristas repetidas (RE) y aristas paralelas (PE), ciclos de uno (L1), dos (L2) o 3 nodos (L3); atributos globales: grado total promedio (ATD), densidad, diámetro, camino más corto promedio (ASP) y coeficiente de agrupación (CC).

De acuerdo a las instrucciones dadas, se esperaba que los sujetos generaran una secuencia lineal, es decir, una secuencia en la que cada palabra correcta fuera seguida por una palabra correcta diferente, sin repeticiones. Un rendimiento correcto en la prueba debería generar como resultado un grafo con un número de nodos (N) igual al de palabras (WC), N-1 aristas, sin aristas paralelas, repetidas o bucles ni componentes fuertemente conecta-

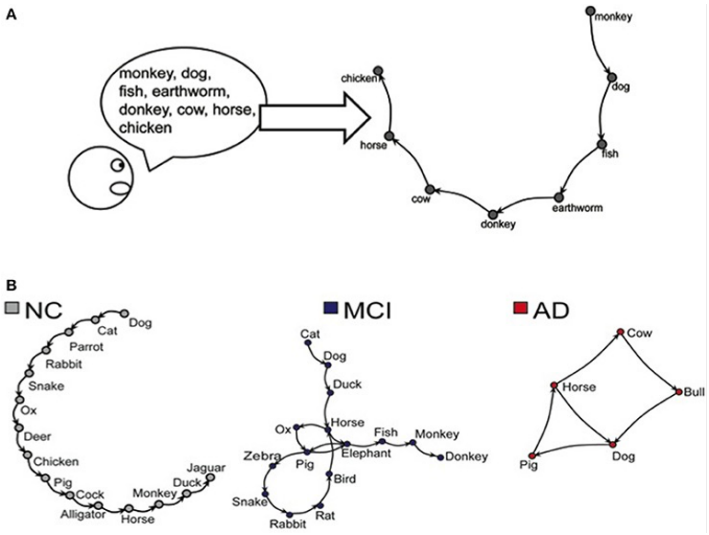


Figura 3.3: (A) Representación de la secuencia de palabras producida en la tarea de Fluidez Verbal Semántica. (B) Representaciones de redes generadas por sujetos NC, MCI y AD durante la tarea de Fluidez Verbal Semántica. [8]

dos, un grado total promedio (ADT) cercano a 2, una densidad y coeficiente de agrupamiento bajos, y un diámetro amplio igual al número de aristas (E).

Con respecto a los resultados del estudio, se encontró que los grupos mostraron diferencias significativas en el desempeño en las Actividades de la Vida Diaria (ADLs), en el estado cognitivo general,y en la generación de palabras correctas y totales, así como en las medidas del grafo de habla, incluyendo conteo de palabras, nodos, aristas, bucles de 3 nodos, diámetro, camino corto promedio y densidad. El grupo NC tuvo un mejor desempeño en las ADLs, obteniendo puntuaciones más altas en el MMSE (Mini Examen del Estado Mental), produjo una red con más nodos, diámetro más grande en la red y menos densa, en comparación con los grupos MCI y AD. El grupo MCI mostró un desempeño intermedio entre los grupos NC y AD en todas las medidas evaluadas.

Por último, la comparación de los cuatro grupos arrojó diferencias significativas en las ADLs, el estado cognitivo general, las palabras totales y correctas producidas, y en el recuento de palabras, los nodos, las aristas, el diámetro, el ASP y la densidad.

Por otro lado, los resultados del clasificador Naïve Bayes mostraron que una selección de SGA correlacionada con el deterioro funcional y cognitivo medi-

do por otros instrumentos, proporcionó un poder de clasificación de bueno a excelente, siendo similar al poder de clasificación del MMSE, o incluso mejor para la distinción entre los grupos NC y MCI. Cuando los SGA se asociaron con el Índice de Lawton (Escala que evalúa el grado de independencia para realizar las actividades de la vida diaria.) o el MMSE, el poder de clasificación aumentó; una combinación de las 3 mediciones proporcionó la máxima calidad de clasificación. En general, la combinación de medidas de grafos y dependencia funcional produjo una clasificación diferencial muy precisa del AD (1.00) y MCI (0.78) contra el grupo NC, y entre el MCI y el AD (0.84).

Asimismo, los resultados del clasificador Naïve Bayes indican que el SGA seleccionado tiene un buen poder de clasificación para el diagnóstico de subtipos de MCI en comparación con el envejecimiento cognitivamente saludable, y también una buena clasificación contra el grupo de demencia. Por otro lado, el SGA proporcionó una clasificación pobre cuando se utilizó para distinguir entre los dos subtipos de MCI. Cuando los SGA se combinaron con el Índice de Lawton, observamos un aumento en el poder de clasificación a través de los cuatro grupos, excepto entre los dos subtipos de MCI.

Por último, la combinación de los SGA con el MMSE mostró menos efectividad en comparación con la combinación con el índice de Lawton; la combinación de estas tres variables apenas mejoró la clasificación más allá de la combinación de SGA y Lawton. Estos resultados indican que la combinación de medidas de grafos y dependencia funcional nuevamente proporciona una buena clasificación entre los tres grupos (AUC = 0.71–0.85), excepto entre los subtipos de MCI (AUC = 0.47).

Con base en este trabajo de investigación los autores concluyen que los resultados demuestran que el SGA difiere significativamente entre los grupos AD, MCI y NC y podría usarse para clasificar a los grupos. Además que los resultados muestran el potencial del análisis gráfico de la tarea de fluidez verbal para discriminar entre estos grupos en la práctica clínica. Destacando, que algunos atributos de SGA podrían indicar la progresión del deterioro cognitivo y el deterioro funcional, como lo demuestran las redes más densas y más pequeñas, con un menor número de nodos, en sujetos con deterioro cognitivo más grave.

Por otra parte, se encontró que la correlación entre el SGA y el MMSE o el

Índice Lawton indica que el SGA está asociado con el estado cognitivo general y el rendimiento funcional, dos medidas clínicas importantes utilizadas en la evaluación geriátrica. En conclusión, con base a los resultados del estudio, se sugiere que los SGA podrían ser una herramienta útil para ayudar en el diagnóstico diferencial entre MCI y AD.

Sin embargo, los resultados delimitan un campo que necesita ser explorado más a fondo en estudios futuros, involucrando la densidad de las redes y la fuerza entre las palabras en la memoria semántica de los adultos mayores con envejecimiento patológico. Otro aspecto que consideran los autores como investigación adicional es la ausencia de diferencia entre los grupos en los atributos de conectividad (LSC, ATD y CC). Esto plantea la hipótesis de que incluso redes muy diferentes pueden compartir una estructura similar de conexiones locales, en la cual una pequeña porción de las palabras están altamente conectadas con otras palabras menos conectadas, manteniendo la integridad de la conexión general de la red.

Finalmente, considerando el análisis de grafos realizado en este estudio, construido a partir de una co-ocurrencia de palabras y basado en el vínculo temporal entre ellas, los autores contemplan que estudios futuros deberían considerar el análisis de escalado multidimensional y de agrupamiento jerárquico.

Estos tipos de análisis representarían la relación entre las variables y las agruparían, mejorando los resultados. También, plantean que los trabajos futuros deberían abordar las diferencias entre los pacientes con MCI y otras condiciones neurológicas en las que los deterioros cognitivos son bastante similares, por ejemplo, la Epilepsia del Lóbulo Temporal, así como la posible asociación entre el análisis de grafos, la neuroimagen y otros instrumentos de diagnóstico. Además, son necesarios estudios longitudinales para evaluar si los SGA pueden ayudar a identificar a los sujetos con MCI con mayor riesgo de progresar a la enfermedad de Alzheimer.

3.6. Estudio del deterioro del lenguaje relacionado con el alzhéimer mediante aprendizaje maquinal

El artículo contempla el uso de técnicas avanzadas de aprendizaje maquinal para analizar el lenguaje espontáneo en varios idiomas y detectar deterioros relacionados con la AD. Los autores proponen identificar cambios especificos en el habla de las personas con AD que, en valoraciones clínicas habituales posiblemente no sean evidentes, con el fin de desarrolar un modelo multiiingüe que aporte evidencia de cómo el Alzheimer influye en el deterioro del lenguaje. Lo cual, puede ayudar en el diagnóstico temprano y supervisar la progresión de la enfermedad a través del tiempo utilizando muestras del habla espontánea, siendo que podría ser un método menos invasivo y más asequible que otros métodos tradicionales actuales.

Para llevar a cabo este trabajo los autores crearon un corpus multilingüe de descripciones de imágenes de habla espontánea en inglés y francés. Las características se organizaron en subgrupos (específicas de la tarea, semánticas, sintácticas, paralingüísticas). Para cada corpus se obtiene un conjunto idéntico de características. Después, se realiza una inspección estadística de correlaciones entre idiomas y pruebas. El planteamiento consiste en que las características que diferencian y modelan el trastorno del proceso lingüistico deben ser significativas en ambos idiomas.

Posteriormente se realiza una clasificación entre todas las características semánticas, sintácticas y paralingüísticas de forma separada en ambos idiomas y en un entorno multilingüe usando procesamiento de lenguaje natural (PLN) y aprendizaje maquinal, la Figura 3.4 muestra el esquema.

Esto se compara con los resultados de clasificación donde solo se usan características de lenguaje generalizable, siendo estas aquellas características semánticas, sintácticas y paralingüísticas que son significativas en ambos idiomas [21]. Este enfoque tiene como objetivo identificar el deterioro del lenguaje relacionado con la AD que se generaliza más allá de un único corpus o idioma y modela los procesos de deterioro del lenguaje clínicamente observable.

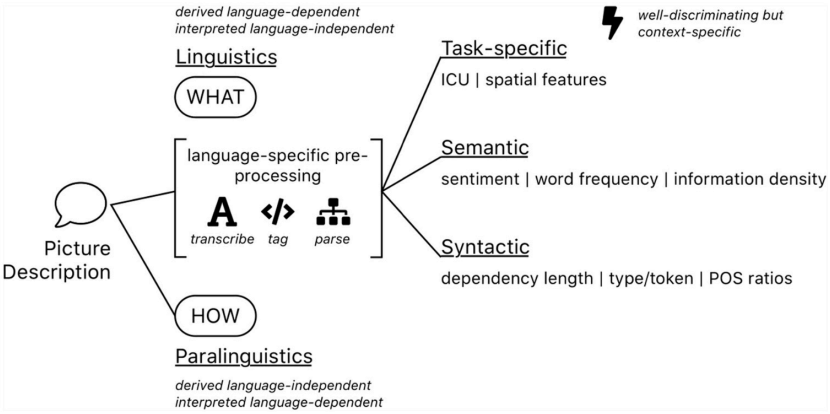


Figura 3.4: Una visión esquemática de los tipos de características que típicamente se extraen de las descripciones de imágenes mediante habla espontánea. [21]

En la realización de este estudio se utilizó una muestra de 154 sujetos extraídos entre dos corpus: ADReSS INTERSPEECH challenge and EIT-Digital ELEMENT project. La muestra correspondiente a ADReSS INTERSPEECH challenge incluye un subconjunto equilibrado (por edad y género) de 53 controles sanos (HC) y 54 pacientes con AD confirmada del DementiaBank. Contiene 106 grabaciones normalizadas y transcripciones anotadas manualmente de la tarea de descripción de la imagen del robo de galletas. Los pacientes fueron evaluados entre 1983 y 1988 en el Programa de Investigación del Alzheimer en la Universidad de Pittsburgh.

Por otra parte, la muestra del corpus EIT-Digital ELEMENT project, contempla 47 participantes que completaron la misma tarea de descripción. Los participantes fueron reclutados para un estudio clínico del proyecto EIT-Digital ELEMENT en Niza, Francia. Las grabaciones se realizaron usando una aplicación automatizada.

En ambas muestras, los participantes describieron una imagen en blanco y negro, conocida como el “Robo de Galletas” del Boston Diagnostic Aphasia Examination. El personal de prueba generalmente no proporciona retroalimentación durante las descripciones de los participantes, a menos que la respuesta inicial del paciente sea muy breve. Posteriormente, las características se dividen en cuatro categorías: semánticas, sintácticas, específicas de la tarea y paralingüísticas.

Para las características de la tarea se definieron unidades de información

(IUs) generales que aparecen en la tarea y se mapearon a un conjunto más amplio de palabras clave sinónimas. En cuanto a las características semánticas utilizan recursos específicos de la tarea, pero modelan la semántica combinando las IUs definidas y sus palabras clave mapeadas en características semánticas globales refinadas. También se calculan características semánticas que no dependen de las definiciones de IU, como la frecuencia de palabras y la riqueza léxica.

En el análisis estadístico inferencial se examinaron características semánticas, sintácticas, y paralingüísticas del lenguaje en pacientes con AD y HC en francés e inglés, observando correlaciones principalmente negativas en características semánticas y específicas de la tarea, indicando peor rendimiento en pacientes con AD. Además, las características paralingüísticas mostraron el comportamiento más específico por idioma y las correlaciones más débiles, especialmente en inglés. Asimismo, un porcentaje de las características de cada categoría mostró significancia estadística antes de la corrección por múltiples comparaciones, pero solo una fracción mantuvo su significancia después de la corrección. En cuanto a los términos de idioma, algunas características fueron significativas solo en uno de los idiomas, mientras que otras lo fueron en ambos.

Con respecto a los experimentos de aprendizaje maquinal se evaluaron modelos de regresión logística (LR), máquinas de vectores de soporte (SVM) y redes de perceptrón multicapa (MLP) usando características del lenguaje generalizables. Para la evaluación del rendimiento de la clasificación, se usa el Área Bajo la Curva de Operación del Receptor (AUC) para cada escenario experimental. También se reportan matrices de confusión para el modelo multilingüe con el conjunto de características del lenguaje generalizable. Una matriz es reportada para la clasificación general y luego el error se desglosa por idioma individual para investigar si el clasificador entrenado multilingüe funciona igualmente en ambos idiomas.

Los modelos mostraron una mejora en su desempeño al utilizar características generalizables en comparación con modelos que utilizaban todas las características del lenguaje, con la SVM y LR mostrando las mayores mejoras en AUC, especialmente en el modelo multilingüe. Además, los modelos mostraron diferentes tasas de error y porcentajes de falsos positivos por idio-

ma, siendo generalmente más balanceadas en francés que en inglés.

Los autores concluyen con base en la metodología propuesta, que el estudio muestra la presencia de un posible deterioro del lenguaje en la AD, observable en el habla espontánea cotidiana y que puede entenderse en un sentido neurocognitivo. El enfoque utilizado demuestra que las características del lenguaje independientes de tareas específicas en el habla deteriorada por la AD apuntan hacia deterioros neurocognitivos del lenguaje. Los hallazgos principales están alineados con la comprensión clínica actual sobre los deterioros del lenguaje relacionados con la demencia por AD; se sugiere un deterioro semántico primario y también se observa evidencia de un deterioro sintáctico más leve, complicado por múltiples otros procesos cognitivos.

Además, los resultados apoyan que el deterioro del lenguaje puede ser medido mediante técnicas de NLP motivadas clínicamente sin sacrificar el rendimiento general. Por lo tanto, se destaca que al abordar la explicabilidad y generalización por diseño en los experimentos de aprendizaje maquinal, la investigación no solo puede generar aplicaciones clínicas eficientes de métodos de NLP para la detección de AD a partir de habla espontánea, sino también resultar en conocimientos clínicamente accionables.

El estudio presenta dos limitaciones principales. Primero, el uso de un conjunto de datos clínicos pequeño plantea varios desafíos, especialmente porque no se dispone de un conjunto de prueba independiente para los datos en francés, lo que podría inflar artificialmente los resultados del modelo de aprendizaje maquinal. Segundo, el rendimiento deficiente de las características paralingüísticas podría estar influido por múltiples factores, como el género, diferencias significativas de edad en la población francesa y la calidad del audio de las grabaciones. Se sabe que la edad y el género afectan los patrones de habla y el rango de tono debido a diferencias anatómicas.

Por otro lado, los autores sugieren que para futuras investigaciones, se explorar más a fondo cómo estos factores afectan la explicabilidad y generalización de las características paralingüísticas y replicar el estudio con más datos y poblaciones emparejadas por edad, género y educación para validar los resultados. Además, consideran que sería útil aplicar esta metodología a otras tareas clínicas que generen habla espontánea para verificar la robustez de los hallazgos en diversos escenarios y podrían considerar otros clasificadores

o técnicas de ajuste para mejorar los resultados de clasificación.

Capítulo 4

Desarrollo: Métodología y Diseño de Experimentos

En este capítulo, se detalla el desarrollo de la metodología descrita en el Capítulo 1 (ver sección 1.4) para llevar a cabo la implementación del sistema de clasificación de deterioro cognitivo mediante el análisis de redes léxico-semánticas en adultos mayores de habla hispana (SIDERLEX) mediante el análisis de conversaciones cotidianas utilizando técnicas de procesamiento de lenguaje natural (NLP, *Natural Languages Processing*) y Aprendizaje Maquinal (ML, *Machine Learning*) y redes léxico-semánticas.

La metodología comprende la recopilación de los datos, el preprocesamiento de los datos, la construcción de redes léxico-semánticas, la agrupación de los datos y los resultados como ilustra la Figura 4.1.

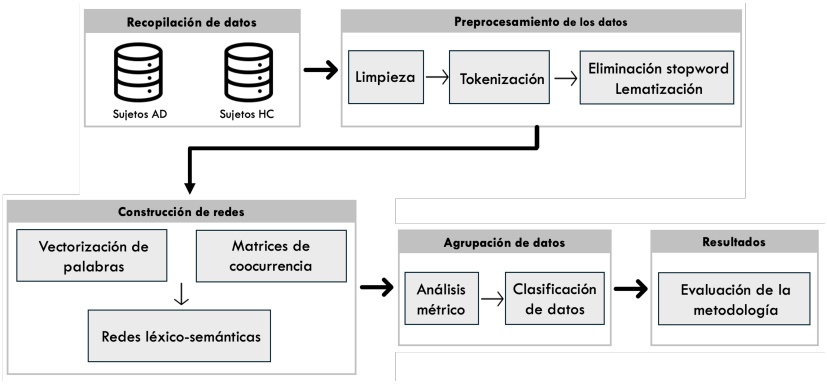


Figura 4.1: Metodología del modelo SIDERLEX.

4.1. Recopilación de datos

Los datos recopilados en esta investigación consisten en conversaciones provenientes de dos grupos de participantes: pacientes diagnosticados con la enfermedad de Alzheimer (AD, *Alzheimer's Disease*) y personas de control (HC, *Control Healthy*) aparentemente sin ninguna afección cognitiva.

Las conversaciones de los pacientes con AD pertenecen al corpus PerLA en español, creado por el Dr. José Luis Pérez Mantero, en Valencia, España. El corpus está conformado por 27 transcripciones de 21 pacientes con AD hispano-cataleños entre los 63 y 85 años. En la Tabla 4.1 se muestran los datos demográficos.

Sujeto	Género	Edad (años)	Nivel afección	Localidad	Fecha	Duración
500-TRM	femenino	74	GDS 4	casa	25/04/2012	20:09
501-CGS	masculino	79	GDS 4	casa	28/04/2012	20:05
502-LML	masculino	71	GDS 4	casa	30/04/2012	19:50
503-FGB	masculino	80	GDS 5	centro de cuidado	02/05/2012	19:58
504-MSS	femenino	72	GDS 5	casa	04/06/2012	19:54
505-JAC	femenino	74	GDS 4	casa	08/05/2012	20:00
506-TAZ	femenino	79	GDS 4	casa	14/05/2012	20:00
507-JSP	femenino	78	GDS 5	casa	29/05/2012	20:02
508-IBM	femenino	75	GDS 4	casa	30/05/2012	19:57
509-MMT	femenino	71	GDS 6	casa	31/05/2012	19:59
510-AAP	femenino	85	GDS 6	casa	01/06/2012	19:56
511-DGD	femenino	75	GDS 5	centro de cuidado	05/06/2012	20:06
512-LGP	femenino	78	GDS 6	centro de cuidado	06/06/2012	20:00
513-ALG	femenino	70	GDS 5	centro de cuidado	07/06/2012	19:56
514-RCB	masculino	67	GDS 4	casa	08/06/2012	20:00
515-AAM	femenino	76	GDS 6	centro de cuidado	11/06/2012	17:07
516-PVN	masculino	78	GDS 6	centro de cuidado	12/06/2012	19:57
517-DGN	femenino	75	GDS 4	casa	18/06/2012	19:58
518-RJC	femenino	76	GDS 5	casa	19/06/2012	19:58
519-JGM	masculino	63	GDS 4	casa	20/06/2012	19:59
520-PUE	femenino	83	GDS 4	casa	26/06/2012	20:00
521-TRM	femenino	75	GDS 4	casa	26/06/2013	20:05
522-MMT	femenino	72	GDS 6	casa	26/06/2013	19:59
523-CGS	masculino	80	GDS 4	casa	03/07/2013	20:06
524-IBM	femenino	76	GDS 5	casa	04/06/2014	20:03
525-FGB	masculino	82	GDS 6	casa	09/06/2014	20:07
526-JGM	masculino	65	GDS 5	casa	11/06/2014	20:04

Tabla 4.1: Información demográfica de los pacientes del corpus PerLA.

De acuerdo a la información de la tabla, el corpus está conformado por 15 pacientes con AD de género femenino entre 70 y 85 años y 6 de género masculino entre 71 y 80 años, los cuales cuentan con una clave alfanumérica única para su identificación. Cabe subrayar que los últimos seis registros pertenecen a pacientes con AD que se les dió seguimiento dentro de un de-

terminado tiempo, lo cuál se detallará más adelante. Las conversaciones se llevaron a cabo en casa de los pacientes con AD y en centros de cuidado en compañía de familiares y/o amigos, teniendo una duración de alrededor de 20 minutos.

Asimismo, se puede observar que los pacientes con AD tienen asignada una categoria de acuerdo con el nivel de afección establecido por la Escala de Deterioro Global (*Global Deterioration Scale*, GDS), donde GDS 4 corresponde a una fase inicial, GDS 5 a una fase moderada y GDS 6 a una etapa avanzada del deterioro. La Tabla 4.2, presenta a los pacientes con AD agrupados según el nivel de afección en el momento de la conversación.

GDS 4 (Inicial)	**GDS 5 (Moderada)**	**GDS 6 (Avanzada)**
500-TRM	503-FGB	509-MMT
501-CGS	504-MSS	510-AAP
502-LML	507-JSP	512-LGP
505-JAC	511-DGD	515-AAM
506-TAZ	513-ALG	516-PVN
508-IBM	518-RJC	522-MMT
514-RCB	524-IBM	525-FGB
517-DGN	526-JGM	
519-JGM		
520-PUE		
521-TRM		
523-CGS		

Tabla 4.2: Información demográfica de los niveles de afección de los pacientes del corpus PerLA.

Como se muestra en la tabla, el corpus PerLA cuenta con 12 registros de pacientes con AD en el nivel GDS 4, 8 registros en el nivel GDS 5 y 7 registros el nivel GDS 6, siendo que los dos últimos registros de cada una de las columnas en la tabla corresponden a registros de seguimiento. Es importante destacar que no se encontró información sobre los criterios utilizados por el investigador para asignar la categoría a cada uno de los pacientes con AD.

Este corpus contempla 6 registros de seguimiento, cuya información consiste en conversaciones posteriores que se realizaron a algunos pacientes con AD después de transcurrido un determinado período de tiempo, como lo muestra la Tabla 4.3.

La tabla detalla el período de tiempo transcurrido entre la primera y segunda conversación de los pacientes con AD, así como el estadio inicial y final de afección. Con base en los datos, se puede ver que los pacientes con AD

Sujeto	Nivel de afección	Período (meses)	Sujeto	Nivel de afección
500-TRM	GDS 4	14	521-TRM	GDS 4
501-CGS	GDS 4	14	523-CGS	GDS 4
508-IBM	GDS 4	24	524-IBM	GDS 5
519-JGM	GDS 4	24	526-JGM	GDS 5
503-FGB	GDS 5	25	525-FGB	GDS 6
509-MMT	GDS 6	13	522-MMT	GDS 6

Tabla 4.3: Seguimiento de los pacientes con AD después de un período de tiempo.

con clave 500-TRM, 501-CGS y 509-MMT (celda gris) se mantuvieron en el mismo nivel de afección después del tiempo transcurrido. En cambio, los pacientes con AD con clave 508-IBM, 519-JGM y 503-FGB (celda amarilla) si manifestaron un cambio en el nivel de afección, transitando a un estadio más avanzado de la enfermedad.

En lo que concierne a los sujetos HC, se consideró llevar a cabo una selección semejante a los pacientes con AD en términos de edad y género. Se realizaron 12 conversaciones, las cuales se desarrollaron en casa de los sujetos HC en compañía de familiares o amigos en la Ciudad de México y en el Estado de México. La Tabla 4.4 contiene los datos demográficos de este grupo.

Sujeto	Género	Edad (años)	Nivel educativo	Ocupación	Localidad	Fecha	Duración
600-SBR	femenino	65	media superior	costura	casa	17/11/2023	20:05
601-MAM	femenino	66	primaria	hogar	casa	23/11/2023	07:00
602-YPG	femenino	58	primaria	ventas	casa	01/12/2023	20:00
603-CCV	masculino	63	licenciatura	comerciante	casa	02/12/2023	20:00
604-ECH	femenino	79	primaria	hogar	casa	08/12/2023	20:39
605-GCH	femenino	77	primaria	hogar	casa	08/12/2023	20:00
606-MAG	masculino	84	primaria	retirado	casa	16/12/2023	20:07
607-GMR	masculino	60	técnico	ventas	restaurante	09/03/2024	20:00
608-LCH	femenino	63	licenciatura	hogar	restaurante	09/03/2024	19:30
609-GCS	masculino	68	primaria	retirado	casa	06/06/2024	20:10
610_LPM	masculino	74	licenciatura	asesorias	casa	28/09/2024	19:54
611-MAH	femenino	68	licenciatura	hogar	casa	28/09/2024	20:19

Tabla 4.4: Información demográfica de los sujetos HC.

En lo que respecta al tiempo de duración de las conversaciones del grupo HC, oscilan entre 7 y 20 minutos, haciendo un símil de las preguntas abiertas que se les realizaron al grupo AD en esos períodos de tiempo.

Las conversaciones de ambos grupos (pacientes con AD y sujetos HC) se obtuvieron a partir de diálogos cotidianos, mediante la realización de una serie de cuestionamientos sobre la familia, actividades cotidianas, pasatiempos, lugar de origen y experiencias personales, con el objetivo de propiciar una conversación natural y espontánea.

Todas las conversaciones se grabaron para posteriormente realizar la transcripción. Para obtener las transcripciones de los sujetos HC se uso una aplicación de Google llamada Pinpoint. Despúes, se verifico cada una de ellas de forma manual. Las transcripciones de ambos grupos se almacenaron garantizando la confidencialidad de los datos personales de los participantes.

4.2. Preprocesamiento de la información

El preprocesamiento de datos implica el análisis de las conversaciones mediante técnicas de minería de datos. El objetivo de la fase dos es garantizar que los datos recolectados sean transformados en un formato apropiado para su análisis posterior, eliminando información no necesaria e inconsistencias que propicien alteraciones en los resultados del estudio. A continuación, se describen los procedimientos empleados en el preprocesamiento de las transcripciones de las conversaciones.

Limpieza de datos

El primer paso en el preprocesamiento consistió en la limpieza de datos, lo cual involucró realizar las siguientes tareas:

- Verificación de las transcripciones: Las conversaciones de los sujetos HC fueron revisadas manualmente para corregir errores tipográficos y de ortografía cometidos durante la transcripción. Por otro lado, las transcripciones de los pacientes con AD se consideraron tal y como fueron proporcionadas para este trabajo de investigación.

  Las transcripciones de las conversaciones se guardaron en archivos de texto. Cada documento contiene un encabezado que incluye el nombre del archivo, fecha, duración, datos personales y acrónimos de los participantes en la conversacion, lo cual permite identificar a cada uno de las personas dentro de la misma, localidad donde se llevaron a cabo las conversaciones y nivel de afección en caso de los pacientes con AD.

  En la Figura 4.2 se presentan dos fragmentos del formato de las conversaciones. Como se puede observar, las transcripciones están estructuradas en forma de diálogo, donde cada línea está etiquetada con el acrónimo de las personas involucradas en la conversación. El primer

extracto (inciso a) es parte de la transcripción de un paciente con AD en etapa inicial (GDS 4), y el otro extracto (inciso b) corresponde a un sujeto HC.

(a)
```
@Location: Valencia, Comunidad Valenciana, Spain
@Situation: Informant's home
@Activities: Colloquial conversation
@Comment: GDS 4
INV: y usted ha trabajado siempre de ama de casa o ha tenido un trabajo fuera de casa ?
TRM: yo (.) sí (.) hombre (.) he tenido .
TRM: pero trabajo→ .
TRM: del Círculo de Lectores <lo he> [>] +/.
INV: <anda> [<] .
TRM: +, lo he llevao cinco años (.) a casa a repartir yyy to(do) eso .
TRM: aquí no (.) en Castellar &*INV:sí .
TRM: ⁅como hemos viví(d)o allí más⁆ .
TRM: y lo de Avón también .
```

(b)
```
@Location: Ciudad de México, Iztapalapa, México
@Situation: Home
@Activities: Colloquial conversation
@Comment: HC
EST:a qué se dedica actualmente?
SBR:pues pues ay me dedico al hogar, me dedico a hacer este con cómo se puede decir artes
plásticas.
SBR:o sea, me encanta todo eso de de actividades culturales y todo lo que es manualidad y eso.
EST:que bien. ¿tiene hijos?
SBR:sí, este dos.
EST:¿que edades tienen?
SBR:uno el mayor tiene 31 años y el más chico tiene 24.
EST:¿viven con usted?
SBR:nada más el chico el soltero.
```

Figura 4.2: Fragmentos de las transcripciones: (a) Pacientes con AD. (b) Sujetos HC.

- Eliminación de información general: Mediante la minería de datos se eliminó, en todas las conversaciones, información y segmentos que no pertenecen a los pacientes con ni a los sujetos HC, con la finalidad de enfocar el análisis lingüístico en los participantes. Como resultado de este proceso se obtuvo un documento en formato JSON, el cual, contiene sólo el texto correspondiente a los pacientes con AD y a los sujetos HC que conforma el corpus de este trabajo de investigación, como muestra la Figura 4.3.

(a)
```
{
    "500-TRM.txt - TRM": [
        "yo sí hombre he tenido",
        "pero trabajo",
        "del círculo de lectores lo he",
        "lo he llevao cinco años a casa a repartir todo eso",
        "aquí no en Castellar",
        "como hemos vivído allí más",
```

(b)
```
{
    "600-SBR.txt - SBR": [
        "pues pues ay me dedico al hogar me dedico a hacer este con cómo se puede decir artes plásticas",
        "o sea me encanta todo eso de actividades culturales y todo lo que es manualidad y eso",
        "sí este dos",
        "uno el mayor tiene 31 años y el más chico tiene 24",
        "nada más el chico el soltero",
```

Figura 4.3: Fragmento de la información contenida en el archivo json: (a) Pacientes AD. (b) Sujetos HC.

El primer fragmento (inciso a) pertenece a uno de los pacientes con AD cuyo acrónimo es TRM y el otro fragmento (inciso b) concierne al sujeto HC con acrónimo SBR. El contenido del archivo sólo comprende las líneas de texto de las conversaciones concernientes a los participantes en la investigación.

Stopwords (Palabras vacías)

En esta fase del preprocesamiento se implementaron dos funciones en lenguaje de programación Python: la primera función es para la obtención de una lista de palabras que a posteriori, se excluyeron del texto, y la segunda función tokeniza y lematiza el corpus obtenido en el paso anterior del preprocesamiento para más adelante, dar paso a la vectorización de las palabras que conforman dicho corpus.

Las stopwords, son palabras que tienen poca o ninguna significancia, es decir, aportan escasa o nula información en el análisis de un texto. Suelen eliminarse del texto durante el preprocesamiento para extraer sólo las palabras que tienen mayor significado y contexto. Algunos ejemplos de stopwords son “a”, “la”, “el”, “yo”. Cada uno de los idiomas puede contar con su propia conjunto de stopwords.

Partiendo del corpus, se obtuvo una lista personalizada de stopwords. Para ello se implemento una función que se ilustra en el Listado 4.1

```
1 def get_stopwords_from_responses(responses, most_common=100):
2 all_text = ' '.join([response for conversation in responses for response_list in
      conversation.values() for response in response_list])
3 all_words = re.findall(expresion_regular, all_text.lower())
4 word_counts = Counter(all_words)
5 stopwords = set(word for word, count in word_counts.most_common(most_common))
6 return stopwords
```

Listado 4.1: Función que permite obtener las 100 palabras más comunes dentro de las conversaciones.

Esta función permite obtener una lista de stopwords específico del corpus, lo cual permite adecuar los stopwords a las características del lenguaje utilizado.

La función recibe dos argumentos: `responses`, este argumento representa una lista de respuestas agrupadas por conversación, las cuales están contenidas en un archivo JSON. El segundo argumento es `most_common`, el cual,

Figura 4.4: Diagrama de bloques de la función `get_stopwords_from_responses`.

establece el número de palabras más comunes denominadas stopwords, cuyo valor en este caso es igual a 100, como se detalla en el diagrama de la Figura 4.4. El primer proceso en la función consiste en iterar sobre cada conversación en `responses` con el fin de unir todas las respuestas en una sola cadena separada por espacios para luego extraer las palabras mediante la expresión regular `r'\b[A-Za-záéíóúÁÉÍÓÚüÜñÑ]+\b'` y generar una lista.

Posteriormente,se crea un objeto `Counter` a partir de la lista de palabras para obtener un diccionario donde las claves son las palabras y los valores son las frecuencias de las mismas en la lista. Con base a las frecuencias se realiza la selección de las 100 palabras más comunes obteniendo como resultado el conjunto de stopwords. La Figura 4.5 ilustra la explicación anterior con un ejemplo de la ejecución de la función `get_stopwords_from_responses`.

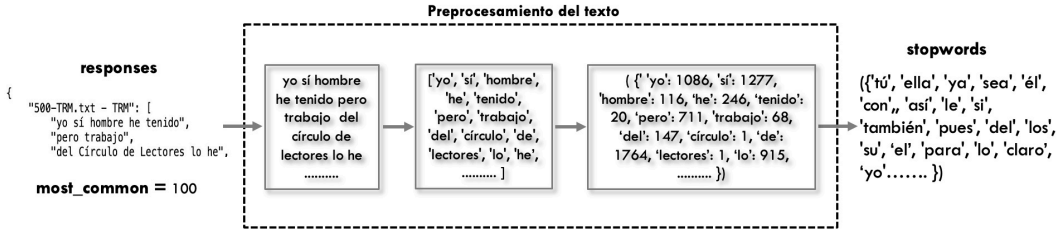


Figura 4.5: Ejemplo de la ejecución de la función `get_stopwords_from_responses`.

Tokenización y lematización

El siguiente paso de esta fase fue normalizar el texto, cuyo proceso consistió en la tokenización y la lematización del corpus para la posterior vectprización. de las palabras. Para entender la importancia de estos conceptos se realiza una breve explicación.

Los tokens son componentes textuales mínimos e independientes que tienen

una sintaxis y semántica definidas. Un párrafo de texto o un documento tiene varios componentes, incluyendo oraciones que pueden desglosarse aún más en cláusulas, frases y palabras. Por lo tanto, la tokenización puede definirse como el proceso de descomponer o dividir datos textuales en componentes más pequeños y significativos llamados tokens [29]

El proceso de lematización consiste en la eliminación de los afijos -que son partes de las palabras que se añaden a una base (raíz)- de las palabras para llegar a una forma base de la raíz, también conocida como lema [29]. Por ejemplo, en la palabra “corriendo”, el sufijo “-iendo” es un afijo que se añade a la raíz “corr-” para formar el gerundio del verbo correr; al eliminar el sufijo y volver a la raíz, la lematización reconoce que la raíz “corr-” xcorresponde al lema “correr”.

En este trabajo se llevó a cabo la tokenización y la lematización del corpus implementando la función que se presenta en el Listado 4.2

```
1 def clean_text(text, stopwords):
2 doc = nlp(text.lower())
3 lemmatized_words = [token.lemma_ for token in doc if token.lemma_ not in stopwords]
4 return ' '.join(lemmatized_words)
```

Listado 4.2: Definición de la función para tokenizar y lematizar el corpus.

Esta función básicamente, permite primero convertir el texto a minúsculas para su posterior tokenización y lematización mediante PLN. A continuación, elimina las palabras contenidas en la lista de stopwords, dando como resultado un texto libre de palabras que no aportan mucho valor semántico. El diagrama de la Figura 4.6 muestra de forma más simple las tareas realizadas por la función.

Figura 4.6: Diagrama de bloques de la función `clean_text`.

De acuerdo al diagrama, la función `clean_text` recibe dos argumentos: `text`, que es el texto a procesar contenido en el archivo JSON (responses), y la lista de `stopwords`, cuyas palabras fueron obtenidas mediante la función

`get_stopwords_from_responses` (ver listado 4.1) que se requieren excluir del texto. En el trabajo efectuado en el bloque del preprocesamiento del texto, se emplea una herramienta de PLN, la cual posibilita, como primer paso, tokenizar el texto, para después, obtener los lemas de cada uno de los *tokens* (palabras). Por último, se realiza el filtrado de los *tokens* lematizados que se encuentran contenidos en la lista de stopwords, obteniendo una cadena de texto que permitirá la posterior vectorización de las palabras del corpus, como lo muestra la Figura 4.7 mediante un ejemplo de la ejecución de la función.

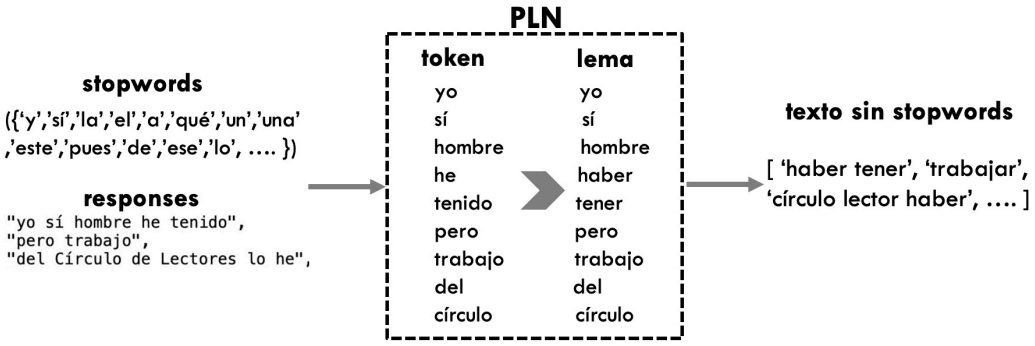


Figura 4.7: Ejemplo de la ejecución de la función `clean_text`.

4.3. Generación de las redes léxico-semánticas

La fase tres reside en la construcción de las redes léxico-semánticas a partir de la vectorización de las palabras que forman el corpus para crear las matrices de coocurrencia y con base en estas, generar dichas redes.

Con el fin de determinar la técnica más adecuada para la vectorización de las palabras del corpus, se exploraron tres enfoques principales: Word2Vec, las matrices de coocurrencia clásicas y el algoritmo TF-IDF (Term Frequency-Inverse Document Frequency). Aunque Word2Vec permite capturar relaciones semánticas profundas mediante el aprendizaje de representaciones distribuidas, su eficacia depende de grandes volúmenes de datos para obtener vectores significativos, lo cual no se ajustaba a las características del corpus analizado. Por su parte, las matrices de coocurrencia tradicionales ofrecen una representación simple basada en la frecuencia conjunta de palabras, pero carecen de mecanismos para ponderar su relevancia dentro del conjunto completo de conversaciones.

En cambio, TF-IDF proporciona una forma eficiente de representar la importancia relativa de las palabras en cada conversación, permitiendo construir redes léxico-semánticas más representativas del uso lingüístico individual. Por estas razones, se seleccionó TF-IDF como método de vectorización para este estudio.

Seleccionado el enfoque TF-IDF (*Term Frequency and Inverse Document Frequency*) como técnica de vectorización, se procedió a aplicarlo al corpus de conversaciones. Este algoritmo permite ponderar la relevancia de cada palabra considerando su frecuencia dentro de un documento y su presencia en el conjunto total. En este estudio, cada renglón de la transcripción fue tratado como un documento individual, mientras que la conversación completa se consideró como la colección de documentos, lo que permitió capturar la importancia relativa de las palabras en el contexto específico de cada participante.

Para vectorizar las palabras del corpus, se utilizó el algoritmo Frecuencia de Término-Frecuencia Inversa de Documento (TF-IDF, *Term Frequency and Inverse Document Frequency*). Este algoritmo permite identificar la importancia y relevancia de las palabras contenidas en un documento en el contexto de una colección de documentos. En este caso, en cada una de las conversaciones, cada renglón es tratado como un documento independiente y por ende, la conversación completa es considerada como la colección de documentos.

El proceso de vectorización de las palabras en este trabajo se realizó implementando el código del Listado 4.3

```
1   # Vectotiza las palabras
2  vectorizer = TfidfVectorizer(stop_words=list(stopwords))
3  X = vectorizer.fit_transform(cleaned_responses)
```

Listado 4.3: Código para la vectorización de palabras.

En la segunda línea del código se uso la clase `TfidfVectorizer` de la libreria scikit-learn, que convierte el conjunto de palabras de las conversaciones del corpus en una matriz de características TF-IDF omitiendo las palabras contenidas en la lista de stopwords. En la siguiente línea, se ajusta el modelo a los datos, para su posterior transformación en una matriz dispersa por conversación. La función `fit_transform` identifica las palabras únicas, asigna un índice a cada una y convierte cada respuesta en `cleaned_responses`

en un vector donde cada índice representa una palabra en el vocabulario aprendido y cada valor es el peso TF-IDF de esa palabra en la respuesta particular.

En esta matriz cada fila representa una respuesta dentro de las conversaciones y cada columna denota un término (palabra) dentro de la conversación como se ilustra en la Tabla 4.5

`cleaned_responses` = ['haber tener', 'trabajar', 'círculo lector haber']

	círculo	**haber**	**lector**	**tener**	**trabajar**
0	0.000000	0.605349	0.000000	0.795961	0.0
1	0.000000	0.000000	0.000000	0.000000	1.0
2	0.622766	0.473630	0.622766	0.000000	0.0

Tabla 4.5: Matriz de dispersión de la lista `cleaned_responses`.

El contenido de la tabla representa la matriz dispersa creada con base en la lista `cleaned_responses`. Esta matriz tiene tres filas y cinco columnas, cuyos renglones corresponden a las respuestas de la conversación y las columnas son las palabras (sin duplicados) que está contenida al menos en una de las respuestas. Los valores de la matriz dispersa indican la importancia de cada palabra dentro de cada respuesta en el contexto de todas las respuestas.

En este caso, los renglones cero, uno y dos conciernen a las respuestas 'haber tener', 'trabajar', 'círculo lector haber', respectivamente. Las columnas son las palabras: 'círculo', 'haber', 'lector', 'tener' y 'trabajar'. Asimismo, los valores indican la importancia relativa de cada palabra en cada respuesta. Por ejemplo, la palabra 'haber' tiene un valor más alto en la primera respuesta en comparación con la tercera. Esto se debe a que, aunque es relevante en la primera respuesta, su importancia disminuye en la tercera debido a su aparición en otra respuesta.

Por otro lado, a la palabra 'tener' se le asigna un valor más alto con respecto a 'haber' en la primera respuesta, ya que 'tener' aparece únicamente en una respuesta (fila cero), lo que la hace más relevante en este contexto que 'haber'. Por lo tanto, las palabras que son frecuentes en una respuesta pero raramente aparecen en las demás, reciben valores altos, lo que indica su relevancia para esa respuesta. En cambio, las palabras que se repiten en

varias respuestas obtienen valores más bajos, lo cual, indica que la palabra es menos importante para esa respuesta. Además, el valor cero asignado a las palabras indica que la palabra no está presente en esa respuesta.

Acorde con la matriz de dispersión de la Tabla 4.5, los vectores correspondientes para las respuestas en `cleaned_responses` serían:

['haber tener']	[0.000000, 0.605349, 0.000000, 0.795961, 0.0]
['trabajar']	[0.000000, 0.000000, 0.000000, 0.000000, 1.0]
['círculo lector haber']	[0.622766, 0.473630, 0.622766, 0.000000, 0.0]

donde los índices cero, uno, dos, tres y cuatro de cada vector están vinculados a las palabras 'círculo', 'haber', 'lector', 'tener' y 'trabajar', respectivamente.

A partir de los resultados obtenidos al aplicar el algoritmo TF-IDF, se crean las matrices de co-ocurrencia del corpus. Una matriz de co-ocurrencia de un corpus es una matriz cuadrada $N \times N$, donde N corresponde al número de palabras únicas en el corpus. Una celda m_{ij} contiene el número de veces que la palabra w_i co-ocurre con la palabra w_j, es decir, las celdas contienen valores numéricos que indican la frecuencia con la que los pares de palabras aparecen juntos dentro de un contexto específico. [20].

En este caso, las matrices de co-ocurrencia se generaron mediante el código del Listado 4.4

```
1       Xc = (X.T * X).todense()
2       matrix = pd.DataFrame(Xc, index=vectorizer.get_feature_names_out(),
    columns=vectorizer.get_feature_names_out())
3
```

Listado 4.4: Código para la creación de las matrices de co-ocurrencia

El renglón uno, contiene el código que permite crear las matrices de co-ocurrencia; para ello, se multiplica la matriz de dispersión `(X)`, que se obtuvo con anterioridad mediante el algoritmo TF-IDF, con su transpuesta `(X.T)`. En estas matrices, cada fila y columna representan una palabra del vocabulario y cada elemento `(i, j)` de la matriz inidica cuántas veces las palabras en la posición `i` aparece junto la palabra en la posición `j` en las mismas respuestas, ponderada por sus valores TF-IDF, es decir, cada valor en la matriz representa la co-ocurrencia ponderada entre dos palabras en las respuestas.

La función `todense()` se utiliza para convertir las matrices dispersas, en

matrices densas, es decir, todos los valores son almacenados, facilitando el procesamiento y análisis de la información. La Tabla 4.6 muestra la matriz de co-ocurrencia, calculada a partir de la tabla 4.5.

	círculo	**haber**	**lector**	**tener**	**trabajar**
círculo	0.387838	0.294960	0.387838	0.000000	0.0
haber	0.294960	0.590772	0.294960	0.481834	0.0
lector	0.387838	0.294960	0.387838	0.000000,	0.0
tener	0.000000	0.481834	0.000000	0.633553	0.0
trabajar	0.000000	0.000000	0.000000	0.000000	1.0

Tabla 4.6: Matriz de co-ocurrencia de la lista `cleaned_responses`.

En esta matriz de co-ocurrencia cada fila y columna de la matriz corresponde a una palabra específica del vocabulario (‘círculo’, ‘haber’, ‘lector’, ‘tener’, ‘trabajar’). Los valores de cada celda representa la co-ocurrencia ponderada entre dos palabras del vocabulario, Por ejemplo, la co-ocurrencia entre ‘haber’ y ‘tener’ es de 0.481834, mientras que entre ‘haber’ y ‘lector’ es de 0.294960. La co-ocurrencia más alta se observa entre las palabras ‘haber’ y ‘tener’. Esto no solo indica que estas palabras aparecen juntas con frecuencia, sino que también sugiere una relación significativa en el contexto de la conversación.

A partir de estas matrices, es posible derivar medidas de similitud semántica, detectar patrones de uso de palabras y crear modelos estadísticos que ayuden a comprender cómo se relacionan las palabras en un corpus de texto. En el contexto de PLN, estas matrices se utilizan principalmente para analizar las relaciones entre palabras dentro de un corpus.

Por último, en el renglón dos del código, se crean DataFrames usando la libreria. `pandas` con base en las matrices de co-ocurrencia, facilitando la interpretación y manipulación de los datos para la generación de las redes léxico-semánticas. Las redes léxico-semánticas son conexiones establecidas entre palabras que representan la estructura de la memoria semántica, el procesamiento del lenguaje y el acceso léxico. La Figura 4.8 es una red léxico-semántica obtenida de la conversación del paciente con AD.

La red léxico-semántica ilustra las conexiones y la estructura semántica del lenguaje utilizado por el paciente AD durante la conversación. Esta visualización permite detectar cambios en la estructura y coherencia del lenguaje.

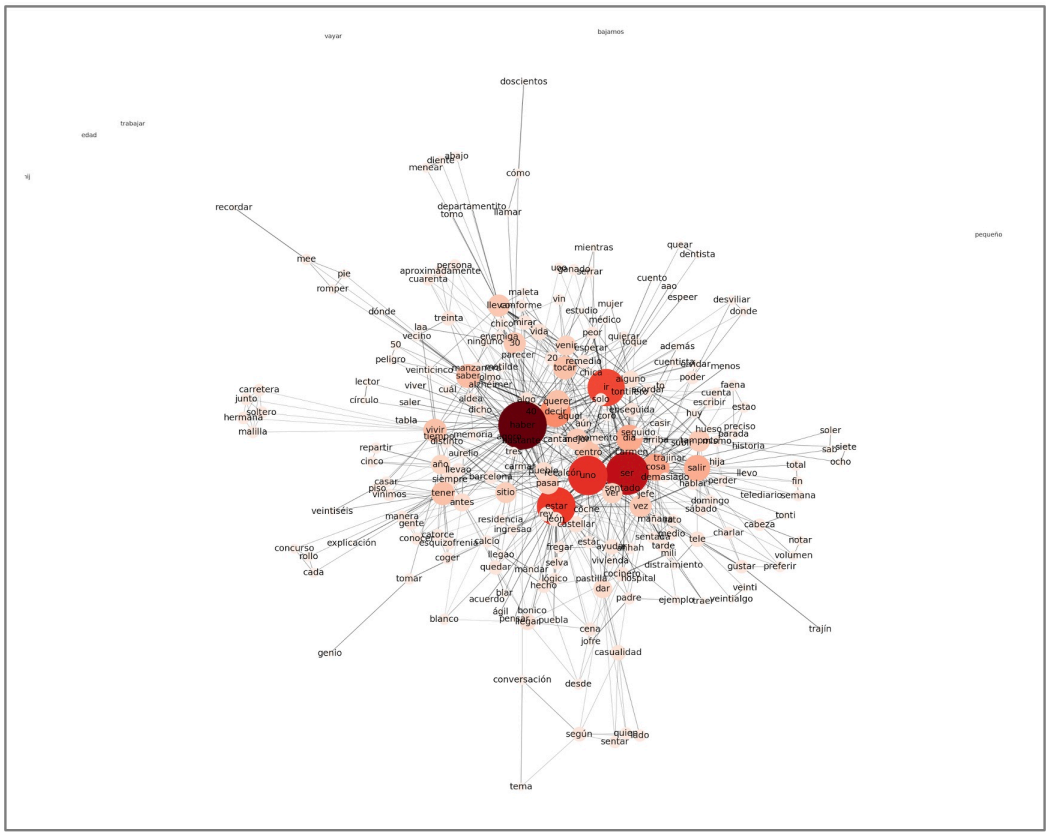


Figura 4.8: Red léxico-semántica de un paciente con AD.

Aunque las redes léxico-semánticas han sido estudiadas relativamente poco en este contexto, la elección de utilizar este recurso en la investigación, se fundamenta en su capacidad para capturar y visualizar patrones sutiles en la organización del lenguaje, que son a menudo indicativos de deterioro cognitivo. Este enfoque no solo ofrece nuevas posibilidades para métodos de diagnóstico menos invasivos y más accesibles, sino que también propicia una evaluación más precisa y temprana de la AD, proporcionando una herramienta fundamental que puede complementar e incluso mejorar los procedimientos tradicionales para el diagnóstico y monitoreo de la progresión de la enfermedad.

4.4. Agrupamiento de datos con K-medias

En esta fase de la investigación se analiza la estructura de las redes léxico-semánticas con base en la teoría de grafos, incorporando métricas orientadas a evaluar sus propiedades más relevantes y preparar los datos para

su posterior categorización. Las métricas consideradas son: centralidad de Intermediación (betweenness centrality), centralidad de cercanía (closeness centrality), centralidad de intermediación de aristas (edge betweenness centrality), densidad del grafo (graph density), diámetro del grafo (graph diameter), excentricidad (eccentricity), coeficiente de agrupamiento (clustering coefficient), transitividad (transitivity), distribución de grado (degree distribution) y asortatividad (Assortativity).

Estas métricas posibilitan la identificación de aspectos relacionados con la organización y coherencia del lenguaje. En particular, podrían facilitar la detectección de pérdida de palabras clave en la comunicación, la evaluación de la fluidez del discurso, la medición de la complejidad y cohesión del lenguaje, y la identificación de tendencias hacia un vocabulario limitado. Este análisis podría constituir una base para realizar un diagnóstico más preciso y temprano de la AD.

El conjunto de datos numéricos que se obtiene a partir de las métricas se emplea posteriormente en pruebas con algoritmos de aprendizaje maquinal —como K-vecinos más cercanos, regresión logística, máquinas de soporte vectorial (SVM), Random Forest, C4.5, red neuronal y tablas de decisión— con el propósito de evaluar su capacidad para discriminar entre los diferentes grupos de pacientes con AD. Para ello, se aplicó un esquema de validación cruzada con un 70 % de los datos para entrenamiento y un 30 % para prueba, cuyos resultados se presentan en términos de exactitud (accuracy) en la Tabla 4.7.

Clasificador	**Exactitud (%)**
K-vecinos más cercanos	56.67
Regresión logística (LR) con regularización L_2	58.97
Máquinas de soporte vectorial (SVM)	56.67
Random Forest	58.68
C4.5	62.50
Red neuronal (TensorFlow)	66.76
Tablas de decisión	60.58

Tabla 4.7: Resultados de exactitud de los algoritmos de aprendizaje maquinal.

En general, los porcentajes de exactitud obtenidos fueron relativamente bajos, lo que indica que los modelos supervisados no lograron una discrimi-

nación robusta entre los grupos. Esta limitación puede atribuirse al tamaño reducido del conjunto de datos, lo que dificulta el aprendizaje de patrones consistentes y afecta la capacidad de generalización de los algoritmos.

Debido a los resultados obtenidos por los algoritmos supervisados, se opta por incorporar un análisis complementario utilizando el algoritmo K-medias, el cual es un modelo no supervisado diseñado para identificar patrones subyacentes y agrupar los datos en función de sus similitudes sin requerir etiquetas predefinidas.

Para llevar a cabo este agrupamiento, el algoritmo K-medias se basa en la idea de que, dado un agrupamiento inicial pero no óptimo, reasigna cada punto a su centro más cercano, actualiza los centros calculando la media de los puntos de cada grupo y repite este proceso hasta cumplir los criterios de convergencia [19].

En el contexto de este estudio, el algoritmo K-medias se utiliza para agrupar a los pacientes con AD y a los sujetos HC según niveles similares de deterioro cognitivo. Previo a la aplicación del algoritmo sobre el conjunto de datos numéricos obtenidos a partir de las métricas de redes complejas, se determinó el número óptimo de agrupamientos (K) mediante el método del codo y el coeficiente de silueta, con el objetivo de garantizar la efectividad del procedimiento en la categorización de los datos.

Cada paciente se representa mediante un vector de características numéricas derivadas de las métricas. La Tabla 4.8 presenta un ejemplo de cinco conversaciones con sus respectivos vectores conformados con algunas métricas, las cuales reflejan propiedades estructurales del lenguaje en el discurso de cada sujeto.

Conversación	Centralidad de intermediación	Densidad del grafo	Coeficiente de agrupamiento	Diámetro del grafo	Excentricidad	Grupo
500-TRM	0.376356	0.103013	0.501602	0.333333	0.364221	GDS 4
507-JSP	0.388420	0.068666	0.567874	0.500000	0.585758	GDS 5
510-AAP	0.400381	0.073427	0.173741	1.000000	1.000000	GDS 6
515-AAM	0.892355	0.270298	0.000000	0.666667	0.788846	GDS 6
600-SBR	0.139578	0.385159	0.895083	0.000000	0.109614	HC

Tabla 4.8: Valores correspondientes a métricas extraídas de las redes léxico-semánticas que conforman los vectores de características empleados en el proceso de agrupamiento mediante el algoritmo K-medias.

El algoritmo K-medias toma estos vectores y agrupa a los pacientes que pre-

sentan valores similares en las métricas seleccionadas. En este ejemplo, el sujeto 600-SBR (HC) presenta valores altos de coeficiente de agrupamiento (0.895083) y densidad (0.385159), así como valores bajos en diámetro y excentricidad, lo que indica una red densa y bien conectada, característica de un lenguaje organizado.

En contraste, los pacientes con AD 510-AAP y 515-AAM (GDS 6) exhiben valores reducidos en coeficiente de agrupamiento y densidad, junto con un incremento en diámetro y excentricidad, lo que sugiere redes más fragmentadas y menor cohesión, propias de un deterioro cognitivo avanzado.

Los resultados de este proceso permitieron identificar diferencias significativas entre los grupos analizados, evidenciando la utilidad del enfoque propuesto. Esta metodología resalta la eficacia de combinar minería de datos, PLN, matrices de co-ocurrencia, redes léxico-semánticas, teoría de redes complejas y aprendizaje maquinal en el análisis del lenguaje, lo que podría contribuir al diagnóstico temprano y a una mejor comprensión de la progresión de la AD.

Capítulo 5

Experimentación y Resultados

Este capítulo presenta la descripción de los procedimientos experimentales realizados en el transcurso de la investigación, así como los resultados obtenidos y su correspondiente análisis. Para examinar el conjunto de datos conformado por transcripciones de conversaciones cotidianas, se emplearon técnicas de procesamiento de lenguaje natural (PLN), minería de datos y aprendizaje maquinal. Estas herramientas facilitaron el procesamiento de la información y la extracción de características lingüísticas y semánticas para la construcción de redes léxico-semánticas que posibilitan la exploración detallada del proceso lingüistico de las personas que integran esta investigación.

5.1. Conjunto de datos

El conjunto de datos que conforman esta investigación consta de un corpus de conversaciones reales y cotidianas en diversos entornos, orientado a identificar diferencias lingüísticas y semánticas entre adultos mayores diagnosticados con la enfermedad de Alzheimer (AD, *Alzheimer's Disease*) y sanos (HC, *Control Healthy*) sin ninguna alteración cognitiva aparente como se mencionó anteriormente en el capitulo 4 (ver sección 4.1).

Las tablas presentadas a continuación contienen la información de todas las personas que integran el corpus de la investigación. Cada individuo está identificado por un acrónimo único que facilita su seguimiento a lo largo del

estudio; los acrónimos entre 500 y 526 corresponden a pacientes con AD, mientras que el resto identifica a sujetos HC.

Además, se incluyen datos relevantes, como la edad de las personas (que oscila entre 63 y 85 años), la duración de las conversaciones (entre 7 y 20 minutos) y, en el caso de los pacientes con AD, el nivel de afección observado según la Escala de Deterioro Global (*Global Deterioration Scale*, GDS), al momento de la conversación, lo que permite distinguir los grados de deterioro cognitivo dentro del grupo afectado. Estos datos son considerados fundamentales, ya que podrían facilitar la identificación de posibles relaciones entre la progresión de la enfermedad y las características lingüísticas observadas.

La Tabla 5.1 muestra los datos de los pacientes con AD en fase GDS 4.

GDS 4			
Sujeto	**Género**	**Edad (años)**	**Duración**
500-TRM	femenino	74	20:09
501-CGS	masculino	79	20:05
502-LML	masculino	71	19:50
505-JAC	femenino	74	20:00
506-TAZ	femenino	79	20:00
508-IBM	femenino	75	19:57
514-RCB	masculino	67	20:00
517-DGN	femenino	75	19:58
519-JGM	masculino	63	19:59
520-PUE	femenino	83	20:00
521-TRM	femenino	75	20:05
523-CGS	masculino	80	20:06

Tabla 5.1: Datos demográficos de los pacientes con AD en nivel de afección GDS 4.

Como se observa, la Tabla 5.1 está integrada por 12 registros, de los cuales los dos últimos (521-TRM y 523-CGS) corresponden a una segunda recopilación de datos realizada 14 meses después de la primera conversación con los pacientes con AD, identificados con los acrónimos 500-TRM y 501-CGS.

La Tabla 5.2 presenta los datos de los pacientes en fase GDS 5, la cual indica un nivel de deterioro cognitivo más avanzado en comparación con la fase GDS4. Como en el caso de la Tabla 5.1, los dos últimos registros (524-IBM y 526-JGM) corresponden a una segunda recopilación de datos, realizada en

este caso 24 meses después de la primera conversación con los pacientes con AD, correspondientes a los acrónimos 508-IBM y 519-JGM.

GDS 5			
Sujeto	**Género**	**Edad (años)**	**Duración**
503-FGB	masculino	80	19:58
504-MSS	femenino	72	19:54
507-JSP	femenino	78	20:02
511-DGD	femenino	75	20:06
513-ALG	femenino	70	19:56
518-RJC	femenino	76	19:58
524-IBM	femenino	76	20:03
526-JGM	masculino	65	20:04

Tabla 5.2: Datos demográficos de los pacientes con AD en nivel de afección GDS 5.

A continuación, se presenta la Tabla 5.3, que incluye los datos de los pacientes en fase GDS 6. Esta fase refleja un nivel de deterioro cognitivo más avanzado que las fases GDS 4 y GDS 5. De la misma forma que en las Tablas 5.1 y 5.2 , los dos últimos registros (522-MMT y 525-FGB) corresponden a una segunda recolección de datos. En este caso, la conversación correspondiente al acrónimo 522-MMT se realizó 13 meses después de la primera (509-MMT), mientras que la conversación correspondiente al acrónimo 525-FGB se llevó a cabo 25 meses después de la primera (503-FGB) con los pacientes con AD.

GDS 6			
Sujeto	**Género**	**Edad (años)**	**Duración**
509-MMT	femenino	71	19:59
510-AAP	femenino	85	19:56
512-LGP	femenino	78	20:00
515-AAM	femenino	76	17:07
516-PVN	masculino	78	19:57
522-MMT	femenino	72	19:59
525-FGB	masculino	82	20:07

Tabla 5.3: Datos demográficos de los pacientes con AD en nivel de afección GDS 6.

Finalmente, en la Tabla 5.4 se presentan los datos de los sujetos HC, quienes conforman el grupo de control de esta investigación. Este grupo, que no presenta alteraciones cognitivas aparentes, se utiliza como referencia para comparar las diferencias lingüísticas y semánticas con los pacientes diagnosticados con Alzheimer en las fases GDS 4, GDS 5 y GDS 6.

HC			
Sujeto	**Género**	**Edad (años)**	**Duración**
600-SBR	femenino	65	20:05
601-MAM	femenino	66	07:00
602-YPG	femenino	58	20:00
603-CCV	masculino	63	20:00
604-ECH	femenino	79	20:39
605-GCH	femenino	77	20:00
606-MAG	masculino	84	20:07
607-GMR	masculino	60	20:00
608-LCH	femenino	63	19:30
609-GCS	masculino	68	20:10
610-LPM	masculino	74	20:00
611-MAH	femenino	68	20:19

Tabla 5.4: Datos demográficos de los sujetos HC.

Es importante señalar que, si bien ambos conjuntos de conversaciones pertenecen a adultos mayores, las conversaciones de los sujetos HC se realizaron en español de México, mientras que las conversaciones de los pacientes con AD fueron registradas en español castellano, específicamente en Valencia, España.

5.2. Vectorización del conjunto de datos

La vectorización del conjunto de datos en esta investigación se realiza con la finalidad de representar las palabras utilizadas en las conversaciones de las personas mediante redes léxico-semánticas, lo que permite un análisis del proceso lingüístico.

Para llevar a cabo la vectorización, en primera lugar, se realiza un preprocesamiento de la información que consiste en el proceso de la limpieza y normalización del texto (ver capitulo 4, sección 4.2). Este proceso tiene como objetivo mejorar la calidad de los datos y, por ende, la precisión de la vectorización.

5.2.1. Aplicación del Algoritmo TF-IDF

El proceso de vectorización mediante el algoritmo TF-IDF se realiza de forma independiente para cada una de las conversaciones, considerando cada

respuesta de la persona como un documento único dentro de un corpus que esta compuesto por todas las respuestas de una única conversación. Esto permite obtener la vectorización individual de cada conversación, con el fin de generar una matriz de co-ocurrencia por individuo para su análisis posterior.

Por lo tanto, los valores calculados con el algoritmo TF-IDF para cada palabra en cada respuesta se organizan en una matriz TF-IDF, donde:

- Cada fila representa una respuesta de una conversación específica.
- Cada columna denota un término relevante del vocabulario dentro de todas las respuetas de esa conversación.

Cada celda contiene el valor TF-IDF de un término en una respuesta específica, reflejando el peso y la relevancia de la palabra en la conversación en relación con todas las respuestas del corpus de esa conversación. De modo que, esta matriz TF-IDF permite entender la importancia de cada palabra, no solo en la respuesta específica, sino también en comparación con otras respuestas en la misma conversación.

5.2.2. Creación de las matrices de co-ocurrencia

A partir de la matriz TF-IDF se forman las matrices de co-ocurrencia, como se explicó en el capítulo anterior (ver capitulo 4, sección 4.3), para cada una de las conversaciones. Cabe recordar, que las matrices resultantes son cuadradas, donde tanto las filas como las columnas denotan las palabras del vocabulario usado en las conversaciones. Además, los valores contenidos en las celdas indican la fuerza de la relación entre las palabras; un valor alto implica que los términos co-ocurren con frecuencia, lo que sugiere una relación semántica o contextual entre ellos.

Esto permite identificar patrones en el uso del lenguaje, como palabras que se utilizan con frecuencia de manera conjunta, lo que puede revelar asociaciones semánticas o contextuales relevantes. Para facilitar este análisis, las matrices de co-ocurrencia se organizan en DataFrames, lo que posibilita la creación de las redes léxico-semánticas.

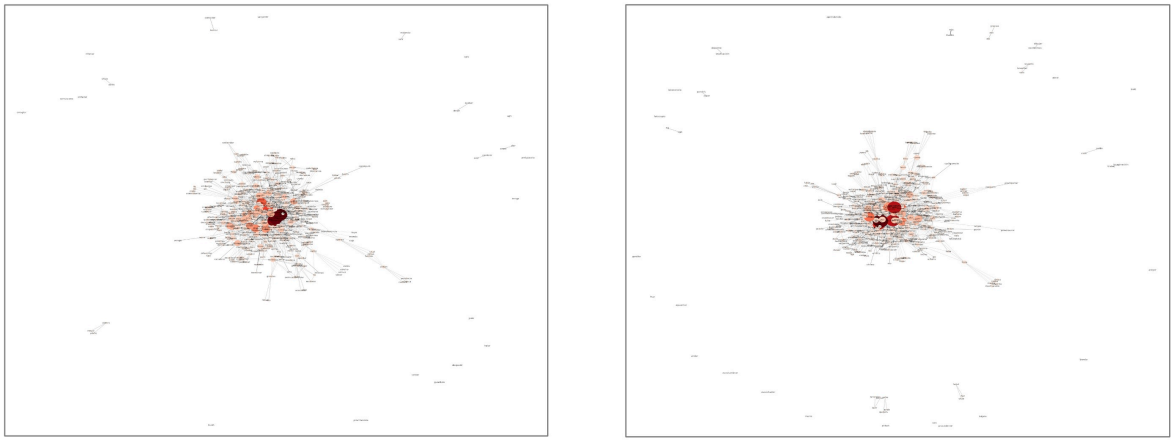
5.3. Generación de las redes léxico-semánticas

Para construir las redes léxico-semánticas de cada conversación, se emplean las matrices de co-ocurrencia obtenidas previamente. En este contexto, cada palabra representada en la matriz de co-ocurrencia se convierte en un nodo, mientras que las aristas, o conexiones entre los nodos, reflejan la fuerza de la relación entre dichas palabras, basada en su frecuencia de co-ocurrencia dentro de las conversaciones.

A partir de estas matrices, se generan las redes léxico-semánticas utilizando las bibliotecas `networkx`, `matplotlib` y `itertools`. Estas herramientas permiten transformar las relaciones numéricas en representaciones gráficas, facilitando tanto la creación como la visualización de las redes. Esto posibilita la exploración y el análisis de las relaciones léxicas presentes en las respuestas de las personas que conforman el estudio.

La exploración y el análisis se realizaron primero en las redes léxico-semánticas correspondientes a los pacientes con AD de cada nivel de afección (GDS 4, GDS 5 y GDS 6), lo que, permite observar cómo varían las características lingüísticas y la estructura de las redes a medida que la afección progresa.

La Figura 5.1 ilustra dos redes léxico-semánticas correspondientes al grupo de pacientes con AD en el nivel de afección GDS 4.



(a) 502-LML (b) 505-JAC

Figura 5.1: Redes léxico-semánticas de pacientes con AD correspondientes al nivel GDS 4.

En general, estas redes léxico-semánticas muestran una estructura visualmente densa en el centro, caracterizada por una concentración de nodos

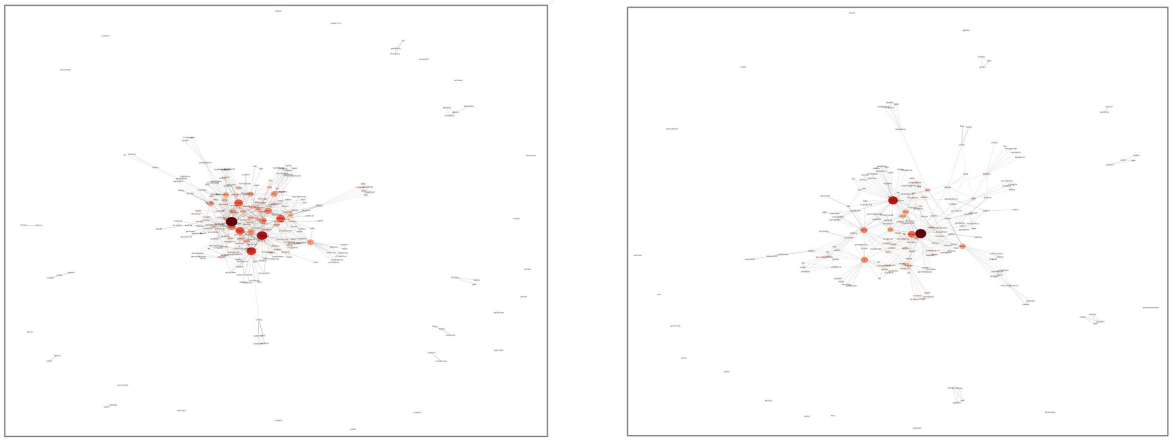
interconectados que reflejan no solo una moderada variedad temática, sino posiblemente también una cierta fluidez en la producción del discurso.

Esta organización se relaciona con valores relativamente altos en el número de nodos, que podrían indicar una mayor diversidad léxica; en el número de aristas, que reflejan una mayor conectividad entre las palabras; y el grado medio, al mostrar cuántas conexiones tiene en promedio cada palabra, sugiere una frecuencia constante de relaciones semánticas por palabra. Los valores correspondientes a estos tres parámetros se presentan en la Tabla 5.5.

Sujeto	Número de nodos	Número de aristas	Grado medio
500-TRM	260	927	7.13
501-CGS	463	1755	7.58
502-LML	421	1775	8.43
505-JAC	414	1682	8.13
506-TAZ	328	1271	7.75
508-IBM	270	1141	8.45
514-RCB	312	1180	7.56
517-DGN	234	819	7.00
519-JGM	176	527	5.99
520-PUE	336	1314	7.82
521-TRM	242	977	8.07
523-CGS	409	1601	7.83

Tabla 5.5: Número de nodos, aristas y grado medio en las redes léxico-semánticas de pacientes con AD en el nivel GDS 4.

No obstante, se identifican excepciones dentro del conjunto de las redes mostradas en la Figura 5.1, como es el caso de los pacientes con los acrónimos 519-JGM y 517-DGN, cuyas redes presentan una estructura menos conectada, especialmente en el primer caso. Esta red cuenta con 176 nodos, 527 aristas y un grado medio de 5.99; mientras que la red 517-DGN presenta 234 nodos, 819 aristas y un grado medio de 7.00 (ver Tabla 5.5). Estos valores son inferiores a los registrados en otros pacientes del estudio, lo que explica la menor densidad y la baja conectividad semántica observadas en sus redes, tal como se ilustra en la Figura 5.2.



(a) 517-DGN (b) 519-JGM

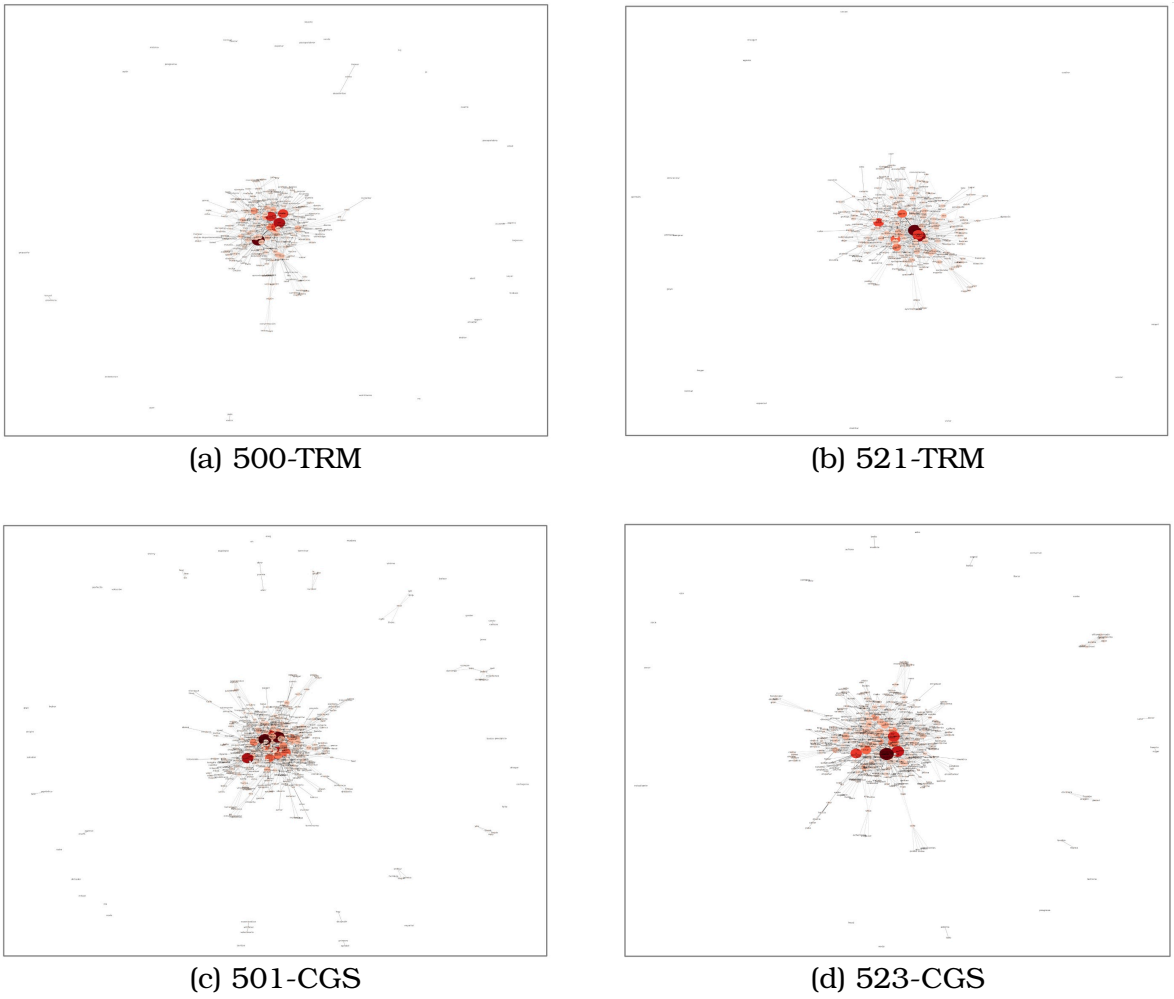
Figura 5.2: Redes léxico-semánticas de pacientes con AD correspondientes al nivel GDS 4.

Por otro lado, la Figura 5.3 muestra las redes léxico-semánticas de los pacientes con AD identificados con los acrónimos 521-TRM y 523-CGS, que corresponden a los casos previamente etiquetados como 500-TRM y 501-CGS, respectivamente.

La comparación visual entre las redes iniciales (véanse las Figuras 5.3(a) y 5.3(c)) y aquellas obtenidas tras un período de 14 meses (véanse las Figuras 5.3(b) y 5.3(d))permite observar una estructura relativamente estable en ambos casos. Los pacientes mantuvieron su nivel de afección en GDS 4 durante la segunda evaluación, lo que sugiere una continuidad en sus características cognitivas y léxico-semánticas a lo largo de dicho período.

Esta estabilidad observada a nivel visual se ve reforzada por los valores estructurales obtenidos en ambas evaluaciones. Los datos, presentados en la Tabla 5.5, respaldan cuantitativamente la consistencia en la organización léxico-semántica a lo largo del tiempo en ambos casos.

Por ejemplo, el paciente 500-TRM presentó en la primera evaluación una red con 260 nodos, 927 aristas y un grado medio de 7.13, mientras que en la segunda evaluación (521-TRM), la red contó con 242 nodos, 977 aristas y un grado medio de 8.07. De forma similar, el paciente 501-CGS pasó de 463 nodos, 1755 aristas y un grado medio de 7.58 a 409 nodos, 1601 aristas y un grado medio de 7.83 en la evaluación posterior (523-CGS).

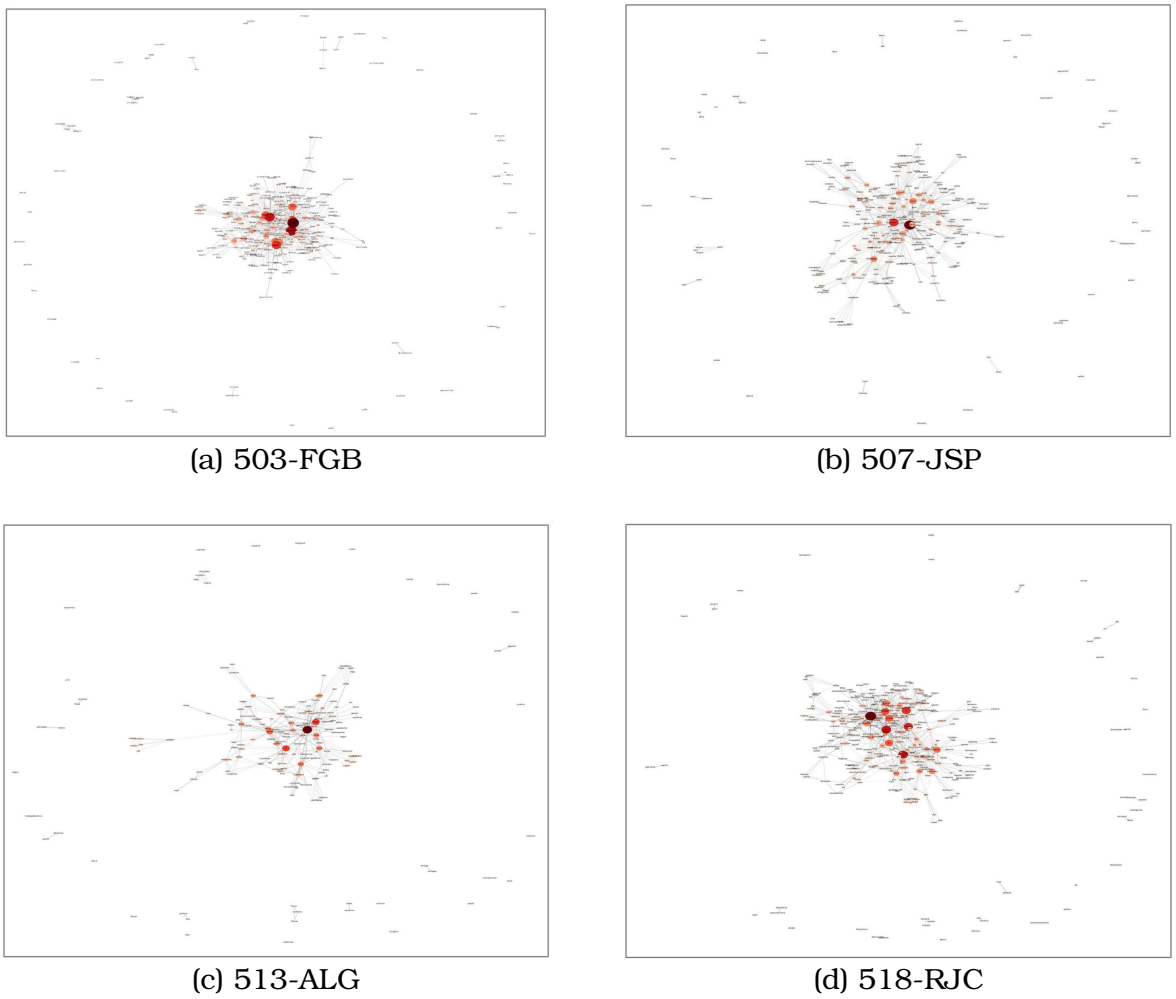


(a) 500-TRM (b) 521-TRM

(c) 501-CGS (d) 523-CGS

Figura 5.3: Redes léxico-semánticas iniciales de los pacientes 500-TRM y 501-CGS, y sus respectivas redes tras una segunda evaluación realizada 14 meses después, identificadas como 521-TRM y 523-CGS.

La Figura 5.4 contiene las representaciones de las redes léxico-semánticas que pertenecen a los pacientes con AD en el nivel GDS 5. Como puede observarse, las redes asociadas a este nivel de afección tienden a evidenciar una reducción en la complejidad y en la interconexión de los términos, lo cual se refleja en una menor densidad. Esta disminución se manifiesta en un número reducido de nodos (palabras) y conexiones (relaciones semánticas), en comparación con las redes de pacientes con AD en el nivel GDS 4.



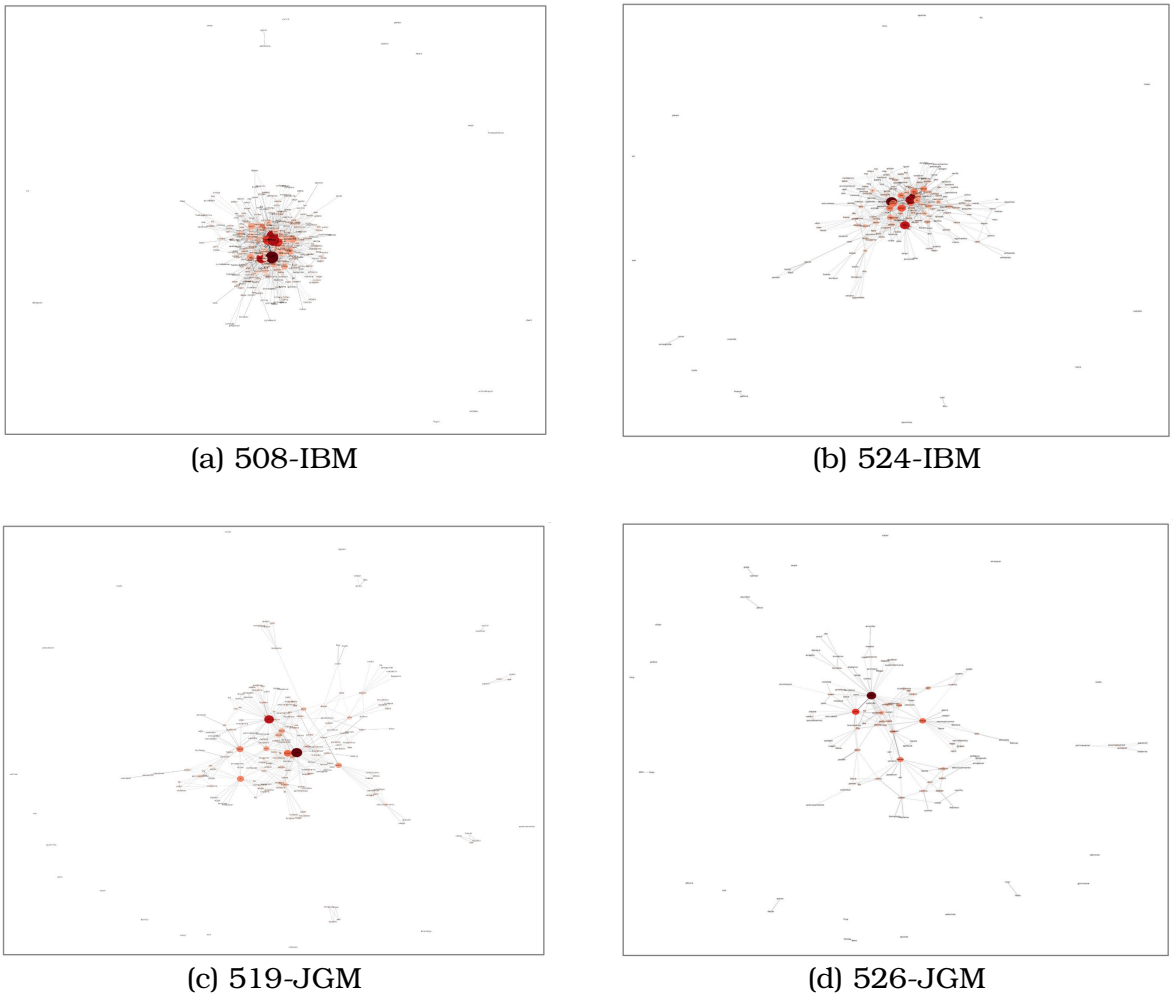
(a) 503-FGB (b) 507-JSP

(c) 513-ALG (d) 518-RJC

Figura 5.4: Redes léxico-semánticas de pacientes con AD correspondientes al nivel GDS 5.

La Figura 5.5 presenta dos casos particulares que permiten observar la variabilidad en la evolución de las redes léxico-semánticas dentro del nivel GDS 5. Se trata de los pacientes con AD identificados con los acrónimos 508-IBM y 519-JGM, evaluados inicialmente en el nivel GDS 4, tras un período de 24 meses, en el que fueron etiquetados en el nivel GDS 5 como 524-IBM y 526-JGM, respectivamente.

En el caso del paciente identificado como 524-IBM, se observa una disminución en la densidad de su red léxico-semántica. Este paciente transitó de la etapa GDS 4 (Figura 5.5(a)) a la etapa GDS 5 (Figura 5.5(b)) tras dicho período, lo que refleja el avance de la afección y su impacto en las propiedades estructurales del lenguaje representadas en la red.



(a) 508-IBM (b) 524-IBM

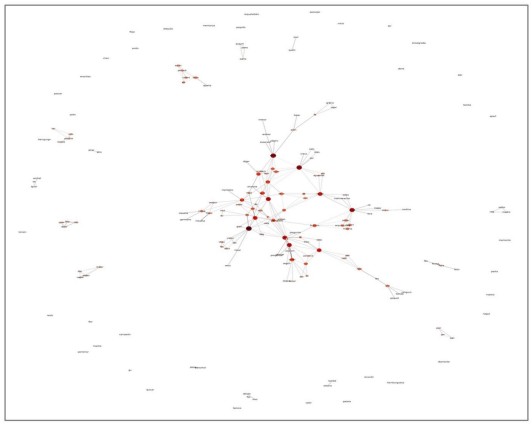
(c) 519-JGM (d) 526-JGM

Figura 5.5: Redes léxico-semánticas iniciales de los pacientes 508-IBM y 519-JGM, y sus respectivas redes tras una segunda evaluación realizada 24 meses después, identificadas como 524-IBM y 526-JGM.

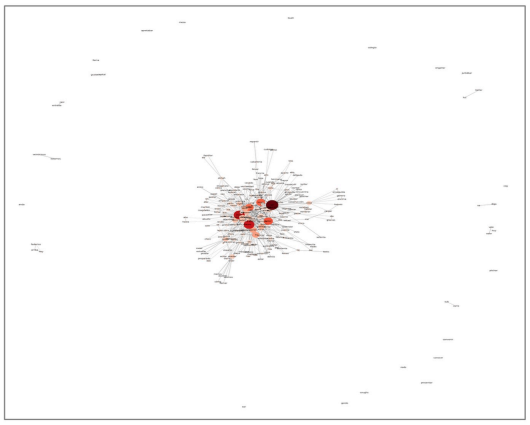
Por otro parte, el paciente 519-JGM (Figura 5.5(c)) como se mencionó anteriormente, fue etiquetado en el nivel de afección GDS 4, a pesar de que su red léxico-semántica presentaba una densidad menor en comparación con las demás redes del mismo nivel. Transcurrido un período de 24 meses, se observó una disminución adicional en la densidad de su red, cuya representación corresponde al acrónimo 526-JGM (Figura 5.5(d)).

En la Figura 5.6 se presenta las redes léxico-semánticas de los pacientes con AD con nivel de afección GDS 6. En estas redes se puede observar una baja densidad en el centro de las redes, debido a la disminución de nodos (palabras) y de conexiones (relaciones semánticas) con respecto a los pacientes

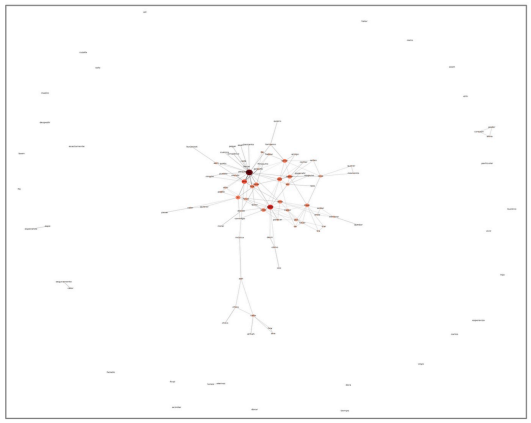
en los niveles GDS 4 y GDS 5.



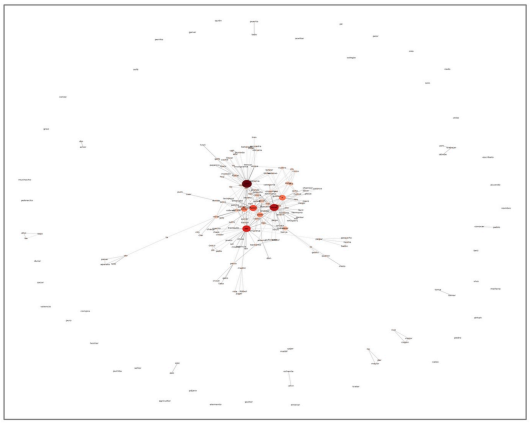
(a) 510-AAP



(b) 512-LGP



(c) 515-AAM



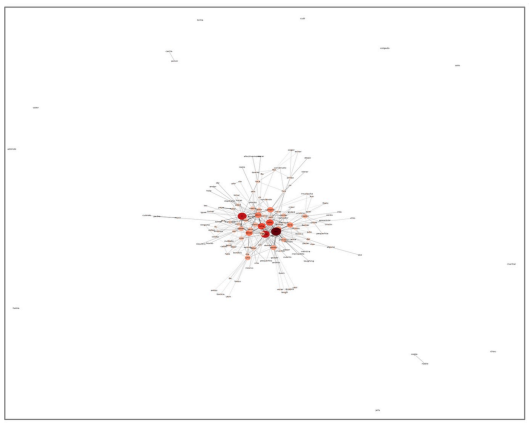
(d) 516-PVN

Figura 5.6: Redes léxico-semánticas de pacientes con AD correspondientes al nivel GDS 6.

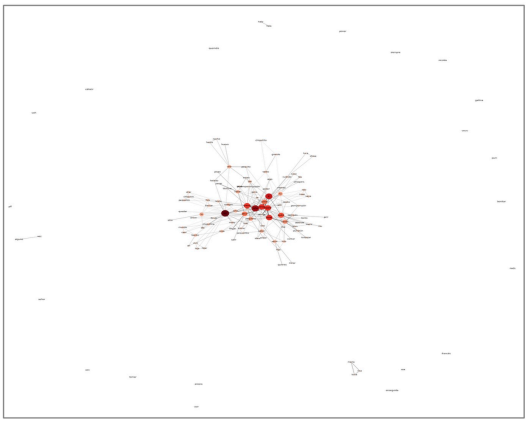
En este nivel GDS 6, el paciente identificado como 522-MMT (previamente 509-MMT, Figura 5.7(a)) mantuvo su nivel de afección tras una segunda entrevista realizada 13 meses después, lo que sugiere una posible estabilización en su progresión cognitiva durante este periodo, como se observa en la Figura 5.7(b).

Por su parte, el paciente con AD identificado como 525-FGB (previamente 503-FGB, Figura 5.7(c)) presentó un avance en su nivel de afección tras una nueva conversación realizada 25 meses después, lo que se refleja en una menor complejidad de su red léxico-semántica como se muestra en la Figura 5.7(d).

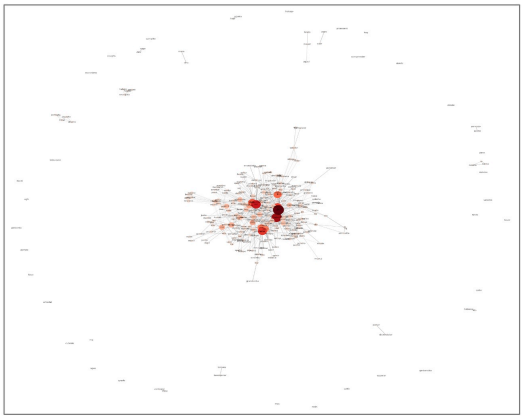
Esta configuración de las redes léxico-semánticas permite analizar los pa-



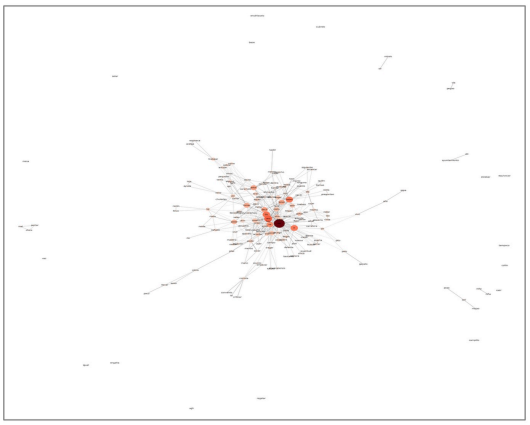
(a) 509-MMT



(b) 522-MMT



(c) 503-FGB



(d) 525-FGB

Figura 5.7: Redes léxico-semánticas iniciales de los pacientes con AD 509-MMT y 503-FGB, junto con sus respectivas redes tras una segunda evaluación realizada 13 y 25 meses después, respectivamente, identificadas como 522-MMT y 525-FGB.

trones de lenguaje mediante la teoría de redes complejas, especificamente aplicando algunas métricas de las redes.

Cabe señalar que se desconocen los criterios específicos utilizados para asignar a los pacientes con AD a los distintos niveles de afección.

5.4. Métricas de los grafos léxico-semánticos de los pacientes con AD

En las investigaciones sobre el deterioro lingüístico en pacientes con AD, las redes léxico-semánticas generadas a partir de las conversaciones representan una herramienta clave para analizar cómo las propiedades del lenguaje se modifica a medida que avanza la enfermedad. Este estudio se centró en realizar dicho análisis mediante las propiedades estructurales de las redes léxico-semánticas, empleando métricas derivadas de la teoría de grafos para evaluar los cambios asociados con la progresión del deterioro cognitivo.

Estas métricas posibilitan cuantificar y describir características importantes de las redes como la densidad, la centralidad de intermediación, el diámetro, el coeficiente de agrupamiento, entre otras. Asimismo, facilitan una perspectiva más detallada sobre las relaciones entre las palabras y cómo estas cambian según el estado cognitivo de los personas que conforman el corpus.

La Tabla 5.6 presenta las métricas normalizadas que se emplearon para hacer el análisis de las redes léxico-semánticas vinculados a los pacientes con AD.

La Tabla 5.6 muestra las métricas estructurales extraídas de las redes léxico-semánticas correspondientes a cada conversación. Los acrónimos utilizados en los encabezados son los siguientes:

- **BC (Betweenness Centrality)**: mide cuántas veces un nodo actúa como puente en los caminos más cortos entre otros nodos.
- **DG (Graph Density)**: representa la proporción de conexiones existentes sobre el total posible en el grafo.
- **CC (Clustering Coefficient)**: cuantifica la tendencia de un nodo a formar clústeres con sus vecinos.
- **CL (Closeness Centrality)**: evalúa la cercanía promedio de un nodo al resto de los nodos de la red.
- **TR (Transitivity)**: mide la probabilidad de que dos vecinos de un nodo estén también conectados entre sí.

Conv.	BC	DG	CC	CL	TR	DI	DD	EX	EB	AS
500-TRM	0.269227	0.328341	0.650741	0.767183	0.105455	0.250000	0.670040	0.101714	0.239076	0.047903
501-CGS	0.000000	0.000000	0.588868	0.641556	0.130627	0.500000	0.782490	0.292734	0.000000	0.263382
502-LML	0.121301	0.108273	0.761479	0.819163	0.090981	0.500000	0.995118	0.489815	0.042679	0.112590
503-FGB	0.244563	0.213129	0.592768	0.555791	0.273219	0.500000	0.571007	0.300512	0.220029	0.174549
504-MSS	0.518256	0.536342	1.000000	0.937050	0.320121	0.250000	0.961762	0.056844	0.309833	0.142982
505-JAC	0.036169	0.096396	0.817176	0.802172	0.000000	0.250000	0.918515	0.191217	0.014009	0.009752
506-TAZ	0.221079	0.215233	0.761973	0.776173	0.129658	0.250000	0.824702	0.215250	0.145489	0.100145
507-JSP	0.356686	0.176307	0.639397	0.398530	0.876630	0.750000	0.526938	0.665099	0.278423	0.587614
508-IBM	0.355430	0.443101	0.629159	0.905595	0.157774	0.250000	1.000000	0.222877	0.202202	0.062454
509-MMT	1.000000	1.000000	0.488514	0.917723	0.358224	0.250000	0.735259	0.153443	0.741110	0.153461
510-AAP	0.261737	0.237070	0.163321	0.000000	1.000000	1.000000	0.000000	1.000000	0.428919	1.000000
511-DGD	0.372138	0.469253	0.534150	0.748147	0.226838	0.250000	0.720633	0.075676	0.294562	0.091489
512-LGP	0.353212	0.359615	0.448744	0.756235	0.035229	0.250000	0.595751	0.219214	0.322837	0.008439
513-ALG	0.377368	0.557254	0.394322	0.415412	0.418006	0.250000	0.360853	0.269181	0.433903	0.337150
514-RCB	0.271000	0.233582	0.806992	0.721944	0.195509	0.500000	0.778271	0.278828	0.182950	0.168569
515-AAM	0.558400	0.827359	0.000000	0.226699	0.432287	0.250000	0.065487	0.163033	0.803031	0.731167
516-PVN	0.146111	0.254135	0.399307	0.223246	0.235956	0.500000	0.113759	0.376445	0.322315	0.364515
517-DGN	0.425933	0.402464	0.716036	0.607719	0.458819	0.500000	0.637378	0.380956	0.332766	0.337609
518-RJC	0.327298	0.148591	0.507287	0.501470	0.484783	0.750000	0.543918	0.559068	0.263823	0.422206
519-JGM	0.576412	0.525793	0.809773	0.528847	0.404095	0.250000	0.384775	0.173814	0.573674	0.238047
520-PUE	0.168403	0.204820	0.874587	0.789035	0.115356	0.250000	0.842542	0.255033	0.112591	0.085791
521-TRM	0.475158	0.504623	0.715205	1.000000	0.135469	0.250000	0.905721	0.210880	0.312624	0.000000
522-MMT	0.780657	0.923398	0.351066	0.538493	0.380262	0.000000	0.330290	0.000000	0.831535	0.381225
523-CGS	0.153980	0.082043	0.680610	0.768927	0.160134	0.500000	0.844396	0.370550	0.086766	0.203349
524-IBM	0.805705	0.799232	0.704260	0.926171	0.311233	0.500000	0.811372	0.372814	0.559696	0.117873
525-FGB	0.784215	0.558732	0.440282	0.647010	0.208914	0.750000	0.424721	0.539069	0.713369	0.181128
526-JGM	0.819582	0.660511	0.379353	0.256561	0.571403	0.750000	0.070838	0.646948	1.000000	0.412667

Tabla 5.6: Métricas obtenidas de las redes léxico-semánticas de los pacientes con AD que conforman el corpus.

- **DI (Diameter)**: es la mayor distancia geodésica entre dos nodos cualesquiera del grafo.
- **DD (Degree Distribution)**: representa la variabilidad de la cantidad de conexiones por nodo.
- **EX (Eccentricity)**: determina la distancia máxima entre un nodo y cualquier otro nodo del grafo.
- **EB (Edge Betweenness)**: similar a BC pero aplicada a las aristas, mide su importancia en la conectividad global.
- **AS (Assortativity)**: evalúa la preferencia de los nodos a conectarse con otros que tienen un grado similar.

El objetivo principal de estas métricas es identificar patrones cuantitativos que diferencien a los grupos de estudio, facilitando la detección de características lingüísticas asociadas con la progresión de la enfermedad. Para tal propósito, es fundamental aplicar técnicas de análisis que permitan reconocer patrones comunes dentro de los datos. En este caso, se empleó el algoritmo de agrupamiento K-medias.

Antes de aplicar el algoritmo K-medias, es esencial analizar la distribución de los datos. Este análisis permite identificar posibles sesgos, valores atípicos y patrones de dispersión que podrían afectar el rendimiento del algoritmo. Para ello, se generaron histogramas de cada una de las métricas incluidas en la Tabla 5.6, lo que facilita la visualización de la distribución de los valores.

Las Figuras 5.8, 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16 y 5.17, muestran los histogramas correspondientes a cada métrica,

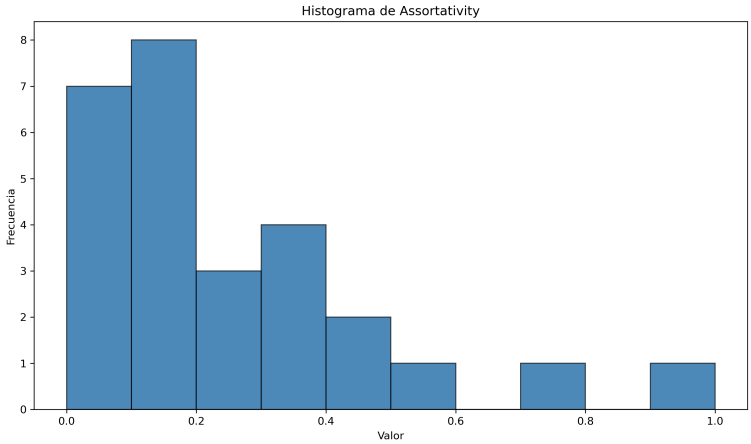


Figura 5.8: Histograma de asortatividad.

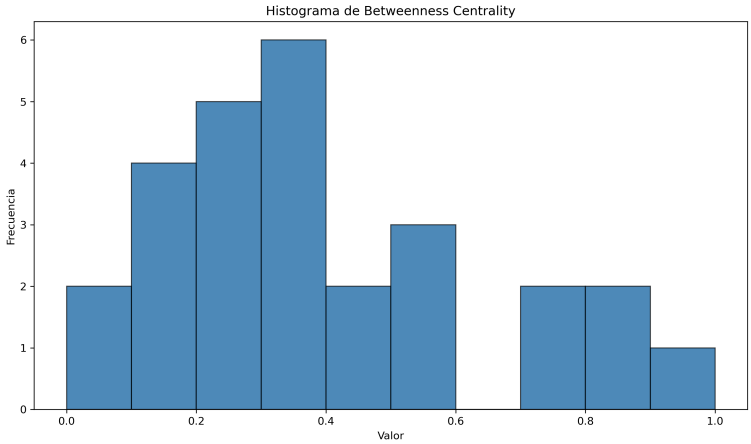


Figura 5.9: Histograma de centralidad de intermediación.

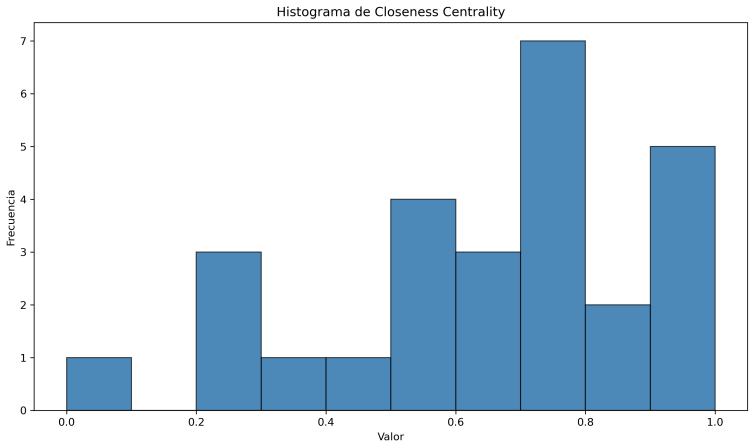


Figura 5.10: Histograma de centralidad de cercania.

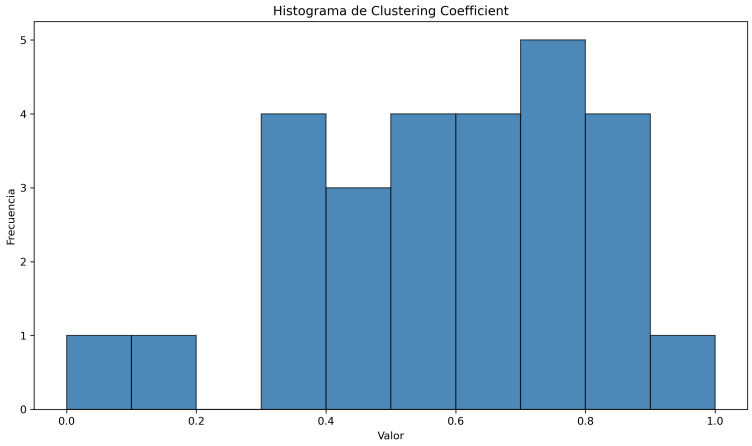


Figura 5.11: Histograma de coeficiente de agrupamiento.

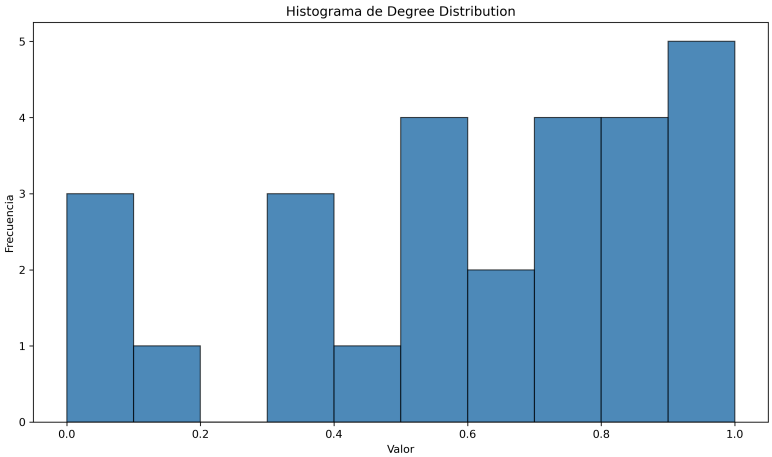


Figura 5.12: Histograma de distribución de grado.

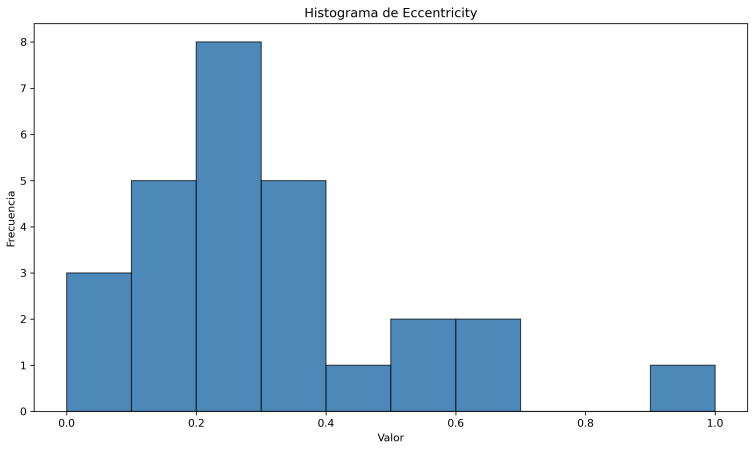


Figura 5.13: Histograma de excentricidad.

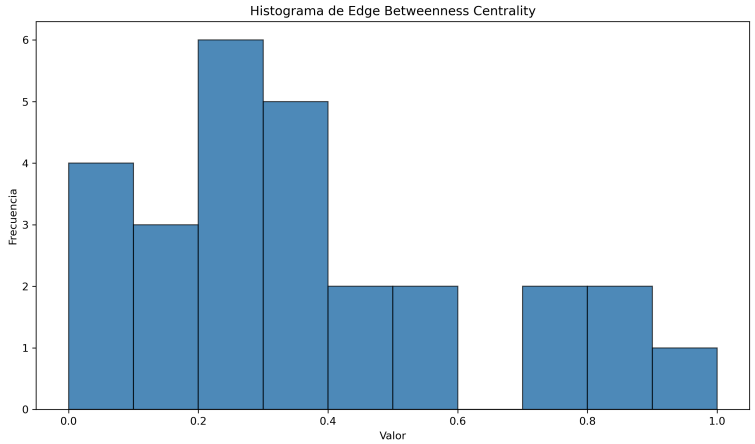


Figura 5.14: Histograma de centralidad de intermediación de aristas.

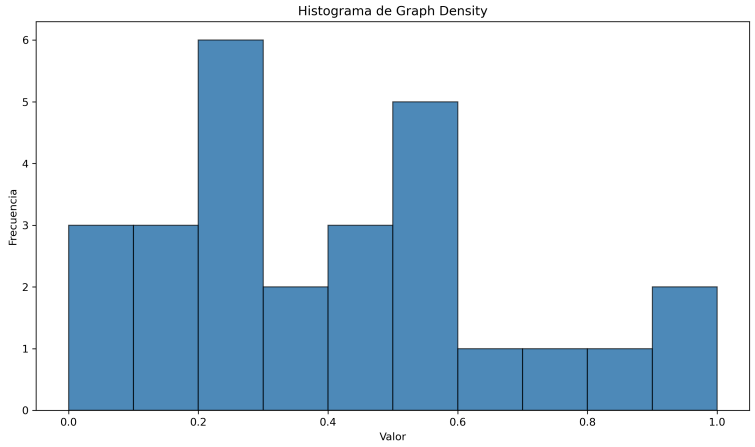


Figura 5.15: Histograma de densidad del grafo.

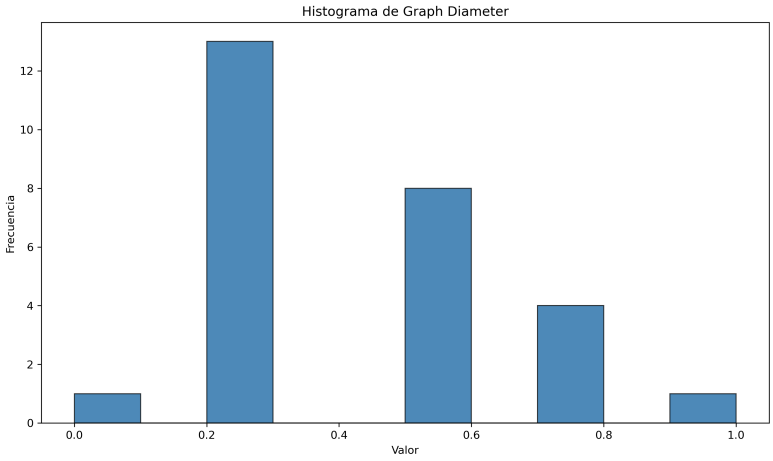


Figura 5.16: Histograma del diámetro del grafo.

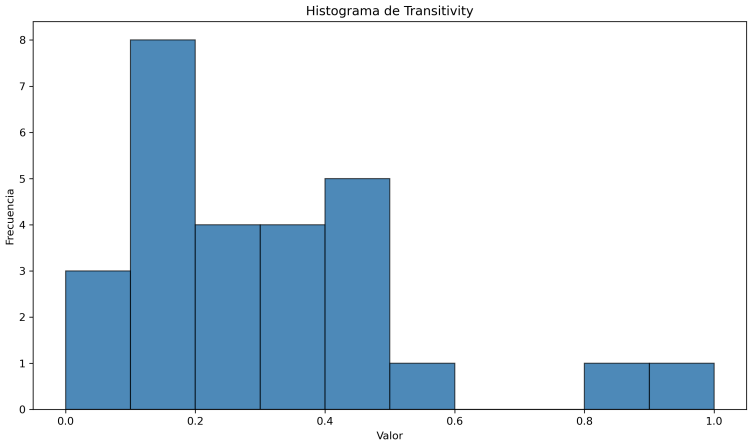


Figura 5.17: Histograma de transitividad.

Con base en el análisis de los histogramas, se puede señalar que las métricas presentan distintos patrones de distribución. Para evaluar estas distribuciones, se analizaron las mediciones de la media, mediana, desviación estándar y asimetría. Se observó que algunas métricas siguen una distribución relativamente normal con sesgos moderados, lo que facilita una segmentación estable, mientras que otras presentan un sesgo fuerte, lo que podría afectar la precisión del agrupamiento.

Entre las métricas con una distribución relativamente normal se encuentran: Betweenness Centrality, Closeness Centrality, Clustering Coefficient, Degree Distribution, Edge Betweenness Centrality, Graph Density y Graph Diameter. No obstante, algunas de estas métricas presentan un sesgo moderado, ya sea hacia la derecha o hacia la izquierda. Por ejemplo, Betweenness Centrality y Graph Density muestran un ligero sesgo hacia la derecha, mientras que Clustering Coefficient y Closeness Centrality presentan un sesgo hacia la izquierda.

En contraste, otras métricas, como Assortativity, Eccentricity y Transitivity, exhiben un sesgo más pronunciado hacia la derecha. Esto se debe a la presencia de valores atípicos significativamente más altos que el resto de los datos. Estos valores extremos pueden influir en la distribución de las métricas y, en algunos casos, puede afectar la estabilidad de la segmentación realizada por el algoritmo de agrupamiento.

5.5. Análisis de los datos de los pacientes con AD mediante K-medias

Tomando en cuenta el anásis de las métricas mediante los histogramas, para aplicar el algoritmo K-medias a los datos, es crucial definir un número adecuado de clústeres que permita capturar correctamente la estructura de los datos y minimizar el impacto de los valores atípicos y poder obtener resultados con una mayor representatividad. Para determinar el número k de clústeres en los que los datos serán agrupados. se emplearon dos métodos: el Método del Codo y el análisis de la Silueta.

Las Figuras 5.18 y 5.19 muestran los valores de k determinando que valor de k es el más óptimo para la creación de clústeres de los datos.

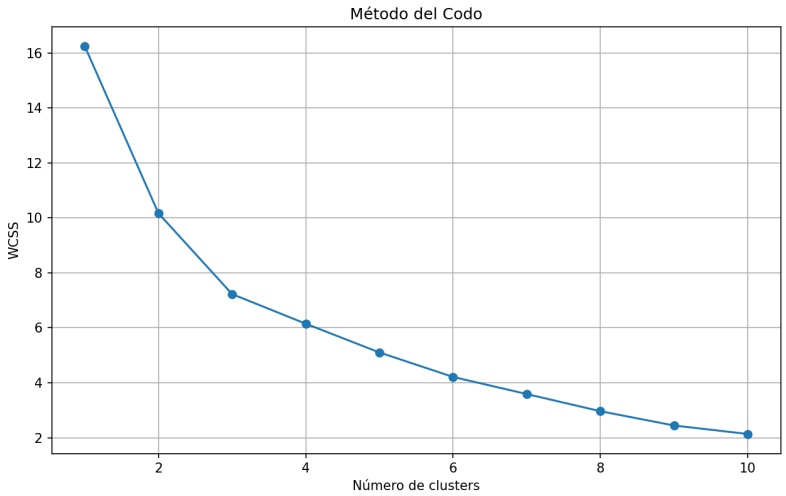


Figura 5.18: Cálculo del número k mediante el método del codo.

Como se puede observar, en la Figura 5.18 se identifica un punto de inflexión en k=3, lo que sugiere que los datos se agrupan de manera más representativa en tres clústeres. Por otro lado, la Figura 5.19 confirma la pertinencia del valor de k=3, ya que, el coeficiente de silueta alcanza su valor más alto en el valor 3. Por lo tanto, ambos métodos coinciden que el número óptimo de clústeres para este análisis es k=3.

Una vez determinado este valor, se procedió a ejecutar el algoritmo K-medias con el propósito de agrupar a los pacientes según el nivel de deterioro cog-

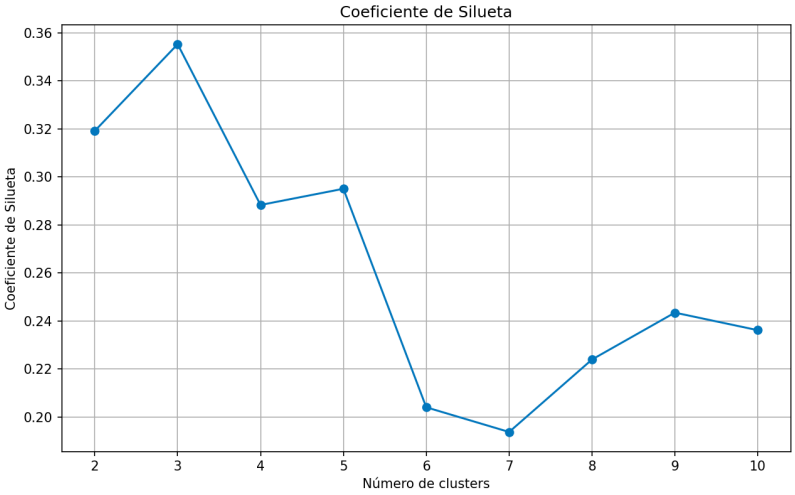


Figura 5.19: Cálculo del número k mediante el coeficiente de silueta.

nitivo, con base en las características estructurales de sus redes léxico-semánticas. El resultado de esta agrupación se presenta en la Figura 5.20.

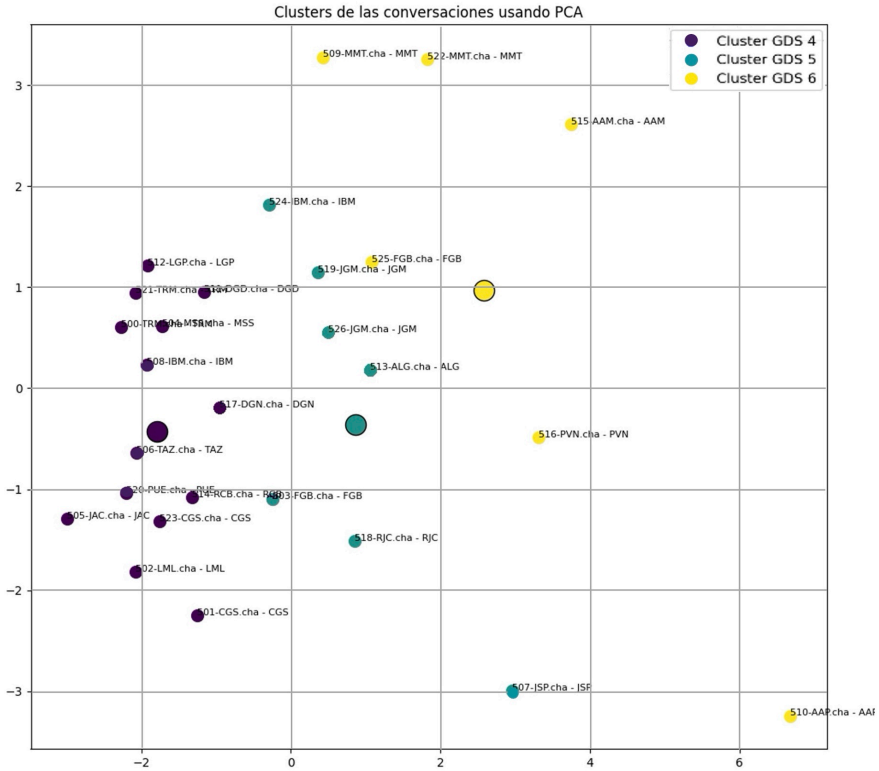


Figura 5.20: Agrupamiento de los datos de los pacientes con AD con el algoritmo K-medias.

La gráfica muestra la distribución de los pacientes en tres grupos, correspondientes a los niveles de deterioro GDS 4, GDS 5 y GDS 6. En términos generales, se observa una separación clara entre los grupos, lo que sugiere que el algoritmo identificó patrones estructurales característicos en las redes léxico-semánticas asociadas a cada nivel de deterioro.

Para evaluar la calidad del agrupamiento generado por el algoritmo K-medias, se calcularon métricas que permiten determinar en qué medida los grupos identificados reflejan la estructura real de los datos. En este estudio, se utilizaron las métricas de exactitud (accuracy) y precisión (precision).

La Tabla 5.7 muestra los resultados obtenidos a partir de las métricas de evaluación, los cuales permiten analizar la capacidad del modelo para agrupar correctamente a los pacientes con AD según los niveles GDS 4, 5 y 6.

Métrica	**Valor obtenido**
Exactitud	85.19 %
Precisión (GDS 4)	91.67 %
Precisión (GDS 5)	75 %
Precisión (GDS 6)	85.71 %

Tabla 5.7: Resultados de exactitud global y precisión por nivel de deterioro cognitivo, obtenidos en la categorización de pacientes con Alzheimer mediante el algoritmo K-medias.

Los resultados obtenidos reflejan una exactitud del 85.19 %, lo que sugiere que el modelo K-medias tuvo un desempeño sólido en la agrupación de los pacientes con AD según su nivel de deterioro cognitivo. No obstante, se observan diferencias en la precisión de los niveles GDS 4 (91.67 %), GDS 5 (75 %) y GDS 6 (85.71 %).

Las diferencias observadas podrían atribuirse a posibles errores de etiquetado en el corpus PerLA, derivados de una subestimación o sobrestimación del nivel de deterioro cognitivo en ciertos pacientes con AD. El análisis de las redes léxico-semánticas indica que algunos individuos presentan patrones lingüísticos más cercanos a los de otros niveles, lo que habría influido en las agrupaciones erróneas generadas por el modelo.

Por ejemplo, se identificó al paciente 512-LGP (ver Figura 5.6(b)), quien fue etiquetado en GDS 6 pero agrupado en GDS 4, lo que sugiere que sus redes léxico-semánticas presentan características más cercanas a etapas tempranas del deterioro. Asimismo, el paciente 519-JGM (ver Figura 5.2(b)) con

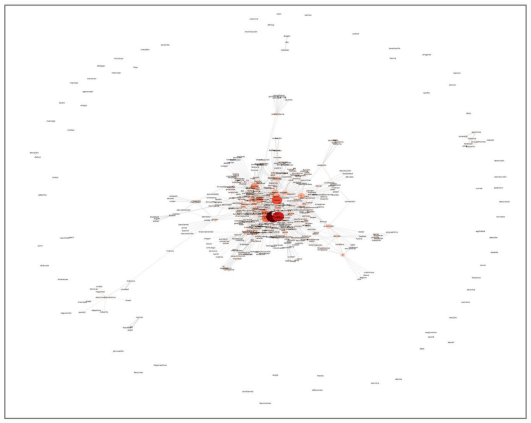
diagnóstico en GDS 4 fue agrupado en GDS 5, lo que podría indicar una progresión más avanzada del deterioro. Estos casos respaldan la hipótesis de que el desempeño del modelo podría ser aún más preciso.

5.6. Generación de las redes léxico-semánticas de los sujetos HC

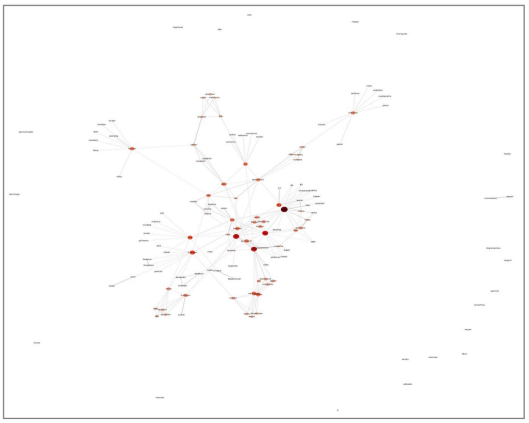
Para complementar este análisis, se generaron las redes léxico-semánticas de los sujetos HC. Esto permite establecer una comparación con las redes de los pacientes con Alzheimer y examinar las diferencias en su organización y estructura. A través de esta exploración, se busca identificar patrones distintivos en la conectividad y cohesión de estas redes, lo que podría proporcionar información relevante sobre los cambios lingüísticos asociados con la enfermedad.

La Figura 5.21 muestra algunas de las redes léxico-semánticas obtenidas de los sujetos HC.

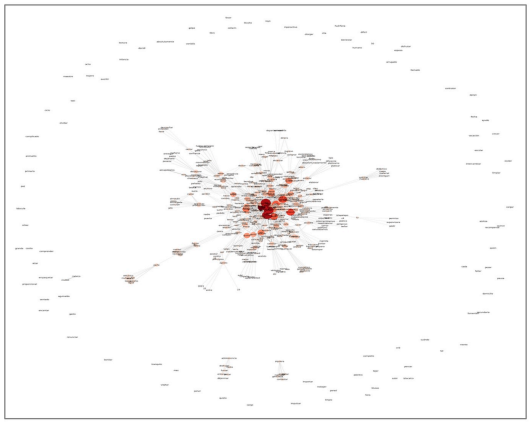
En las imágenes de las redes léxico-semánticas de los sujetos HC se identifican algunas configuraciones estructurales que, a simple vista, presentan ciertas similitudes con las redes de los pacientes con Alzheimer en los distintos niveles de afección. En particular, la red correspondiente al sujeto 601-MAM (ver Figura 5.21(b)) parece asemejarse a las redes de pacientes con AD en el nivel GDS 6, ya que principalmente se observa una menor densidad en comparación con las otras redes de los sujetos HC.



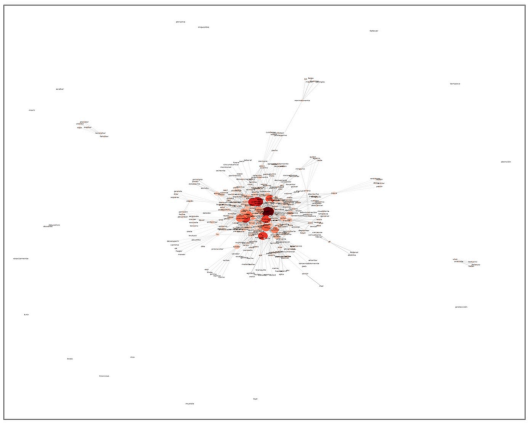
(a) 600-SBR



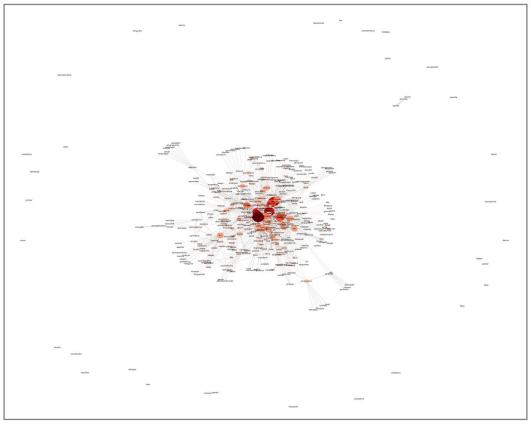
(b) 601-MAM



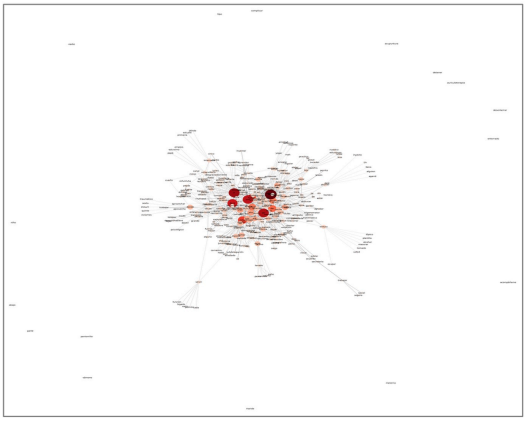
(c) 602-YPG



(d) 608-LCH



(e) 609-GCS



(f) 611-MAH

Figura 5.21: Redes léxico-semánticas de sujetos HC.

5.7. Métricas de los grafos de las redes léxico-semánticas del corpus

Para evaluar con precisión estas similitudes, se realizó un análisis de las redes incorporando las métricas de los sujetos HC, con el fin de evaluar su relación con los grupos previamente identificados en los pacientes con Alzheimer. Este análisis permite cuantificar e identificar patrones estructurales que podrían no ser evidentes a simple vista. Los resultados de las métricas se presentan en la Tabla 5.8.

Conversación	Centralidad de intermediación	Densidad del grafo	Coeficiente de agrupamiento	Centralidad de cercania	Transitividad	Diámetro del grafo	Distribución de grado	Excentricidad	Centralidad de intermediación de aristas	Asortatividad
500-TRM	0.376356	0.103013	0.501602	0.434232	0.064553	0.333333	0.041450	0.364221	0.379929	0.082522
501-CGS	0.165015	0.000000	0.488024	0.355014	0.068612	0.500000	0.050107	0.500557	0.200501	0.263531
502-LML	0.227945	0.355549	0.615275	0.477867	0.035053	0.500000	0.064429	0.619351	0.219334	0.115298
503-FGB	0.344295	0.069676	0.481533	0.335828	0.125639	0.666667	0.036845	0.793096	0.355768	0.189586
504-MSS	0.486959	0.178960	0.796005	0.508143	0.201465	0.333333	0.063185	0.299922	0.382631	0.180909
505-JAC	0.182132	0.029825	0.649022	0.460375	0.000000	0.166667	0.057879	0.282368	0.208879	0.036836
506-TAZ	0.341392	0.075645	0.613255	0.483668	0.059714	0.333333	0.056473	0.359700	0.304850	0.088391
507-JSP	0.388420	0.068666	0.567874	0.303441	0.295721	0.500000	0.039767	0.585758	0.363311	0.369292
508-IBM	0.426262	0.145661	0.479163	0.540257	0.092021	0.333333	0.065059	0.326522	0.343394	0.074676
509-MMT	0.855770	0.301467	0.341651	0.488695	0.207936	0.333333	0.043102	0.396224	0.709861	0.220514
510-AAP	0.400381	0.073427	0.173741	0.000000	0.526861	1.000000	0.000000	1.000000	0.533842	1.000000
511-DGD	0.423172	0.135492	0.389933	0.393208	0.132409	0.333333	0.041339	0.344718	0.409959	0.163189
512-LGP	0.411522	0.103258	0.305055	0.388980	0.033322	0.333333	0.032827	0.425841	0.433752	0.100225
513-ALG	0.455784	0.176484	0.321996	0.229379	0.236690	0.333333	0.022482	0.475999	0.513435	0.369557
514-RCB	0.383161	0.082376	0.671017	0.457150	0.095053	0.333333	0.054157	0.466228	0.331439	0.150529
515-AAM	0.892355	0.270298	0.000000	0.133446	0.332464	0.666667	0.005208	0.788846	1.000000	0.781833
516-PVN	0.379074	0.080071	0.255506	0.177109	0.077549	0.500000	0.006833	0.528372	0.524953	0.265821
517-DGN	0.434743	0.133365	0.586445	0.339083	0.248971	0.333333	0.042466	0.492849	0.401911	0.344324
518-RJC	0.355327	0.059777	0.398166	0.315399	0.255766	0.500000	0.041925	0.640072	0.336296	0.423090
519-JGM	0.457085	0.171727	0.669167	0.271820	0.219518	0.333333	0.023898	0.385672	0.514143	0.250795
520-PUE	0.276688	0.058382	0.640874	0.422268	0.074548	0.333333	0.048778	0.373325	0.284290	0.145135
521-TRM	0.508456	0.159658	0.575450	0.529364	0.142585	0.333333	0.057038	0.432459	0.414365	0.128699
522-MMT	0.803736	0.298189	0.312536	0.324022	0.230598	0.166667	0.022432	0.298762	0.816622	0.420079
523-CGS	0.284291	0.027828	0.605950	0.483185	0.062373	0.500000	0.055566	0.522109	0.265569	0.182311
524-IBM	0.730988	0.235420	0.506922	0.471323	0.149824	0.500000	0.045020	0.564828	0.602848	0.173604
525-FGB	0.561773	0.173456	0.333056	0.305040	0.156301	0.333333	0.024954	0.410180	0.584035	0.264273
526-JGM	0.588425	0.198901	0.184259	0.120773	0.255377	0.333333	0.000746	0.422302	0.785023	0.463177
600-SBR	0.139578	0.385159	0.895083	0.821353	0.344932	0.000000	0.275016	0.109614	0.070518	0.018599
601-MAM	1.000000	0.794829	0.856699	0.650390	1.000000	0.333333	0.150479	0.354497	0.424685	0.566570
602-YPG	0.063192	0.287704	0.893910	0.842526	0.286601	0.166667	0.321619	0.104881	0.040834	0.025239
603-CCV	0.164494	0.372959	0.975698	0.800370	0.457399	0.166667	0.268995	0.213950	0.076104	0.080747
604-ECH	0.152039	0.260951	0.848862	0.807068	0.218406	0.000000	0.223653	0.093005	0.085306	0.012105
605-GCH	0.200288	0.893924	0.983105	0.931852	0.829936	0.000000	0.450398	0.016206	0.055907	0.104901
606-MAG	0.000000	1.000000	1.000000	1.000000	0.804431	0.000000	1.000000	0.000000	0.000000	0.036255
607-GMR	0.100697	0.211426	0.912189	0.740992	0.286809	0.166667	0.225126	0.209633	0.068096	0.078871
608-LCH	0.200796	0.365614	0.856801	0.799866	0.331276	0.166667	0.244112	0.146465	0.093069	0.011303
609-GCS	0.172012	0.231769	0.910642	0.765715	0.254620	0.166667	0.204343	0.225140	0.095845	0.026617
610-LPM	0.185200	0.308927	0.862161	0.721294	0.365137	0.166667	0.209870	0.145978	0.100134	0.112604
611-MAH	0.227343	0.291751	0.874898	0.759113	0.290139	0.333333	0.194979	0.505941	0.119026	0.000000

Tabla 5.8: Métricas obtenidas de las redes léxico-semánticas de los pacientes con AD y sujetos HC que conforman el corpus.

A partir de estos datos, es posible identificar diferencias clave entre los grupos. Por ejemplo, se observa que la mayoria de las redes de los pacientes con AD presentan una menor densidad, menor coeficiente de agrupamiento y mayor diámetro, Sin embargo, algunas métricas de ciertos sujetos sanos presentan valores que se asemejan a los encontrados en pacientes con niveles de deterioro GDS 5 o GDS 6.

Para analizar con mayor profundidad la estructura de las redes léxico-semánticas y determinar si los sujetos HC presentan patrones diferenciados o similitudes con los grupos previamente identificados en los pacientes con AD, se procedió a aplicar el algoritmo K-medias a los datos presentados en la Tabla 5.8.

5.8. Análisis de los datos del corpus mediante el algoritmo K-medias

Este análisis posibilitará determinar si los sujetos HC forman un grupo homogéneo o si, por el contrario, se pueden distinguir subgrupos con características estructurales particulares. Los resultados obtenidos a partir de este análisis permitirán evaluar la relación entre los sujetos HC y los pacientes con AD previamente identificados, proporcionando una perspectiva cuantitativa sobre la organización léxico-semántica y su posible vinculación con el deterioro cognitivo.

Para ello, se emplearon nuevamente los métodos de codo y coeficiente de silueta con el fin de determinar una segmentación adecuada de los grupos en función de las características estructurales de las redes léxico-semánticas. Las gráficas correspondientes se encuentran en las Figuras 5.22 y 5.23, respectivamente.

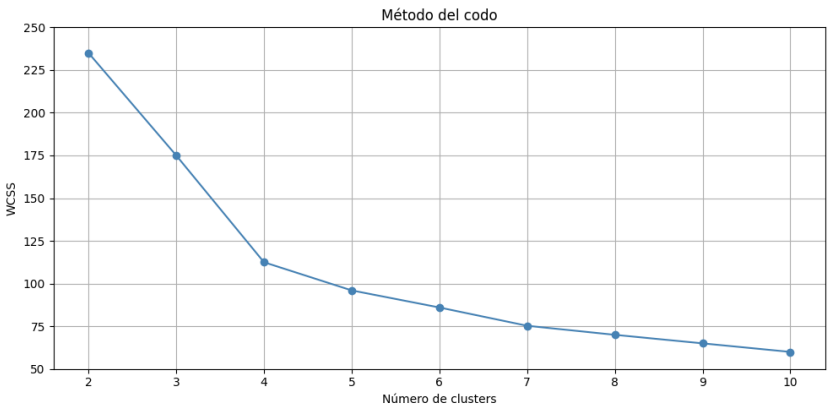


Figura 5.22: Cálculo del número k mediante el método del codo.

En este caso, la Figura 5.22 presenta un punto de inflexión más pronunciado en k=4, por lo que se considera como el valor más óptimo. Del mismo modo,

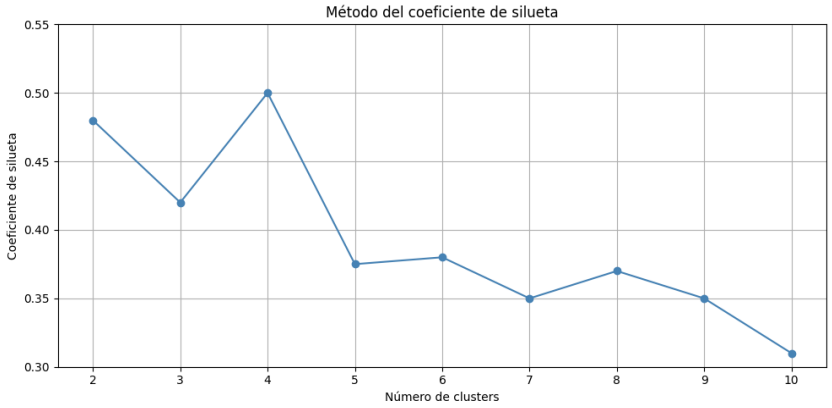


Figura 5.23: Cálculo del número k mediante el coeficiente de silueta.

en la Figura 5.23 el coeficiente de silueta alcanza su valor máximo en k=4, lo que sugiere que la forma más adecuada de agrupar los datos es mediante una segmentación en 4 grupos.

Con base en los resultados obtenidos a partir del análisis del método del codo y el coeficiente de silueta, se procede a aplicar el algoritmo K-medias para agrupar las redes léxico-semánticas en cuatro grupos. El resultado de este agrupamiento se presenta en la Figura 5.24, donde cada punto corresponde a una red léxico-semántica representada mediante un vector de características estructurales, y su asignación a uno de los grupos identificados.

En este caso, se puede observar en la Figura 5.24, una separación clara entre los sujetos HC y los pacientes con AD. Los sujetos HC se agruparon en el clúster 0 (representado en color amarillo), mientras que los pacientes con AD se distribuyen en los clústeres 4, 5 y 6. Esta distribución sugiere que las métricas estructurales extraídas de las redes léxico-semánticas permiten distinguir con precisión entre ambos grupos.

No obstante, entre los pacientes con AD, la separación entre los grupos no es tan clara, ya que algunos puntos correspondientes a diferentes niveles de deterioro cognitivo aparecen cercanos entre sí. Esto podría reflejar similitudes estructurales en las redes de los pacientes con niveles de afectación GDS 5 y GDS 6.

Para profundizar en el análisis del rendimiento del modelo, se calcularon métricas cuantitativas que permiten evaluar su capacidad de agrupamiento.

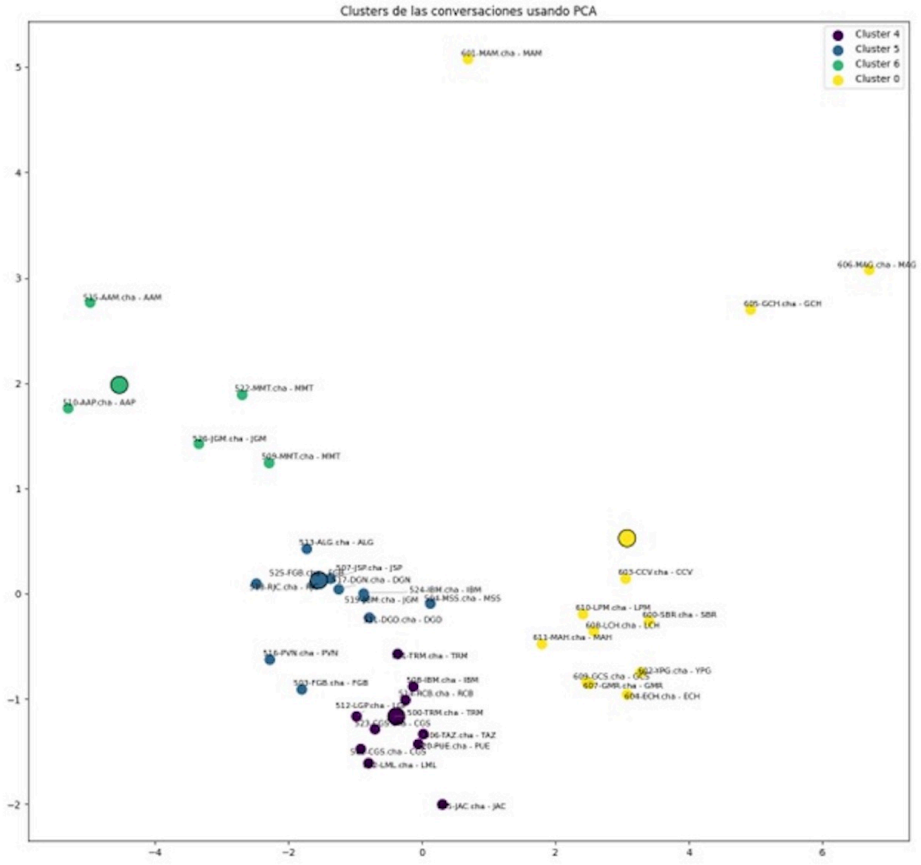


Figura 5.24: Agrupamiento de los datos de los pacientes con AD y sujetos HC con el algoritmo K-medias.

La Tabla 5.9 presenta los resultados de exactitud global y precisión por clase, los cuales ofrecen una visión más detallada sobre qué tan efectivamente el modelo logró distinguir entre los distintos niveles de deterioro cognitivo (GDS 4, 5 y 6) y el grupo de sujetos HC, a partir de las características estructurales de sus redes léxico-semánticas.

Métrica	**Valor obtenido**
Exactitud	84.61 %
Precisión (GDS 4)	83.34 %
Precisión (GDS 5)	87.50 %
Precisión (GDS 6)	71.42 %
Precisión (HC)	100 %

Tabla 5.9: Resultados de exactitud global y precisión obtenidos por el modelo K-medias en la categorización de los pacientes con AD según su nivel de deterioro cognitivo (GDS 4, GDS 5 y GDS 6) y los sujetos HC.

Los resultados muestran una exactitud global del 84.61 %, lo que indica que el modelo logró asignar correctamente la mayoría de las redes léxico-semánticas a sus respectivos grupos. Este valor sugiere que, en general, las métricas estructurales empleadas permiten representar de manera significativa los patrones lingüísticos asociados a cada nivel de deterioro cognitivo y al grupo de sujetos sanos.

En lo que respecta a la precisión por grupo, el modelo obtuvo un 100 % de precisión para el grupo de sujetos HC. Este resultado indica que, en términos generales, las redes correspondientes a personas sanas presentan propiedades estructurales distintivas y consistentes, en contraste con las de los pacientes con AD. Dicha precisión confirma la eficacia del modelo para identificar el grupo sano sin ambigüedades.

En el caso del grupo GDS 4, la precisión alcanzada fue de 83.34 %, lo cual sugiere que el modelo logró identificar de manera adecuada la mayoría de las redes asociadas a pacientes en etapas iniciales del deterioro. Este desempeño refleja que, aunque hay algunas similitudes con otros niveles, las redes en esta etapa aún conservan una estructura diferenciada.

Para el grupo GDS 5, la precisión fue ligeramente superior, con un valor de 87.50 %, indicando que el modelo también puede reconocer con buen nivel de acierto las redes correspondientes a esta fase intermedia del deterioro cognitivo.

Por otro lado, el grupo GDS 6 obtuvo una precisión del 71.42 %, la más baja entre los niveles evaluados. Este resultado podría deberse a la mayor similitud estructural entre las redes léxico-semánticas correspondientes a las etapas avanzadas del deterioro cognitivo, lo que dificulta una delimitación precisa entre grupos. Asimismo, no puede descartarse la posibilidad de errores de etiquetado en el corpus PerLA, especialmente considerando la complejidad inherente al diagnóstico clínico y a la progresión individual de la enfermedad, factores que podrían haber influido en los resultados del agrupamiento.

Los resultados confirman que el modelo logra una clara separación entre los pacientes con AD y los sujetos HC, lo que señala la utilidad de las métricas estructurales para capturar diferencias significativas en sus redes léxico-

semánticas. Además, el desempeño en la categorización de los niveles de deterioro cognitivo dentro del grupo con AD fue particularmente sólido en las etapas intermedias. No obstante, se identificó una menor capacidad para diferenciar las etapas más avanzadas, posiblemente debido a similitudes estructurales entre las redes o a imprecisiones en la asignación diagnóstica original, factores que podrían haber influido en la calidad del agrupamiento.

Capítulo 6

Conclusiones y trabajo futuro

6.1. Conclusiones

El diagnóstico de la enfermedad de Alzheimer en una etapa temprana continúa siendo una labor desafiante, ya que los primeros signos pueden ser sutiles y difíciles de distinguir de los efectos normales del envejecimiento. Uno de estos signos es la alteración del proceso lingüístico, un indicador significativo, dado que el deterioro semántico puede manifestarse antes que otros síntomas cognitivos evidentes. En este contexto, el estudio del deterioro lingüístico surge como un enfoque valioso para la detección temprana de la enfermedad. Identificar estos cambios en las fases iniciales de la enfermedad podría contribuir al desarrollo de estrategias de intervención más eficaces, dirigidas a retrasar el impacto del Alzheimer y preservar la calidad de vida de los pacientes.

Con este propósito, la presente investigación se orientó al análisis de las redes léxico-semánticas generadas a partir de matrices de coocurrencia, utilizando un repositorio de conversaciones con pacientes que padecen la enfermedad de Alzheimer (AD) y sujetos sanos (HC). Posteriormente, se evaluó la estructura de estas redes mediante métricas basadas en la teoría de grafos y técnicas de agrupamiento, con la finalidad de detectar de forma temprana el deterioro cognitivo asociado a la enfermedad de Alzheimer en adultos mayores de habla hispana.

Como resultado de este análisis, se obtuvieron hallazgos que reflejan varia-

ciones significativas en la estructura de las redes léxico-semánticas de los pacientes con AD y sujetos HC.

Por un lado, la agrupación de los pacientes con AD en los niveles de deterioro cognitivo GDS 4, 5 y 6 presentó una exactitud del 85.19 %, lo que indica que el modelo logró asignar correctamente a una proporción significativa de los pacientes. Sin embargo, la distinción entre los niveles más avanzados de afección presentó algunas dificultades. Mientras que GDS 4 fue identificado con alta precisión (91.67 %)„ los niveles GDS 5 y GDS 6 obtuvieron valores inferiores (75.67 % y 85.71 %, respectivamente).

Una posible explicación de estos resultados es que, si bien las métricas de las redes complejas utilizadas permiten capturar diferencias estructurales en las redes léxico-semánticas de los pacientes con AD, la agrupación no fue completamente precisa. Esto podría estar relacionado con la metodología empleada para etiquetar a los pacientes según sus conversaciones dentro de los niveles GDS, ya que no se cuenta con información detallada sobre los criterios utilizados en dicha asignación. En consecuencia, la posible imprecisión en el etiquetado inicial de los niveles de deterioro podría haber influido en los resultados obtenidos por el modelo.

Por otro lado, al integrar a los sujetos HC en el análisis, se observó una segmentación clara entre pacientes con enfermedad de Alzheimer (AD) y personas sanas. El modelo alcanzó una exactitud global del 84.61 % y una precisión del 100 % para el grupo HC, lo cual indica que las métricas estructurales de las redes léxico-semánticas capturan eficazmente las diferencias lingüísticas asociadas al deterioro cognitivo. Estos resultados refuerzan la hipótesis de que el deterioro semántico impacta de manera significativa en la organización de las redes léxico-semánticas, generando patrones distinguibles entre ambos grupos. No obstante, los valores de precisión para los niveles GDS 4 (83.34 %), GDS 5 (87.50 %) y GDS 6 (71.42 %) sugieren que la separación entre los distintos grados de afección es más compleja, posiblemente debido a la progresiva superposición de características estructurales conforme avanza la enfermedad, así como a posibles inconsistencias en la asignación diagnóstica.

En general, los resultados de este estudio revelan que el análisis de redes léxico-semánticas, sustentado en métricas estructurales y técnicas de agru-

pamiento, constituye una herramienta eficaz para identificar la reducción en la conectividad semántica, la disminución en la cohesión entre conceptos y la fragmentación progresiva de la red lingüística en pacientes con AD. Esta aproximación no solo permite diferenciar entre sujetos HC y pacientes con AD, sino también examinar cómo se modifica la estructura de las redes semánticas conforme avanza el deterioro cognitivo.

6.2. Trabajo futuro

Dado que el tamaño del corpus es una de las limitaciones del estudio, futuros trabajos podrían centrarse en ampliar la muestra de sujetos de habla hispana, incluyendo mayor diversidad en términos de edad, nivel educativo y contexto sociocultural. Asimismo, sería relevante evaluar el impacto de las diferencias lingüísticas entre el español de Valencia y el español de México en la construcción de redes léxico-semánticas, esto permitiría mejorar la robustez del modelo.

Por otro lado, también sería importante incorporar nuevas métricas de redes complejas para evaluar si es posible mejorar la segmentación por niveles de deterioro cognitivo (GDS 4, 5 y 6), ya que no fue completamente precisa. Esto subraya la necesidad de explorar métricas adicionales de teoría de modelos de redes complejas que permitan distinguir con mayor precisión los distintos estadios de la enfermedad, especialmente en los niveles intermedios de deterioro, donde la segmentación presentó mayores dificultades.

Además, un enfoque futuro relevante sería realizar un análisis longitudinal, en el que se estudie la evolución de las redes léxico-semánticas en los mismos pacientes con AD a lo largo del tiempo. Esto permitiría evaluar si los cambios en la conectividad semántica pueden predecir la progresión del deterioro cognitivo en fases tempranas.

Finalmente, sería interesante aplicar este enfoque a otras enfermedades neurodegenerativas, como la demencia frontotemporal, el Parkinson o la demencia vascular. Este tipo de análisis podría permitir la identificación de patrones diferenciadores entre distintas enfermedades, facilitando la creación de agrupaciones específicas basadas en las características lingüísticas propias de cada patología.

De esta manera, se podría evaluar si los modelos de segmentación pueden no solo distinguir entre sujetos HC y pacientes con AD, sino también entre distintos tipos de demencia, proporcionando una herramienta diagnóstica más precisa.

Estas líneas de investigación tienen el potencial de contribuir al desarrollo de estrategias más efectivas para la detección y seguimiento del deterioro cognitivo a partir del análisis del proceso lingüistico.

Referencias

[1] Demencia: datos y cifras. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/dementia

[2] Demencia: datos y cifras. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] S. Adhikari, S. Thapa, U. Naseem, P. Singh, H. Huo, G. Bharathy, and M. Prasad, "Exploiting linguistic information from nepali transcripts for early detection of alzheimer's disease using natural language processing and machine learning techniques," *International Journal of Human-Computer Studies*, vol. 160, p. 102761, 2022.

[4] A. Baddeley, "Working memory," *Current biology*, vol. 20, no. 4, pp. R136–R140, 2010.

[5] J. B. Barrón-Martínez and N. Arias-Trejo, "Word association norms in mexican spanish," *The Spanish journal of psychology*, vol. 17, p. E90, 2014.

[6] J. B. Barrón-Martínez, V. Mijangos, N. Arias-Trejo, and G. Bel-Enguix, *Métodos para explorar las redes léxicas desde la infancia hasta la etapa adulta*. Instituto de Ingeniería, 2021, ch. Maquetado Relaciones entre palabras en la niñez y el envejecimiento, pp. 81–104.

[7] C. A. Benavides-Caro, "Deterioro cognitivo en el adulto mayor," *Revista Mexicana de anestesiología*, vol. 40, no. 2, pp. 107–112, 2017.

[8] L. Bertola, N. B. Mota, M. Copelli, T. Rivero, B. S. Diniz, M. A. Romano-Silva, S. Ribeiro, and L. F. Malloy-Diniz, "Graph analysis of verbal

fluency test discriminate between patients with alzheimer's disease, mild cognitive impairment and normal elderly controls," *Frontiers in aging neuroscience*, vol. 6, p. 185, 2014.

[9] J. Borge-Holthoefer and A. Arenas, "Semantic networks: Structure and dynamics," *Entropy*, vol. 12, no. 5, pp. 1264–1302, 2010.

[10] J. F. Cecato, J. E. Martinelli, R. Izbicki, M. S. Yassuda, and I. Aprahamian, "A subtest analysis of the montreal cognitive assessment (moca): which subtests can best discriminate between healthy controls, mild cognitive impairment and alzheimer's disease?" *International psychogeriatrics*, vol. 28, no. 5, pp. 825–832, 2016.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] M. Coscia, "The atlas for the aspiring network scientist," 2021.

[13] J. Eisenstein, "Natural language processing," *Jacob Eisenstein*, 2018.

[14] P. D. Emmady, C. Schoo, and P. Tadi, "Major neurocognitive disorder (dementia)," in *StatPearls [Internet]*. StatPearls Publishing, 2022.

[15] P. Flach, *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press, 2012.

[16] M. A. Flores-Coronado, E. M. Vargas García, A. Minto García, G. Bel-Enguix, and N. Arias-Trejo, *Redes léxico-semánticas en el envejecimiento típico*. Facultad de Ingeniería, 2021, ch. Maquetado Relaciones entre palabras en la niñez y el envejecimiento, pp. 110–135.

[17] M. Grossman and P. Koenig, "Semantic memory," *Encyclopedia of Cognitive Science. Academic Press, San Diego*, 2001.

[18] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[19] X. Jin and J. Han, "K-means clustering," *Encyclopedia of machine learning*, pp. 563–564, 2011.

[20] J. Lin, "Scalable language processing algorithms for the masses: A case study in computing word co-occurrence matrices with mapreduce,"

in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 419–428.

[21] H. Lindsay, J. Tröger, and A. König, "Language impairment in alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning," *Frontiers in aging neuroscience*, vol. 13, p. 642033, 2021.

[22] S. Mårdh, K. Nägga, and S. Samuelsson, "A longitudinal study of semantic memory impairment in patients with alzheimer's disease," *Cortex*, vol. 49, no. 2, pp. 528–533, 2013.

[23] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[24] M. Newman, *Networks: An Introduction*. Oxford university press, 2010.

[25] A. Nikolaev, E. Higby, J. Hyun, and S. Ashaie, "Lexical decision task for studying written word recognition in adults with and without dementia or mild cognitive impairment," *JoVE (Journal of Visualized Experiments)*, no. 148, p. e59753, 2019.

[26] J. L. Pérez Mantero, "Interacción y predictividad: Los intercambios conversacionales con hablantes con demencia tipo alzhéimer," *revista de investigación Lingüística*, vol. 17, pp. 97–118, 2014.

[27] J. L. Pérez Mantero *et al.*, "El déficit lingüístico en personas con demencia de tipo alzhéimer: breve estado de la cuestión," 2012.

[28] R. L.-H. Sánchez, *Psicología del lenguaje*. Ediciones Pirámide, 2003.

[29] D. Sarkar, *Text analytics with python*. Springer, 2016, vol. 2.

[30] D. Saumier and H. Chertkow, "Semantic memory," *Current neurology and neuroscience reports*, vol. 2, no. 6, pp. 516–522, 2002.

[31] S. Shaji, J. F. A. Ronickom, A. K. Ramaniharan, and R. Swaminathan, "Study on the effect of extreme learning machine and its variants in differentiating alzheimer conditions from selective regions of brain mr images," *Expert Systems with Applications*, vol. 209, p. 118250, 2022.

[32] C. L. Stopford, J. C. Thompson, D. Neary, A. M. Richardson, and J. S. Snowden, "Working memory, attention, and executive function in alzheimer's disease and frontotemporal dementia," *Cortex*, vol. 48, no. 4, pp. 429–446, 2012.

[33] TalkBank Project, "Dementiabank pitt corpus," https://dementia.talkbank.org/access/English/Pitt.html, 2025, accessed: 2025-07-18.

[34] A. Villarejo and V. Puertas-Martín, "Utilidad de los test breves en el cribado de demencia," *Neurología*, vol. 26, no. 7, pp. 425–433, 2011.

[35] S. Wasserman and K. Faust, "Social network analysis: Methods and applications," 1994.

[36] L. M. Wright, M. De Marco, and A. Venneri, "A graph theory approach to clarifying aging and disease related changes in cognitive networks," *Frontiers in aging neuroscience*, vol. 13, p. 676618, 2021.

[37] Z. Yao, H. Wang, W. Yan, Z. Wang, W. Zhang, Z. Wang, and G. Zhang, "Artificial intelligence-based diagnosis of alzheimer's disease with brain mri images," *European Journal of Radiology*, p. 110934, 2023.

[38] R. A. L. Zunini, F. Knoefel, C. Lord, F. Dzuali, M. Breau, L. Sweet, R. Goubran, and V. Taler, "Event-related potentials elicited during working memory are altered in mild cognitive impairment," *International Journal of Psychophysiology*, vol. 109, pp. 1–8, 2016.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00116
Matrícula: 2222800233

Sistema de clasificación de deterioro cognitivo mediante el análisis de redes léxico-semánticas en adultos mayores de habla hispana (SIDERLEX)

En la Ciudad de México, se presentaron a las 12:00 horas del día 7 del mes de agosto del año 2025 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DRA. GEMMA BEL ENGUIX
DR. RENE MACKINNEY ROMERO
DR. BENJAMIN MORENO MONTIEL

Bajo la Presidencia de la primera y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

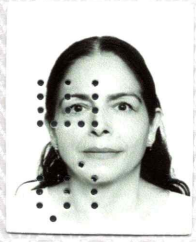
MAESTRA EN CIENCIAS (CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN)

DE: MARIA ANGELICA GUTIERREZ LOZANO

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, la presidenta del jurado comunicó a la interesada el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.



MARIA ANGELICA GUTIERREZ LOZANO
ALUMNA

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

Roman Linares Romero
DR. ROMAN LINARES ROMERO

PRESIDENTA

DRA. GEMMA BEL ENGUIX

VOCAL

DR. RENE MACKINNEY ROMERO

SECRETARIO

DR. BENJAMIN MORENO MONTIEL