Estudio Longitudinal del Lenguaje Escrito para Detectar Deterioro Cognitivo

Oliver Raúl Martínez González
PCyTI
UAM-I
CDMX, México
cbi2243801534@xanum.uam.mx

Ricardo Marcelín Jiménez
dept. Ingeniería Eléctrica, CBI
UAM-I
CDMX, México
calu@xanum.uam.mx

Óscar Yáñez Suárez
dept. Ingeniería Eléctrica, CBI
UAM-I
CDMX, México
oyanez@izt.uam.mx

***Abstract*—El presente es un estudio longitudinal enfocado en el uso del Procesamiento del Lenguaje Natural, análisis sintáctico y gramatical, así como teoría de grafos para identificar deterioro cognitivo en la enfermedad de Alzheimer, analizando escritos a lo largo del tiempo. Se busca establecer una nueva metodología de diagnóstico clínico automatizada que pueda detectar la sintomatología de forma temprana y remota. El análisis se centra en las novelas de dos escritoras: Jean Iris Murdoch y Phyllis Dorothy James, utilizando sus carreras literarias como ventanas temporales para modelar el cambio lingüístico. La metodología implica el preprocesamiento del texto, la tokenización, la extracción de medidas léxicas y sintácticas como la distancia de dependencia promedio, el análisis de similitud y agrupación de novelas. Los resultados preliminares sugieren que la proporción de ciertas categorías gramaticales se vuelve errática en las obras de Murdoch, mientras que las de James permanecen estables en un rango, ofreciendo evidencia del potencial de estos cambios como marcadores digitales de la progresión de la enfermedad.**

***Index Terms*—NLP, deterioro cognitivo, Alzheimer, estilometría, grafos**

I. Introducción

La demencia es un problema de salud pública, datos proporcionados por la Organización Mundial de la Salud (OMS) indican que es la 7ma causa de muerte y una de las principales causas de discapacidad y dependencia entre las personas de edad en el mundo, la enfermedad de Alzheimer (AD) conforma entre el 60 y 70% de los casos [19]. La mortalidad por AD es un problema de salud pública en México con tendencia creciente, especialmente entre mujeres y adultos mayores [7].

Se estima que en el mundo cada año se dan alrededor de 10 millones de casos nuevos, las herramientas para diagnóstico clínico para AD van desde las que evalúan de manera fisiológica (electroencefalograma, resonancia magnética (RM), RM funcional, tomografía computarizada y por emisión de positrones), pruebas de evaluación mental que toman en cuenta edad, escolaridad, memoria, lenguaje, funciones ejecutivas, habilidades visuales y motrices (fluencia verbal semántica (animales), Mini-Cog^{TM}, Mini-Examen del Estado Mental (MMSE), Evaluación Cognitiva de Montreal, por mencionar algunas [20]), también hay pruebas bioquímicas (acumulación de la proteína beta amiloide) y genéticas (APOE4), todo este tipo de pruebas requieren de personal especializado y capacitado para aplicarlas, y la presencia del sujeto al que se le aplica el estudio, sin embargo nos enfrentamos a dos problemas: 1) el contacto con el médico suele ser en etapas avanzadas de la enfermedad que por desgracia es crónico degenerativa y actualmente no tiene cura; 2) la disponibilidad de equipo y personal en los hospitales para evaluar de manera fisiológica, bioquímica y genética suele ser de alta demanda ya que varias de las herramientas no sólo se utilizan para diagnosticar este tipo de enfermedad, por lo tanto es necesario encontrar una nueva metodología para detectar en su etapa temprana la sintomatología de AD. Se propone la exploración de una nueva herramienta de diagnóstico clínico automatizada que analice texto y pueda identificar patrones de deterioro, la cual pueda aplicarse de forma remota, sin necesidad de la presencia del sujeto de estudio, simplemente analizando sus escritos digitales por ejemplo: artículos científicos, correo, conversaciones en aplicaciones de redes sociales, libros como es el caso de este estudio, o cualquier texto digitalizado.

Agradeciendo al Posgrado en Ciencias y Tecnologías de la Información de la UAM-I y a mis asesores anexos en las referencias sin los cuales no hubiera sido posible la escritura de este artículo.

II. Antecedentes

Diferentes autores han estudiado los cambios que sufre el lenguaje escrito. Se sabe que el AD afecta las capacidades de comunicación de quienes lo padecen, ya sea de forma oral o escrita, es por ello que nos proponemos analizar el lenguaje escrito de dos autores en sus obras a lo largo de sus vidas, uno que se sabe tuvo AD en su última etapa de vida, Iris Murdoch (IM), y otro que no, Dorothy James (PDJ). Mediante procesamiento del lenguaje natural (NLP), medidas léxico sintácticas, gramaticales, estadística y teoría de grafos buscaremos patrones en el lenguaje que puedan indicar la evolución y posible comienzo de la enfermedad.

En [22] analizaron cuatro libros de IM (a quien se le diagnóstico y comprobó post mortem la presencia de AD), "Under the Net" (1954), "The Sea, The Sea" (1978), "The Green Knight" (1994) y "Jackson's Dilemma" (1995). Los libros fueron escaneados, digitalizados como imágenes TIFF y posteriormente convertidos a texto usando Tesseract un programa de reconocimiento óptico de caracteres (OCR) disponible gratuitamente. Ellos extrajeron 20 pasajes al azar no contiguos, evitando seleccionar diálogo. Para medir la complejidad gramatical de las oraciones en

inglés, implementaron tres enfoques computacionales que utilizan el analizador sintáctico de Stanford ([13]), que produce una representación de árbol de las oraciones de entrada. Utilizaron dos de los enfoques de puntuación de la complejidad gramatical que se basan en contar el número de ramas y la profundidad del elemento léxico. El tercer método es el de puntuación de Yngve [26].

Como lo mencionan en [10], el estudio Nun[1], utilizó medidas de complejidad gramatical y densidad de ideas para cuantificar el contenido sintáctico y semántico de oraciones escritas. En su estudio sobre las obras de IM optaron por eliminar el formato de los textos, excepto los saltos de línea y párrafo, los pasajes de diálogos o citas directas los identificaron como caracteres especiales. Con esto construyeron concordancias de dos tipos: (i) se utilizó el software Concordance. para transformar los textos completos en listas de palabras alfabéticas, mostrando la frecuencia de cada palabra por tipo y mostrando el contexto de cada suceso y (ii) el mismo software se configuró para seleccionar muestras aleatorias de 100 palabras de cada libro, para producir listas alfabéticas similares.

En [14], analizaron a tres escritores británicos, la escritora y dramaturga Agatha Mary Clarissa Miller mejor conocida como Agatha Christie, quien se sospecha padecía alguna enfermedad neurodegenerativa en su última etapa de vida, no diagnosticada, y la escritora británica de novelas policiacas PDJ, quien fue el patrón de una escritora con envejecimiento "normal" y también a la escritora y filósofa IM. Utilizaron los textos completos de quince a veinte novelas por cada una de las escritoras anteriormente mencionadas. Buscaron signos tempranos de demencia en las obras, centrándose en marcadores léxicos y sintácticos: tamaño del vocabulario, repetición, especificidad de las palabras, déficit de clases de palabras, rellenos, complejidad gramatical y el uso de la voz pasiva. Los textos requirieron varios niveles de procesamiento, por ejemplo secuencias de palabras no lematizadas, secuencias lematizadas, etiquetado del discurso y análisis sintácticos completos. Separaron los signos de puntuación y clíticos[2] de tokens de palabras a las que están adjuntos. Lematizaron las palabras con el método de Murphy de Natural Language Toolkit (NLTK [18]) y WordNet. Determinaron los límites de las oraciones con un algoritmo y generaron un árbol para cada oración, utilizando el analizador Charniak [17].

En [15] proponen la distancia de dependencia (DD) de una oración como una medida en la dificultad de comprensión de un lenguaje, la DD puede definirse como la cantidad de palabras que separan a una palabra de su gobernante. Es aceptado en lingüística teórica y computacional que la estructura sintáctica de una oración consiste en la dependencia entre palabras: es una relación binaria (arista dirigida) entre 2 unidades lingüísticas, usualmente asimétrica, donde una unidad actúa como gobernante y la otra como dependiente (la arista va del gobernante al dependiente), se clasifica en término de relaciones gramaticales generales y se muestra convencionalmente encima de la arista como se muestra en la Fig. 1. Uno de sus hallazgos más representativos es que los idiomas humanos tienden a un umbral de distancia de dependencia promedio (MDD) minimizado y que se encuentra dentro de la capacidad de la memoria de trabajo. El cálculo de la MDD se puede ver en la Fig. 2.

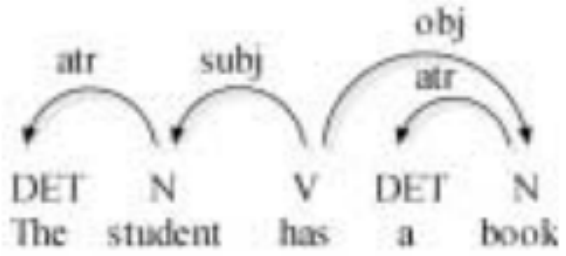

Fig. 1. Estructura de dependencia de la oración "The student has a book", imagen tomada de [15].

Fig. 2. Cálculo de la MDD a traves de la DD, imagen tomada de [15].

A. Cambios en el lenguaje escrito cuando se padece una enfermedad neurodegenerativa

Diferentes autores [3]–[5], [8]–[11], [16], [23] nos dan evidencias del cambio que sufre el lenguaje en una persona que padece algún tipo de demencia.

Como se puede apreciar en la "Guía de Consulta de los criterios diagnósticos del DSM-5™", en la Tabla 1 si observamos el dominio de Lenguaje de la página 323 donde nos muestran ejemplos de síntomas u observaciones en personas con deterioro cognitivo leve, como pueden ser, omisiones sutiles o usos incorrectos de artículos, preposiciones, verbos auxiliares, etc. [1].

En [8] menciona cómo un grupo de investigadores de Londres realizó un estudio de narrativa, comparando el estilo de las obras de IM "Jackson's Dilemma" (última obra), "Under the Net" (primer obra) y "The Sea, The Sea"(una de sus mejores obras):

> ... "La conclusión a la que se llegó fue que aunque en su última novela la estructura y la gramática

[1](citado de [10]) ..."Fue posible gracias a la cooperación de miembros de una comunidad religiosa cerrada, que se inscribieron en un estudio prospectivo sobre incidentes de demencia y pusieron a disposición un archivo de documentos escritos que datan del momento de la entrada de cada sujeto en la orden [23]".

[2]Palabra átona que se apoya fonéticamente en la palabra tónica que la precede o sigue, formando grupo acentual con ella.

permanecían invariables, su vocabulario disminuye y su lenguaje era más simple [10]".

El mismo autor [8] nos menciona cómo personas con diferencia entre raza, clase social o capacidades intelectuales son víctimas de la enfermedad de AD, hay datos que indican que entre mayor sea la reserva cognitiva de una persona, existe menor probabilidad de sufrir una enfermedad de deterioro cognitivo, sin embargo esto no nos exenta de padecerlas, pues del mismo modo se ha comprobado que el deterioro cognitivo se da de una forma más lenta hasta alcanzar el punto donde es evidente la presencia de éste:

> ..."cuanto mayor es el nivel educativo, menor es la probabilidad de sufrir demencia en edades avanzadas [8]".

En [22] en su análisis sobre las obras de IM encontraron patrones de deterioro cognitivo mediante medidas computarizadas de complejidad sintáctica. Sus resultados de la mediciones de longitud media de las oraciones y el número de cláusulas por oración son similares a los obtenidos en [5], [10] . También hacen evidente la necesidad de normas basadas en la edad y la educación. Además encontraron diferencias en la complejidad entre los primeros escritos con los de su última etapa

> ..."la disminución en la complejidad gramatical observada entre 1994 y 1995 excede la tasa de cambio entre 1954 y 1978, o entre 1978 y 1994, lo que indica una aceleración que puede ser atribuible a los efectos del AD más que de envejecimiento "normal" [22]".

En el artículo [10], los resultados sugieren un enriquecimiento del vocabulario disponible entre las etapas temprana y media de la carrera de la escritora IM, seguido de un relativo empobrecimiento durante la composición del trabajo final que se sabe ya contaba con una etapa avanzada de AD. La evidencia de la disponibilidad de un vocabulario más restringido durante la redacción del trabajo final fue proporcionada por: (i) El menor número de tipos de palabras únicos en relación con el recuento total de palabras, y (ii) La menor tasa de aumento de esta proporción en muestras incrementales sucesivas, en "Jackson's Dilemma" (que fue el último libro en escribir antes del diagnóstico de AD avanzado post mortem) en comparación con los dos trabajos anteriores ("Under the Net" y "The Sea, The Sea"). Ambas observaciones implican una mayor tasa de repetición de palabras ya utilizadas en el libro final y una mayor tasa de introducción de nuevas palabras en los dos trabajos anteriores. También hay que tener en cuenta como sugiere el artículo que la cantidad de verbos sobre sustantivo es mayor con un deterioro cognitivo de AD.

En [14], citando a diversos autores resaltan que:

> ..."La patología del AD probablemente comienza muchos años y tal vez décadas antes de la aparición de los síntomas." Personas que envejecen de forma saludable no están exentas de padecer enfermedades neurodegenerativas...
>
> "en un envejecimiento saludable, el vocabulario aumenta durante la edad adulta media, pero luego puede comenzar a disminuir"...
>
> "La complejidad sintáctica del lenguaje, definida por medidas como cláusulas por enunciado, disminuye con la edad tanto en el lenguaje hablado como escrito"...
>
> "En la demencia, el vocabulario disminuye mucho más rápidamente, especialmente el uso de palabras de baja frecuencia y más específicas".

Además nos sugieren también que el uso de verbos en lugar de sustantivos es frecuente en una enfermedad neurodegenerativa del mismo modo que [10]. Por otro lado nos da una recopilación de conclusiones a las que llegaron diferentes autores sobre el comportamiento del lenguaje en enfermedades de deterioro cognitivo...

> ..."las repeticiones léxicas aumentan"....
>
> ..."las ideas de expresiones anteriores a menudo las repiten con las ..."mismas palabras"...
>
> ...y en su trabajo encontraron que ..."las palabras de relleno (por ejemplo: um, ah) y las disfluencias aumentan"...

En [21] estudian si la MDD está asociada con niveles de deterioro en pacientes de una clínica, encontraron que una MDD corta indica menor carga cognitiva y sintaxis más simple, por otro lado una MDD más grande sugiere mayor complejidad en la oración, por lo que la MDD se podría utilizar como marcador clínico para detectar deterioro cognitivo.

En [16] hicieron una detección generalizable del deterioro del lenguaje en AD a partir del habla espontánea, utilizando aprendizaje automático multilingüe, evaluaron la semántica, sintáctica, elementos paralingüísticos[3] en los idiomas inglés y francés. Personas con AD tienen promedios más bajos que el grupo control en las características semánticas, así mismo en el uso de sustantivos y preposiciones. La reducción del uso de preposiciones se encontró que es independiente del idioma.

Un estudio longitudinal en idioma alemán [4], muestra que el uso de pronombres comienza a cambiar una década antes del diagnóstico de AD. Hay un aumento en el uso de pronombre "D"[4] para referirse a personas cercanas.

[3] componentes no verbales que acompañan a las palabras y transmiten información adicional, como el tono de voz, el ritmo, el volumen, las pausas y otros aspectos sonoros

[4] pronombres personales en caso dativo: mir (a mí), dir (a ti), ihm (a él/ello, masculino), ihr (a ella), ihm (a ello, neutro), uns (a nosotros/nosotras), euch (a vosotros/vosotras), ihnen (a ellos/ellas, masculino/neutro, femenino, plural), y Ihnen (a usted/ustedes).

Disminución en el uso de pronombre impersonal "man"[5], así como en el pronombre posicional "das"[6].

En [11] investigaron grabaciones de 109 pacientes de enfermedades neurodegenerativas, incluyendo AD, demencia frontotemporal y el deterioro cognitivo vascular, encontraron que el uso de vocabulario y sintaxis más sencillos (palabras más cortas y menos frases proposicionales) está correlacionado con el deterioro cognitivo, sospechan de los resultados que la reducción de sustantivos va acompañado del aumento de pronombres, los pacientes con mayor deterioro cognitivo tienden a usar sustantivos comunes y de alta frecuencia. Además agregan que las evaluaciones del habla son automatizadas, rápidas y pueden administrarse de manera remota y con alta frecuencia, lo que las convierte en herramientas valiosas para el monitoreo del deterioro cognitivo a nivel individual.

En [5] hallaron diferencias lingüísticas entre un grupo de AD-temprano y un grupo control (NC) en la diversidad léxica, complejidad sintáctica y disfluencia del lenguaje, así como una longitud media de la oración más corta, mayor proporción de pausas largas, así como usar palabras de mayor frecuencia y sustituir sustantivos concretos[7] por pronombres ("eso" o "esto") para mantener un habla fluida, los clasificadores lograron una precisión de hasta 88% en la distinción entre AD-temprano y NC. El modelo de máquina de soporte vectorial (SVM) 93% al combinar características lingüísticas con biomarcadores. La relación estandarizada de captación (SUVR)[8], el volumen del hipocampo, la longitud media de oración y la ratio de pausas largas fueron identificados como características cruciales, en el orden en que se mencionaron. El estudio se hizo con 80 participantes del hospital Cardinal Tien en Teipéi, Taiwán, con 48 personas con AD-temprano, incluyendo deterioro cognitivo leve amnésico y AD-leve y el resto fueron NC. Confirmaron los diagnósticos con biomarcadores y los participantes fueron monitoreados durante 2 años.

III. Hipótesis

- El declive cognitivo tiene un impacto en la riqueza lingüística.
- La riqueza lingüística de una persona puede modelarse a través de sus escritos.
- La pérdida de la riqueza lingüística se puede ver reflejada en los cambios de los razgos lingüísticos.

[5]"uno" o "se" (en español impersonal). Se utiliza para referirse a una persona de manera general, sin especificar quién aún en personas cercanas (familia), o para expresar una acción que se realiza de forma impersonal.

[6]puede funcionar como artículo definido neutro ("el/la/lo" para sustantivos neutros) o como pronombre demostrativo neutro ("eso").

[7]palabras que nombran objetos, personas, lugares o animales que se pueden percibir con los cinco sentidos.

[8]relación que se utiliza en imágenes de tomografía por emisión de positrones para comparar la captación de un radiofármaco en una región de interés con la captación en una región de referencia, permitiendo cuantificar y comparar la actividad metabólica en diferentes áreas del cuerpo.

IV. Metodología

Nuestro objeto de estudio son dos destacadas escritoras, una con novelas predominantemente filosóficas, IM, y otra se movía en el género de las novelas policiacas, PDJ, se consiguieron la mayor cantidad de obras posibles, omitiendo algunas debido a que se salian de su propio género, como autobiografías o reflexiones filosóficas, en total alrededor de 20 novelas cada una.

Jean Iris Murdoch (IM), nacida el 15 de julio de 1919 en Irlanda, a temprana edad se mudó a Inglaterra, estudió en Oxford e hizo un posgrado en Cambridge, fue poeta, escritora, filósofa y profesora, influenciada por filósofos como Platón, Freud, Simone Weil y Sartre, y por los novelistas ingleses y rusos del siglo XIX. A la edad de 35 años comenzó su vida de escritora pública con su novela Under the Net, en 1987 fue nombrada Dama Comandante de la Orden del Imperio britanico, fue premiada por varias obras incluidas The Sea, the Sea, su última novela fue Jackson's Dilemma en 1995, año donde se le detectó la enfermedad de AD y muere 4 años después en 1999 a los 79 años de edad, al morir su esposo dona a la ciencia su cuerpo y en la autopsia se revela un AD avanzando, es reconocida por no utilizar editores ni permitir edición en sus textos lo que los convierte en trazas valiosas de la evolución de la enfermedad en el lenguaje escrito.

Phyllis Dorothy James (PDJ) nacida el 3 de agosto de 1920 en Inglaterra, estudió en Cambridge, trabajó como administradora de seguridad social y posteriormente como funcionaria pública del ministerio del interior, fue escritora y política, sus novelas eran de policías, algunas de las cuales tuvieron adaptación exitosas en el cine, comenzó su vida de escritora publica a los 43 años, fue miembro de la real sociedad de literatura, miembro de la sociedad de artes Royal, recibió premios por algunas de sus obras literarias, muere en el 2014 a la edad de 94 años de edad.

Para este estudio, en los textos realizamos 4 pasos generales los cuales podemos describir a grandes rasgos en:

1) Preprocesamiento, preparación de documentos.
2) Procesamiento, tokenizar el texto hasta lograr un token básico que usualmente son las palabras para poder etiquetarlas y lematizar como posible medida de reducción de dimensión.
3) Generación de medidas, e.g. estadísticas como conteo normalizado de ocurrencia de las palabras, conteo de oraciones, longitud de oraciones, tamaño de vocabulario, conteo normalizado de ocurrencia de las etiquetas. Incluso podemos formar grafos con los textos, si pensamos las oraciones como subgrafos, las palabras como nodos, podemos definir diferente tipos de aristas para capturar la información.
4) Análisis, hallar y explicar patrones presentes en un lenguaje con deterioro cognitivo.

A continuación daremos una explicación con más detalle de cada uno los pasos.

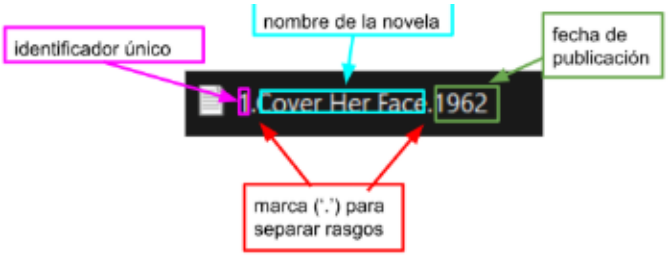

Fig. 3. Formato para el nombre de archivo de cada novela.

A. 1) Preprocesamiento

El preprocesamiento consiste básicamente en obtener las novelas digitalizadas en formato .txt, a cada texto se le debe eliminar prólogo, epílogo, índice y datos extras que pueda contener, dado que usualmente estas novelas se dividen en capítulos colocamos una marca manual (“####”) al final de cada capítulo a excepción del último, cada nombre de archivo de las novelas tiene un formato especial para poder trabajar de manera eficaz, el formato se puede apreciar en la Fig. 3, consta de 3 partes separadas por un par de signos de puntuación, 1) un identificador único el cual tiene la información del orden cronológico, 2) nombre de la novela y 3) fecha de publicación. Todos los archivos de un autor se encuentran en una misma carpeta y usando la librería os[9] podemos leer todos los archivos contenidos, filtrar los tengan formato .txt, y ordenarlos cronológicamente.

B. 2) Procesamiento

Se utilizó el lenguaje de programación python y bibliotecas del tipo open source, lo primero es leer el archivo para obtener el texto, limpiar el texto, es decir, eliminar caracteres no deseados validando únicamente los de interés, e.g. convertir todo a caracteres a formato ascii para evitar problemas de formato, expandir palabras en caso de que están contraídas, separar signos de puntuación de palabras, etc. Una vez que se tiene un texto limpio podemos recuperar cada capítulo con la marca manual utilizada y el método .split(“####”)[10], cada capitulo debemos tokenizarlo en oraciones, a su vez cada oración debe ser tokenizada en unidades, i.e. palabras. Una vez tokenizado por completo el texto conviene etiquetarlo y si se desea lematizarlo para reducir dimensión, en nuestro caso vamos a probar modificando lo menos posible el texto asi que dejaremos las contracciones de las palabras y probaremos sin lematizar primero. Existen diferentes herramientas que se pueden utilizar para lograr lo anterior descrito, para nuestro caso dado que una de las cosas que nos interesa es calcular las DD para obtener la MDD, utilizaremos spaCy [24], el cual es una herramienta robusta que nos facilitara las tareas anteriores, a grandes razgos podemos decir que spacy tiene 3 clases, doc, spam y token, la unidad básica es el token, despues siguen los spam es una secuencia de tokens, podemos verla como las oraciones del texto y por último el documento que contiene todas las oraciones o la clase doc. Dado que vamos a procesar varias novelas haremos una copia de los objetos de spacy para tratar de aligerar la informacion almacenada y no se sature la memoria ya que esperamos que cada capitulo de cada novela sea un doc el cual tiene como atributo las oraciones, cada oracion es una secuencia de tokens, y cada token tiene diferentes atributos. Nuestro bosquejo de clases quedaria de la siguiente forma: una **clase Token** que contiene los siguientes atributos “is_stop” identifica si es una palabra vacía (stopword), “is_punct” identifica si es un signo de puntuación, “is_space” identifica si es una espacio en blanco, “is_sent_start” identifica si es el token de incio de oración, “lemma_” el lema del token, “pos_” etiqueta parte del discurso (POS), “tag_” categoría gramatical (TAG), “dep_” tipo de dependencia, “text” texto del token, “head” gobernador de dependencia, “i” identificador único en el doc. Una **clase Oración** que tendrá 3 atributos que son “tokens” que contiene la secuencia de objetos tipo Token, “root” que contiene el verbo principal de la oración y “i” que contiene el identificador único en el documento. La **clase Capítulo** la cual contiene la lista de objetos tipo Oración. La **clase Libro** que contiene una lista de objetos tipo Capítulo. Por último será la **clase Colección de Autor** que contendrá 2+n atributos, donde “n” es el total de novelas, tiene un atributo “novels” que contiene el nombre de las novelas, otro “dates” que contiene la fecha de publicación, ambos ordenados cronologicamente, y por último los “n” restantes atributos serán los nombre de sus novelas, donde cada uno de estos atributos contendrá un objeto tipo Libro con su novela correspondiente. Separamos el diálogo de la narrativa, la oración era catalogada como diálogo si iniciaba con un signo de puntuación (i.e. comillas simples, dobles o guión), o si el verbo raíz de la oración se encontraba en una lista de verb dicendi (verbos del habla), el resto de oraciones era catalogado como narrativa.

Hasta aquí hemos cubierto los pasos 1) y 2), ahora vamos a obtener medidas cuantificables. Varios autores concuerdan en el hecho de que se simplifica el vocabulario y/o las oraciones [5], [8], [10], [11], [14], [21], [22]. En diferentes estudios sugieren que los pacientes de enfermedades neurodegenerativas sufren un cambios en sus categorías gramaticales, en [1] mencionan un sutil cambio en artículos, preposiciones y verbos, en [22] encontraron un mayor cambio en las categorías gramaticales en la última etapa de IM. En [10] encontraron menor introducción y mayor repetición en las palabras de su última obra en comparación a las anteriores, así como cambios en la proporción sujeto/verbo, este último hallazgo lo comparten [14]. En [16] encontraron cambios en los sustantivos y preposiciones. [4] muestran que el cambio en los pronombres ocurre décadas antes de que aparezca la sintomatología, por otro lado [11] menciona que el aumento en el uso de pronombres viene acompañado de una disminución en los sustantivos y [5] confirman cambios en los pronombres y sustantivos.

[9]Disponible al 11/21/2025 en la documentación os
[10]Disponible al 28/11/2025 en la documentación split

C. 3) Generación de medidas

Con los datos anteriores parece razonable comenzar las medidas obteniendo el conteo de ocurrencia de las palabras normalizado (TF), de esta manera podemos comprobar el tamaño del vocabulario, también podemos comprobar el tamaño de las palabras, obtener la DD de cada palabra y la MDD de la oración, un tamaño de palabras más corto y una MDD corta sin lugar a dudas expresa simplicidad en léxico y oraciones.
Otra medida a tener en consideración es el conteo y longitud de las oraciones, una disminución en la cantidad y tamaño de las oraciones, así como disminución de la MDD es un indicador de simplicidad léxica.
Algunos autores sugieren que en AD los últimos recuerdos perdidos son aquellas historias que comparten más y por tanto están más reforzadas, traspasando este pensamiento al lenguaje las últimas palabras en olvidar podrían ser las más utilizadas, en lingüística computacional las stopword son famosas por ser palabras que (depende el estudio) no aportan información, se puede decir que son palabras de uso frecuente las cuales hace una función de ser conectores entre las palabras pero sin embargo no tienen relevancia en el o los temas que puede contener la oración y/o documento. Sin embargo dado que estas palabras son de uso frecuente son excelentes candidatos para ser de las últimas palabras que pueda olvidar una persona que sufre AD, es por ello que también nos proponemos hacer un conteo de proporción de palabras tipo stopword en las oraciones, puede ser que en oraciones de la misma longitud exista mayor número de palabras tipo stopword en un lenguaje con AD.
Haremos también un conteo de palabras nuevas y repetidas en un determinado número de oraciones ya que algunos autores sugieren que en AD es menor el número de las palabras nuevas y mayor el de las repetidas.
Podemos también hacer un conteo normalizado de etiquetas POS (pos_), TAG(tag_), dependencia (dep_) y lema de las palabras (lemma_). Debido a que varios autores encontraron cambios en las categorías gramaticales, modelamos categoria por categoria POS y TAG para ver los patrones que se forman tratando de encontrar el patrón del deterioro.
Podríamos también generar un grafo por novela (o inclusive por un determinado número de oraciones, e.g. capítulos), donde el vocabulario son los nodos, tomando cada oración como un subgrafo podemos definir diferentes tipos de aristas, e.g. cada oración es un subgrafo completo, otra forma sería que cada subgrafo fuera dirigido de forma que la primer palabra apunte a la segunda, la segunda a la tercera y así sucesivamente hasta la penúltima que apunta a la última, otra forma es que cada subgrafo sea como el de la Fig. 1, salvo que cada arista sería mixta, i.e. además de tener su categoría de dependencia tendría el valor de la DD que se muestra debajo de las palabras en la Fig. 2.
Con todos los subgrafos formados por cada una de las oraciones podemos superponerlos en los nodos comunes para formar el grafo final, posteriormente podríamos sacar las medidas como orden, tamaño, diámetro, longitud de trayectoria promedio, grado, coeficiente de agrupamiento local y global. Generalizando la idea de los grafos y aprovechando las etiquetas podríamos generar hipergrafos, donde los nodos seguirían siendo el vocabulario y las hiperaristas serán cada una de las categorias (i.e. pos_, tag_, dep_) y obtendremos el orden, cardinalidad de cada hiperarista y los nodos comunes entre las hiperaristas, tratanto de encontrar patrones del deterioro en los cambios de las propiedades y medidas de los grafos.

D. 4) Análisis

Proponemos que cada autor sea su propio grupo control a lo largo del tiempo, posteriormente comparar resultados entre IM y PDJ, remarcando las diferencias que posiblemente se deban a la enfermedad de AD. Los conteos normalizados son distribuciones de probabilidad de ocurrencia en el texto. Dado que tenemos diferentes textos tenemos estas distribuciones en una ventana de tiempo dado, podemos comparar estas distribuciones para saber qué tanto fueron cambiando a lo largo del tiempo, existen varias medidas de similitud entre textos, una de las más utilizadas [2], [6] es la **distancia coseno**, ec. 1.

$$distancia\_coseno = 1 - similitud\_coseno \tag{1}$$

Se basa en la similitud coseno ec. 2 entre dos vectores, i.e. el coseno del ángulo entre estos vectores.

$$similitud\_coseno = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \, \|\vec{v}\|} \tag{2}$$

Dado que trabajamos con distribuciones de probabilidad y no existe algún estándar para obtener la similitud entre textos, utilizaremos otro método de similitud que es la **distancia Jensen Shannon (JS)** ec. 3 que se basa en la **divergencia Jensen Shannon (JSD)** ec. 4 la cual se basa en la **divergencia de Kullback (D_{KL})** ec. 5.
Dadas dos distribuciones de probabilidad (P, Q).

$$distancia\_JS(P,Q) = \sqrt{JSD(P||Q)} \tag{3}$$

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{4}$$

donde $M = \frac{P+Q}{2}$.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) * \log \frac{P(x)}{Q(x)} \tag{5}$$

Despues de comparar la similitud entre los textos podemos hacer un análisis de componentes principales (PCA) para redimensionar nuestros vectores de alta dimension y reducirlos a dos o tres dimensiones y poder visualizar un agrupamiento en los datos. Tambien podríamos aplicar un agrupamiento de aprendizaje automático (ML) como K-medias o DBSCAN. Para estos procesos de agrupamiento se utilizarán bibliotecas ya diseñadas en python de scikit-learn [25].

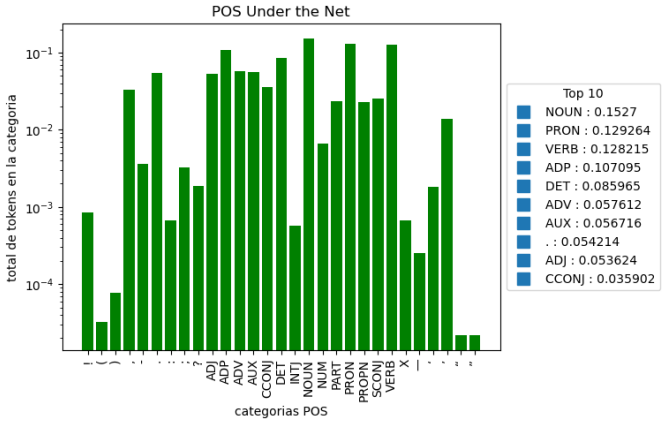

Fig. 4. Distribución de etiquetas POS en la narrativa de IM.

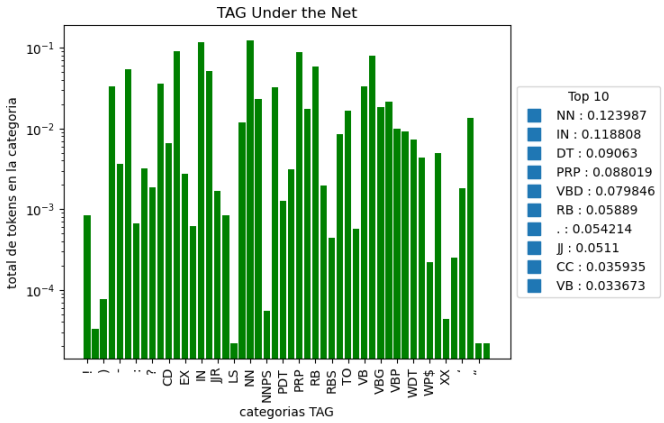

Fig. 5. Distribución de etiquetas TAG en la narrativa de IM.

V. Resultados preeliminares

En la Fig. 4 podemos ver la distribución de las categorías POS en la primer novela de IM, Under the Net.

En la la Fig. 5 podemos ver la distribución de las categorías TAG en la primer novela de IM, Under the Net.

En la Figs. 8 y 9 se puede apreciar la matriz de similitud Jensen-Shannon de IM y PDJ, respectivamente.

En la Figs. 10 y 11 se puede apreciar la matriz coseno de similitud de IM y PDJ, respectivamente.

En la Figs. 12 se puede apreciar el análisis de las 3 primeras componentes principales (PCA) en las novelas de IM y PDJ.

VI. Conclusiones

En las Figs. 4 y 5 podemos apreciar la distribución de ocurrencia de las categorías POS y TAG en la novela Under de Net de IM. En el eje x podemos apreciar las categorías y en el eje y el conteo normalizado, debido a que las categorías con mayor ocurrencia eran mucho mayores a las demás, el resto no se apreciaba correctamente por lo que se cambió el eje y, se grafica en escala logarítmica para una mejor apreciación de todas las categorías. De esta misma manera podemos encontrar la distribución para cada una de las novelas de IM y PDJ, una vez que contamos con las distribuciones para cada una de las novelas, podemos fijarnos solo en algunas de las categorías, i.e. sujeto, adposiciones y determinantes (de las categorías POS, Fig. 6), resulta que estas categorías disminuyeron a los largo del tiempo en IM y en PDJ se mantuvieron estables a lo largo del tiempo, por otro lado las categorías adjetivos y adverbios se comportan de manera errática en IM y estable en PD, esto cambios en las categorías podría ser atribuido a la enfermedad de AD. Por otro lado, si nos fijamos en la mayoría de las categorías (TAG) de sujeto y verbo, Fig. 7 podemos observar que en el caso de IM se comportan de forma errática y para PDJ se comportan de manera ordenada, lo que podría atribuirse a la desarrollo de la enfermedad de AD.

Tanto para las matrices de similitud como para el PCA, los resultados mostrados corresponden a las distribuciones de categorías TAG. En las Figs. 8 y 9 podemos apreciar la matriz Jensen-Shannon de IM y PDJ, respectivamente, las cuales muestran qué tanto se parecen las distribuciones de las categorías en las diferentes obras de cada autor. Para ambas se puede apreciar que se divide en 2 etapas, sin embargo para IM el cambio ocurre a los 47 años y para PDJ a los 74 años, este adelanto en el tiempo puede atribuirse a a enfermedad de AD. Por otro lado si nos fijamos en la matriz de similitud coseno en IM Fig. 10 podemos ver estas 2 etapas y además se puede apreciar que a partir de los 61 años de edad, las novelas cada vez se parecen más a la última que se sabe escribió con AD, pudiendo indicar que el AD fue evolucionando con mayor rapidez en esta etapa. En la Fig.. 11 que corresponde a la matriz de similitud coseno para PDJ podemos ver que la similitud entre sus textos permanece constante a excepción de la novela escrita a los 79 años, "Time to Be in Earnest", al identificarla podemos ver que no se trata de una novela policiaca sino de una autobiografía y es por ello que tiene menor similitud con las demás, por lo que hay que verificar todas las novelas tanto IM como PDJ y corroborar que se trate de novelas del mismo tipo, de manera que los cambios en las categorías sean atribuidos a la enfermedad de AD y no al cambio en el tema en las novelas. Una vez verificada las novelas el siguiente paso puede ser explorar el cambio en las categorías a lo largo de una novela, de ser cierto que las categorías en el deterioro se comportan de forma errática tendríamos que ver cómo en las primeras novelas de IM sus categorías se comportan de forma estable y conforme avanza la enfermedad estas categorías se vuelven cada vez más erráticas hasta el punto de ser evidentes.

Con el fin de reducir la dimensión y agrupar los datos de tal manera que podamos ver las novelas como puntos en el espacio, aplicamos PCA y nos quedamos con la información de las 3 primeras componentes las cuales conservan la mayor parte de información de las distribuciones. En las Figs. 12 y **??** podemos observar los PCAs de IM y PDJ, respectivamente. Una de las diferencias más notorias es que para PDJ las novelas están concentradas en un lugar en el espacio, siendo

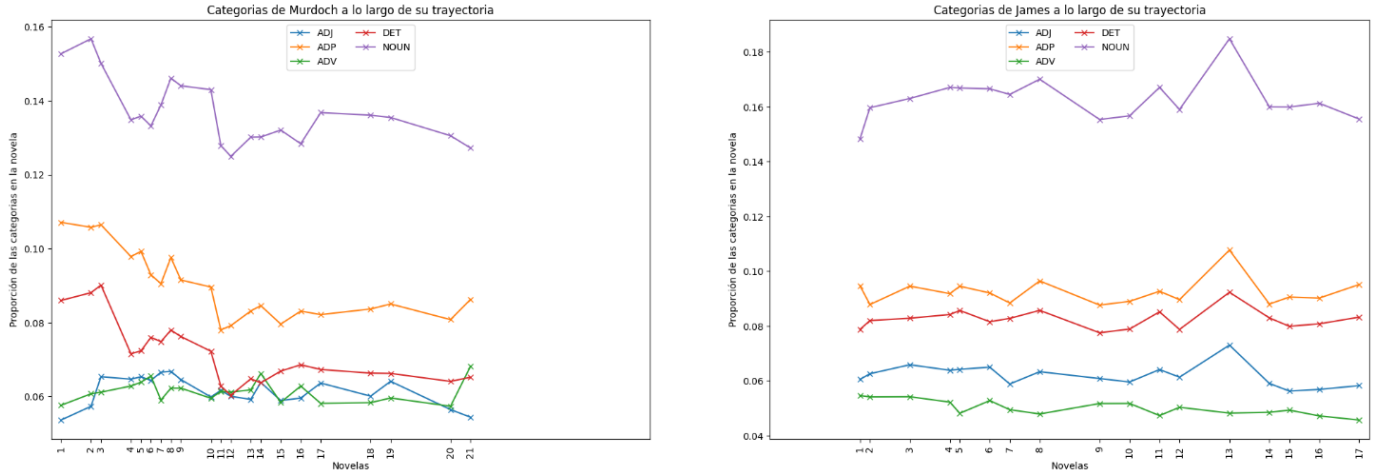

Fig. 6. Distribución de etiquetas POS (adverbios, sujetos, adposiciones y determinantes) a lo largo de las novelas, en la tabla I se desglosa el nombre de la novela, edad del autor y fecha de publicación de la novela para cada identificador.

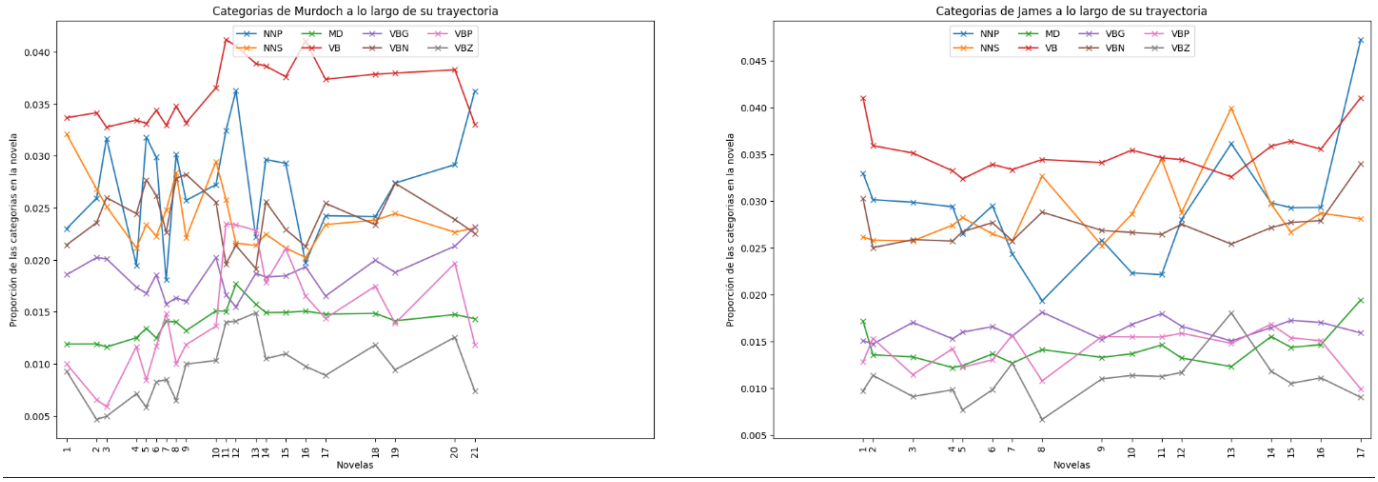

Fig. 7. Distribución de categorias TAG (sujeto y verbo) a lo largo de las novelas, en la tabla II se desglosa el nombre de la novela, edad del autor y fecha de publicación de la novela para cada identificador..

congruente con que sus categorías se mantienen en un rango estable durante su producción literaria en el tiempo, por otro lado el PCA de IM muestra mayor dispersión. Con lo visto anterior podemos confirmar y dar evidencia del cambio que sufren las categorías gramaticales cuando se sufre una enfermedad de AD, estos cambios podrían convertirse en marcadores digitales los cuales puedan detectar los cambios desde las primeras etapas de la enfermedad de tal manera que el contacto con el médico se haga de forma oportuna para poder mejorar la calidad de vida de las personas que tengan este tipo de padecimientos.

REFERENCES

[1] American Psychiatric Association, "Guía de Consulta de los criterios diagnósticos del DSM-5™". Asociación Americana de Psiquiatría, Burg Translations, Inc., Chicago (EEUU), Arlington, VA., p. 323, nota. Trastornos neurodegenerativos, Tabla 1, dominio: Lenguaje (lenguaje expresivo, 2013.

[2] J. Atkinson Abutridy, "Analítica textual, introducción a la ciencia y aplicación del análisis de información no estructurada". Alfaomega–Marcombo, 2023.

[3] C. Brown, T. Snodgrass, S.J. Kemper, R. Herman, M.A. Covington, "Automatic measurement of propositional idea density from part-of-speech tagging". Springer, vol. 40, núm. 2, pp. 540-545, 2008.

[4] D. Bittner, C. Frankenberg, J. Schröder, "Changes in Pronoun Use a Decade before Clinical Diagnosis of Alzheimer's Dementia—Linguistic Contexts Suggest Problems in Perspective-Taking", Brain Sciences, vol. 12, p. 121, doi. 103390/brainsci12010121, 2022.

[5] C.J. Chou, C.T. Chang, Y.N. Chang, C.Y Lee, Y.F. Chuang, Y.L. Chiu, W.L. Liang, Y.M. Fan, Y.C. Liu, "Screening for early Alzheimer's disease: enhancing diagnosis with linguistic features and biomarkers". Frontiers in Aging Neuroscience, vol. 16. doi. 10.3390/brainsci12010121, 2024.

[6] S. Dipanjan, "Text analytics with python: a practical real-world approach to gaining actionable insights from your data". Apress, doi= 10.1007/978-1-4842-2388-8, 2016.

[7] A.J. Celis-De la Rosa, C.E. Cabrera-Pivaral, M.G.L. Báez-Báez, A. Celis-Orozco, G. Gabriel-Ortiz, M.A. Zavala-González, "Mortalidad por enfermedad de Alzheimer en México de 1980 a 2014", Gaceta de México, vol. 154, pp. 550, 2018, doi 10.24875/GMM.18003361.

[8] M. Soláns García, "Navegando en la oscuridad: Iris Murdoch y la enfer-

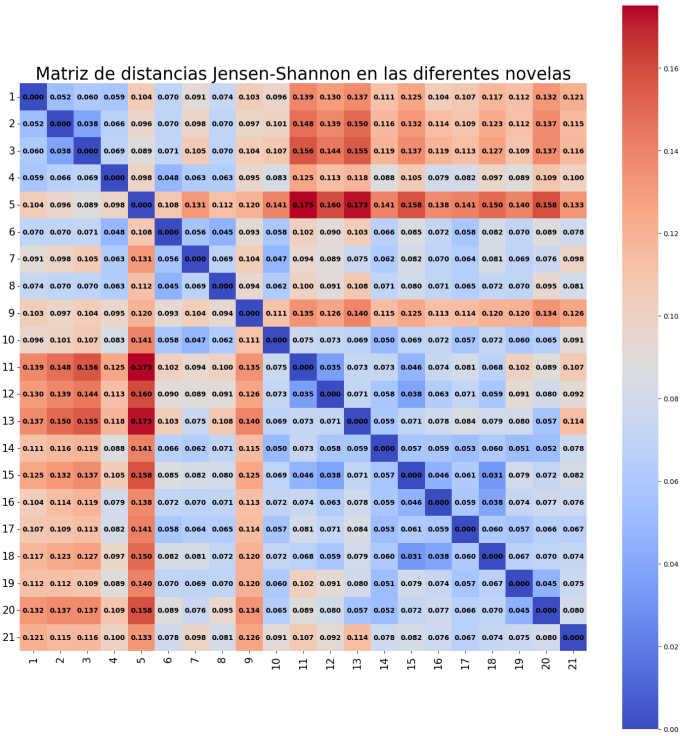

Fig. 8. Matriz de similitud Jensen-Shannon para los novelas de IM.

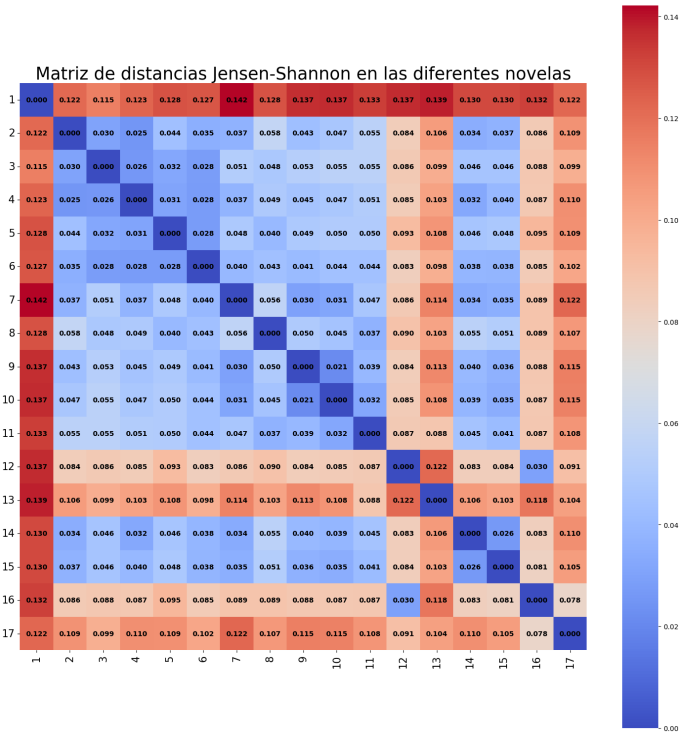

Fig. 9. Matriz de similitud Jensen-Shannon para los novelas de PDJ.

medad de Alzheimer", Universidad Nacional de Educación a Distancia, Revista Signa, núm. 23, pp. 203–230, 2014.

[9] P. García, N. Jimeno, "Evaluación de la escritura en pacientes con enfermedad de Alzheimer con deterioro cognitivo ligero o moderado". Universidad de Valladolid, Facultad de Medicina, Trabajo de fin de grado, 2016.

[10] P. Garrard, L.M. Maloney, J. R. Hodges, K. Patterson, "The effects of very early Alzheimer's disease on the characteristics of writing by a

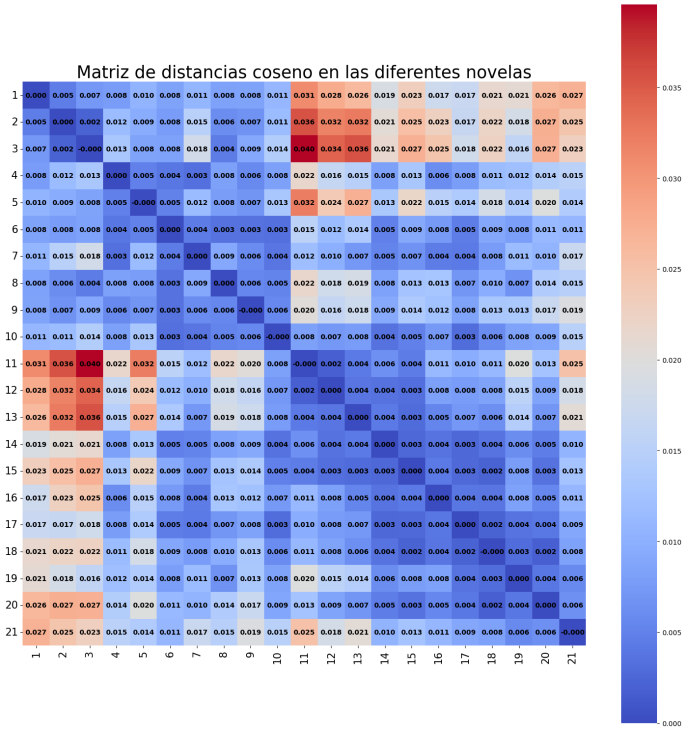

Fig. 10. Matriz de similitud coseno para los novelas de IM, señaladas con su edad y fecha de publicación.

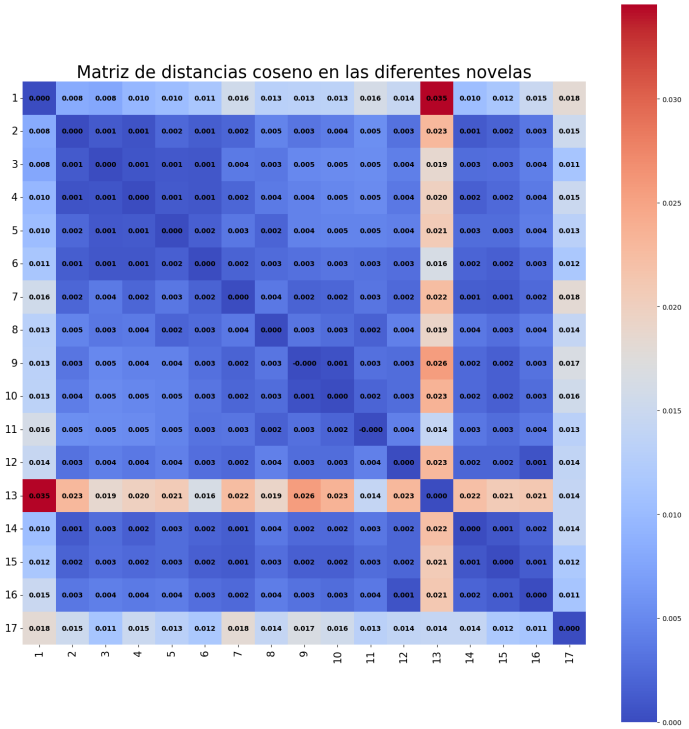

Fig. 11. Matriz de similitud coseno para los novelas de PDJ, señaladas con su edad y fecha de publicación.

renowned author", Oxford university press, Brain, vol. 128, núm. 2, pp. 250-260, 2005.

[11] M. Gumus, M. Koo, C.M. Studzinski, A. Bhan, J. Robin, S.E. Black, "Linguistic changes in neurodegenerative diseases relate to clinical symptoms". Frontiers in Neurology, vol. 15, doi.10.3389/fneur.2024.1373341, 2024.

[12] Instituto Mexicano del Seguro Social (IMSS), "Incrementan enfermedades neurodegenerativas; hay que detectarlas a tiempo, recomienda

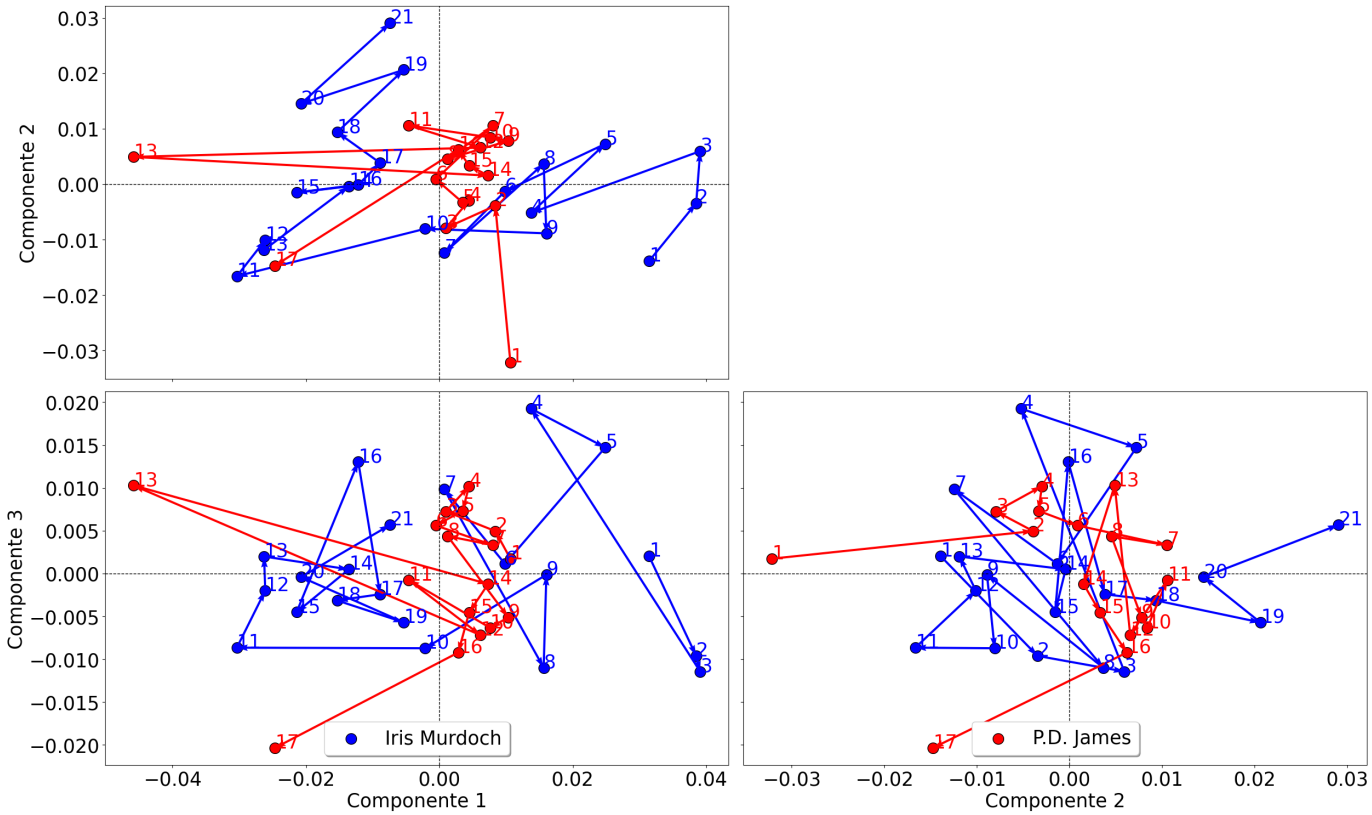

Fig. 12. PCA en 3 dimensiones

el IMSS', comunicación social, 19/06/2017.

[13] D. Klein, C. Manning, "Parsing and hypergraphs", Springer, new developments in parsing technology, pp. 351-372, 2005.

[14] X. Le, I. Lancashire, G. Hirst, R. Jokel, "Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists", the european association for digital humanities, Literary and linguistic computing, vol. 26, núm. 4, pp. 435-461, 2011.

[15] H. Liu, "Dependency distance as a metric of language comprehension difficulty", Instituto de Comunicación Lingüística Aplicada, Universidad de China, Beijing, Revista de ciencias cognitivas, núm. 9, pp. 159-191, 2008.

[16] H. Lindsay, J. Tröger, 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, doi. 10.3389/fnagi.2020.642033, 2021.

[17] D. McClosky, E. Charniak, M. Johnson, "Effective self-training for parsing". actas de conferencia sobre tecnología del lenguaje humano de NAACL, pp.152-159, 2006.

[18] S. Bird, E. Klein, E. Loper, "Natural language processing with Python: analyzing text with the natural language toolkit", O'Reilly Media, Inc., 2009.

[19] Organización Mundial de la Salud (OMS), "Demencia", noticias ONU marzo de 2025.

[20] Instituto nacional de geriatría, "Guía de instrumentos de evaluación de la capacidad funcional". Secretaria de Salud, 2022.

[21] F. Sand Aronsson, M. Kuhlmann, V. Jelic, P. Östberg, "Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis". Aphasiology, vol. 35(7), pp. 900-913, 2020.

[22] S. Pakhomov, D. Chacon, M. Wicklund, J. Gundel, "Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing", Springer. Behavior research methods, vol. 43, pp. 136-144, 2011.

[23] D.A. Snowdon, C.L. Tully, C.D. Smith, K.P. Riley, W.R. Markesbery, "Serum folate and the severity of atrophy of the neocortex in Alzheimer disease: findings from the Nun study", Oxford university press, the American journal of clinical nutrition, vol. 71, núm. 4, pp. 993-998, 2000.

[24] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python". doi: 10.5281zenodo.1212303, 2020.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine Learning in Python". Journal of machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[26] V.H. Yngve, "A model and an hypothesis for language structure", JSTOR, Proceedings of the American philosophical society, vol. 104, núm. 5, pp. 44-466, 1960.

Id	Nombre	Publicación	Edad
1	Under the Net	1954	35
2	The Sandcastle	1957	38
3	The Bell	1958	39
4	A Severed Head	1961	42
5	An Unofficial Rose	1962	43
6	The Unicorn	1963	44
7	The Italian Girl	1964	45
8	The Red and the Green	1965	46
9	The Time of the Angels	1966	47
10	Bruno's Dream	1969	50
11	A Fairly Honorable Defeat	1970	51
12	An Accidental Man	1971	52
13	The Black Prince	1973	54
14	The Sacred and Profane Love Machine	1974	55
15	Henry and Cato	1976	57
16	The Sea, the Sea	1978	59
17	Nuns And Soldiers	1980	61
18	The Good Apprentice	1985	66
19	The Book and the Brotherhood	1987	68
20	The Green Knight	1993	74
21	Jackson's Dilemma	1995	76

TABLE I
NOVELAS DE IM

Id	Nombre	Publicación	Edad
1	Cover Her Face	1962	42
2	A Mind to Murder	1963	43
3	Unnatural Causes	1967	47
4	Shroud for a Nightingale	1971	51
5	An Unsuitable Job for a Woman	1972	52
6	The Black Tower	1975	55
7	Death of an Expert Witness	1977	57
8	Innocent Blood	1980	60
9	A Taste for Death	1986	66
10	Devices and Desires	1989	69
11	The Children of Men	1992	72
12	Original Sin	1994	74
13	Time to Be in Earnest	1999	79
14	The Murder Room	2003	83
15	The Lighthouse	2005	85
16	The Private Patient	2008	88
17	Death Comes to Pemberley	2012	92

TABLE II
NOVELAS DE PDJ