UNIVERSITAT DE BARCELONA

Grau de Lingüística

Treball de Fi de Grau

Curs 2019-2020

REFLEJOS DE MUJER:
EL ETALECTO EN RELATOS AUTOBIOGRÁFICOS ESCRITOS POR MUJERES Y SU APLICACIÓN A LA LINGÜÍSTICA FORENSE

Ariadna Grau Magallón

TUTORES:

Montserrat Nofre i Sheila Queralt

Barcelona, 12 de juny de 2020

RESUMEN

El presente trabajo es un estudio del uso del lenguaje desde la perspectiva de la lingüística forense. A partir de un corpus ya existente, basado en relatos autobiográficos de mujeres, se examinan características lingüísticas que puedan ser discriminantes para la elaboración de perfiles lingüísticos. En concreto, se realiza una búsqueda de rasgos lingüísticos que puedan resultar útiles para diferenciar entre diferentes rangos de edad. Además, este análisis se lleva a cabo mediante herramientas que ponen a nuestra disposición las Tecnologías de la Información y la Comunicación.

Palabras clave: lingüística forense, corpus, mujer, perfiles lingüísticos, etalecto.

ABSTRACT

The present investigation is a study of the use of language from the perspective of forensic linguistics. Based on an existing corpus, constituted by autobiographical women stories, linguistic features are examined in order to be useful to create linguistic profiles. Specifically, we carry out a search of linguistic features to discriminate between different age ranges. In addition, this analysis is carried out using tools that Information and Communication Technologies offer to us.

Keywords: forensic linguistics, corpus, women, linguistic profile, agelect.

AGRADECIMIENTOS

Para empezar, quiero agradecer a mi tutora, Montse Nofre, su energía, su entusiasmo y su generosidad, sin los que este trabajo no hubiese llegado tan lejos.

También quiero dar las gracias a la gran lingüista forense, Sheila Queralt, por su tiempo, sus opiniones y su ayuda como cotutora.

No quiero olvidarme de reconocer la atención de mi amiga Aina Fernández, por su ayuda con los cálculos estadísticos de este trabajo.

Por último, doy las gracias a mi familia por su apoyo incondicional, y también a todas aquellas personas que han aportado su granito de arena en este trabajo.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Figura 16: Gráfico de dispersión de las variables seleccionadas (*bichunks*) entre los distintos grup.

1. INTRODUCCIÓN

La presente investigación se enmarca dentro del ámbito de la lingüística forense, la lingüística de corpus y la lingüística computacional. El objetivo general es estudiar los rasgos lingüísticos de grupo que pueden resultar útiles para la creación de un perfil lingüístico.

La construcción de un perfil lingüístico se basa en el análisis de los rasgos lingüísticos de un texto que permiten caracterizar distintas variables sociolingüísticas, como pueden ser el sexo, la edad o la procedencia del autor. Dichos perfiles pueden ayudar a acotar los posibles sospechosos en una investigación policial cuando hay un escrito como centro del caso, por ejemplo.

En este trabajo final de grado, nos hemos centrado en identificar las características lingüísticas discriminantes para rangos de edad en mujeres y comprobar su utilidad en la elaboración de perfiles lingüísticos.

En concreto, analizamos relatos autobiográficos escritos solamente por mujeres (extraídos del corpus Biodigithum[1], del cual hablaremos extensamente en el apartado cuatro), para identificar rasgos que permitan discriminar entre edades y, así, clasificar los textos según las franjas de edad.

Hemos realizado un estudio con perspectiva de género y, por ello, solo se analizan las voces de un centenar de mujeres que nos han transmitido a través de sus relatos todos aquellos sentimientos e ideas que configuran su personalidad. Concretamente, se estudian textos autobiográficos de mujeres que explican situaciones como el maltrato, la infidelidad, el abuso, la adicción, la anorexia, la depresión, la misoginia, la violencia o situaciones traumáticas, así como también temas como el afán de superación, los sueños, el amor, etc.

El análisis de los distintos textos pretende determinar la distribución de las variables lingüísticas (de diferentes tipos: morfológicas, léxicas, sintácticas, etc.) según la edad y establecer cuáles poseen un potencial discriminante mayor para determinar el rango de edad.

[1] http://stel3.ub.edu/biodigithum/queesbiodigithum.php

El proyecto se estructura en seis capítulos, el primero de los cuales es esta introducción. El segundo capítulo está dedicado a explicar brevemente el marco teórico y el estado de la cuestión que se aborda; en él definimos, a grandes rasgos, las disciplinas en las que se enmarca el presente trabajo, como son la lingüística forense, la lingüística de corpus y la lingüística computacional. A continuación, exponemos los objetivos y propósitos de la presente investigación y de la hipótesis que planteamos como punto de partida en el capítulo tercero. En el capítulo cuarto se muestra la metodología utilizada, es decir, cuál es el corpus analizado, qué variables se han tenido en cuenta para el análisis y qué herramientas se han utilizado para la codificación morfosintáctica, la extracción de datos cuantitativos y las pruebas estadísticas. En el quinto capítulo, se presentan los resultados, agrupados en diversos subapartados, correspondientes a la tipología de las variables analizadas en cada momento. Para acabar, exponemos las conclusiones derivadas de nuestra investigación, con el objetivo de comprobar si la hipótesis inicial queda verificada. Además, explicamos las contribuciones de nuestro estudio, las limitaciones y mencionamos futuras líneas de trabajo posibles.

2. ESTADO DE LA CUESTIÓN

En este apartado se abordan todas aquellas cuestiones generales referentes a las disciplinas de la lingüística forense, la lingüística de corpus, la lingüística computacional y la sociolingüística, para contextualizar nuestro trabajo.

2.1. La lingüística forense

La lingüística forense es una disciplina que, tal y como define la Asociación Internacional de Lingüistas Forenses en su página web (IAFL, por sus siglas en inglés), "covers all the áreas where law and language intersect".

Crystal (1997) propone una definición más restrictiva, ya que para él la disciplina consiste en "the use of linguistics to investigate crimes in which language data form part of the evidence". Hay que tener en cuenta, sin embargo, que esta definición solo puede aplicarse en aquellos casos en los que se utiliza el análisis del lenguaje como prueba.

Olsson (2010, p.4) considera que la lingüística forense es "the apply of linguistic knowledge and techniques to the language implicated in (i) legal cases or proceedings or (ii) private disputes between parties which may at a later stage result in legal action of some kind being taken".

Coulthard y Johnson (2010, prefacio) definen la lingüística forense como "the study of language and the law, covering topics from legal language and courtroom discourse to plagiarism. It also concerns the applied (forensic) linguist who is involved in providing evidence, as an expert, for the defense and prosecution, in areas as diverse as blackmail, trademarks and warning labels".

Así pues, la lingüística forense es la disciplina relacionada con la aplicación de conocimientos lingüísticos, con la finalidad de evaluar y analizar las características lingüísticas de una muestra oral o escrita para ser aportada como prueba en un proceso judicial. Dicha tarea puede ser el análisis de un texto o de una prueba oral en un contexto forense. El propósito que se persigue con la realización del análisis es poder aportar datos relevantes en una investigación policial o judicial.

Es importante mencionar la intersección de la lingüística forense con muchas otras disciplinas, es decir, no es independiente, sino multidimensional e interdisciplinar. De hecho, la primera definición citada ya nos deja claro que se trata de una disciplina

polifacética. Por lo tanto, podemos afirmar que la lingüística forense mantiene una estrecha relación con otras disciplinas como la sociolingüística, la lingüística de corpus o la lingüística computacional. Pero también dentro del campo del derecho, la lingüística forense incorpora diversas disciplinas de los ámbitos jurídico y judicial.

Tal y como explican Garayzábal, Jiménez y Reigosa (2014, p. 31), la disciplina, en España, es un área emergente de investigación con escasa presencia de profesionales formados en el análisis de datos de carácter lingüístico, por el momento.

Además, mientras que en otros países se dispone de diferentes corpus para trabajar, en España no se dispone de tales bases de datos a causa del tipo de legislación nacional. Este hecho también supone una dificultad para el avance de la disciplina en este país.

En el libro *Fundamentos de la lingüística forense*, Garayzábal, Queralt y Reigosa(2019, p. 21), hacen referencia al estado de la disciplina en España y se afirma que los beneficiarios últimos de los estudios lingüísticos en el ámbito forense continúan, en la mayoría de los casos, ignorando cómo pueden hacer uso de estos análisis periciales en las distintas áreas de la lingüística forense. No obstante, cada vez hay más interés en obtener información acerca de las posibilidades que este tipo de pericias pueden aportar en una investigación.

2.1.1. El lenguaje evidencial

Según Gibbons y Turell (2008, p. 1) las áreas de la lingüística forense son las siguientes: el lenguaje jurídico, es decir, los textos legales y la legislación en general; el lenguaje judicial, que hace referencia al discurso legal de los actores judiciales y entrevistas policiales y, por último, el lenguaje probatorio o evidencial, basado en el uso de las pruebas lingüísticas. Concretamente, Gibbons y Turell explican que las áreas son "the written language of the law, particularly the language of legislation; spoken legal discourse, particularly the language of court proceedings and police questioning; the social justice issues that emerge from the written and spoken language of the law; the provision of linguistic evidence".

En el presente trabajo, nos centramos en el lenguaje probatorio o evidencial. Por ello, dedicaremos este apartado únicamente a dicha área.

Gibbons (1994, p. 320) explica que la evidencia se puede centrar en la producción; en este sentido, se tratará de determinar quién ha podido producir un determinado texto y, para ello, se utilizará una metodología comparativa y un análisis contrastivo, que permitirán descartar o confirmar la autoría de un texto.

Más específicamente, cuando se habla de lenguaje evidencial, se hace referencia a las características lingüísticas que pueden constituir una prueba en una investigación, es decir, a los datos lingüísticos que permitan identificar quién es el autor de un texto o elaborar un perfil lingüístico que acote la lista de sospechosos en un caso policial.

Cicres y Gavaldá (2014, p. 62) explican que "al seu torn, les finalitats evidencials consisteixen a aportar evidencia lingüística durant el procés (judicial). Aquestes evidències tenen a veure majoritàriament amb la comparació de textos orals o escrits amb l'objectiu d'assistir el jutge o tribunal en la determinació de l'existència de plagi, o bé en la determinació o atribució de l'autoria d'un text escrit o d'una gravació oral".

Tal y como se presenta en el libro *Fundamentos de la lingüística Forense* (2019, p. 23), McMenamin, dentro del área de lenguaje evidencial (lengua escrita) incluye: la determinación de autoría de textos dubitados, la detección de plagio, el perfil lingüístico, el análisis automático de textos y la ambigüedad textual.

Entre los propuestos por McMenamin, nuestra investigación profundizará en el tema de los perfiles lingüísticos, pues el objetivo del trabajo es poder identificar rasgos lingüísticos en la elaboración de perfiles lingüísticos a partir de textos autobiográficos escritos por mujeres y agrupados por franjas de edad.

2.1.2. Construcción de perfiles lingüísticos

Según Isabel Picornell (2014, p. 80), la elaboración del perfil de un autor "se basa en los rasgos lingüísticos del texto, de los que se infieren género, edad, nivel educativo, cultural y lingüístico. Este análisis puede resultar de ayuda para orientar el curso de una investigación cuando no existen sospechosos", es decir, cuando los primeros textos o producciones de habla son anónimos. Este perfil se construye con el fin de poder reducir la lista de posibles autores de dichos textos.

Los expertos de la Unidad de Análisis de Conducta (BAU; por su nombre en inglés, *Behavioral Analysis Unit)* del FBI (Fitzgerald, 2007, p.7) proponen una definición de

perfil lingüístico desde un punto de vista más pragmático: “Linguistic Profile, requires advanced training and experience in the fields of criminal behavior and forensic linguistics. This section incorporates assessments based on a behavioral and linguistic analysis of the communication, including indications of capabilities, commitment, deception, biographical information (e.g., sex, age, race/ethnicity, education, and native language), and, if necessary, the level of threat (e.g., low, moderate, or high)”.

Una vez reducida la lista de sospechosos, se puede realizar una comparación entre los textos dubitados (elemento de la prueba del cual se tiene dudas respecto a su autor) y los indubitados (conjunto de textos de los cuales se conoce el autor: son los textos obtenidos de los diferentes sospechosos) con el objetivo de realizar un análisis de las semejanzas y diferencias entre ellos.

Nini (2019, p.39) explica que “two ways of doing authorship profiling have emerged from forensic case work and research: (1) analysis of salient linguistic markers, and (2) analysis of writing style. The first type of profiling is the application of sociolinguistic knowledge on a case by case basis to extract ad hoc linguistic features that are markers of a certain demographic background. In contrast, the second type of profiling consists in the analysis of the stylistic variation exhibited by the text as a whole. This analysis often involves the study of the frequency with which certain features are used, like the study of register variation and takes as the unit of analysis the text itself”. En este trabajo, emplearemos ambas formas de análisis para elaborar los perfiles lingüísticos.

2.1.3. Sociolingüística

Tal y como ya se ha comentado anteriormente, la lingüística forense es interdisciplinar, es decir, trabaja junto con muchas otras disciplinas. Una de ellas es la sociolingüística.

La sociolingüística entra en juego en la elaboración de perfiles lingüísticos, ya que las variables que se pueden inferir de un texto mediante su lenguaje son de tipo sociolingüístico: sexo, edad, procedencia del autor, entre otras.

En el *libro Fundamentos de la Lingüística Forense* (2019, p. 37) se explican estas variables de la siguiente manera: “La variedad atiende a distintos rasgos que dan cuenta de la dinámica de la variación lingüística y constituyen lo que se conoce como factores extralingüísticos o sociolingüísticos: geográfico (dialecto), social (sociolecto), funcional (estilo), especializado (jerga), individual (idiolecto)”.

Por último, también se habla de que cada uno de los modos en que se presenta una variable se denomina variante lingüística. Un ejemplo de variantes lingüísticas de una misma palabra podría ser el siguiente: [maðríd], [maðríθ], [maðrít], [maðrí].

2.2. La lingüística computacional

La lingüística computacional es otra de las materias que también abarca el campo de la disciplina forense, ya que tiene como objetivo crear sistemas lingüísticos de inteligencia artificial que imiten las capacidades humanas y, así, facilitar nuestras tareas.

Cualquier trabajo de Procesamiento del Lenguaje Natural (PLN) inicia su estudio con un corpus. En nuestra investigación se usa el corpus Biodigithum (del cual hablaremos en profundidad en el apartado cuatro) para realizar un análisis de los diferentes textos que nos permita detectar rasgos relevantes para la elaboración de perfiles lingüísticos.

Mediante el uso de diferentes aplicaciones informáticas se puede realizar un análisis del corpus objeto de estudio y, de esta manera, obtener datos lingüísticos y estadísticos relevantes que ayuden a clasificar los textos del corpus.

Sousa-Silva (2018, p. 120) explica que "large volumes of data make it virtually imposible for linguists to manually process and analyse the data quickly and accurately. Therefore, they usually resort to the use of computational tools. Such analysis is can be heavily computational, i.e. it can be conducted with no or very little human intervention, or computer-assisted, in which computational tools and techniques are used as an aid to the manual analysis, e.g. in searching words or phrases, or comparing some textual elements against a reference corpus or tagging a text, among others. The use of computational linguistics in forensic contexts has become so indispensable".

2.2.1. Estilometría

Para definir la estilometría, primero hay que prestar atención a la definición de estilo. Según McMenamin (2010, p. 488), "style in written language reflects both a writer's conscious response to the requirements of genre and context as well as the result of his or her unconscious and habituated choices of the grammatical elements acquired through the long-term experiential process of writing. Style is in part, then, the sum of the recurrent choices the writer makes in the process of writing".

La estilometría es una disciplina usada por los expertos forenses con el fin de poder definir características discriminantes en un texto a partir del estilo. Esta disciplina se respalda en técnicas cuantitativas, tal y como apunta Chaski (2005, p. 2), "stylometry is quantitative and computational, focusing on readily computable and countable language features". Estas características cuantificables en un texto podrían ser la longitud de las palabras, la longitud de las oraciones o las frecuencias de las palabras, por ejemplo.

Además, mediante las técnicas estilométricas también se puede examinar la riqueza léxica o la frecuencia de aparición de ciertas palabras en un texto. Pavelec et al. (2009, p. 2445) coinciden con este punto de vista y afirman que "the literature shows that several stylometric features that have been applied include various measures of vocabulary richness and lexical repetition based on word frequency distributions".

En un artículo de Ainsworth y Juola (2019, p. 1161) se afirma que "when looking for clues, linguistic analysts examine systematic language variation on many levels. These systematic language usage patterns were called "style markers", and analysis based on style markers became "forensic stylistics".Los patrones de puntuación, los tipos de estructura, el lenguaje formal o informal pueden ser marcadores que reflejen un idiolecto.

2.3. La lingüística de corpus

Según la Asociación Española de Lingüística de Corpus (AELinCo), "por medio de este ámbito de la lingüística se procede al análisis de textos (orales o escritos) a la vez que se adopta como metodología de trabajo el uso de medios computacionales para presentar datos, de modo que estos, una vez sistematizados y ordenados, puedan servir como base empírica para alcanzar conclusiones pertinentes". El tratamiento y el análisis del corpus se realizan a partir de la forma escrita, incluyendo en este campo las transcripciones de los corpus orales.

Sousa-Silva (2018, p. 123) apoya el punto de vista anterior, al explicar que "applied linguists have since the 1980s relied on corpora for research and practice. In order to make assumptions about linguistic events and language use, linguists usually rely on large volumes of spoken and/or written linguistic data that have been produced as a

result of communication in context: a corpus. [...] A corpus needs to be available in electronic form so that it can be processed by a computer".

Insistiendo en el tratamiento automático del corpus, es importante mencionar que, a través de los diferentes métodos de análisis de corpus textuales, se pueden analizar todos los niveles de la lengua (ortográfico, léxico, sintáctico, etc.) utilizando distintas herramientas informáticas. Una vez realizado el análisis, se pueden examinar los datos obtenidos y, de esta manera, observar diferentes ejemplos de los fenómenos lingüísticos objeto de estudio y valorar su adecuación y representatividad en referencia al tema objeto de investigación.

Por medio de esta disciplina, pues, se puede llevar a cabo una comparación entre los distintos textos que forman un corpus, o entre distintos corpus, y evaluar los datos obtenidos. En esta evaluación, y mediante cálculos estadísticos, se pretende obtener la validación de nuestra hipótesis en referencia a los rasgos identificativos del autor o autores de los textos. Es decir, si un rasgo, por ejemplo, es muy frecuente entre los distintos textos o, por lo contrario, es muy poco común. Con ello, podremos formalizar algunas conclusiones en cuanto al fenómeno que analizamos.

En resumen, cabe decir que, en cuanto a la relación de esta disciplina con la lingüística forense, la lingüística de corpus es un enfoque metodológico que puede ser de gran utilidad "en la determinación de autoría y el plagio de textos", como apunta Renouf (1987, p. 1).

3. HIPÓTESIS Y OBJETIVOS

La pregunta de investigación que queremos responder con este estudio es la siguiente: ¿Existen rasgos lingüísticos que nos permitan acotar la franja de edad del autor (en este caso, autora) de un texto?

La hipótesis de partida es que es posible identificar rasgos discriminantes significativos desde un punto de vista estadístico en cuanto a la variación lingüística por franjas de edad (etalecto, traducción del término inglés *agelect* tomada de Silva-Corvalán (2001, p. 101)).

Para validar nuestra hipótesis, hemos realizado un análisis lingüístico de los textos, principalmente en dos niveles, el léxico y el morfosintáctico. Este análisis se realizará con diferentes herramientas lingüísticas, puesto que otro de los objetivos que perseguimos es la aplicación de las herramientas propias de la lingüística de corpus y la lingüística computacional y el aprovechamiento de su potencia en cuanto al procesamiento de datos y el cálculo estadístico.

El objetivo principal de este trabajo es determinar, a partir de un corpus ya existente, posibles rasgos lingüísticos distintivos según franjas de edad, que contribuyan a la construcción de perfiles lingüísticos de mujeres en relatos autobiográficos.

Un objetivo secundario del trabajo es reivindicar el corpus Biodigithum como un corpus relevante desde un punto de vista cualitativo y cuantitativo, que nos permite dar voz y hacer visibles a mujeres anónimas, contemporáneas de nuestras madres y abuelas, que decidieron escribir sobre sus vivencias y a las que no se debe seguir manteniendo en silencio.

4. METODOLOGÍA

En este punto tratamos sobre el origen del corpus, los distintos tipos de procesamiento al que lo hemos sometido, las variables que hemos considerado susceptibles de análisis y las herramientas utilizadas para las distintas fases del análisis, tanto la del estudio propiamente lingüístico como la de los cálculos estadísticos.

4.1. El corpus

El primer paso en nuestra investigación fue la elección de un corpus concreto sobre el cual realizamos el procesamiento, la extracción de datos cuantitativos y los cálculos estadísticos que requiriera la validación de la hipótesis inicial. se ha realizado todo el trabajo de análisis lingüístico posterior.

De entre las opciones que teníamos a nuestro alcance, por sus especiales características y contenido, convinimos que la construcción de un corpus a partir de la base de datos Biodigithum era, sin duda, la mejor elección.

La base de datos Biodigithum está depositada en la Unitat d'Estudis Biogràfics (UEB) de la Universitat de Barcelona, y está formada por textos de tema autobiográfico presentados a las distintas ediciones de un certamen literario (cinco en total, entre 1998 y 2002). Tal como explica la propia web de Biodigithum,"participaron en él un total de 9.729 autores (un 70 % de mujeres, 30 % de hombres)". En las primeras ediciones solo se permitía la participación de mujeres, y la propuesta de permitir la participación de hombres no se incluyó en las bases del concurso hasta sus dos últimas ediciones.

Cabe decir que el concurso formaba parte de una estrategia publicitaria para el lanzamiento de una nueva línea de perfumes, por lo que el certamen se bautizó con el nombre de dicho producto. Colaboraron en su convocatoria la edición española de la revista *Marie Claire*, dirigida a un público mayoritariamente femenino, y la UEB, como entidad experta en el estudio de relatos autobiográficos. En el anuncio del premio, insertado en el Boletín de la Unidad de Estudios Biográficos,se explica que "esta firma (Calvin Klein) y *Marie Claire* quieren conocer las contradicciones, promocionando un premio que rescate sus experiencias, sus sentimientos, el peso de la memoria".

Todos los textos debían estar escritos en lengua española y tener una extensión de entre tres y siete folios. El jurado que valoró los textos, ciertamente, estaba integrado por profesionales de prestigio.

La profesora Anna Caballé, directora en ese momento y todavía hoy de la UEB concedió en 2016 una entrevista al Diario de Cádiz (15 de octubre de 2016), en la cual explicaba la constitución del fondo Biodigithum, "que recopila unos diez mil textos autobiográficos de otras tantas mujeres" y que han servido de base para la realización de diferentes estudios desde la perspectiva de la psicología, la sociología, la antropología o la pragmática. Se centraron en los relatos de mujeres porque, en palabras de Anna Caballé, "la mayor parte de la producción de textos personales se han perdido, quemado, destruido… En España, entre las guerras, el miedo al qué dirán o que a la hora de casarse las mujeres destruían todo lo relacionado con su vida pasada, es muchísimo lo que se ha perdido. Es una literatura que apenas se ha valorado". Así pues, mediante este estudio, se pretende volver a alzar la voz de todas esas mujeres y conceder la importancia que merecen a sus reflexiones y a su uso del lenguaje.

Los textos seleccionados para el presente estudio tienen una temática diversa, con un nivel de carga emocional diferente en función del tema tratado (desde el maltrato al afán de superación). Hay que decir que la clasificación de los temas la hemos tomado directamente de la base de datos de Biodigithum.

En cuanto a la selección de textos, el criterio principal ha sido geográfico: hemos escogido aquellos cuya autora procede de una comunidad monolingüe para evitar, de esta manera, las posibles interferencias de otras lenguas en el uso del español.

Por la misma razón, hemos dejado fuera del estudio los textos que proceden una comunidad autónoma con un dialecto meridional del español (Andalucía).

Así pues, los textos recopilados han sido escritos en lo que hemos considerado una variedad del español mínimamente interferida por otras lenguas. En principio, solo habíamos previsto seleccionar textos de las siguientes comunidades autónomas: Aragón, Castilla-La Mancha, Castilla y León, La Rioja, Madrid y Murcia. No obstante, necesitábamos más textos para completar una muestra suficientemente representativa del total del corpus y, dado que solo teníamos a nuestra disposición aproximadamente una décima parte del corpus, decidimos ampliar el área geográfica, añadiendo en

algunos casos comunidades bilingües, a modo de excepción. Por consiguiente, se incluyeron textos de autores procedentes de Asturias, Extremadura, Navarra, Tenerife y Valencia, aunque la cantidad incorporada fue mucho menor que la del resto de zonas.

En la base de datos de Biodigithum, los textos aparecen clasificados en cuatro rangos de edad: menos de 20 años, entre 20 y 35 años, entre 35 y 50 años y más de 50 años. En nuestro caso, decidimos eliminar la primera categoría, es decir, la de menos de 20 años, porque es la menos productiva del conjunto. Moreno (1996, p. 264), por ejemplo, en su artículo, usa las mismas franjas de edad que el presente trabajo.

Se han seleccionado 30 textos para cada uno de los rangos de edad estudiados. La cantidad de textos a tratar se estableció según un criterio estadístico, puesto que es la muestra mínima suficiente para que los resultados obtenidos sean estadísticamente relevantes. Por lo tanto, nuestra investigación se basará en un total de 90 textos.

A continuación, se muestra un resumen de la distribución del corpus en nuestro trabajo.

RANGO EDAD	EXTENSIÓN (Nº PALABRAS)	PROVINCIAS	TEMÁTICAS
20-35	49.652	Badajoz, Burgos, Ciudad Real, Huesca, Madrid, Murcia, Oviedo, Palencia, Soria, Valladolid, Zamora y Zaragoza.	Drogas, reflexión, infidelidad, amistad, depresión, anorexia, maternidad, separación, malos tratos, infancia, enfermedad, suicidio, amor y trastornos
35-50	56.781	Badajoz, Burgos, Ciudad Real, Logroño, Madrid, Murcia, Teruel, Toledo, Valladolid y Zaragoza.	Maternidad, desengaño, amor, enfermedad, crisis, separación, muerte, depresión, viajes, infancia, infidelidad, abusos, anorexia, familia, drogas, guerra, malos tratos y adolescencia.
+50	59.262	Alicante, Asturias, Huesca, León, Madrid, Navarra, Soria, Tenerife, Toledo, Valladolid y Zaragoza.	Guerra, amor, pobreza, familia, desengaño, misoginia, infancia, represión, muerte, universidad, amistad, ama de casa, anorexia, crisis, depresión, abusos , asesinato, enfermedad, soledad, trastorno y franquismo.

Tabla 1: Resumen de la distribución del corpus de estudio.

4.2. Tratamiento del corpus

Los textos de Biodigithum están almacenados en un servidor del Servei de Tecnologia Lingüística (STeL) de la Universitat de Barcelona en formato de imagen .jpeg. Los 90 textos seleccionados se descargaron y se codificaron con el número de convocatoria (de la I a la V) seguido del número originalmente asignado por los organizadores del concurso. Mediante un programa OCR (*OpticalCharacterRecognition*), se procedió a convertir las imágenes en texto plano (formato .txt) mediante el programa Image Reader de Microsoft.

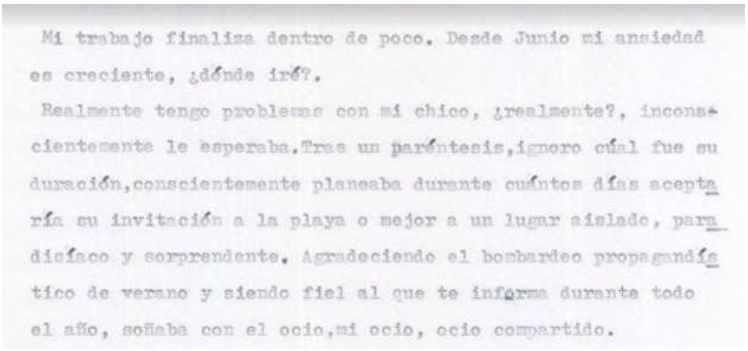
Mi trabajo finaliza dentro de poco. Desde Junio mi ansiedad
es creciente, ¿dónde iré?.
Realmente tengo problemas con mi chico, ¿realmente?, inconscientemente le esperaba.Tras un paréntesis,ignoro cúal fue su duración,conscientemente planeaba durante cuántos días aceptaría su invitación a la playa o mejor a un lugar aislado, paradisíaco y sorprendente. Agradeciendo el bombardeo propagandístico de verano y siendo fiel al que te informa durante todo el año, soñaba con el ocio,mi ocio, ocio compartido.

Figura 1: Texto original.

Mi trabajo finaliza dentro de poco. Deste Junio ni ansiedad
en creciente, ¿dónde 1767,

Realmente tengo problemas com mi chico, ¿realmente?, inconse
efentenente le esperaba, Ts un paréntеoio,icnoro cÁal fue su
auración,conscientenente planeaba durante cufntos afas acepta
Tía en invitación a la playa o mejor a un Mngar efelado, pera.
diofaco y sorprendente, Arradeciendo el bombardeo propacndfe
tico de verano y siendo fiel al que te inform durante todo
el año, soñabe com el ocio,mi octo, ocio comeartido.

Figura 2: Transcripción con OCR antes de su revisión.

Tras acabar el cambio de formato, se revisaron los textos uno a uno de forma manual, con el objetivo de garantizar que el texto con el que trabajaríamos fuera idéntico al original, incluyendo los errores ortográficos y de tipografía. Esta revisión es una tarea

que hay que realizar obligatoriamente, puesto que el programa OCR siempre ofrece un porcentaje de error, que aumenta en función del estado del papel, su conservación, los tipos utilizados, etc., como puede observarse a simple vista comparando las Figuras 1 y 2. Así, ciertas grafías pueden dar lugar a confusiones y errores en el reconocimiento, como es el caso de *í* y *f* o *n* y *u*, por poner un ejemplo. Esta corrección de errores se ha llevado a cabo comparando el texto resultante de la conversión con Image Reader con la imagen .jpeg descargada directamente de la base de datos Biodigithum.

Es importante mencionar que algunos de los textos preseleccionados tuvieron que ser eliminados a causa de su ininteligibilidad y fueron sustituidos por otros.

Todos los textos resultantes se almacenaron en formato .txt, que es el requerido por la gran mayoría de programas informáticos de análisis de datos textuales.

4.3. Variables

Las variables lingüísticas propuestas para estudiar se clasifican en tres grupos: las variables con respecto a los caracteres, las léxicas y las morfosintácticas. Para la detección de rasgos lingüísticos, hemos seguido la propuesta de Stamatatos (2008, p. 540), la cual se muestra en la Figura 3.

TABLE 1. Types of stylometric features together with computational tools and resources required for their measurement (brackets indicate optional tools).

Features		Required tools and resources
Lexical	Token-based (word length, sentence length, etc.)	Tokenizer, [Sentence splitter]
	Vocabulary richness	Tokenizer
	Word frequencies	Tokenizer, [Stemmer, Lemmatizer]
	Word *n*-grams	Tokenizer
	Errors	Tokenizer, Orthographic spell checker
Character	Character types (letters, digits, etc.)	Character dictionary
	Character *n*-grams (fixed-length)	-
	Character *n*-grams (variable-length)	Feature selector
	Compression methods	Text compression tool
Syntactic	Part-of-Speech	Tokenizer, Sentence splitter, POS tagger
	Chunks	Tokenizer, Sentence splitter, [POS tagger], Text chunker
	Sentence and phrase structure	Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser
	Rewrite rules frequencies	Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser
	Errors	Tokenizer, Sentence splitter, Syntactic spell checker
Semantic	Synonyms	Tokenizer, [POS tagger], Thesaurus
	Semantic dependencies	Tokenizer, Sentence splitter, POS tagger, Text Chunker, Partial parser, Semantic parser
	Functional	Tokenizer, Sentence splitter, POS tagger, Specialized dictionaries
Application-specific	Structural	HTML parser, Specialized parsers
	Content-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries
	Language-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries

Figura 3: Propuesta de análisis de funciones lingüísticas (Stamatatos, 2008).

En la tabla 2, detallamos nuestra propuesta de variables a analizar que, además de basarse en la clasificación de Stamatatos, también toma variables del estudio de Nini (2019), ya que comparte algunas de sus características.

	Variable
Análisis de caracteres	Longitud de palabras
Análisis léxico	Recuento de palabras por texto
	Recuento de *tokens* y *types*
	Recuento de *tokens* sin *stop-words*
	Cálculo de TTR y MATTR.
	Cálculo de n-gramas de *tokens* (bigramas y trigramas)
	Cálculo de n-lemas (bigramas y trigramas)
Análisis morfosintáctico	Número de oraciones
	Número de estructuras de subordinación
	Frecuencia de aparición de categorías gramaticales
	Recuento de puntuación
	Recuento de c*hunks*

Tabla 2: Resumen de las variables a analizar en este trabajo.

En el primer grupo, que hace referencia al análisis de los caracteres, hemos hecho un recuento, para cada uno de los textos, de la longitud de las palabras (en caracteres). Este cálculo es importante, ya que, como dice Stamatatos (2008, p. 541), "according to this family of measures, a text is viewed as a mere sequence of characters. That way, various character-level measures can be defined, including alphabetic characters count, digit characters count, uppercase and lowercase characters count, letter frequencies, punctuation marks count, etc. (…) This type of information is easily available for any natural language and corpus and it has been proven to be quite useful to quantify the writing style".

En cuanto al segundo grupo, hemos calculado el número de palabras por texto, el número de *tokens* y *types*, es decir, el número de formas y ocurrencias de cada texto, y su riqueza léxica en términos de TTR (*type-token ratio*) y MATTR, ambos con normalización de frecuencias.

También en este primer análisis, hemos creado una lista de la frecuencia de *tokens* de cada texto en su versión completa y otra tras eliminar las *stop-words*, es decir, aquellas palabras que no aportan ningún significado (como por ejemplo, artículos, preposiciones, entre otros). Stamatatos (2008, p. 540-ss) habla de las *stop-words* cuando afirma: "(…) note that such words are usually excluded from the feature set of the topic-based text classification methods since they do not carry any semantic information and they are usually called 'function' words".

Por último, hemos calculado los bigramas y trigramas de los *tokens* de cada grupo, es decir, combinaciones de dos y de tres palabras que aparecen en el texto, acompañados por su frecuencia de aparición. Hemos llevado a cabo la misma tarea para crear una lista de bigramas y trigramas de los lemas de cada grupo.

En relación con el análisis morfosintáctico, hemos realizado un recuento del número de oraciones por grupo y del número de estructuras de subordinación (oraciones de relativo, completivas y adverbiales). Además, hemos efectuado estudios frecuenciales sobre la aparición de las categorías gramaticales por grupos, aplicando diversos filtros y también sobre las combinaciones de estas mismas categorías gramaticales, a las que llamaremos *chunks*,[2] siguiendo el modelo de Stamatatos (2008, p. 540-ss).El autor, hablando sobre este aspecto (2008, p. 542-543), refiere algún ejemplo: "Another attempt to exploit syntactic information was proposed by Stamatatos, et al.(2000; 2001).They used an NLP tool able to detect sentence and chunk (i.e., phrases) boundaries in unrestricted Modern Greek text. For example, the first sentence of this paragraph would be analyzed as following: NP [*Another attempt*] VP [*to exploit*] NP [*syntactic information*] VP [*was proposed*] PP [*by Stamatatos et al. (2000)."*

Además, Sidorov (2019, p. 15) también habla de estas estructuras en su libro, en el que explica que "traditional n-grams are sequences of textual elements (words, lemmas, POS tags, etc.) in the order of their appearance in a text. Traditional n-grams represent syntagmatic information, and they are widely and successfully used in various computational linguistics tasks." En este punto, debemos decir que seguimos fielmente el método de trabajo de Cicres y Queralt (2019). Por último, en este mismo nivel, hemos llevado a cabo un recuento de la puntuación de cada grupo.

[2] Utilizamos este término para distinguir estas agrupaciones de categorías sintácticas (bichunks, trichunks) de las secuencias de palabras (bigramas o trigramas) o de lemas (bilemas o trilemas).

En el caso del estudio de las categorías gramaticales, hemos observado su frecuencia según distintas perspectivas. En concreto, los verbos, la puntuación y los pronombres han sido examinados con más detalle, atendiendo a su tipología (en el caso de pronombres y puntuación) o el tiempo (en el caso de los verbos).

En el caso del cálculo de *chunks* más frecuentes, es decir, la combinación de categorías que más suele aparecer en cada texto, las frecuencias se han calculado sin tener en cuenta los subtipos que engloba cada categoría (género, número, tiempo…), tal y como se puede observar en el siguiente ejemplo. Para ello, ha sido necesario realizar un proceso de simplificación de etiquetas, como ya se ha hecho anteriormente en otros proyectos, como el de Cicres y Queralt (2019), con el fin de agrupar todas las combinaciones que solo se diferenciaban por los ya mencionados subtipos. Es decir, se han agrupado, por ejemplo, las combinaciones ASF + NCSF (artículo singular femenino seguido de nombre común singular femenino) y ASM + NCSM (Artículo singular masculino seguido de nombre común singular masculino).

Etiqueta	**Descripción etiqueta**	**Frecuencia**
AS + NCS	Artículo singular + nombre común singular	13858
DS + NCS	Determinante singular + nombre común singular	2104
NCS + SP	Nombre común singular + preposición	5328
SP + AS	Preposición + artículo singular	7916
NCS + Fc	Nombre común singular + coma	1364

Tabla 3: Ejemplo del cálculo de frecuencias en *chunks*.

Una vez realizada esta simplificación, hemos seleccionado las 10 combinaciones más frecuentes y hemos hecho un recuento texto a texto de cada una de estas combinaciones para poder obtener su frecuencia y, de esta manera, poder construir la matriz de datos en SPSS[3], como se explicará en el siguiente apartado.

4.4. Herramientas

[3] https://www.ibm.com/es-es/analytics/spss-statistics-software

Durante el proceso de trabajo que estamos describiendo, se han utilizado diversas herramientas informáticas con finalidades distintas.

En primer lugar, para realizar los cálculos referentes al nivel léxico, utilizamos el programa *AntConc*[4], un conjunto de herramientas para el análisis de datos textuales y la elaboración de concordancias. *AntConc*, tal y como se explica en algunos estudios, como el de Muchnik-Rozanov y Tsybulsky (2020, p. 215), fue diseñado por Anthony Laurence para ser usado como instrumento de análisis textual en el aula. Mediante esta herramienta, se pueden generar concordancias, obtener la frecuencia de palabras y palabras clave, generar *clusters*(segmentos repetidos, conjuntos de palabras que suelen aparecer juntas), analizar el léxico y crear diagramas de distribución de palabras.

En cuanto al generador de concordancias de AntConc, Kunnanets, Levchenko y Hadzalo (2018, p. 145) explican que "Concordance Tool provides searches and results in KWIC format (keyword in context). This allows to analyze what words and phrases are commonly used in the text body".

En segundo lugar, para realizar el análisis del nivel morfosintáctico, se procesaron los textos de cada grupo con el programa *FreeLing*[5], que, tal y como explica Padró (2011, p. 13), es "una librería de código abierto para el procesamiento multilingüe automático, que proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas". Aplicando el etiquetado automático del programa obtuvimos una salida verticalizada del texto en forma de columnas correspondientes a la forma, el lema y la categoría gramatical de cada palabra (*token*) del texto. Las categorías están representadas por etiquetas que siguen el estándar EAGLES[6], asignadas tras un proceso previo de desambiguación. En la propia web del proyecto EAGLES, se explica que "el analizador morfológico para el castellano utiliza un conjunto de etiquetas para representar la información morfológica de las palabras. Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas". Se trata de etiquetas de tipo PoS (*Part of Speech*).

[4] http://www.laurenceanthony.net/software/antconc/
[5] http://nlp.lsi.upc.edu/freeling/index.php/node/1
[6] https://www.cs.upc.edu/~nlp/tools/parole-sp.html

Hemos utilizado la línea de comandos del entorno *Unix,* un sistema operativo portable, multitarea y multiusuario para obtener otro tipo de resultados: *chunks*, bilemas y trilemas de los grupos correspondientes a cada rango de edad, junto con su frecuencia de aparición.

A continuación, se muestra la tabla con las herramientas utilizadas según la variable.

	Variable	**Herramienta**
Análisis de caracteres	Longitud de palabras	Excel
Análisis léxico	Recuento de palabras por texto	AntConc
	Recuento de *tokens* y *types*	AntConc
	Recuento de *tokens* sin *stop-words*	AntConc
	Cálculo de TTR y MATTR.	AntConc y Lancaster Stats Tool
	Cálculo de n-gramas de *tokens* (bigramas y trigramas)	AntConc y Unix
	Cálculo de n-lemas (bigramas y trigramas)	Freeling y Unix
Análisis morfosintáctico	Número de oraciones	Freeling y Excel
	Número de estructuras de subordinación	Freeling y Excel
	Frecuencia de aparición de categorías gramaticales	Freeling
	Recuento de puntuación	Freeling y Excel
	Recuento de c*hunks*	Freeling y Unix

Tabla 4: Resumen de las variables a analizar y la herramienta con la que se realizará dicho análisis.

4.5. Análisis estadístico

La tabulación de los datos obtenidos mediante las diversas herramientas utilizadas en hojas de cálculo Excel nos permitió aplicar filtros, realizar cálculos rápidos aplicando fórmulas y ordenar los datos. A partir de aquí creamos la matriz de datos que se analizaría posteriormente con SPSS. SPSS es, según su propia web indica, "una plataforma de software que ofrece un análisis estadístico avanzado, una amplia biblioteca de algoritmos de *machine learning*, análisis de texto, extensibilidad de código abierto, integración con *big data* y un fácil despliegue en las aplicaciones".

Todos los datos recopilados hasta ese momento aparecen en forma de cifras absolutas, excepto algunos porcentajes y los cálculos TTR y MATTR. Pero para poder realizar un análisis científico y validar nuestra hipótesis procedimos a su normalización y a la aplicación de diversas pruebas que nos ayudaran a confirmar la posibilidad de encontrar rasgos lingüísticos discriminantes entre franjas de edad.

En cuanto a la riqueza léxica, hay que decir que el cálculo TTR presenta dependencia respecto a la extensión de los textos. Como dice Stamatatos (2008, p.540), "unfortunately, the vocabulary size heavily depends on text-length (as the text-length increases, the vocabulary also increases, quickly at the beginning and then more and more slowly). Various functions have been proposed to achieve stability over text-length". En este caso, dado que, aun siendo todos ellos textos breves, la longitud era distinta, optamos por una medida que normalizara este cálculo. Inicialmente, llevamos a cabo un cálculo normalizado con base 1000 ($STTR_{1000}$) pero, más adelante, optamos por utilizar la alternativa MATTR (*Moving Average type/token ratio*) porque, tal y como explica Brezina (2018, p. 58), "instead of dividing the text into successive non-overlapping segments, MATTR uses and overlapping window smoothly moving through the text; for each window position the TTR of the text inside the window is calculated and then the mean value of the TTRs obtained in this way is computed. MATTR is thus a more robust measure of lexical richness than STTR because it takes into account all possible segmentations of the text". Este último cálculo lo obtuvimos gracias a la página Lancaster Stats Tool online y, más concretamente, a la herramienta ToolBox.

En la mencionada página web, se explica que "the website provides practical support for the analysis of corpus data using a range of statistical techniques. "ToolBox, en

concreto, es una caja de herramientas que permite realizar diferentes tipos de tareas, como estudios frecuenciales, elaboración de colocaciones o cálculos sobre correlaciones, agrupaciones y factores.

En cuanto al uso del programa estadístico SPSS para normalizar los cálculos, elaboramos tres tablas de datos para estudiar las diferentes variables: la longitud de las palabras (por caracteres), la frecuencia de las categorías gramaticales (en la que se incluye las variables que hacen referencia al estudio de número de tokens, types y MATTR) y la frecuencia de *chunks* compuestos por dos y tres categorías.

En el caso del cálculo de la longitud de las palabras, realizamos una tabla de datos con las variables de grupo (tres rangos de edad a estudiar), número de texto, número de *tokens* para cada texto, el número de palabras que tienen de uno hasta doce caracteres (clasificadas por separado) en cada texto y acompañada de otra variable con el porcentaje, también para cada texto.

En relación con el cálculo de la frecuencia de categorías, hemos creado una tabla de datos similar a la anterior que contiene datos tanto léxicos como morfosintácticos. Esta vez, la tabla contiene las variables de grupo, número de texto, número de *tokens* en cada texto, número de *types* en cada texto, cálculo de la MATTR, número de oraciones subordinadas en cada texto y el número de veces que aparece una determinada categoría en cada texto acompañada por su porcentaje.

El grupo de categorías incluidas en este análisis está formado por: adjetivos, adverbios (repartidos entre adverbios de negación y el resto), artículos, conjunciones (distinguiendo entre conjunciones de coordinación y subordinación), determinantes, nombres comunes, nombres propios, pronombres (clasificados en dos grupos, relativos y el resto) y verbos (entre los cuales hemos diferenciado los verbos principales, los verbos auxiliares y las formas del verbo ser, distinguiendo en todos los casos si se trata de formas personales o no personales). En el análisis también se han incluido ciertos tipos de puntuación: punto, punto y coma, coma, dos puntos y puntos suspensivos.

Por último, teniendo en cuenta el cálculo de *chunks*, realizamos la matriz de datos con las variables de grupo, número de texto, número de *bichunks* totales en cada grupo (utilizamos este término para designar las combinaciones de dos categorías gramaticales y diferenciarlo así de los *trichunks*, los cuales incluyen tres categorías) y el número de

veces que aparece determinada combinación en cada texto, acompañada también de su porcentaje. Se añadieron las mismas variables en el caso de los *trichunks*.

Una vez completas las matrices de datos, procedimos al análisis estadístico, en el que, en primer lugar, realizamos un cálculo de la normalidad de las variables. Para ello, utilizamos la prueba de Kolmogorov-Smirnov de bondad de ajuste a una distribución normal. Cuando el valor de significación es superior a .05, asumimos que los datos se distribuyen normalmente; en caso contrario, consideramos que la distribución no se ajusta a la normalidad.

Después de calcular la normalidad, se comprobó la diferencia de medias entre los diferentes grupos para cada variable. En el caso de las variables cuyos valores se ajustaban a la distribución normal, se utilizó el análisis de variancias (ANOVA) de un factor; en el caso de las variables cuyos datos no se ajustaban a la normalidad, se optó por una alternativa no paramétrica (prueba H de Kruskal-Wallis).

El análisis de variancias, tal y como explican Cicres y Queralt (2019, p. 61), resulta de utilidad para establecer las variantes que muestran diferencias significativas en cuanto a la clasificación de textos en los grupos preestablecidos; además las pruebas *post hoc* nos muestran las diferencias de grupos dos a dos.

Por último, realizamos un análisis discriminante para cada uno de los tres cálculos. Cicres y Queralt(2019, p. 61) hablan del doble objetivo de un análisis discriminante: por una parte, identifica las funciones que pueden servir para diferenciar dos o más grupos y, por la otra, puede asignar a nuevos casos su clasificación a uno de los grupos. Tenemos una descripción concisa y muy clara de cómo se realiza un análisis discriminante y de la interpretación de los datos que ofrece en la página web del profesor Llopis, *La estadística: una orquesta hecha instrumento*.

5. ANÁLISIS DE LOS RESULTADOS

En el presente apartado presentaremos los resultados obtenidos por cada variable y los describiremos con el fin de poder determinar si dichas variables resultan útiles para discriminar entre los diferentes grupos.

5.1. Longitud de palabras

En referencia a la longitud de las palabras, una vez creada la matriz de datos con una muestra de N=90, comprobamos el ajuste a la distribución normal mediante la prueba de bondad de ajuste de Kolmogorov-Smirnov.

A continuación, contrastamos las medias con distribución normal (palabras de tres, siete y ocho caracteres) con la función ANOVA de un factor. Los resultados muestran que ninguna de las tres variables es significativa.

En cuanto a las variables que no presentan distribución normal, realizamos el contraste mediante la prueba H de Kruskal-Wallis. El resultado obtenido mostró que tampoco hay ninguna variable significativa para discriminar.

5.2. Recuento de tokens, types y MATTR

Según la prueba H de Kruskal-Wallis, el número de tokens de un texto no nos permite discriminar entre grupos, ya que la diferencia de medias arrojó un valor p = 0,869 por lo que no había diferencias estadísticamente significativas.

A continuación, se puede observar un gráfico que muestra el número de *tokens* según el grupo.

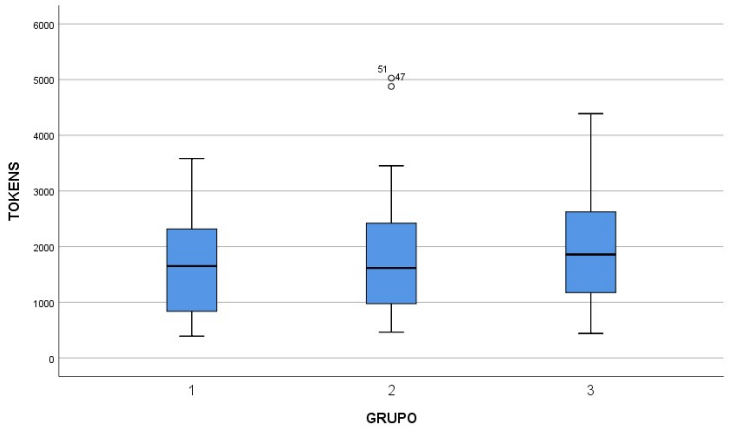

Figura 4: Recuento de tokens.

Tal y como se puede observar en la Figura 4, la mediana es más alta en el tercer grupoque en el resto (1650.50, 1615.00 y 1858.50 para los grupos 1, 2 y 3, respectivamente). El rango del número de *tokens* es más o menos igual en los dos primeros grupos pero mucho más amplio en el tercero. En concreto, los valores mínimos son 391, 464 y 440 para los grupos 1, 2 y 3, respectivamente. El valor máximo es de 3581 para el grupo compuesto por mujeres más jóvenes, de 3453 para las mujeres de 35 a 50 años y de 4388 para las mujeres de edad más avanzada.

Así pues, concluimos que el grupo 3, las mujeres de edad más avanzada, difieren de los dos grupos de mujeres más jóvenes porque presentan un rango de *tokens* más amplio en los valores máximos, utilizan entre 800 y 900 *tokens* más, lo que se traduce en que las mujeres de más de 50 años escriben textos más largos.

En relación con el cálculo de *types*, comprobamos que no es una variable significativa estadísticamente. Aun así, con un diagrama de cajas se pueden observar diferencias cualitativas entre los grupos.

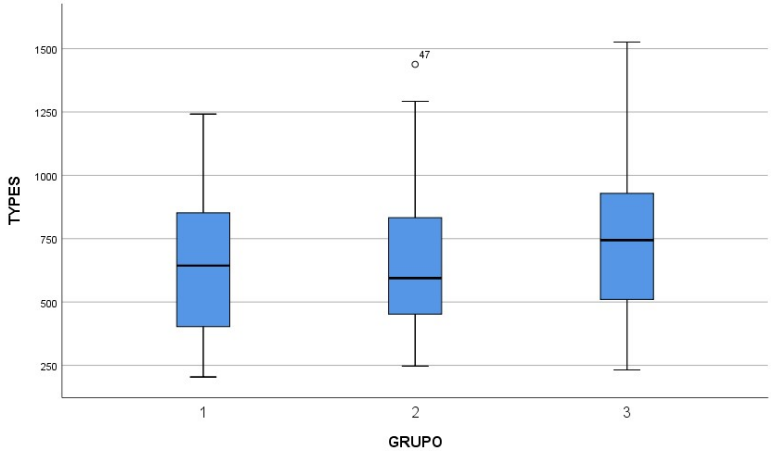

Figura 5: Recuento de *types*.

En la Figura 5, se puede observar que la mediana del tercer grupo es más elevada respecto a las del grupo 1 y 2 (643.50, 654.04 y 744.00 para los grupos 1, 2 y 3, respectivamente). Además, también hay diferencia en cuanto al rango del número de *types*, ya que este es bastante más alto en el tercer grupo que en los dos primeros. Específicamente, los valores mínimos son bastante similares en los tres grupos, ya que

son de 204 para el primer grupo, de 247 para el segundo y 232 para el tercero. Sin embargo, en cuanto al valor máximo se refiere, se puede ver claramente la distancia que separa los grupos 1 y 2 respecto del 3, ya que los valores máximos para estos dos primeros grupos son de 1242 y 1292, mientras que para el tercero es de 1526.

En este caso, también es el tercer grupo el que difiere de los demás, ya que el rango de *types* es más amplio en los valores máximos, utilizan entre 150 y 200 *types* más, por lo que podemos concluir que hay una mayor diversidad en el uso del léxico en el grupo de las mujeres de edad superior a 50 años que en los dos grupos de menor edad.

Por lo que concierne al cálculo de MATTR, es decir, el cálculo de riqueza léxica normalizado, la ANOVA de un factor no ofreció resultados significativos ($F = .886$, $p > .05$). No obstante, si observamos el siguiente gráfico, concluimos que esta variable sí sirve para discriminar entre grupos, ya que se aprecia que la media de riqueza léxica es mayor en el grupo 3 ($MATTR_{500} = .4946$), es decir, en el grupo formado por mujeres de más de 50 años, mientras que el grupo constituido por mujeres de entre 35 y 50 años es el que es el que muestra menor riqueza ($MATTR_{500} = .4850$).

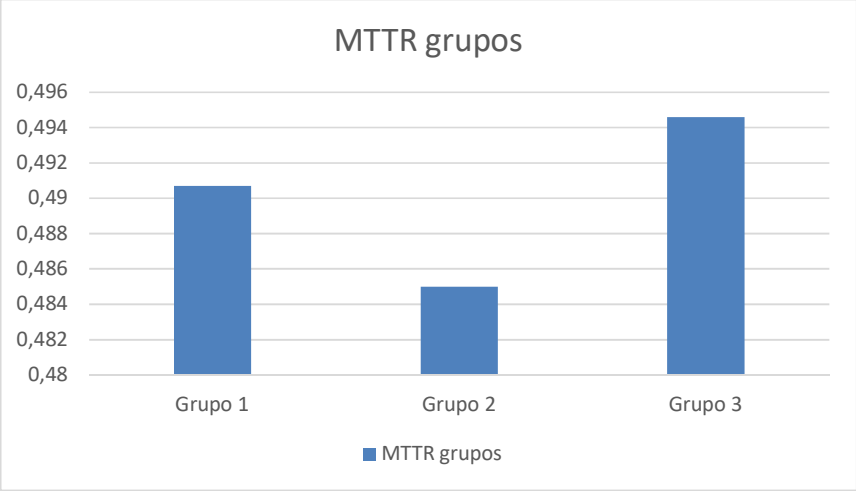

Figura 6: Cálculo de MATTR.

5.3. Tokens de mayor frecuencia.

En consideración a la frecuencia de *tokens*, a continuación, se muestra un gráfico que resume cuáles son los 10 más frecuentes para cada grupo.

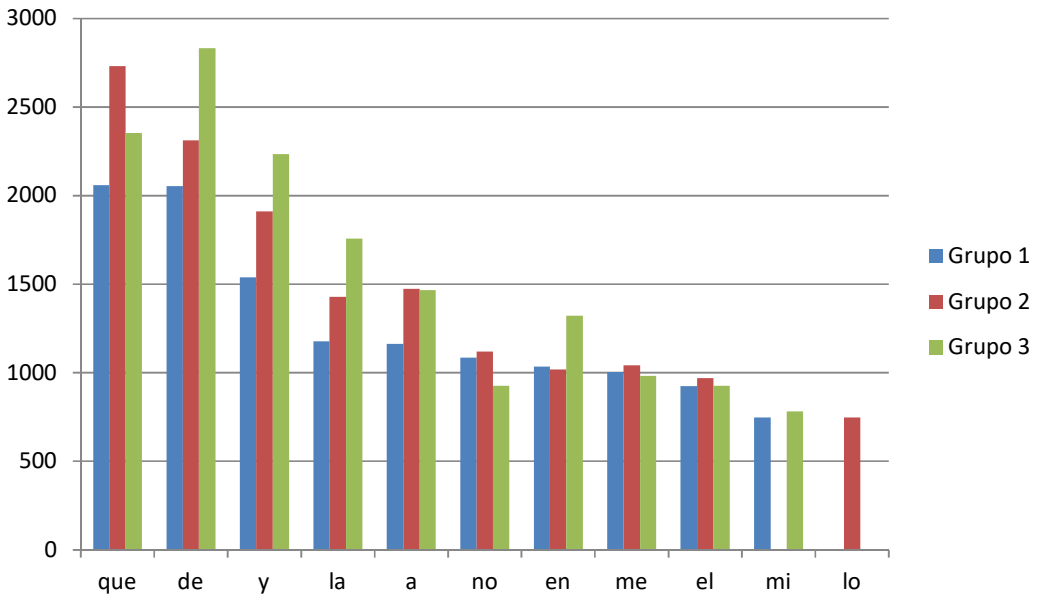

Figura 7: Tokens más frecuentes para cada grupo.

Tal y como se puede observar en el gráfico, "que" es el *token* más frecuente para las mujeres más jóvenes, mientras que para las de más avanzada edad es "de". En cuanto al *token* menos frecuente, no coinciden los tres grupos, sino que para el primero y el tercero es "mi" y para el segundo es el pronombre "lo".

Hemos llevado a cabo el mismo cálculo para la frecuencia de *tokens* una vez eliminadas las *stop-words*.[7]

[7] Lista extraída de https://github.com/stopwords-iso/stopwords-es

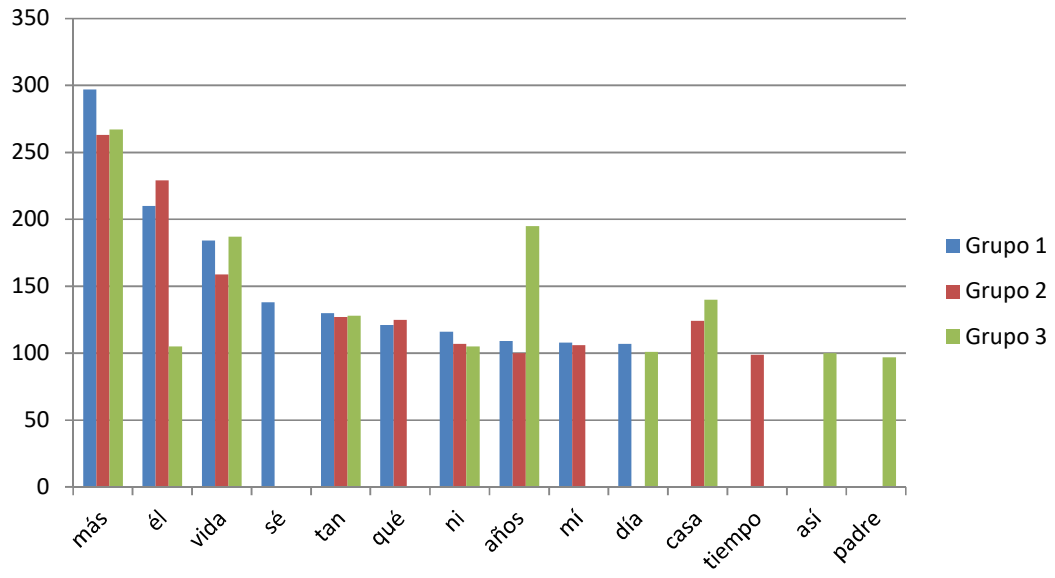

Figura 8: Tokens más frecuentes de cada grupo una vez eliminadas las *stop-words*.

En el caso de *tokens* sin *stop-words* observamos que "más" es el más frecuente para los tres grupos de edad. Sin embargo, notamos bastantes diferencias en el grupo compuesto por las mujeres mayores de 50 años respecto a los demás, ya que presenta dos *tokens* que no aparecen en los otros grupos, "así" y "padre". Además, destaca la gran frecuencia de la palabra "años".

5.4. Bigramas y trigramas de palabras

En cuanto a la frecuencia de bigramas y trigramas de palabras, estos datos también los hemos recopilado en forma de gráficos. En la Figura 9 aparecen los 10 bigramas más frecuentes para cada grupo.

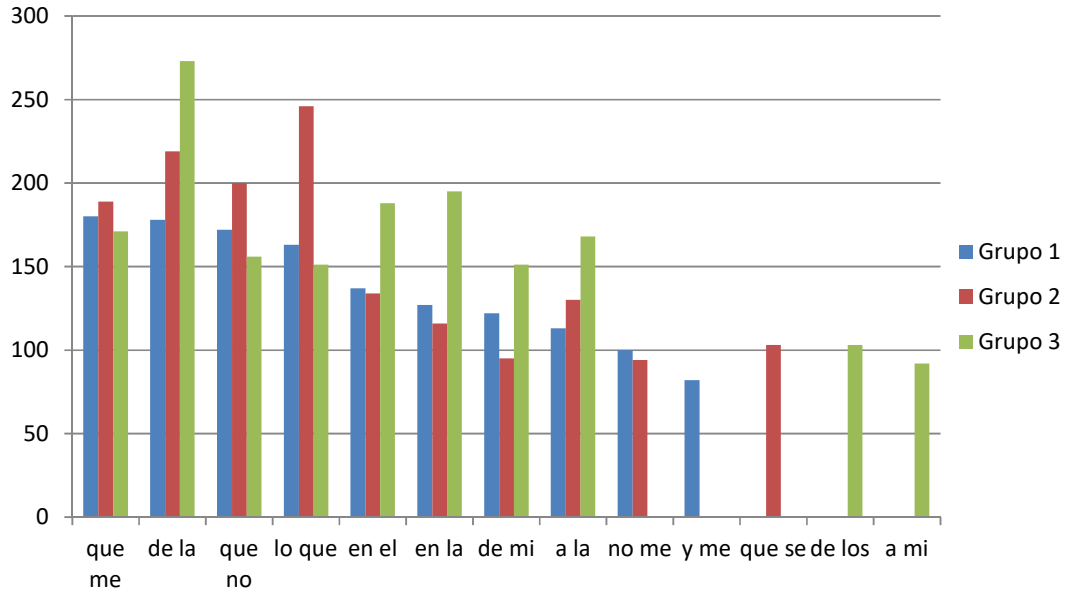

Figura 9: Bigramas más frecuentes para cada grupo.

En el caso de los bigramas, se pueden apreciar muchas más diferencias entre grupos, que en el caso de los *tokens* y, de nuevo, las diferencias se hacen más patentes en el grupo de mujeres de mayor edad, ya que este cuenta con dos bigramas que no aparecen en el resto de grupos: “de los” y “a mi”. También de este tercer grupo, llama la atención la alta frecuencia del bigrama “de la”.

En relación con los trigramas, los resultados se pueden ver a continuación.

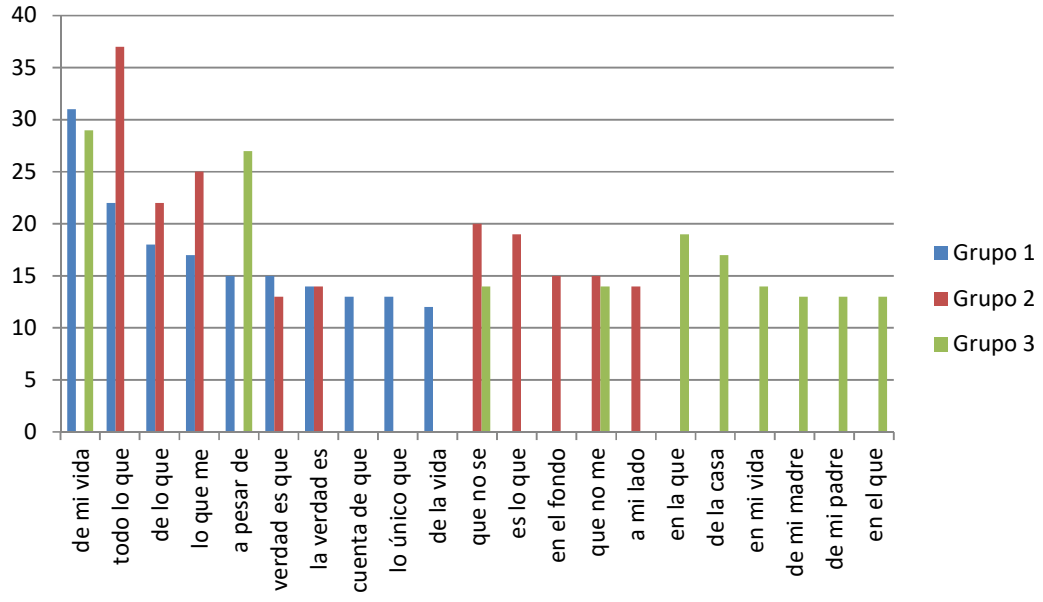

Figura 10: Trigramas más frecuentes para cada grupo.

En el caso de los trigramas, las diferencias son notables entre todos los grupos, aunque también destaca más el grupo de mujeres mayores de 50 años, porque es el que posee más trigramas diferentes al resto de grupos (“en la que”, “de la casa”, “en mi vida”, “de mi madre”, “de mi padre” y “en el que”). Como se puede observar rápidamente en la Figura 15, no hay ningún trigrama compartido por los 3 grupos.

5.5. Bilemas y trilemas

A continuación, se muestra un gráfico en el que se representan los 10 bilemas más frecuentes de los tres grupos.

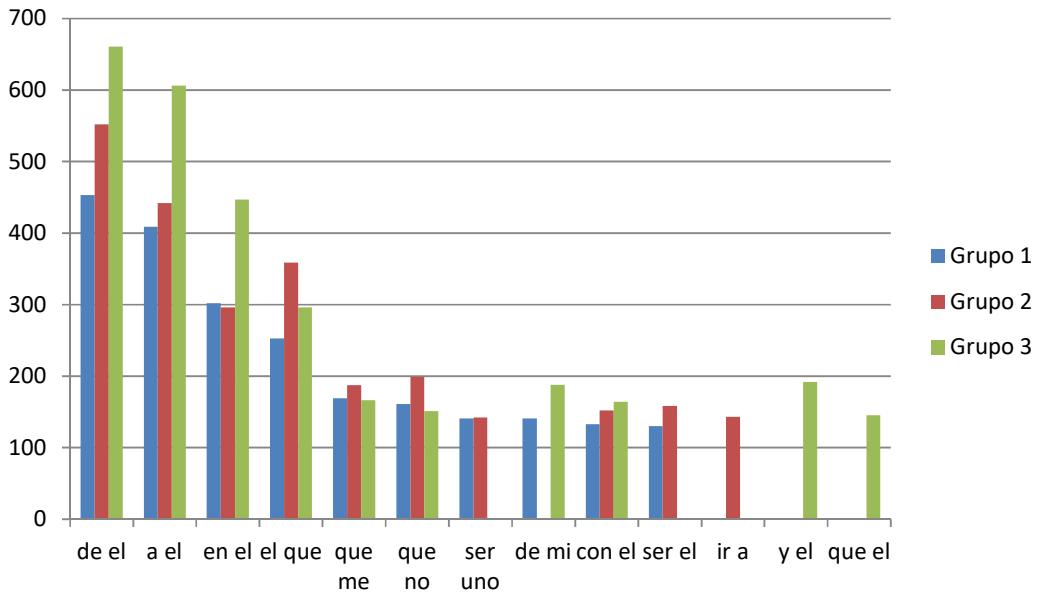

Figura 11: Bilemas más frecuentes para cada grupo.

En este caso, el bilema "de el" es el que obtiene mayor frecuencia en los tres grupos, seguido de "a el". Otra vez, el grupo compuesto por las mujeres de edad más avanzada es el que más se diferencia, ya que tiene dos bilemas propios ("y el" y "que el") que no aparecen en el resto de grupos.

En cuanto a los trilemas, también se presentan seguidamente las tablas con los resultados.

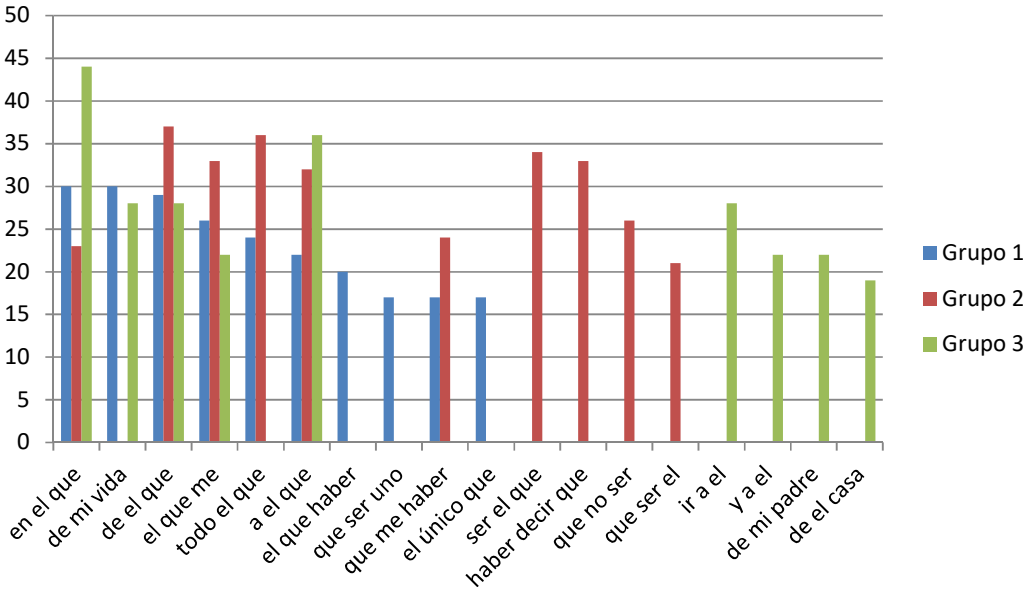

Figura 12: Trilemas más frecuentes para cada grupo.

Al igual que con los trigramas de palabras, también hay mucha diferencia entre grupos en los trilemas, ya que cada grupo tiene varios que no contienen los otros grupos, tal y como se ve claramente en el gráfico.

5.6. Cálculos sobre el número de oraciones y el número de estructuras subordinadas

En relación con los cálculos del número de oraciones por texto y el número de estructuras subordinadas por texto, según la prueba Kruskal-Wallis no se dan diferencias significativas entre los grupos ($p > .05$). Así pues, esta variable no servirá para identificar el rango de edad al que pertenece un autor.

5.7. Frecuencia de aparición de categorías gramaticales

En referencia a la frecuencia de aparición de categorías gramaticales, se han analizado las siguientes categorías: el adjetivo, el adverbio, el adverbio de negación, el artículo, la conjunción coordinada, la conjunción subordinada, el determinante, la interjección, el nombre común, el nombre propio, la preposición, el relativo, el verbo auxiliar, el verbo principal y el verbo ser.

Según la prueba Kruskal-Wallis, la variable que indica el porcentaje de aparición de la categoría referente a la interjección muestra diferencias significativas en sus medias

entre los distintos grupos, ya que presenta una significación de p = .018. Mediante las medias calculadas, se ha comprobado que el grupo formado por mujeres de 20 a 35 años tiene tendencia a utilizar más interjecciones que los grupos formados por mujeres de entre 35 y 50 años y por mujeres de más de 50 años.

5.8. Recuento de puntuación

Por lo que concierne al recuento de puntuación, la variable correspondiente a los dos puntos presenta una distribución que, según la prueba Kruskal-Wallis, permite discriminar entre los tres rangos de edad, ya que presenta un valor de significación de p = .001.

Después de estudiar la media del uso de este signo de puntuación, hemos podido concluir que las mujeres mayores de 50 años emplean con más frecuencia los dos puntos que las mujeres de menor edad.

5.9. Bichunks y trichunks

En relación con la frecuencia de los *bichunks*, a continuación, se muestra una tabla con las variables que hemos analizado, que son las más frecuentes.

BICHUNK	**DESCRIPCIÓN ETIQUETA**
DS_NCS	Determinante singular + nombre común singular
AS_NCS	Artículo singular + nombre común singular
NCS_SP	Nombre común singular + preposición
SP_AS	Preposición + artículo singular
SP_DS	Preposición + determinante singular
NCS_Fc	Nombre común singular + coma
SP_VI	Preposición + verbo en infinitivo
NCS_Fp	Nombre común singular + punto
CC_SP	Conjunción coordinada + preposición
SP_NCS	Preposición + nombre común singular

Tabla 5: Resumen de las etiquetas de *bichunks*.

Según la prueba Kruskal-Wallis, las medias de los valores de todas las variables son significativamente diferentes entre los distintos grupos.

Mediante el cálculo de diferencias entre medias, hemos comprobado que el grupo formado por las mujeres con una edad entre los 20 y los 35 años presentan una tendencia a utilizar, más que las mujeres mayores, los siguientes *bichunks*: DS_NCS, AS_NCS, NCS_SP, SP_AS, SP_DS, SP_VI, NCS_Fp y SP_NCS. En cambio, las mujeres de entre 35 y 50 años utilizan los *bichunks* NCS_Fc y CC_SP con una frecuencia mayor que el resto de mujeres.

En el proceso del análisis discriminante, aparecen dos variables como posibles discriminadoras: la variable que representa el *bichunk* formado por nombre común singular seguido de punto (NCS_Fp) y el *bichunk* compuesto por preposición y verbo en infinitivo (SP_VI). Esta información se puede observar en la Figura 13 que proporciona el programa SPSS.

Variables entradas/eliminadas[a,b,c,d]

Paso	Entrada	Lambda de Wilks							
						F exacta			
		Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	Perc_NCS_Fp	,688	1	2	87,000	19,708	2	87,000	,000
2	Perc_SP_VI	,264	2	2	87,000	40,729	4	172,000	,000

En cada paso, se entra la variable que minimiza la lambda de Wilks global.

a. El número máximo de pasos es 20.

b. La F mínima parcial para entrar es 3.84.

c. La F máxima parcial para eliminar es 2.71.

d. El nivel F, la tolerancia o VIN no suficiente para un cálculo adicional.

Figura 13: Variables seleccionadas (*bichunks*).

Tal y como se puede ver, las dos variables presentan una significación menor a .05, por lo que concluimos que nos indican la existencia de potencia discriminadora entre los distintos grupos.

En la tabla de autovalores, que se muestra en la Figura 23, se nos informa de la existencia de dos funciones discriminantes. También se proporciona la estimación de la varianza que cada función es capaz de explicar (F1 = 98,5 %, F2 = 1,5 %).

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	2,647[a]	98,5	98,5	,852
2	,040[a]	1,5	100,0	,195

a. Se utilizaron las primeras 2 funciones discriminantes canónicas en el análisis.

Figura 14: Autovalores de las variables seleccionadas (*bichunks*)

Para acabar, examinamos una última tabla, la referente a los resultados de clasificación. En dicha tabla podemos observar que el 91,1 % de los casos han sido pronosticados correctamente y podemos ver las asignaciones que hace a cada grupo, todas correctas en el grupo 2. En el caso de la validación cruzada, el resultado de clasificación correcta es de 88,9 %.

Resultados de clasificación[a,c]

			Pertenencia a grupos pronosticada			
		GRUP	1	2	3	Total
Original	Recuento	1	25	5	0	30
		2	0	30	0	30
		3	0	3	27	30
	%	1	83,3	16,7	,0	100,0
		2	,0	100,0	,0	100,0
		3	,0	10,0	90,0	100,0
Validación cruzada[b]	Recuento	1	25	5	0	30
		2	0	30	0	30
		3	0	5	25	30
	%	1	83,3	16,7	,0	100,0
		2	,0	100,0	,0	100,0
		3	,0	16,7	83,3	100,0

a. 91,1% de casos agrupados originales clasificados correctamente.

b. La validación cruzada se ha realizado sólo para aquellos casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas de todos los casos distintos a dicho caso.

c. 88,9% de casos agrupados validados de forma cruzada clasificados correctamente.

Figura 15: Resultados de clasificación de las variables seleccionadas (*bichunks*).

A continuación, se muestran los resultados a modo de gráfico de dispersión, para ver la clara diferencia entre los tres grupos. En este caso, la función 1 acumula el mayor valor para discriminar. Se observa como el centroide del grupo 3 se sitúa en valores negativos, el grupo 2 en el valor 0 y el grupo 1 en valores positivos.

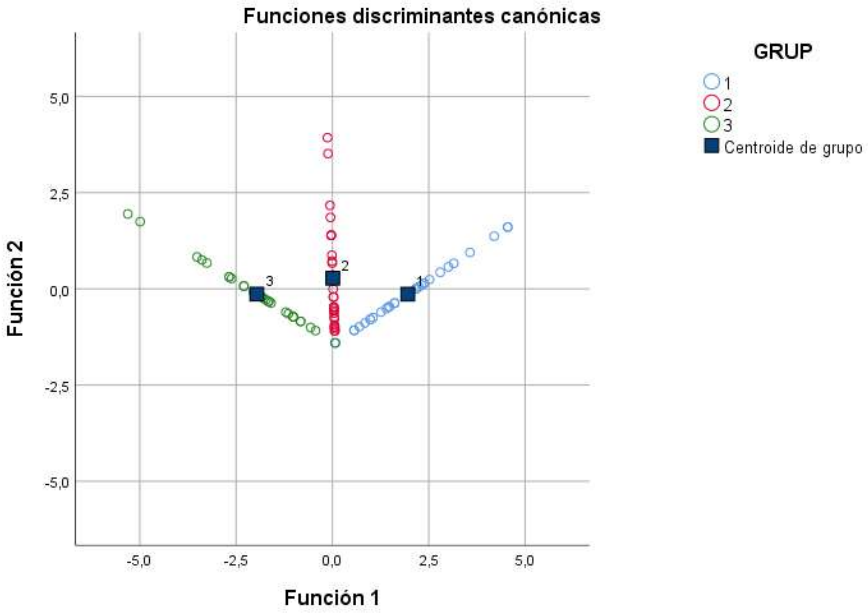

Figura 16: Gráfico de dispersión de las variables seleccionadas (*bichunks*) entre los distintos grupos

Después de examinar la diferencia entre medias de estas dos variables, hemos extraído la conclusión de que las mujeres de una edad entre 20 y 35 años son las que usan más frecuentemente las combinaciones de categorías gramaticales compuestas por nombre común seguido de punto (por ejemplo, "respiración.") y preposición seguida de verbo en infinitivo (por ejemplo, "de vivir"). Por el contrario, las mujeres mayores a 50 años las utilizan con muy baja frecuencia.

Para acabar, en cuanto a los *trichunks*, a continuación, se muestra una tabla con las variables que hemos analizado.

TRICHUNK	DESCRIPCIÓN ETIQUETA
AP_NCP_SP	Artículo plural + nombre común plural + preposición
AS_NCS_CC	Artículo singular + nombre común singular + conjunción coordinada
AS_NCS_Fc	Artículo singular + nombre común singular + coma
AS_NCS_Fp	Artículo singular + nombre común singular + punto
AS_NCS_SP	Artículo singular + nombre común singular + preposición
DS_NCS_Fc	Determinante singular + nombre común singular + coma
DS_NCS_Fp	Determinante singular + nombre común singular + punto
DS_NCS_SP	Determinante singular + nombre común singular + preposición
NCS_Fc_CC	Nombre común singular + coma + conjunción coordinada
NCS_SP_AS	Nombre común singular + preposición + artículo singular

Tabla 6: Resumen de las etiquetas de *trichunks*.

Según la prueba H de Kruskal-Wallis, ninguna de las variables es significativa, hecho que corrobora el análisis discriminante.

A continuación, en la tabla 7, se muestra entre qué grupos discriminan las variables significativas descritas.

	Grupo 1-2	Grupo 2-3	Grupo 1-3
Número de tokens		✓	✓
Número de types		✓	✓
MATTR		✓	✓
Interjección	✓		✓
Dos puntos		✓	✓
AS_NCS	✓		✓
CC_SP	✓	✓	
DS_NCS	✓		✓
NCS_Fc	✓	✓	
NCS_Fp	✓		✓
NCS_SP	✓		✓
SP_AS	✓		✓
SP_DS	✓		✓
SP_NCS	✓		✓
SP_VI	✓		✓

Tabla 7: Resumen de las variables discriminantes entre grupos.

Tal y como se puede apreciar, el primer grupo, formado por las mujeres más jóvenes, se diferencia de los dos últimos por la categoría que designa la interjección y por los *bichunks* AS_NCS, DS_NCS, NCS_Fp, NCS_SP, SP_AS, SP_DS, SP_NCS y SP_VI.

En cambio, el segundo grupo se diferencia del resto por los bichunks CC_SP y NCS_Fc.

En cuanto al tercer grupo, este se diferencia de los dos primeros en el caso de las variables referentes al número de *tokens*, número de *types*, MATTR y los dos puntos.

6. CONCLUSIONES

Una vez expuestos los resultados del análisis, se presentan, en este apartado, las conclusiones que se han podido extraer, teniendo en cuenta los objetivos que persigue el trabajo y la pregunta planteada al inicio de este.

Así pues, se ha podido comprobar, mediante el análisis de las variables propuestas a estudiar, que la hipótesis que presentamos se cumple y que la respuesta a nuestra pregunta inicial es afirmativa. Es decir, existen ciertos rasgos lingüísticos en textos producidos por mujeres que nos permiten clasificarlos por franjas de edad y, por lo tanto, son susceptibles de ser utilizados en una pericial lingüística para la construcción de un perfil lingüístico que determine la edad.

Cabe destacar que no todas las variables propuestas han mostrado potencial discriminatorio. En un primer lugar, se detallan las variables que, según la muestra analizada en este estudio, no muestran diferencias significativas entre los rangos de edad: la longitud de las palabras, el número de oraciones, el número de estructuras subordinadas y los *trichunks*. Ante estos resultados, hay que tener presente la opinión de Nini (2019, p. 53): "The understanding of the underlying linguistic patterns responsible for the predictions is a pre-requisite for forensic authorship profiling because, ultimately, the evidence analysed is linguistic and not statistical. Therefore, although computational methods can and should be employed to aid the analysis, this must not be done at the expense of the underlying linguistic explanations, which should remain the primary focus within forensic linguistics".

En segundo lugar, se han determinado distintas variables que poseen potencial discriminante entre los grupos de edad y que, por tanto, verifican nuestra hipótesis de partida. Estas variables son: el número de *tokens*, el número de *types*, el cálculo de MATTR, la categoría referente a la interjección, el signo de puntuación correspondiente a los dos puntos y los *bichunks*.

En relación con el número de *tokens*, las mujeres mayores de 50 años son las que presentan un rango más amplio y, por tanto, las que escriben textos más largos, mientras que el grupo de mujeres de 35 a 50 años, son las que presentan el menor rango de *tokens*. Es importante añadir que el *token* "que" es el más frecuente entre los grupos 1 y 2, es decir, el de mujeres más jóvenes y "de" es el más habitual para el tercer grupo.

Por lo que concierne al número de *types*, el grupo que presenta un rango más amplio también es el formado por mujeres de edad más avanzada y esto se puede traducir en que este grupo es el que utiliza un léxico más variado. Por lo contrario, las mujeres más jóvenes, las del grupo 1, son las que menos repertorio de vocabulario presentan,

En cuanto al cálculo de MATTR, hemos comprobado que las mujeres mayores de 50 años son las que más riqueza léxica presentan, ya que el rango es el más amplio.

Respecto a la frecuencia de aparición de categorías gramaticales, la categoría referente a la interjección, es la que permite discriminar entre grupos, porque muestra que las mujeres más jóvenes (entre 20 y 35 años) tienen tendencia a usarla con mucha más frecuencia que las mujeres más mayores.

En consideración a los signos de puntuación, las mujeres de edad más avanzada, es decir, las del tercer grupo, se distinguen de las demás por hacer un uso más habitual de los dos puntos.

Referente a la combinación de categorías, los *bichunks* permiten discriminar entre grupos, ya que, después de analizas los resultados, hemos llegado a la conclusión que las mujeres de una edad entre 20 y 35 años, suelen usar más habitualmente los siguientes: DS_NCS, AS_NCS, NCS_SP, SP_AS, SP_DS, SP_VI, NCS_Fp y SP_NCS. En cambio, las mujeres de entre 35 y 50 años utilizan los *bichunks* NCS_Fc y CC_SP con una frecuencia mayor que el resto de mujeres. Cabe mencionar que el análisis discriminante destaca los *bichunks* correspondientes a nombre común singular seguido de punto (NCS_Fp) y a preposición seguida de verbo infinitivo (SP_VI) que han conseguido clasificar correctamente el 91,1 % de los casos.

Tal y como se puede comprobar, el análisis mediante *bichunks* presenta un nivel alto de precisión por lo que son las variables más acertadas para su posible aplicación en el ámbito forense.

Estos resultados concuerdan con los hallazgos de trabajos previos como el de Spassova y Turell (2007) que informan del alto potencial discriminatorio de los *chunks* como marcas de identificación. Cicres y Queralt (2019, p. 55) explican que “the use of n-grams has yielded very successful results in the determination of the authorship of written texts within the field of forensic linguistics”.

Creemos que los resultados aportados corroboran la utilidad de los análisis estilométricos aunque también se muestran las limitaciones de dichos análisis, dado el alto número de variables que no han resultado distintivas entre grupos de edad. Por este motivo, se puede concluir que los análisis estilométricos pueden ser de ayuda, pero no pueden suponer la totalidad del análisis para la construcción de perfiles lingüísticos en contextos forenses.

Respecto a las limitaciones con las que se ha encontrado este proyecto, cabe mencionar que con el análisis de más textos sería posible afinar los resultados. Además, debido a la falta de algunos textos para poder realizar el análisis, hemos tenido que añadir relatos procedentes de zonas geográficas que en un principio no estaban previstas, pero que no han distorsionado el resultado.

Finalmente, otra de las contribuciones de este trabajo es la reivindicación de explotar corpus creados para otros fines. El corpus Biodigitum ha sido susceptible de este análisis lingüístico forense, una perspectiva distinta para lo que fue diseñado. Cabe añadir la relevancia de explotar el corpus para distintos fines puesto que se caracteriza por textos producidos por mujeres (clasificado además, por edad y procedencia) y podría ayudar a conocer mucho más el lenguaje de las mujeres mediante estudios futuros.

Así pues, como futuras líneas de trabajo, se podría utilizar este mismo corpus para ampliar el análisis etalectal con más textos y también realizar un análisis lingüístico que confirme el potencial discriminatorio de otra variable sociolingüística, como puede ser la procedencia del autor del texto.

REFERENCIAS BIBLIOGRÁFICAS

AELINCO. Asociación Española de Lingüística de Corpus. Recuperado de: http://www.aelinco.es/es

Ainsworth, J., y Juola, P. (2019). *Who Wrote This?: Modern Forensic Authorship Analysis as a Model for Valid Forensic Science*. *Washington University Law Review*, 96(5). Recuperado de https://wustllawreview.org/essays/who-wrote-this-modern-forensic-authorship-analysis-as-a-model-for-valid-forensic-science/

BIODIGITHUM: corpus de escritos autobiográficos en lengua española (1998 - 2003) – Escritos autobiográficos. Recuperado de: http://stel3.ub.edu/biodigithum/queesbiodigithum.php

Boletín de la Unidad de Estudios Biográficos (1998, 3). Recuperado de: https://revistes.ub.edu/index.php/bueb/index

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*, Cambridge: Cambridge University Press.

Cantos, P. (2013). *Statistical methods in language and linguistic research*. Oakville, CT: Equinox.

Chaski, C. E. (2005). Who´s at the Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*. Spring, Vol. 4, Issue 1. Recuperado de https://dblp.org/db/journals/ijde/ijde4.html

Cicres, J., y Gavaldà, N. (2014). La lingüística forense: la llengua com a evidència, *Revista de Llengua I Dret*, núm. 61, p.60-71. Recuperado de http://www.raco.cat/index.php/RLD/article/download/279951/367624

Cicres, J., y Queralt, S. (2019). An n-gram based approach to the automatic classification of schoolchildren's writing. *Vial. Vigo international journal of applied linguistics*, 16, 53-80.

Coulthard, M. (2010). Forensic Linguistics: the application of language description in legal contexts. *Langage et société*, 132(2).

Coulthard, M., y Johnson, A. (2010). *The Routledge Handbook of Forensic Linguistics*. London: Routledge.

David Crystal's Introduction to Language. (2020). Recuperado de: http://cw.routledge.com/textbooks/9780415602679/dc-glossary.asp

Eckert, P. (1998). Age as a sociolinguistic variable. En *The Handbook of Sociolinguistics* (pp. 151-167). Blackwell Publishing.

Eckert, P., y McConnell-Ginet, S. (1999). New generalizations and explanations in language and gender research. *Language in society*, *28*(2), 185-201. Recuperado de https://web.stanford.edu/~eckert/PDF/LinS1999.pdf

El sistema operativo Unix (2001). Recuperado de: https://www2.eii.uva.es/jmzama/material/UNIX-apuntes.pdf

Etiquetas EAGLES. Recuperado de: https://www.cs.upc.edu/~nlp/tools/parole-sp.html

Fitzgerald, James R. (2007). The FBI's Communicated Threat Assessment Database. History, Design and Implementation. *FBI Law Enforcement Bulletin*. Recuperado de https://leb.fbi.gov/file-repository/archives/feb07leb.pdf/view

ForensicLinguistics – International Association of Forensic Linguistics. Recuperado de: https://www.iafl.org/forensic-linguistics/

Garayzábal, E., Jiménez, M., y Reigosa, M. (2014). *Lingüística forense: la lingüística en el ámbito legal y policial.* Madrid: Euphonía Ediciones.

Garayzábal, E., Queralt, S,. y Reigosa, M. (2019) *Fundamentos de la lingüística forense*. Madrid: Síntesis.

Gibbons, J. (1994). *Language and the law*. New York: Routledge

Gibbons, J., y Turell, M. T. (2008). *Dimensions of Forensic Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Herrera-Soler, H., Martínez Arias, R., y Amengual, M. (2011). *Estadística aplicada a la investigación lingüística*. Madrird: EOS Universitaria.

International Association for Forensic Phonetics and Acoustics (IAFL). Recuperado de: https://www.iafpa.net/

Kunanets, N., Levchenko, O., y Hadzalo, A. (2018). The Application of AntConc Concordancer in Linguistic Researches, En*2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, (p. 144-147). Lviv, Ukraine: Institut of Electric and Electronic Engineers-IEEE. Recuperado de https://ieeexplore.ieee.org/document/8526591

Lowie, W. y Seton, B. (2013). *Essential statistics for applied lingüístics*. Basingstoke, NY: Palgrave Macmillan.

McMenamin, G.R. (2010). Forensic stylistics: Theory and practice of forensic stylistics. En *The Routledge Handbook of Forensic Linguistics.* Oxon: Routledge.

Moreno, F. (1996). Metodología del "Proyecto para el Estudio Sociolingüístico del Español de España y de América". *Lingüística*, 8, p. 257-287.

Muchnik-Rozano, Y., y Tsybulsky, D. (2020). Linguistic analysis of science teachers' narratives using AntConc software. En Kennedy, E., y Qian, Y. (eds.), *Advancing Educational Research With Emerging Technology* (pp. 211-230). Hershey, PA: IGI Global. Recuperado de https://www.igi-global.com/chapter/linguistic-analysis-of-science-teachers-narratives-using-antconc-software/240392

Nini, A. (2015). *Authorship Profiling in a Forensic Context*. Tesis doctoral. Aston University, Birmingham, UK. Recuperado de https://niniandrea.files.wordpress.com/2016/12/authorship_profiling_in_a_forensic_conte.pdf

Nini, A. (2019). Developing forensic authorship profiling. *Language and Law/Linguagem e Direito* Vol. 5(2), p. 38-58. Recuperado de http://ojs.letras.up.pt/index.php/LLLD/article/view/6116/5758

Olsson, J. (2008). *Forensic Linguistics: second edition*. London: Continuum.

Pavelec, D., Oliveira, L.S., Justino, E., Nobre Neto, F.D., y Batista, L.V. (2009). Compression and Stylometry for author identification. *Proceedings of the 2009 International Joint Conference of Neural Networks.* Atlanta, GA: Institute of Electric and Electronic Engineers-IEEE, p. 2445-2450. Recuperado de https://ieeexplore.ieee.org/document/5178675

Philbrick, Frederick A. (1949). *Language and the law: the semantics of forensic English*. New York: The Macmilan Company.

Renouf, A. (1987). Lexical Resolution. En Meijis, W. *Corpus Linguistics and Beyond*. Brill/Rodopi; Open Humanities Press; Eiditions Rodopi B.V.; Brill Academic Publishers.

Servei de tecnología lingüística. Recuperdo de: http://stel.ub.edu/el-servei

Sidorov, G. (2019). *Syntactic n-grams in computational linguistics*. Springer International Publishing.

Silva-Corvalán, C. (2001).*Sociolingüística y pragmática del español*. Georgetown: University Press.

Sousa-Silva, R. (2018). Computational forensic linguistics: an overview of computational applications in forensics contexts. *Language and Law / Linguagem e Direito*, Vol. 5(2), p.118-143. Recuperado de http://ojs.letras.up.pt/index.php/LLLD/article/view/6120

Spassova, M.S., y Turell, M.T. (2007). The use of morpho-syntactically annotated tag sequences as markers of authorship. En Turell, M.T., Cicres, J., y Spassova, M. S. (eds.) (2007). *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006*. Barcelona: DOCUMENTA UNIVERSITARIA. págs. 229-237.

Stamatatos, E. (2008). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538-556. Recuperado de https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001

Statistics in Corpus Liguistics: Lancaster Stats Tools online. Recuperado de: http://corpora.lancs.ac.uk/stats/index.php

Svartvik, J. (1968). *The Evans statements: A case for forensic linguistics*. Göteborg: Elander.

Vera, P. (2016). La literatura diarística protagoniza las VI Jornadas de la Fundación Ory. *Diario de Cádiz*. Recuperado de: https://www.diariodecadiz.es/ocio/protagoniza-VI-Jornadas-FundacionOry_0_1072392869.html

Wikipedia. Anna Caballé Masforroll. Recuperado de: https://es.wikipedia.org/wiki/Anna_Caball%C3%A9_Masforroll

UNIVERSITAT DE BARCELONA

Coordinació d'Estudis
Facultat de Filologia

Gran Via
de les Corts Catalanes, 585
08007 Barcelona

Tel. +34 934 035 594
fil-coord@ub.edu
www.ub.edu

Declaració d'autoria

Amb aquest escrit declaro que sóc l'autor/autora original d'aquest treball i que no he emprat per a la seva elaboració cap altra font, incloses fonts d'Internet i altres mitjans electrònics, a part de les indicades. En el treball he assenyalat com a tals totes les citacions, literals o de contingut, que procedeixen d'altres obres. Tinc coneixement que d'altra manera, i segons el que s'indica a l'article 18, del capítol 5 de les Normes reguladores de l'avaluació i de la qualificació dels aprenentatges de la UB, l'avaluació comporta la qualificació de "Suspens".

Barcelona, a 20 de maig de 2020

Signatura: