El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español

Maria Stefanova Spassova

TESI DOCTORAL UPF / 2009

DIRECTOR DE LA TESI

Dra. Maria Teresa Turell Julià

(Institut Universitari de Lingüística Aplicada)

UNIVERSITAT POMPEU FABRA

Dipòsit Legal:
ISBN:

A Barcel y Brontë

Agradecimientos[1]

Elaborar aquesta tesi doctoral ha estat per a mi una empresa similar a la que va haver de dur a terme aquell personatge de la mitologia grega a qui van castigar a empènyer una enorme pedra costa amunt, només per veure-la caure en arribar al cim. A diferència de Sísif, he tingut la sort que cada vegada que no em sentia amb prou forces per continuar endavant, hi ha hagut algú darrere meu que m'empenyia a fer-ho. A aquestes persones els vull donar el meu agraïment més sincer.

A la persona a qui li he d'agrair més és a la meva directora de tesi, la doctora Turell, que m'ha dirigit amb molta paciència i professionalitat durants tots aquests anys, recolzant-me malgrat la meva tossuderia, i donant-me sempre el seu suport.

Al doctor Harald Baayen, que em va "enganxar" a l'estudi dels n-grames durant l'estada que vaig fer a la Radboud Universiteit de Nimega (Països Baixos), que va resultar molt fructífera.

Al doctor Tim Grant, que em va acollir a l'Aston University i em va permetre col·laborar en la seva investigació.

A les meves companyes i companys de l'IULA, especialment del grup UVAL i del ForensicLab, per haver-me ajudat, sobretot en les darreres setmanes de nervis i tensió. Un reconeixement especial per a la Núria i l'Amor, sense les quals probablement mai hagués posat un punt i final a aquesta tesi.

Al Jesús, que sempre ha fet d'intermediari entre els meus problemes informàtics i jo. També per haver aconseguit que aquesta tesi sigui tan bonica, estèticament parlant.

A Miroslav Kyosev, que ha desenvolupat el programa Legolas 2.0 i m'ha estalviat moltíssims anys de treballs manuals en l'extracció de les variables d'anàlisi.

[1] Esta tesis se ha beneficiado de la beca de formación de profesorado universitario (FPU) del Ministerio de Educación.

Al Jordi Vivaldi, que ha atès amb molta amabilitat totes les meves consultes sobre el processament del corpus i sobre molts altres temes.

Al Rogelio Nazar, que m'ha socorregut en els meus intents frustrats per programar.

A la Universitat Pompeu Fabra i a l'IULA, per haver-me acollit i haver-me finançat els primers anys de la meva recerca.

I, finalment, als meus pares, sense el suport dels quals aquesta experiència no hauria passat de ser un somni.

Resumen

El objetivo principal de esta tesis es evaluar el potencial discriminatorio de los n-gramas – esto es, combinaciones de secuencias de categorías gramaticales- como posibles marcas de autoría para los fines de la comparación forense de textos escritos en español. La tesis se centra en dos tipos específicos de n-gramas: los bigramas y los trigramas.

Las principales hipótesis de la tesis son, por un lado, que los n-gramas poseen un potencial discriminatorio alto en el análisis de producciones escritas por diferentes autores (variación inter autor). Por otro lado, que la frecuencia de los n-gramas no varía de forma significativa entre las producciones escritas del mismo individuo en el transcurso del tiempo (variación intra autor).

La evaluación del potencial discriminatorio de los n-gramas se ha llevado a cabo en dos corpus diferentes: a) un corpus general de la lengua española; y b) un corpus de casos forenses reales.

Los resultados han indicado que los dos tipos de n-gramas tienen un potencial discriminatorio alto cuando se aplican a los dos corpus. Además, se ha demostrado que la frecuencia de los n-gramas no varía significativamente entre textos escritos producidos por el mismo autor en un intervalo temporal inferior a 20 años.

Abstract

The main objective of this dissertation is to evaluate the discriminatory capacity of n-grams – i.e. combinations of sequences of parts of speech- as potential markers of authorship for the purposes of the forensic comparison of Spanish written texts. The focus is on two particular types of n-grams, namely bigrams and trigrams.

The principle hypotheses of the present dissertation are, on the one hand, that n-grams have a high potential to discriminate between the written productions of different authors (inter author variation). On the other hand, it is also hypothesized that the frequency of n-grams

does not vary significantly between different writings of the same author over a period of time (intra author variation).

The evaluation of the discriminatory capacity of n-grams was carried out in two different corpora: a) a general corpus of the Spanish language; and b) a corpus of real forensic cases.

Results indicate that both types of n-grams have a high discriminatory potential when applied to both corpora. Moreover, it is demonstrated that the frequency of n-grams does not vary significantly between texts produced by the same author within a time-span of less than 20 years.

Índice

Anexos (en CD-Rom adjunto)

Lista de cuadros

Lista de tablas

Lista de gráficos

Introducción

El presente trabajo ha surgido de la necesidad y el reto que tiene planteados la lingüística forense - en particular, en su rama de la comparación lingüística forense de textos escritos a efectos de hacer el trabajo atribución forense de autoría mucho más fiable - de poder contar con posibles marcas de autoría y técnicas metodológicamente y teóricamente sostenidas que faciliten el trabajo de los peritos lingüistas en la compleja tarea de expresar una opinión experta fiable sobre la probabilidad de que uno o varios documentos anónimos cuya autoría se cuestiona hayan sido producidos o no por la misma persona de la que se sabe con certeza que ha escrito otro conjunto de textos disponibles usados como base de la comparación.

Pese a que la investigación en el área de la comparación forense de textos escritos tiene más de medio siglo de historia, lo cierto es que esta investigación ha dado pocos frutos en cuanto a la definición de una metodología estándar a seguir en el análisis lingüístico forense y la delimitación de unidades de uso de la lengua cuya eficacia y precisión produzca resultados válidos y fiables, independientemente de las características del corpus de textos al que se aplique. En parte la problemática de esta cuestión proviene de la dificultad que entraña la falta de conocimientos sobre las particularidades o las idiosincrasias, en este caso, del estilo idiolectal escrito de toda la población de usuarios de una lengua (Coulthard, 1994: 31, citado por Grant, 2007), o como se ha venido definiendo en la disciplina,

la *referencia de distribución poblacional* (Turell, en prensa) o *base-rate knowledge* (Grant, 2007). Un motivo no de menos peso, sin embargo, es, por un lado, el hecho de que muy pocos estudios se dedican a investigar el comportamiento de la variables candidatas a marca identificativa en cuanto a la variación que se produce en su uso entre diferentes individuos y dentro del mismo individuo bajo el efecto de factores que podrían influir en el resultado final de una pericia basada en dicha marca. Este es un paso muy importante para demostrar su utilidad forense. Por otro, lado las marcas y los métodos que se emplean raras veces se evalúan en un corpus de textos forenses de casos reales, sobre todo por la no existencia o imposibilidad de acceder a recursos de este tipo.

En esta tesis doctoral reprendemos el estudio de las secuencias de categorías gramaticales o n-gramas, iniciado en el proyecto de tesis defendido en 2006, replicándolo con un corpus de análisis más amplio y con nuevos experimentos llevados a cabo. El enfoque central del trabajo recae en la evaluación de los n-gramas tanto en un corpus general de la lengua española objeto de análisis, como en un corpus de casos forenses reales. Esta evaluación comprende también la estimación del efecto de factores sobre todo ligados a las características del texto, como son la extensión, el género y el tiempo de medición en su producción.

En esta breve introducción al tema, en el que profundizaremos a lo largo de los capítulos que siguen, cabe aclarar una cuestión terminológica que puede resultar confusa por la diversidad de

nombres que se han dado a la tarea de comparación lingüística forense en función de su finalidad. Desde los comienzos de la disciplina hasta hace poco se ha hablado de identificación de autor, determinación y atribución de autoría sin hacer distinción alguna entre lo que significa cada uno de estos términos. La diferencia entre ellos radica ante todo en las implicaciones del contexto de análisis de cada caso forense real concreto. La identificación de autor (también especificada como el establecimiento de perfiles lingüísticos) supone la presencia de un único texto de autor desconocido a partir del análisis lingüístico forense del cual se pretende dar pistas sobre quien podría ser su autor real. En atribución y determinación, sin embargo, además de los textos anónimos o dubitados, se dispone de uno o varios textos escritos por el posible o posibles sospechosos. La finalidad en ambos contextos de análisis es establecer que los dos tipos de textos no se han producido de forma independiente; en contextos de determinación de autoría los posibles autores pueden ser más que uno y se debe determinar cuál de ellos es el más probable, mientras que en atribución de autoría, el sospechoso es solo uno y se debe determinar si existen pruebas lingüísticas suficientes para atribuir la autoría del texto o textos dubitados a dicho individuo. En esta tesis se usa el término de atribución de autoría para referirse a ambos contextos, aunque la marca objeto de estudio que se analiza también puede tener aplicación en la determinación de autoría, ya que el procedimiento que se lleva a cabo en el análisis es el mismo. Más recientemente (French y Harrison 2007; Rose y Morrison 2008), se prefiere usar el término más general y más neutro de comparación

lingüística, por ser menos categórico y más adecuado, puesto que es al juez, y no al perito, a quien corresponde emitir sentencia.

Respecto al ámbito de estudio, cabe decir que esta tesis doctoral se enmarca en la lingüística aplicada y en la tarea de peritaje lingüístico forense que se realiza en el ForensciLab, el laboratorio de Lingüística Forense del Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra.

A fin de desarrollar el tema que es objeto de estudio en esta tesis, el volumen que se presenta en esta introducción se estructura en seis capítulos más. El capítulo 1, incluido en la primera parte, trata de la fundamentación teórica del tema objeto de estudio y el capítulo 2 plantea los antecedentes del mismo. En la segunda parte de la tesis, que incluye el estudio empírico, el capítulo 3 presenta la fundamentación metodológica (variables, objetivos, hipótesis, corpus y propuesta analítica); el capítulo 4 desarrolla y discute los resultados de los estudios intra autor; el capítulo 5 presenta y discute los resultados de los estudios inter autor y, finalmente, en el capítulo 6 de conclusiones se sistematizan las diferentes conclusiones a las que los resultados de los diferentes experimentos y estudios han permitido llegar.

1. FUNDAMENTACIÓN TEÓRICA

La atribución de autoría ha recorrido un largo camino desde la época en que clérigos y estudiosos literarios practicaban el análisis lingüístico forense con fines puramente exploratorios sobre textos bíblicos e históricos. En la actualidad, los lingüistas forenses aplican sus conocimientos científicos sobre la lengua para tratar pruebas orales y escritas y actuar como expertos en pleitos civiles y criminales. El estatus actual de esta disciplina requiere el desarrollo de una metodología respaldada por teorías lingüísticas para asegurar, por un lado, el reconocimiento científico de su práctica ante la comunidad científica y los organismos responsables de la administración de la justicia y, por otro, la fiabilidad de los informes periciales que se presentan en cada caso.

La teoría de la variación lingüística (Labov, 1972) es la que aporta la base metodológica y conceptual que precisa la atribución de autoría para crear su propio marco teórico. En sus postulados se fundamentan las tres premisas principales que los lingüistas forenses asumen como válidas al abordar los problemas de autoría discutida o dubitada. Dichas premisas se pueden resumir de la siguiente manera:

Premisa 1. Cada usuario nativo de una lengua tiene su propia versión individual e distintiva de la lengua que habla y escribe, su idiolecto (Coulthard, 2004)

Premisa 2. El idiolecto se manifiesta en las selecciones idiosincrásicas que hace el individuo en los distintos niveles lingüísticos al producir un texto: fonético, léxico, sintáctico, semántico. (Halliday et al., 1964: 75)

Premisa 3. El individuo es constante en sus selecciones y estas pueden ser detectadas tanto en sus producciones orales como escritas.

La argumentación teórica de estas premisas, que representan el punto de partida de esta tesis, constituyen el marco teórico de la investigación. La explicación de dichas premisas, que presentamos a continuación, se hará siguiendo fundamentalmente la teoría de la variación lingüística, junto con otras disciplinas lingüísticas.

1.1 Teoría de la variación lingüística

Uno de los postulados claves de la teoría de la variación lingüística, en la que se apoya esta tesis para su justificación teórica, está vinculado con el estudio del papel de los factores sociales y socioculturales que propician la existencia de la variabilidad de sistemas y estructuras lingüísticas y la diversidad de su uso en el proceso de la comunicación. Los conceptos teóricos y metodológicos de la teoría variacionista surgen en respuesta a la necesidad de una revisión de los postulados tradicionales que se formulan en la lingüística estructural y la lingüística generativa

sobre el carácter de la variación lingüística y la evolución del lenguaje (Labov, 1963; Hymes, 1972).

En los tratados estructuralistas (Saussure, 1990) el lenguaje se representa como un sistema de elementos lingüísticos destinados a ocupar determinadas posiciones en la cadena que forman sus componentes y a entrar en determinadas relaciones internas excluyendo otras posibles combinaciones. El número de los elementos y las variantes de combinación es fijo e inalterable en la estructura de la lengua. En virtud de estas premisas, la variación en la lingüística estructural se explica como consecuencia contextual del significado expresado en el uso concreto del lenguaje.

Los generativistas, a su vez, relacionan el fenómeno de la variación lingüística con la competencia gramatical del hablante. Según la Gramática Generativa Transformacional (GGT), conocida también como la Teoría Estándar de Chomsky (1965), esta competencia se manifiesta en la aplicación de determinadas reglas y principios que integran la gramática interna de cada individuo y que le permiten crear un ilimitado número de oraciones gramaticalmente correctas. Estas reglas, que determinan las posibles opciones combinatorias y el número de constituyentes de la oración que se produce tras su activación en el proceso mental de la producción lingüística (reglas sintagmáticas) y que construyen el enunciado en su forma oral o escrita (reglas transformacionales), son categóricas. De ahí que, desde la perspectiva del generativismo, el lenguaje sea concebido como el resultado del efecto de factores internos que rigen la

selección de una regla frente a otra, hecho que explicaría la variación lingüística.

En contraste con estos dos modelos teóricos elaborados por las corrientes estructuralistas y generativistas para atender las cuestiones de lengua y variación, la vertiente variacionista basa su teoría en el concepto de la lengua concebida como un producto de la interacción social (Labov, 1963). La premisa principal de la que parten sus postulados implica, en palabras de Turell:

> [...] no només que l'estructura està configurada pels universals de la cognició, de la memòria i de la lògica humanes, sinó també que una llengua és estructuralment marcada pel seu ús a la societat, o dit d'una altra manera, que és un producte del món social que ens envolta.
>
> (Turell, 1995: 18)

La variación es una propiedad inherente a todas las lenguas (Labov, 1983), que se rige no sólo por factores internos estrictamente lingüísticos —tal como dictan los tratados estructuralistas y generativistas—, sino también por factores externos (sociales, estilísticos, geográficos, etc.), que actúan de forma independiente:

> [...] si es prescindeix de o es canvia un factor intern, els canvis de comportament apareixen en els altres factors interns, però els factors externs no canvien; si es prescindeix

> de o es canvia un factor extern, els altres factors externs canvien, però els interns es mantenen igual.
>
> (Turell, 2003: 6)

Los factores externos son de carácter variable. Por consiguiente, las reglas generativistas que mencionábamos antes comparten y están restringidas por ellos. Eso quiere decir que la ocurrencia de una regla (llamada variante) u otra, o bien la ocurrencia de una opción de realización de la regla (llamada covariante) u otra, no se da a causa de una selección aleatoria, sino que está condicionada por factores externos. La definición del concepto de variación lingüística como un fenómeno inherente multilingüístico que se realiza de forma libre, pero estructurada, que sostiene la escuela laboviana, máxima representante de la corriente variacionista, es la que se adopta como una de las premisas clave en lingüística forense.

La piedra angular de la investigación en atribución de autoría es la presencia de factores externos que, según la teoría variacionista, propician la variación lingüística a diferentes niveles entre las comunidades de hablantes de una lengua. El análisis cualitativo y cuantitativo de la variación es lo que permite la distinción entre los miembros de dos comunidades diferentes que no comparten los mismos rasgos lingüísticos.

En lo que atañe a los fines de la disciplina cabe destacar sobre todo la heterogeneidad en el uso del lenguaje que se observa a nivel geográfico, social e individual. La variación lingüística a nivel

geográfico se manifiesta en el hecho de que cada región o comunidad lingüística ha desarrollado sus propias variantes de la lengua (dialectos, subdialectos, versiones de subdialectos). A nivel social, la variación lingüística suele estar motivada por factores externos como el género, la edad, el grupo étnico y el nivel educativo de la comunidad lingüística, que influyen en el uso de la lengua propio de cada grupo dentro de la comunidad y también en el del individuo como miembro de este grupo. La variación lingüística individual, por último, tiene que ver con la fisiología del hablante, su ideología y sus hábitos lingüísticos, que se plasman en su lenguaje individual o idiolecto.

1.2 El lenguaje individual

En el panorama de las corrientes lingüísticas estructuralista y funcionalista que tratan el lenguaje en su aspecto social y funcional, el idiolecto del usuario de una lengua ocupa un lugar central. Esta concentración del interés en el individuo por parte de los sociolingüistas y los variacionistas se debe al hecho de que el lenguaje individual o idiolecto refleja las características específicas de la comunidad o grupo lingüístico al que pertenece su usuario. LePage (1968: 192) sugiere que la representividad que se atribuye al idiolecto podría explicarse como consecuencia del afán del hablante de una lengua de imitar y asimilar el comportamiento verbal de una comunidad con la que desea identificarse.

Siguiendo esta línea de interpretación, podemos suponer que el individuo "equipa" su repertorio lingüístico con las producciones lingüísticas —sobre todo orales, pero también escritas— a las que ha sido expuesto en la interacción comunicativa con los demás usuarios de la lengua de su entorno para luego reproducirlas en sus propios enunciados. De ahí que el análisis del habla de uno o varios individuos pertenecientes a la comunidad objeto de estudio, en el que se detectan y se escrutan los rasgos idiosincrásicos de su lenguaje, hace viable la descripción de la variante lingüística (dialecto, subdialecto, etc.) de la comunidad. La representividad del idiolecto es una de las características que lo convierten en concepto transcendental en la incipiente teoría lingüística forense. Si podemos inducir conclusiones sobre un colectivo lingüístico, es lógico que el método inverso nos permita identificar un individuo como miembro de este colectivo a partir de nuestros conocimientos sobre la variante lingüística del grupo. Por esta razón, la fonética forense y la atribución de autoría adoptan la metodología establecida por los sociolingüistas dialectólogos y la aplican en el análisis lingüístico forense (acústico, léxico o sintáctico) de grabaciones o textos escritos anónimos, con el fin de crear perfiles lingüísticos de identificación o de discriminar entre posibles autores[2].

[2] El capítulo 5 de esta tesis trata sobre la atribución de autoría usando la variedad lingüística como criterio de clasificación, por lo que remitimos a su lectura para una exposición más detallada del tema. Para una introducción al mismo tema en fonética forense, véase Cicres (2007).

De acuerdo con lo que hemos comentado hasta aquí, el lenguaje individual es el resultado del contacto del individuo con el resto de miembros de su entorno lingüístico a lo largo de su vida; es el conocimiento adquirido en la interacción social. Pero además y sin dejar de ser un fenómeno originado por la sociedad, el lenguaje es también "una cosa social manejada individualmente" (Cohen, 1973: 271):

> El lenguaje suministrado al individuo por el grupo social, es empleado por aquél con más o menos particularidades, en primer lugar su propia voz, que le hace reconocible, igual que sus rasgos y su aspecto general.
>
> (Cohen, 1973: 59)

Las particularidades de las que habla Cohen y, en las que estriba la noción de la unicidad del lenguaje individual, son causadas por aspectos tanto físicos (la constitución del aparato articulatorio, por ejemplo, en el caso de la voz), como culturales (las creencias personales y la experiencia vital lingüística) del hablante. Ambos aspectos influyen en el idiolecto de cada individuo en distintos grados, de modo que no existen dos personas que compartan los mismos rasgos idiolectales en su manera de hablar o escribir. Aunque el lenguaje del grupo o comunidad lingüística del que forma parte el individuo determina su idiolecto, la superposición de rasgos dialectales es demasiado fragmentaria como para eliminar aquellos que son estrictamente individuales (McMenamin, 1993: 52). Cada usuario nativo de una lengua posee una versión del

lenguaje única y diferente a la del resto de hablantes de la misma lengua (Coulthard, 2007a: 161).

Por lo que los lingüistas han podido averiguar mediante numerosos estudios sincrónicos y diacrónicos, la lengua es una materia abstracta en estado de cambio constante y la única muestra y prueba de su existencia es el lenguaje individual. El idiolecto, a diferencia de la lengua viva, es susceptible de cuantificación, y su análisis ha permitido desde antaño llegar a conclusiones y formular suposiciones respecto su carácter general y los procesos de adquisición y producción del lenguaje.

Aparte de servir a las finalidades descriptivas de la lengua, algunas disciplinas abren nuevas vías de investigación en las que el estudio del idiolecto encuentra otras aplicaciones. En psicología forense, por ejemplo, el idiolecto se toma como una forma original de expresion de la individualidad y su análisis puede aportar información sobre el estado ánimico, emocional e incluso mental de una persona (perfil psicológico). Por su lado la lingüística forense en su rama de autoría estudia el idiolecto desde la perspectiva de su aplicación en la resolución de cuestiones de atribución de autoría, identificación de autor y elaboración de perfíles lingüísticos de posibles autores de un texto anónimo o cuya autoría se discute. En su trabajo en este ámbito los lingüistas forenses sostienen el supuesto de que la unicidad lingüística de cada idiolecto se manifiesta mediante elecciones distintivas e idiosincrásicas del repertorio de la lengua en la producción del habla y de la escritura.

Este supuesto, que tiene su respaldo teórico en la variación lingüística, implica que un individuo puede ser identificado a partir de su voz y de sus escritos (Coulthard, 2007a: 161).

El análisis lingüístico forense que hemos llevado a cabo en la investigación de esta tesis usa las producciones lingüísticas escritas para detectar, analizar y clasificar las estructuras sintácticas más frecuentes en los idiolectos de una serie de escritores hispanohablantes que pueden ser usadas como marcas de autoría en la lengua española escrita[3].

Este objetivo hace patente la necesidad de dedicar unas líneas de esta tesis doctoral a la lengua escrita. En los próximos apartados abordaremos, por un lado, el tema de la escritura desde diversas perspectivas, que permiten entender su naturaleza en contraste con el habla y, por otro, comentaremos los procesos adquisitivos y psicolingüísticos de la lengua que explican la unicidad y la idiosincrasia de las producciones lingüísticas escritas.

1.3 Lengua escrita y lengua oral

En términos generales, la lengua hablada y la lengua escrita se pueden definir en lingüística como las dos realizaciones factibles del código lingüístico que se transmiten a través del lenguaje. La relación que se establece entre los dos sistemas, no obstante, ha

[3] Para una descripción más exhaustiva de la metodología de análisis aplicada en la tesis consulte el capítulo 3.

dado pie a una ardua discusión entre los lingüistas, que mantienen posiciones opuestas en cuanto a su carácter. La aproximaciones que se adoptan en la investigación del problema son tres y atribuyen la primogeneidad a uno u otro sistema aportando diferentes argumentos de justificación. La aproximación más extendida entre los lingüistas sostiene que la escritura depende del habla. La lengua escrita, conforme lo que sostienen los partidarios de esta posición, no es más que una mera representación gráfica de la lengua hablada que adquiere su significado vía el componente fonológico del lenguaje (Scinto, 1986: 31). En contraste con esta concepción fonocéntrica, la escuela glosemática de Hjelmslev (1974) defiende la posición totalmente opuesta de que no existe ningún tipo de relación entre el habla y la escritura, sino que se trata de dos sistemas lingüísticos que funcionan de manera independiente, como dos manifestaciones aisladas del código lingüístico subyacente. La tercera y última aproximación adoptada por los lingüistas del Círculo de Praga, en la que se apoya esta tesis, sostiene la interdependencia de lo escrito y lo oral. Esta aproximación concibe los dos sistemas de habla y escritura en un estado permanente de vínculo simbiótico de mutua dependencia en el que la lengua escrita deriva de la lengua hablada y depende de ella en su evolución y uso continuo. Habla y escritura están ligadas por su estructura y cumplen fines complementarios, pero independientes, en la comunidad lingüística (Scinto, 1986: 32).

La conexión estructural, comenta Lentin (1996: 146), se explica por el hecho de que tanto la lengua hablada como la escrita están

regidas por el mismo sistema sintáctico contenido en el código lingüístico. Cada una representa un conjunto de variantes enunciativas correctas y posibles de realización en una lengua que el individuo hablante usa en sus enunciaciones orales y escritas de acuerdo con las necesidades comunicativas y los contextos discursivos. En algunos contextos se puede dar la "intersección" de los dos sistemas escrito y oral. Cuando esto ocurre las enunciaciones que se encuentran en el espacio hipotético de intersección pueden desempeñar la función de formas orales y escritas. Es decir, el individuo puede emplear ciertas locuciones propias del discurso escrito en su habla y viceversa.

La descripción de Lentin del binomio oralidad y escritura alude a una relación entre lo hablado y lo escrito vistos desde una perspectiva de variación y no de oposición. Biber (1991) realiza un extenso trabajo empírico sobre el carácter de esta variación y los posibles rasgos lingüísticos (léxicos y gramaticales) que permiten establecer la diferencia entre lengua hablada y lengua escrita en inglés[4] basándose en criterios formales y funcionales[5]. La conclusión que extrae es la siguiente:

> There's no absolute difference between speech and writing in English; rather there are several dimensions of variation,

[4] Biber repite el mismo procedimiento aplicado al inglés en otras lenguas en un estudio miltidimensional posterior. (1995)
[5] Criterios relacionados, por un lado, con las características lingüísticas de los rasgos y, por otro, con la función comunicativa que cumplen.

> and particular types of speech and writing are more or less similar with respect to each dimension.
>
> Biber (1991:199)

Las dimensiones de las que habla Biber representan los parámetros que, según el investigador, son explicativos de la variación. Cada dimensión se relaciona con un grupo de rasgos lingüísticos y con determinado género y rasgos contextuales. No obstante, no hay muestra aparente de que exista un único rasgo que pueda evidenciar de manera global la distinción entre oralidad y escritura (Castellà, 2004: 32).

Los comentarios de Halliday respecto a la oposición errónea de lengua escrita y lengua hablada, apuntan a la complejidad lingüística como un rasgo básico que podría servir para diferenciarlas. Oral y escrito representan dos modos del discurso en un continuo, que tienen tipos de complejidad diferentes que se manifiestan de manera propia en cada modo (Halliday, 1987). Así, el modo oral está marcado por su complejidad gramatical, mientras que el modo escrito se caracteriza por su alta densidad léxica (Halliday, 1979: 49).

La preferencias por patrones de organización léxicos y de organización gramatical son también característicos del modo escrito y oral, respectivamente, según afirma Castellà (2004) para el catalán, después de estudiar con gran detalle la complejidad lingüística de la oralidad y la escritura de esta lengua. Aunque

nosotros trabajamos con la lengua española, creemos que podemos generalizar las conclusiones a las que llega Castellà, ya que se trata de lenguas románicas que comparten muchos rasgos. He aquí, de manera muy resumida, algunas de las conclusiones referentes a la propiedades características de la lengua escrita en contraste con la hablada a las que llega Castellà, y que nos son útiles para nuestros fines:

- mayor variación léxica
- uso predominante de nombres y complementos nominales
- cláusulas largas compuestas por estructuras sintagmáticas más complejas
- uso preferente de la subordinación, etc.

Conocer estas diferencias de la lengua escrita respecto a la oral facilita la búsqueda de posibles marcas de identificación de autor en la investigación en el campo de la atribución de autoría forense puesto que permite centrar la identificación y selección de rasgos idiosincrásicos de la escritura en aquellas unidades que podrían serlo por su grado, mayor o menor, de recurrencia[6].

1.4 Unicidad e idiosincrasia de las producciones lingüísticas

La hipótesis de partida, que esperamos confirmar con nuestra investigación en el campo de la atribución de autoría, es que las

[6] Los criterios de selección de marcas identificativas se describen en el capítulo 2.

producciones escritas de cada individuo son únicas e irrepetibles. Esto quiere decir que a pesar de que puede haber ciertas similitudes entre los textos creados por distintos individuos debido a que estos pertenecen al mismo género textual, o usuarios de la misma variante lingüística (véase capítulo 5), o incluso, en casos de imitación el estilo, nadie es capaz de reproducir de manera exacta los rasgos idiosincrásicos del idiolecto de un individuo haciendo un uso idéntico de su lenguaje.

Para corroborar esta hipótesis nos amparamos, en primer lugar, en la teoría de la variación lingüística, cuyos conceptos ya hemos detallado en el punto 1.1 de este capítulo; y, en segundo lugar, en los principios de adquisición de la lengua y de la escritura y los modelos teóricos de la psicolingüística que explican los procesos mentales que están detrás de la producción lingüística que explicaremos a continuación.

a) La unicidad y la idiosincrasia lingüística desde la perspectiva de las teorías de adquisición de primeras lenguas

Desde una edad muy temprana la mayoría de los niños, siempre que no padezcan ningún tipo de minusvalía física que se lo impida, muestran una facilidad sorprendente de aprendizaje de la lengua materna, avanzando rápidamente en el transcurso de pocos años hasta alcanzar la capacidad adulta. Aunque esta capacidad para el

lenguaje así como su desarrollo siguen siendo un enigma, no escasean las teorías sobre la adquisición del lenguaje. Aquí prestaremos atención a las que, entre la gran variedad de teorías elaboradas para explicar el fenómeno de la adquisición lingüística, permiten reforzar la hipótesis del carácter idiosincrásico del proceso de aprendizaje y de la capacidad lingüística del usuario de la lengua, que se puede observar a través de sus manifestaciones orales y escritas.

En primer lugar, la teoría innatista, que tiene su origen en la teoría chomskiana de la Rección y el Ligamiento (GBT)*[7], y defiende el carácter innato de la capacidad lingüística. Conforme con lo postulado por esta teoría, el proceso de adquisición del lenguaje es un proceso interactivo en que los principios de la Gramática Universal, que constituyen el conocimiento lingüistico innato del individuo, interactúan con el conocimiento que le proporciona su entorno lingüístico, y dicha interacción da como resultado la formación de una particular gramática del individuo (Nyams, 1986). Según Chomsky (1965), esta interacción entre lo conocido, es decir, el conocimiento innato, y lo nuevo, el imput lingüístico externo, se maneja mediante un mecanismo mental subyacente, llamado ***dispositivo de adquisición lingüística***, que "se dispara" y activa el proceso de adquisición en el momento del primer contacto del niño con la lengua. Gracias a este dispositivo un niño es capaz, incluso

[7] Recordemos del capítulo 1.1 que la teoría generativista defiende la existencia de una gramática universal que contiene una serie de principios y parámetros válidos para todas las lenguas. Cada principio está asociado con un conjunto de valores que marcan los límites dentro de los cuales la gramática puede variar con respecto a dicho principio.

cuando el input lingüistico que recibe es imperfecto, de derivar las reglas gramaticales adecuadas que le permiten crear oraciones bien estructuradas en la lengua diana[8] y determinar la manera en la que deben usarse y entenderse (Berko Gleason y Bernstein-Ratner, 2001: 407).

La relevancia de esta teoría en esta tesis es fundamental porque contempla la adquisición de la capacidad sintáctica. Y saber cómo el individuo aprende a usar la sintaxis nos permitirá demostrar el carácter idiosincrásico de las secuencias de categorías gramaticales[9], objeto de estudio de nuestro trabajo. Si la teoría de Chomsky sobre la adquisición del lenguaje fuera cierta, ya que por ahora no se han encontrado muestras físicas que lo demuestren, y aceptáramos que todos los seres humanos nacen ya "programados" para el lenguaje y el individuo solo ha de "configurar" el programa mental para que su funcionamiento sea el óptimo, hemos de suponer que la configuración, al igual que su rendimiento, serán distintas en cada persona, por sus diferencias individuales intrínsecas.

En segundo lugar, la teoría sociocultural de la adquisición del lenguaje habla de un proceso de aprendizaje en el que el niño desarrolla su destreza lingüística a través de la interacción con la gente de su entorno, y no solo empleando su capacidad innata de descodificar y hacer propia la gramática del lenguaje adulto (Bruner, 1985). Según los teóricos interaccionistas la medición

[8] Término que se usa en adquisición lingüística para designar la lengua que se está aprendiendo. En inglés *target language.*
[9] Las variables de análisis se describen en la sección 1 del capítulo 4.

lingüística de los padres y las demás personas adultas cercanas al niño tiene un efecto crucial en la fase temprana del proceso de adquisición del lenguaje. Dichos estudios sugieren que el modo de hablar que usan los adultos en sus interacciones lingüísticas con los niños les facilita la descodificación y por lo consiguiente el aprendizaje de la lengua materna (Hirsh-Pasek et al., 1987). Además, los padres tienden a corregir las expresiones agramaticales que producen sus hijos estimulando así su adquisición lingüística (Bernstein Ratner, 1987; 1993). El enfoque de la teoría sociocultural, a diferencia de otras teorías, por ejemplo la ya citada teoría innatista, se centra más en los aspectos pragmáticos de la adquisición del lenguaje y concibe la capacidad para el aprendizaje de la lengua materna como una habilidad que la sociedad facilita al individuo mediante el contacto social entre hablantes y la interacción comunicativa. En este sentido, esta teoría comparte la postura de la corriente variacionista en cuanto a la naturaleza del lenguaje como un producto social que viene marcado por los rasgos propios de la comunidad de hablantes y por su portador como usuario de la lengua.

Según estas dos teorías, el proceso de adquisición, que consiste en el desarrollo de la capacidad de la producción lingüística, progresa bajo el efecto de los estímulos externos que percibe el mecanismo de aprendizaje innato en cada ser humano. Estos estímulos provocan a su vez que los individuos alcancen diferentes grados de competencia y destreza lingüísticas, que se pueden percibir en la

riqueza léxica y la creatividad sintáctica de sus enunciados. A este hecho se debe, en gran parte, la idiosincrasia lingüística.

Aprender una lengua dada siempre se relaciona en primer lugar con la adquisición de la capacidad de hablarla, o dicho de otro modo, con el desarrollo de la habilidad de expresar ideas mediante la combinación de palabras en oraciones complejas, contextualmente adecuadas y gramaticalmente correctas, usando la propia voz. Aunque en las sociedades modernas con un nivel avanzado de alfabetización las habilidades de leer y escribir se conciben como dos aspectos de la competencia lingüística de igual importancia, persiste la idea de la primacía del habla en el proceso de aprendizaje. Esto se debe, en primer lugar, a que la comunicación por excelencia es oral, y en segundo lugar, a que la primera forma de expresión y comunicación que conoce y adopta el ser humano de manera inconsciente es también la oral. La escritura se va introduciendo en el individuo en una etapa más tardía y, a diferencia de la lengua oral, su adquisición no es un proceso involuntario sino producto de una enseñanza específica (Perera, 1986: 494).

Al contrario de lo que ocurre en la competencia oral, la capacidad de saber escribir no se adquiere úncamente a través de la experiencia del individuo. Pero aunque su adquisición viene guiada y resulta de la enseñanza, en su proceso de adquisición también influyen factores que afectan al desarrollo de un estilo idiosincrásico de escribir. Entre los factores que se interponen en la

formación del idiolecto escrito cabe resaltar el efecto de la lengua hablada, el de la lectura y el de la enseñanza escolar.

Para algunos autores (Kroll, 1981; Garton y Pratt, 1989; Kress, 1994), el aprendizaje de la escritura es un proceso de transición de la lengua oral a la lengua escrita que transcurre en el hablante en distintas fases. A grandes rasgos, estas fases comprenden el período desde los pasos iniciales en la lengua escrita, cuando se aprenden los aspectos puramente físicos y gradualmente se empiezan a producir los primeros textos que copian la estructura de la lengua hablada, hasta llegar a nivel de competencia en que se singularizan los dos modos, oral y escrito, con su particular forma de organización sintáctica. La última fase de la transición, en la que se puede hablar de una madurez gramatical en la expresión tanto oral como escrita, suele ser alcanzada por una minoría de escritores. Según un estudio realizado sobre la lengua inglesa por Cantor y Rubin (1981), en muchas personas que finalizan sus estudios secundarios la diferenciación entre habla y escritura no se completa del todo y su escritos pueden contener locuciones y construcciones características de su modo de hablar. Esto puede ocurrir incluso cuando se trata de individuos que han avanzado más en su formación como escritores, si se encuentran en estado emocional de alteración o circunstancias similares.

Al mismo tiempo la escritura está afectada por otro factor. Se trata del factor de la lectura. Perera (1986: 516) señala que el individuo aprende la gramática característica de la lengua escrita por medio de

la lectura, y no del habla, ya que las personas que escriben mejor suelen ser las que confirman leer más (Stotsky, 1983). Las lecturas obligatorias en la época de estudio y las de pasatiempos que entretienen el individuo en las etapas tempranas del aprendizaje y a lo largo de su vida influyen en su escritura. Tampoco es de menospreciar el efecto de la educación escolar, que modifica el idiolecto escrito a resultas del empeño de los profesores de inculcar la escritura correcta por medio de estrictas reglas de redacción.

Según el estudio de Gundlach (1981) realizado sobre el tema, a lo largo del proceso de adquisición de la escritura, bajo el efecto de los factores que hemos destacado, el individuo desarrolla ciertas preferencias a determinadas estructuras oracionales y construcciones de su lexicón sintáctico y, en consecuencia, tiende a incurrir en su uso con frecuencia y de modo idiosincrásico. Esta tendencia se contempla también a nivel léxico en el empleo de algunas palabras y formas verbales del vocabulario del individuo.

De todo lo comentado en cuanto a la idiosincrasia de las producciones lingüísticas, vista desde la perspectiva teórica de los estudios de adquisición de primeras lenguas, concluimos que el proceso de aprendizaje es un proceso de individualización. Es así tanto en lo que se refiere al habla como a la escritura ya que en el desarrollo de ambas habilidades lingüísticas influyen diversos factores que condicionan el hecho de que cada individuo alcance distintos grados de madurez en su escritos o en su expresión oral y que emplee habla y escritura en un modo particular e idiosincrásico.

b) La unicidad y la idiosincrasia lingüística desde la perspectiva de los modelos de producción lingüística

La idiosincrasia de las producciones lingüísticas de un individuo puede explicarse también con los modelos teóricos de la psicolingüística (Garett, 1975; Bock, 1982; Levelt, 1993; Bock y Levelt, 1994). Elaborados a partir del análisis de datos sobre errores del habla, estos modelos ofrecen diferentes descripciones del procesamiento psicolingüístico que tiene lugar en la transformación de una idea conceptual en enunciado, pero todos ellos coinciden en distinguir entre dos estadios distintos en los que transcurre su codificación gramatical.

La codificación gramatical, según la definen Hartsuiker et al. (1999: 129), consiste en traducir la representación conceptual del enunciado que está por producirse en una secuencia bien organizada y jerárquicamente estructurada de constituyentes. En el primer estadio, conocido como estadio funcional, se recuperan los elementos léxicos y se les asigna su categoría sintáctica y, en el segundo, a partir de la información recabada en el primer estadio, se formula la estructura de la oración. Durante el proceso en que transcurren estos dos estadios, el hablante debe elegir entre diferentes unidades léxicas y estructuras sintácticas y entre las distintas variantes de orden oracional que permite la lengua y que están almacenadas en su lexicón.

Según Halliday (1985), las idiosincrasias en la actividad lingüística del usuario se relacionan con esta opcionalidad de realización lingüística de una idea o mensaje. Por tanto, la producción de enunciados está ligada a la constante selección de elementos oracionales entre diversas opciones pragmáticamente sinónimas del inventario lingüístico del usuario de la lengua. Por un lado, cuando la situación comunicativa, es decir, el momento concreto de formulación del código emitido, rige la elección, el individuo tiende a concurrir en las mismas opciones. Por otro lado, el carácter idiosincrásico de las construcciones y estructuras sintácticas, de interés en esta tesis doctoral, se puede explicar a partir del fenómeno de la **facilitación sintáctica** [*syntacic priming*], también llamado **persistencia sintáctica** [*syntactic persistence*]. Este fenómeno es propio tanto de la producción oral como de la escrita y se manifiesta en la tendencia a recurrir al uso de la misma estructura en frases consecutivas.

Los estudios de Bock (1986) y Bock y Loebell (1990) indican que la persistencia sintáctica no se debe a la repetición del mismo léxico o de la correspondencia temática entre la frase facilitada y la frase diana, sino a la estructura misma de la frase que ha sido producida con anterioridad. La reiteración de una estructura puede estar asociada a la realización cognitiva de las reglas sintácticas por medio de procesos sintácticos que se activan en la codificación gramatical. Una vez activado este proceso, en la formulación de una frase su efecto puede persistir y así aumentar la posibilidad de repetir la misma estructura en la producción de frases consecutivas

(Bock, 1986; Hartsuiker, 1999). Dado que cada proceso corresponde a una de las varias opciones posibles dentro de la lengua para transformar un mensaje en una estructura sintáctica, el proceso se mantendrá activo, —es decir, el individuo continuará empleando la misma estructura—, en todos los casos en que esta sea aceptable, y sin que intercalar estructuras de otro tipo influya en la persistencia lingüística[10], hasta que use una estructura alternativa (Branigan et al., 1999).

La relevancia de este fenómeno psicolingüístico en atribución de autoría radica en el hecho de que argumenta la hipótesis de trabajo según la cual en la producción lingüística se puede dar el uso frecuente y marcado de una determinada combinación de elementos sintácticos y objetiva la tarea de detección de rasgos idiosincrásicos de esta índole.

En este capítulo hemos resumido los conceptos e ideas que proponen diferentes teorías lingüísticas como la variación, la adquisición del lenguaje y la psicolingüística, y que pueden aplicarse en lingüística forense y, concretamente, en la atribución forense de autoría. En primer lugar, la teoría de la variación lingüística que postula la existencia del idiolecto; en segundo lugar, las teorías innatista y sociocultural que sugieren que el proceso de adquisición de la lengua es un proceso de individualización

[10] La facilitación sintáctica puede perdurar en la producción lingüística del hablante o el escritor incluso cuando se da la aparición de estructuras de otro tipo entre la frase facilitada y la frase diana, según demuestra el estudio de Bock y Griffin (2000).

lingüística, y por último, los modelos teóricos de la psicolingüística que prueban la reiteración de estructuras lingüísticas en el habla y en la escritura de los individuos.

2. ANTECEDENTES

2.1 La base conceptual y el estado de la cuestión en atribución de autoría

El tema de la atribución de autoría de textos dubitados a partir de marcas de identificación sintácticas es un tema incipiente y apenas explorado en otras lenguas, sobre todo en inglés, y sin precedentes para el español. Asimismo es un tema de gran futuro en la Lingüística Forense por lo revolucionario e innovador que puede resultar el descubrimiento de marcas independientes de la pragmática del uso contextual. En los apartados siguientes definiremos los conceptos fundamentales de la línea de investigación sobre atribución de autoría y las pautas de trabajo, y expondremos en breve los estudios que consideramos más relevantes en este campo de cara a la presente tesis doctoral.

a) Conceptos fundamentales en atribución de autoría

La atribución de autoría es la rama de la Lingüística Forense que emplea el conocimiento lingüístico en el análisis de textos que constituyen pruebas lingüísticas con el fin de determinar si su autoría se puede atribuir o no a un sospechoso concreto en la

investigación de un crimen. Esta es la definición que mejor describe el área de estudio en su faceta actual aunque en el aspecto práctico nada ha cambiado desde que su objeto de análisis ha pasado, con los años, de los textos literarios y religiosos de autor anónimo a las pruebas lingüísticas.

La técnica clave en la metodología de trabajo sigue siendo la comparación de los escritos cuya autoría se cuestiona (**textos dubitados** [*disputed texts*]) con otros cuya autoría no se cuestiona (**textos indubitados** [*undisputed texts*]). Comparar, sin embargo, es el paso final de un largo escrutinio del contenido léxico y sintáctico de los textos en el que se mide, calcula y clasifica cada unidad para encontrar posibles rasgos estilísticos distintivos de autor, es decir, **marcas identificativas**[11].

El análisis de atribución de autoría forense está sujeto a determinadas condiciones que tienen que ser observadas para asegurar la validez[12] de los resultados. Bailey (1979: 7) las especifica de la forma siguiente:

> In my view, there are at least three rules that define the circumstances necessary for forensic authorship attribution:
>
> 1. that the number of putative authors constitute a well-defined set;

[11] Para una definición exhaustiva de este concepto, consúltese el apartado 2.2
[12] Véase p.37 para la definición del término

2. that there be a sufficient quantity of attested and disputed samples to reflect the linguistic habits of each candidate;
3. that the compared texts be commensurable.

En el mismo artículo, Bailey (1979: 9) no sólo hace hincapié en las condiciones necesarias para la atribución correcta, sino que también reseña los parámetros que deben tener los textos indubitados para que el efecto de la variación estilística no influya en el análisis forense:

> An autor of personal documents (such as diary or journal) may well adopt a style markedly different from that used in writings for audience for others. Similarly, the same subject matter may call forth different styles on different occasions, while distinct registers may variously encourage or inhibit the personal mark of style. For these reasons, the documents available to establish the style of autor-candidates must resemble the disputed text as much as possible, not only in mode but in audience, register purpose, and time of composition as well.

Desafortunadamente, cumplir con estos requisitos no siempre es posible. Muchas veces el lingüista forense recibe un corpus de comparación de documentos que no comparten las características del documento o del conjunto de documentos dubitados que especifica Bailey. En tales ocasiones, si se cuenta con un número y

una extensión de los textos aceptable, que posibilita el cotejo, la realización del peritaje lingüístico forense depende de la presencia de marcas identificativas.

Por lo tanto, otra condición muy importante en atribución de autoría es que las marcas identificativas se manifiesten en todas las producciones escritas del sujeto tanto a nivel de **variación inter autor** como a nivel de **variación intra autor**[13]. Dicho en otras palabras, para que la identificación de autor sea factible, estos rasgos distintivos deben ser detectables en el estilo del autor aún cuando éste tiende a variar en su manera de escribir de un documento a otro (Chaski 1997: 17).[14]

Sin embargo, el mero hecho de constatar la presencia o la ausencia de una posible marca de identificación no aporta nada al análisis, si no se la estudia dentro del contexto de la lengua viva. Las aseveraciones respecto a la idiosincrasia de un elemento lingüístico, que el perito lingüista forense detecta gracias a su intuición deductiva, se deben respaldar por una investigación del comportamiento lingüístico de la marca candidata definido por la gramática y observado en un corpus de la lengua. Para que los resultados sean admisibles, cualquier conclusión sobre la autoría del texto o los textos anónimos se debe fundamentar en la

[13] La variación inter autor se refiere a las diferencias estilísticas en los textos escritos por individuos diferentes y la variación intra autor en las que se observan en los escritos de una misma persona. En el capítulo 4 comentaremos estos dos conceptos en mayor detalle.

[14] Este es también otro de los criterios de selección de marcas identificativas que contemplaremos en el apartado *b)* de este capítulo.

cuantificación de los datos de cada variable o marca (análisis estadístico) y en el escrutinio de sus propiedades y funciones lingüísticas (análisis calificativo).

Junto a las pautas directas que los lingüistas forenses están elaborando en la actualidad para establecer un procedimiento común de trabajo en atribución de autoría, el peritaje de los documentos dubitados debe cumplir con dos normas básicas. En primer lugar, se tiene que fijar cuál es el cometido específico del análisis. Es decir, si se espera poder atribuir la autoría a un solo autor, a uno dentro de un grupo de posibles autores o a un tercer autor desconocido. En cuanto a esta cuestión, es importante informarse con antelación sobre la posible implicación de más de un autor en la creación del texto. En el contexto forense se asume (por regla) que el documento dubitado es resultado de la producción lingüística de una sola persona que debe ser identificada entre una serie de posibles autores. No obstante, con frecuencia no podemos saber con certeza si el texto no presenta otro tipo de autoría. Las circunstancias de creación de un texto escrito, según Love (2001: 32-50), permiten establecer cinco tipos de autoría, a saber: colaborativa, precursora, ejecutiva, declarativa y de revisión. Cuando un escrito ha sido producido íntegramente por la misma persona, hablamos de **autoría ejecutiva**. Las revisiones y ediciones posteriores del original hechas por terceras personas califican el texto final como de **autoría colaborativa**. Un texto se puede clasificar como el producto de **autoría de revisión** si el mismo autor participa en su edición. Los discursos políticos son un ejemplo

excelente de **autoría declarativa**. Los textos que pronuncian los representantes de los distintos partidos raras veces han sido escritos por ellos, pero son redactados de tal modo que reflejen su ideología y el público pueda asimilarlos como suyos. Y por último, la **autoría precursora** se da cuando en la elaboración de un texto el autor usa partes de textos de otros autores incorporándolas directamente o reformulando su contenido.

En segundo lugar, hemos de tener presente siempre el hecho de que la lingüística, a diferencia de otras disciplinas, no es una ciencia exacta. Lo mismo es válido para la lingüística forense y los dictámenes que informan sobre los resultados de la atribución de la autoría de las pruebas lingüísticas. Las interpretaciones equívocas de los datos pueden tener implicaciones graves para el acusado si las pruebas son aceptadas y llegan a presentarse ante un jurado. A propósito de esto Coulthard (1994: 31) declara:

> (...) linguistic evidence is currently more appropriate for the defence, where the need is to show 'reasonable doubt', than for the prosecution, where the need is to demonstrate 'beyond reasonable doubt'.

Los peritos lingüistas forenses evitan expresarse con seguridad absoluta aún cuando están convencidos de la certeza de sus hallazgos. A consecuencia de esto, en atribución de autoría se ha adoptado la práctica de presentar los resultados en una **escala de**

grados de probabilidad [*probability scale*] como la que encontramos en Coulthard (2007b: 45) (Tabla 1)[15].

Tabla 1. *Escala de grados de probabilidad, según Coulthard (2007b)*

Most Positive
5 'I personally feel *quite satisfied* that X is the author'
4 'It is in my view *very likely* that X is the author'
3 'It is in my view *likely* that X is the author'
2 'It is in my view *fairly likely* that X is the author'
1 'It is in my view *rather more likely than not* that X is the author'
0 'It is in my view *possible* that X is the author'
-1 'It is in my view *rather more likely than not* that X is *not* the author'
-2 'It is in my view *fairly likely* that X is *not* the author'
-3 'It is in my view *likely* that X is *not* the author'
-4 'It is in my view *very likely* that X is *not* the author'
-5 'I personally feel *quite satisfied* that X is the *not* author'
Most Negative

El uso de escalas de probabilidad no esta unificado entre los lingüistas forenses[16], lo que crea un serio problema de interpretación (Coulthard, 2007b) y favorece algunas críticas sobre su objetividad (Broeders, 1999: 237). La solución de este problema podría encontrarse en la normalización de una escala de probabilidades para toda la comunidad de lingüistas forenses.

[15] Esta escala ha sido adaptada por Coulthard de la que emplea la International Assotiation of Forensic Phonetics and Acoustics.
[16] El número y las denominaciones de cada grado varia según el perito

b) Las marcas identificativas: características y criterios de selección

Hasta aquí nos hemos dedicado a describir las condiciones para el desarrollo del buen peritaje lingüístico forense. La clave del análisis lingüístico forense, sin embargo, está en las marcas identificativas, su localización y selección.

Dar una definición precisa y exhaustiva del concepto de marca identificativa no resulta tarea fácil pues cualquier elemento lingüístico de la escritura de un individuo podría ser una marca cuando manifiesta determinadas características. Bailey (1979: 10) recomienda la aplicación de los principios para distinguir los dialectos del inglés establecidos por Labov (1963) como criterios generales de selección de marcas identificativas. Conforme a dichos criterios, las marcas han de poseer las características siguientes:

- prominencia
- distribución
- independencia relativa del control consciente
- uso frecuente
- cuantificación fácil

La **prominencia** [*saliency*] es un concepto prestado de la Lingüística de Corpus para denominar las palabras que destacan estadísticamente cuando se compara un subcorpus con otro

subcorpus o un subcorpus con un corpus completo (Abecassis, 2002). El corpus es una de las herramientas de trabajo primordiales en atribución de autoría puesto que es una muestra de la lengua viva. Su consulta es una práctica frecuente entre los lingüistas forenses, ante todo con el fin de sacar conclusiones acerca de la relevancia de la marca identificativa. En atribución de autoría no nos limitamos solamente a averiguar si las unidades léxicas son prominentes, como puede desprenderse de la definición de Abecassis citada, sino también si las marcas sintácticas o de cualquier otro carácter discursivo responden a esta característica en el contexto del documento estudiado.

La **independencia relativa del control consciente** tiene que ver con el grado en el que el usuario de una lengua es consciente y capaz de controlar sus selecciones en el proceso de la producción lingüística. Cuanto menos manipulable por el control consciente del individuo se muestra el uso de una unidad lingüística, tanto mayores son sus posibilidades de ser marca identificativa.

El concepto de **distribución** tiene que ver con el número de ocurrencias de la marca en cada texto. Para la realización del análisis forense interesa que la marca aparezca en todos los documentos disponibles, tanto dubitados como indubitados, y con una **frecuencia** que permita su cuantificación.

La **cuantificación fácil** alude a aquellas unidades que no necesitan un tratamiento especial o un análisis previo que posibilite su cálculo.

La aplicabilidad de algunas marcas identificativas, sobre todo las léxicas, en casos reales se cuestiona, pues se podría considerar que los textos delictivos como las cartas de amenaza, por poner un ejemplo, son resultado de una producción lingüística premeditada y menos espontánea. Es de suponer que el autor ha tenido tiempo para reflexionar sobre el texto de la carta y de hacer lo posible para disimular su estilo o imitar el de otra persona para no ser reconocido. En este sentido se recomienda buscar marcas que, aunque no se muestren completamente independientes de una planificación previa, por lo menos estén sujetas a un **grado de control consciente bajo**.

En los estudios recientes (Chaski, 2005; Turell, 2004) se habla también de evaluar el potencial de un rasgo idiosincrásico como marca de identificación en relación a su **marcadez** [*markedness*]. El término fue acuñado en los años 30 del siglo pasado en el Círculo de Praga por Jacobson y Trubetzkoy. En los estudios de identificación de autor, viene a caracterizar las formas lingüísticas que se desvían en su uso (en el idiolecto del usuario de una lengua concreta) de las normas establecidas por la gramática, o como las define Jacobson (Jacobson en Turell, 2004) "las que aportan información más específica que la expresada por la forma no marcada".

Grant y Baker (2001:71-73) en su publicación en respuesta a Chaski (2001) discuten la trascendencia de otros dos conceptos: **fiabilidad** [*reliability*] y **validez** [*validity*], entre los criterios de identificación de marcas en atribución de autoría. Siguiendo la definición que ofrecen Grant y Baker de ambos conceptos podemos concluir que la fiabilidad y la validez de una marca identificativa dependen en sumo grado de la técnica que implementa dicha marca. Ellos dicen que la **fiabilidad** de una marca queda comprobada cuando, al aplicar la técnica a textos del mismo autor se producen los mismos resultados, especialmente si estos resultados son similares a los obtenidos mediante otros métodos de identificación de autor. Su **validez**, en cambio, tiene que ver con la subjetividad de la técnica. No es válida una marca que adopta una técnica que se centra en aspectos del texto irrelevantes para la atribución de la autoría o, en otras palabras, que pretende determinar otras propiedades de éste, como por ejemplo, el contexto sociolingüístico en el que ha surgido.

2.2 Estudios de atribución de autoría

Desde sus inicios, la investigación en atribución de autoría se ha ido encaminando hacía un método de análisis lingüístico forense fiable y válido que corresponda a las necesidades prácticas que presenta la casuística del trabajo en autoría. De acuerdo con este objetivo y los parámetros de cada estudio, estos se han realizado generalmente en torno a tres líneas de experimentación exploratoria:

- con datos de textos literarios;
- con datos de texto de producción inducida;
- con datos de textos de casos reales.

Entre estos tres prevalecen los trabajos elaborados a partir del análisis de datos extraídos de textos literarios por ser, a diferencia de las pruebas documentales de casos reales, de acceso no restringido y por la relativa libertad que brindan al investigador de recrear o simular contextos de análisis propios del trabajo en atribución de autoría de diversa complejidad y de traer a la luz distintos problemas pendientes de solución en la actualidad. En este sentido no se pueden menospreciar los estudios realizados por Holmes (1992, 1994), Holmes et al. (1995, 2001), Burrows (1987, 1989, 2002, 2003) o Hoover (2001, 2002, 2003) , entre otros, que indagan sobre la aptitud de algunos elementos lingüísticos para desempeñar la función de marcas identificativas y plantean temas de futuras discusiones, que comentaremos en mayor detalle en los próximos apartados.

La segunda línea de experimentación exploratoria más seguida, por el número de trabajos producidos en el ámbito de autoría, correspondería a los estudios llevados a cabo a partir de datos procedentes de simulación de casos reales. Los trabajos experimentales de este perfil que han tenido difusión científica son pocos, hecho que se explica por un lado por la necesidad de financiación continua de la que los proyectos de esta índole suelen

carecer y, por otro, por la visión negativa que se mantiene respecto a la validez de los resultados obtenidos mediante este método de recogida de datos en los círculos del gremio forense. Un ejemplo excelente de este tipo de estudios, que no deja de ser meritorio a pesar de las serias críticas que ha recibido por su metodología, es la monografía de Chaski (2001). En ella la autora ofrece una clasificación de las técnicas de análisis lingüístico forense más extendidas en este campo junto con la evaluación de su potencial discriminatorio y su eficacia. Volveremos sobre este estudio en los capítulos dedicados a los principales grupos de marcas de identificación (véase apartado 2a de este capítulo).

Por último cabe mencionar los estudios que parten de pruebas lingüísticas de casos reales. Los pocos que han llegado a publicarse se han convertido en punto de referencia de toda la comunidad de lingüistas forenses, entre ellos el de Svartvik (1968), que marca el nacimiento de la lingüística forense con su análisis lingüístico de los testimonios del caso Evans. De gran valor para el avance en lingüística forense en general y la atribución de autoría en particular son también las contribuciones de Coulthard (entre ella la más importante es quizás la del caso Derek Bentley) (1993, 2005a) que reflejan sus intervenciones como perito lingüista forense en varios casos de autoría discutida.

A pesar de ser una disciplina nueva y, en palabras de Coulthard (2005b) "todavía en mantillas", la atribución de autoría cuenta con un volumen considerable de publicaciones que representan aquella

piedra angular que sirve a todo investigador en los comienzos de sus pesquisas.

a) Estudios de atribución de autoría mediante marcas léxicas

En la revisión de la bibliografía sobre atribución de autoría que ofrecemos en los apartados siguientes resumimos aquellos trabajos que consideramos más relevantes y de mayor interés científico dentro de la tipología antes especificada y en relación al objeto de estudio de la presente tesis doctoral. Los resúmenes están organizados según la marca que se explora en cada estudio y siguiendo dentro de lo posible la cronología de su publicación.

En el desarrollo de la práctica lingüística forense ha prevalecido de forma contundente la aplicación de las técnicas basadas en las unidades léxicas o, como las denominamos en atribución de autoría, las marcas léxicas. Gran parte de los estudios de los que se tiene constancia apuestan por diferentes aspectos del vocabulario empleado en los textos de análisis como posibles marcas de identificación. Entre los más extendidos podemos enumerar la riqueza léxica (Wools y Coulthard, 1998), la longitud de las palabras (Mosteller y Wallace, 1964; Smith, 1983), la distribución léxica (marca estrechamente relacionada con los fenómenos de hapax legomena y hapax dislegomena (Smith, 1987)), las palabras funcionales (Elegard, 1962; Burrows, 1987, 1989; 2002; Sánchez Pol et al. 2005), las locuciones (Hoover, 2001) entre otros.

El mayor número de estudios se ha dedicado a las palabras funcionales, que han tenido una considerable repercusión en los avances de la disciplina. Este hecho se debe a tres principales características propias de este tipo de marcas: en primer lugar, a su alta frecuencia de uso, que facilita su cuantificación y posterior análisis estadístico; en segundo lugar, a su papel gramatical, que restringe el control consciente en su producción por parte del sujeto-autor, lo cual presupone un indicio claro de su aptitud como marcas identificativas. Además, el contenido semántico de las palabras funcionales está limitado a la relación gramatical o a la propiedad genérica que designan y no suelen estar ligadas al contexto concreto, lo que es indicio claro de su pertenencia al grupo de rasgos de estilo autorial.

Elegard (1962) es un de los primeros en considerar la aplicación de las palabras funcionales como parámetro para tratar un tema de autoría en su estudio pionero sobre "*The Junius Letters*", una serie de panfletos políticos publicados en Inglaterra entre 1769 y 1772. Como punto de partida en su análisis, Elegard toma la lista de las palabras de mayor y menor frecuencia de uso con el fin de establecer el ratio de disimilitud de cada una respecto a las observadas en la obra escrita de otros autores de la época y así poder aplicar su valor como criterio discriminante.

Esta línea de trabajo será retomada por Mosteller y Wallace unos años después (1964) en su famoso estudio sobre la autoría de una colección de artículos publicados de forma anónima por J. Jay, J.

Madison y A. Hamilton en el período 1787-1988, y conocida bajo el nombre *The Federalist Papers*. La discusión que se produce acerca de estos *Papers* estriba en la incógnita del posible autor de doce de los artículos cuya autoría se disputa entre Hamilton y Madison. Para dar una respuesta, Mosteller y Wallace usan como medida discriminante las frecuencias relativas de algunas de las palabras funcionales más recurrentes y concluyen que el autor de los artículos "dubitados" es Madison. Este trabajo marca un avance significativo en la aplicación de la metodología basada en las palabras funcionales en el sentido que introduce por primera vez métodos de análisis estadísticos y enfoca el problema de la autoría discutida desde un punto de vista más pragmático.

Durante las dos últimas décadas del siglo XX aparecen los estudios independientes de Burrows (1987, 1989; 2002; 2003) y Holmes (1992, 1995, 2001), que marcan una nueva etapa en la evolución metodológica del análisis lingüístico forense de autoría por medio de las marcas léxicas. Estos autores introducen algunos métodos más sofisticados de análisis estadístico como el análisis multivariante, el análisis clúster y el análisis de componentes principales (ACP), que permiten manejar un volumen extenso de datos. Con el análisis de componentes principales de las palabras más comunes usadas en los diálogos de los personajes en las novelas de Jane Austen, Burrows (1987) logra demostrar que es posible además de diferenciar entre los idiolectos creados por la escritora, trazar la trayectoria de desarrollo de cada idiolecto a lo largo de la novela. Holmes (1992, 1995, 2001), por su parte,

confirma la eficacia de los análisis multivariante al detectar y cuantificar los rasgos idiosincrásicos léxicos del estilo de un autor como procedimiento de identificación autorial.

Los resultados alentadores de los estudios citados estimularán el comienzo de lo que actualmente constituye un continuo de experimentos que investigan el potencial discriminatorio de las palabras funcionales y sus limitaciones. Argamon et al. (2005: 1) discuten la validez de las palabras funcionales como marcas de identificación frente a otros candidatas léxicos y atribuyen su poder discriminante al uso de muchos textos de gran extensión:

> (…) even simple language modeling techniques can greatly improve in effectiveness when larger quantities of data are applied.

Según demuestra su estudio, el potencial discriminatorio de las palabras funcionales disminuye a medida que se reduce el tamaño de los documentos analizados. A propósito de su aptitud como marca, estos autores expresan algunas dudas respecto a su naturaleza minimalista y su capacidad de plasmar las cuestiones estilísticas subyacentes y afirman que para mejorar el rendimiento de la técnica es preciso recurrir al uso de unidades más complejas. Su artículo hace hincapié en el hecho de que los estudios deben intentar proporcionar información relativa a las oscilaciones que se dan en los niveles porcentuales de atribución correcta en el análisis

de los mismos textos al reducir y aumentar su tamaño[17]. Como ejemplo de estudio con este tipo de diseño experimental Argamon et al. refieren al lector al trabajo de Hoover (2001).

Los estudios de Hoover (2001; 2002; 2003) van más allá de las palabras funcionales en la búsqueda de elementos léxicos de carácter idiosincrásico que permitan hacer la distinción entre textos de diferentes autores. Este autor es partidario de la idea de que el uso de las palabras funcionales, siendo éstas unidades mínimas, no está tan estrechamente ligado al estilo, y en cambio sí lo están las construcciones formadas por varios elementos y de mayor carga semántica. Hoover (2003) contrasta la eficacia de las palabras funcionales de mayor frecuencia con la de las locuciones más recurrentes en escenarios diferentes[18]. Hemos de notar aquí que el término *locución* no es empleado por el autor en su sentido directo, sino para designar las secuencias de palabras contiguas o palabras que ocurren en proximidad de la una con la otra[19]. Al alternar el número de variables consideradas (de 50 a 800) y el número de componentes en el caso de las locuciones en el análisis multivariante, Hoover concluye que aunque las palabras funcionales

[17] Sobre este tema véase también Zhao y Zobel (2005).
[18] Previamente Morton (1978) condujo un experimento estilométrico similar en el que se emplearon tres tipos de locuciones: las de estructura formada por dos componentes en orden consecutivo, las de parejas proporcionales y las de patrones posicionales de las palabras funcionales dentro de la oración. Para una revisión del estudio véase Totty et al. (1987).
[19] Por ejemplo, la frase en ingles *An awfully long and boaring story* daría, de acuerdo con la fragmentación bidireccional de Hoover y si se toma un lapso de 2 palabras paritiendo de la palabra "and", las siguentes locuciones *and long* y *and boaring*. El recuento de las locuciones se realiza mediante el programa TACT de la University of Toronto.

manifiestan un potencial discriminatorio significativo, las locuciones lo superan en el análisis de textos de tamaño reducido como también en el análisis de determinados grupos de textos. Volveremos a comentar el diseño experimental y el procedimiento de análisis que usa Hoover en el capítulo sobre Metodología ya que ambos elementos metodológicos han sido transcendentales para desarrollar el modelo de la técnica de análisis propuesta en esta tesis.

A nivel mundial, puesto que la Lingüística Forense se ha asentado en los países anglosajones, es lógico que predomine la literatura sobre atribución de autoría en y sobre la lengua inglesa. Los estudios realizados sobre autoría en otros idiomas mediante palabras funcionales u otras marcas son muy escasos. Para la lengua española y en relación a las marcas léxicas, concretamente aquellos que usan las palabras más frecuentes, podemos mencionar el de Sánchez Pol et al. (2005). El objetivo de su trabajo es medir la similitud intra e inter autor para determinar si dicha similitud es mayor dentro de los textos de un autor o no. Para ello adoptan una fórmula matemática, propia de la investigación en genética, que mide la distancia por parejas entre las secuencias formadas por las 5 palabras más frecuentes en cada texto de análisis. A cada elemento de la secuencia se le asigna una puntuación diferente, de 5 a 20 según la posición que ocupa dentro de la secuencia, para luego aplicar la fórmula y calcular el valor del **grado de similitud**[20]. Su

[20] Cicres (2007) emplea el mismo término para referirse a la distancia entre las muestras de voz del mismo autor en su análisis de la variación intra autor en tiempo real.

análisis no solo muestra que las palabras funcionales son discriminantes en el contexto de la lengua española, sino que también confirma la viabilidad de un método que emplea secuencias de palabras como marcas de identificación.

b) Estudios de atribución de autoría mediante marcas sintácticas

El estudio de los rasgos idiosincrásicos de la sintaxis idiolectal en cualquier lengua supone la costosa labor de su selección y cuantificación para el análisis posterior. Esta circunstancia ha implicado que las construcciones y las estructuras sintácticas como posibles marcas de autoría hayan resultado un objeto de estudio poco atrayente. Sin embargo, hubo quien se decidiera a aceptar el reto de pasar largas horas leyendo y analizando una frase tras otra sin la ayuda de ningún tipo de herramienta informática similar a las de las que disponemos hoy en día. De hecho, el primer estudio que dio nombre a la Lingüística Forense como disciplina y puso sus cimientos fue de uno de estos pioneros incansables. Resulta imposible hablar de Lingüística Forense y de atribución de autoría sin mencionar el nombre de Jan Svartvik y su análisis lingüístico de los testimonios del caso Evans. Es el paso inicial que abre el camino a casi medio siglo de historia en lingüística forense y que, pese al transcurso del tiempo, mantiene su actualidad como una de las publicaciones más importantes de esta disciplina. Curiosamente también se trata del primer estudio que usa marcas sintácticas para

la atribución de autoría, motivo por el cual no podemos dejar de mencionarlo.

Antes de pasar a la descripción del análisis que efectuó Svartvik juzgamos oportuno dar algunos detalles acerca del caso tratado en el estudio. Creemos que de esta manera se podrá recrear una visión más nítida de la envergadura del problema al que Svartvik decidió enfrentarse "armado" solo de su conocimiento experto en lengua inglesa.

El caso llegó a manos del autor cuando los familiares del Timothy Evans se dirigieron a él con la petición de revisar las únicas pruebas inculpatorias existentes. Evans era un joven condenado a pena de muerte por doble homicidio y ejecutado años antes por sus declaraciones ante la policía, que fueron presentadas como pruebas lingüísticas en el proceso judicial. Se trata de los textos de las cuatro declaraciones de Evans (MT1, MT2, NH1, NH2)[21] que relatan los acontecimientos relacionados con el crimen. La primera y principal circunstancia desfavorable de cara al peritaje lingüístico forense radica en el hecho de que Evans era analfabeto y las declaraciones que hizo fueron transcritas en su integridad por una tercera persona diferente en cada interrogatorio. Otra dificultad del trabajo proviene de la ausencia de documentos indubitados, es decir, de muestras de la producción lingüística espontánea de Evans para el análisis contrastivo. No obstante, en la revisión de los textos, Svartvik observa una serie de rasgos y diferencias estilísticas en

[21] MT – Merthyr Tydfil Police Station; NH – Notting Hill Police Station

algunos de los documentos, pero sobre todo en el texto de la declaración donde Evans se confiesa culpable de los asesinatos (NH2). La inconsistencia lingüística respecto a las demás declaraciones y la complejidad oracional atípica e inusual en el habla de una persona analfabeta y de inteligencia limitada captan la atención de Svartvik y despiertan su interés.

Para minimizar el efecto negativo de los diferentes estándares de edición de que fueron objeto las pruebas lingüísticas, Svartvik decide analizar cada texto por separado. En el caso de la declaración que jugó un papel decisivo en la acusación (NH2) este lingüista examina cada una de sus partes integrantes (a, b, c) contrastándolas con las de MT2 que según dijo Evans durante el proceso, era íntegramente cierta. En la opinión de Svartvik, la manera más adecuada de proceder ante un problema de autoría discutida de este tipo es definir el grado en que cada texto es internamente coherente con respecto a sus rasgos lingüísticos. Dado que la longitud oracional y la variedad de frases no son marcas fiables en un corpus tan pequeño, Svartvik otorga dicho papel a los esquemas combinatorios de las frases y su función dentro de la oración.

El análisis cuantitativo de las cláusulas, clasificadas en seis grupos (A, B, C, D, E, F)[22] según sus características sintácticas revela, por un lado, una diferencia destacable en la distribución de las frases de

[22] Para más detalles acerca de la clasificación y las especificaciones de cada grupo, véase Svartvik (1968: 31- 38).

tipo B[23] en el fragmento (b) de NH2, en comparación con (a) y (c) del mismo texto. Es muy llamativa la repetición de la estructura pospuesta con "then", típica de los informes policiales (Coulthard, 1993, 2005a; Wools y Coulthard, 1998: 33-37). Por otro lado, el análisis estádistico muestra que la sección (b) de NH2 guarda muy pocas similitudes con el resto de fragmentos y textos. Ante estos resultados Svartvik termina concluyendo que aunque la validez de su interpretación no sea contundente, la incongruencia en el uso de las cláusulas entre las declaraciones de Evans es suficientemente notable, como para que puedan ser atribuidas al efecto de factores como la situación lingüística en la que fueron producidos los textos y el hecho de no ser escritos por el propio Evans.

Aunque el trabajo de Svartvik sobre las pruebas lingüísticas del caso Evans es muy sugerente en cuanto a líneas de investigación, éste no tuvo la misma acogida de la que disfrutan los estudios sobre marcas léxicas. La investigación sobre marcas sintácticas se quedó prácticamente estancada durante varias décadas y no sería hasta la llegada de los medios informáticos, que permitieran el tratamiento automático de los textos y la extracción de unidades tanto léxicas como sintácticas, que se reiniciaría la búsqueda de identificadores de carácter sintáctico para los fines de la atribución de autoría forense. De este período es el artículo de Baayen et al. (1996), convertido en referencia clave para este tipo de estudios.

[23] Son cláusulas que se unen por medio de un conector movible, es decir, que puede aparecer tanto en posición inicial como en posición final. Conectores de este tipo en inglés son "then" y "also".

Baayen et al. estudian el potencial discriminatorio de las reglas de reescritura[24] y su aplicación con fines forenses. El experimento que llevan a cabo se sirve de los textos del subcorpus de ciencia ficción del corpus Nimega[25] que contiene textos de diversos géneros. De este subcorpus se tomaron para el análisis dos textos de novelas policíacas (de aproximadamente 20.000 palabras) limitando así el experimento a un solo tipo textual. Los autores justifican esta decisión metodológica con los resultados de su estudio piloto sobre el corpus completo. Su análisis reveló que los textos de un mismo autor que pertenecen al mismo género literario difieren más estilísticamente entre sí que los textos del mismo género escritos por autores diferentes.

Para la anotación sintáctica de ambos textos se empleó el sistema de anotación semi-automática TOSCA (Tools for Syntactic Corpus Análisis). En el proceso de anotación el sistema asigna dos etiquetas a cada constituyente de la oración de análisis: una que describe sus propiedades sintácticas (categoría y función) y otra, las semánticas (atributos). Puesto que el objetivo de Baayen et al. era utilizar las mismas técnicas y software que se aplican en el trabajo con unidades léxicas, los árboles sintácticos resultantes del análisis fueron reducidos a reglas de reescritura y luego transformadas en pseudo-palabras. Además para designar los límites de las pseudo-frase se introdujeron separadores de unidades de texto [*text unit separador*].

[24] La combinación del núcleo de la frase y sus constituyentes inmediatos.
[25] Consulte Keulen (1986) para mayores detalles sobre este corpus.

Con el propósito de evaluar la técnica de las reglas de reescritura, Baayen et al. fraccionaron los dos textos de análisis en 14 muestras indubitadas y 16 muestras dubitadas de manera que cada fracción de texto o muestra tuviera aproximadamente la misma extensión. A continuación emplearon el ACP[26] con los datos de ocurrencia de las 50 reglas de reescritura más frecuentes en una prueba de simulación de caso de autoría discutida y compararon los resultados con los de una prueba paralela de los datos de las 50 palabras funcionales de mayor frecuencia de uso en las 30 muestras.

Los resultados de su experimento demuestran que el uso de las reglas de reescritura, o en otras palabras, de las construcciones sintácticas, es más constante que el de las palabras funcionales. Asimismo señalan que las reglas de mayor y menor frecuencia destacan por un alto potencial discriminatorio en la prueba de simulación. No obstante, el tamaño de los textos que se analizaron sugiere que la aplicabilidad de la técnica se confina a documentos de similar extensión. Por lo tanto, para corroborar su eficacia es recomendable replicar el experimento con textos cortos y, como comentan Hirst y Feiguina (2007), también sustituir el sistema de anotación por alguno más moderno y de mayor precisión que el *parser* TOSCA.

[26] Análisis de los componentes principales. Método de análisis estadístico propulsado por Burrows (1989) que mide la densidad léxica.

De los estudios que exploran la sintaxis como fuente de posibles marcadores identificativos hemos de destacar también los publicados por Chaski (1997, 2001). Chaski (2001) ofrece un repaso de las técnicas que se usan actualmente en el análisis lingüístico forense en los Estados Unidos. Este repaso incluye también la evaluación empírica de cada una de las técnicas, que consiste en determinar si éstas discriminan correctamente entre los textos escritos por diferentes autores y si la prueba de agrupación [*clustering*] sitúa el texto pseudoanónimo[27] en el grupo de textos de su autor real.

Para la realización de este estudio, la autora cuenta con la financiación del gobierno norteamericano, lo que le permite crear un corpus de análisis que recoge las producciones lingüísticas escritas de cuatro sujetos que comparten las mismas características sociolingüísticas: edad, género, grupo étnico, dialecto y nivel educativo. Chaski pretendía imitar las circunstancias que se dan en el trabajo en un caso real de atribución de autoría de modo que de la totalidad de textos obtenidos mediante el método de inducción se seleccionaron para el estudio solo aquellos cuya extensión rondaba el límite de palabras susceptible de análisis. Así pues, el texto más corto en el corpus tiene 93 palabras y el más largo, 556. El número de textos para cada individuo en el experimento se fijó en tres. Los participantes en el experimento escribieron sobre 10 temas ideados de manera que incitaran el uso del dialecto por un lado, y el uso del lenguaje formal por otro, y así asegurar la variedad de géneros.

[27] Texto cuya autoría se conoce pero se ha ocultado para los fines de la prueba.

Con la finalidad de probar la hipótesis de que las estructuras sintácticas abstractas pueden diferenciar e identificar el autor de un texto, Chaski aplica el test estadístico de la chi-cuadrada[28]. Los resultados confirman que la técnica utilizada discrimina entre autores, distinguiendo entre el texto pseudoanónimo y asignándolo al autor correcto. El margen de error de la atribución mediante esta marca es igual a *0*[29]. La autora reconoce que si bien estos resultados son muy alentadores, es imprescindible replicar el estudio antes de divulgar ese dato. Esto no impide, sin embargo, que Grant y Baker (2001) le dirijan críticas muy severas en cuanto a la metodología aplicada y a la validez de las conclusiones que extrae de este experimento y de los demás que forman parte del estudio. Su crítica se fundamenta, por una parte, en el hecho que Chaski generaliza los resultados de la experimentación a partir de un corpus muy homogéneo que no es representativo de toda la población; por otra parte, en que no establece la fiabilidad de las marcas identificativas del estudio fuera del contexto de su corpus de análisis. Y, por último, en que falta una explicación de la variación intra e inter autor que justifique su conclusión en el sentido de que determinadas marcas son fiables. A pesar de las carencias metodológicas que presenta el estudio de Chaski, éste no deja de ser un buen ejemplo

[28] "El test de la chi-cuadrada permite contrastar la hipótesis de que los dos criterios de clasificación utilizados (las dos variables categóricas) son independientes. Para ello, compara las frecuencias observadas (las frecuencias de hecho obtenidas) con las frecuencias esperadas (las frecuencias que teóricamente debería haber en cada casilla si los dos criterios de clasificación fueran independientes)"(Pardo y Ruiz, 2002: 228).

[29] El margen de error en el estudio se calcula a partir del porcentaje de casos en los que la técnica examinada fracasa en diferenciar entre los textos de los diferentes autores y no asigna el texto dubitado correctamente.

de cómo debería desarrollarse la investigación en atribución de autoría.

Otro estudio de atribución de autoría que se basa en los rasgos sintácticos como posibles marcas de identificación de autor es el de Stamatatos et al. (2001). El método que proponen estos investigadores es completamente automático y no requiere ningún tipo de preparación previa de los textos, ni intervención humana. A diferencia del de Baayen et al. (1996) el método de Stamatatos et al. no genera árboles sintácticos a partir del texto procesado, sino que se limita a detectar las frases constituyentes, aquellas que no se solapan, y a definir su función oracional. Este proceso, denominado *multiple-pass parsing,* se lleva a cabo mediante una herramienta llamada SCBD (Sentence and Chunk Boundaries Detector). En cada paso del *parsing* se analiza una parte de la oración a partir de los resultados de los pasos anteriores del procesamiento y se almacena información para los pasos siguientes.

Para su estudio, Stamatatos et al. recuperaron de la web 300 artículos periodísticos escritos en lengua griega que fueron proporcionalmente divididos en un corpus de entrenamiento y un corpus de prueba, de forma que el segundo representara la mitad del primero.

El SCBD devolvió 22 rasgos estilísticos (media de palabras por frase nominal, ratio de frases nominales respecto el resto de tipos de

frases o *chunks*, etc.)[30], al analizar conjuntamente los textos del corpus de entrenamiento y el corpus de prueba. Los rasgos o marcas representan en su mayoría los valores resultantes del cálculo del ratio o la media que se da entre las variables de análisis.

El análisis automático de Stamatatos et al. revela que frente a las marcas léxicas, su variable se mantiene mucho más estable como factor discriminante respecto a las reducciones en el tamaño de los textos analizados. La principal carencia de la técnica es la dependencia de la aplicación del método de la herramienta SCBD, diseñada específicamente para el griego.

c) Estudios de atribución de autoría mediante n-gramas

Finalmente entre los estudios sobre atribución de autoría se encuentra el grupo de trabajos en el que se incluye la investigación que hemos llevado a cabo en esta tesis doctoral. El grupo en cuestión abarca los estudios que indagan sobre la viabilidad y la aplicabilidad de los n-gramas como marca distintiva. El n-grama se puede definir como la secuencia de *n* componentes. Estos componentes pueden ser unidades de diversa índole, dependiendo del nivel extralingüístico o lingüístico al que se observan. Pueden representar combinaciones de palabras, carácteres alfanuméricos,

[30] Véase Stamatatos et al. (2001) para más información sobre las marcas y su precisión.

signos de otra tipología, etc. Dentro del grupo de n-gramas lingüísticos se pueden divisar de manera muy general dos tipos de n-gramas que han sido objeto de exploración en la rama de la atribución forense de autoría: n-gramas de letras y n-gramas de categorías sintácticas. En el marco del primer tipo se contemplan los estudios de Kjell (1994), Forsyth y Holmes (1996) y Keselj et al. (2003), que consideran las frecuencias relativas de los bigramas de letras, y los de Khmelev (2000) y Khmelev y Tweedie (2001), que elaboran su método de atribución de autoría en base al modelo de cadenas de Markov. Aunque las letras del alfabeto son representativas de cada lengua o familia de lenguas, no dejan de ser meros signos de ortografía que fuera del contexto de la palabra que forman, no poseen carácter lingüístico ninguno. Por consiguiente, a nuestro parecer, los métodos que acabamos de citar difícilmente pueden ser clasificados como lingüísticos, sino más bien como matemáticos. Aun así, hemos juzgado oportuno mencionarlos aquí para dar una visión más amplia de las distintas perspectivas del estudio de los n-gramas como marcas distintivas. Por todo lo expuesto no nos detendremos en su descripción, pero sí en la de los n-gramas del segundo tipo, los de categorías sintácticas.

Antes de adentrarnos en el tema del estudio de los n-gramas de categorías sintácticas como identificadores de autoría, conviene dar una definición de los n-gramas más específica y descriptiva que resalte su carácter lingüistico:

> A gram is defined to be a permutation of words – a sequence – that varies in length and the words pertain to the defined syntactic categories. The length of the permutation is defined as an n-gram, where the n is an integer value indicating the number of words in the gram.
>
> (Diab et al., 1998: 2)

El estudio de Diab et al. (1998) encabeza los trabajos en autoría que experimentan con n-gramas de categorías sintácticas para determinar hasta qué punto su aplicación como marcas identificativas es viable. Diab y sus colegas usan la correlación entre las frecuencias de los gramas de uno a seis constituyentes con el objetivo de probar sus hipótesis, mediante la realización de tres experimentos consecutivos:

a) La viabilidad del método a nivel intra e inter autor. Es decir, si el grado de correlación calculado a partir de las frecuencias de cada n-grama entre los textos del mismo autor y los textos de diferentes autores, es alto o significativo.

b) La aptitud de los n-gramas de discriminar entre diferentes autores.

c) La atribución de la autoría del texto de *Funeral Elegy* a su supuesto autor, William Shakespeare.

Para ello se sirven de los textos completos de cinco obras teatrales que pertenecen a dos épocas diferentes y géneros distintos[31]. Los autores de las piezas escogidas son William Shakespeare, Thomas Middleton y Nicholas Wardigo.

El método de Diab et al. tiene tres fases. En la primera, todos los textos son fragmentados en secuencias de palabras delimitadas por signos de puntuación en orden consecutivo, desde el principio hasta el final de cada oración. La segunda fase implica la anotación sintáctica de cada integrante de la secuencia con una única etiqueta de categoría gramatical y la conversión de ésta en un código numérico. La última fase consiste en la extracción de los n-gramas y el cálculo del coeficiente de correlación entre los diferentes textos usados en la investigación.

A lo sumo, los resultados de los experimentos confirman la presencia de una correlación significativa entre los textos del mismo autor que permite discriminarlos de los textos de otros autores. La simplicidad del diccionario y el número limitado de textos analizados, sin embargo, hacen patente la necesidad de continuar la investigación en esta línea puesto que, como los mismos autores del estudio coinciden en comentar, un corpus más amplio y la aplicación de un *tagger* más complejo podrían influir sin duda en el mejor rendimiento de la técnica.

[31] Para obtener mayor número de textos, Diab et al. dividen tres de las piezas en dos muestras separadas. La diversidad de géneros se explica por la intención de los autores de escoger las obras más representativas del estilo de cada dramaturgo.

Spassova (2006) y Spassova y Turell (2007) informan sobre el potencial discriminatorio de los n-gramas de categorías sintácticas en textos escritos en español como marcas de identificación, evaluado mediante pruebas estadísticas de autoría[32]. Los parámetros de sus experimentos y sus resultados se discutirán en más detalle en el apartado 0.00 por lo que de momento nos limitaremos a apuntar algunas disparidades metodológicas entre el estudio de 2007 y el de Diab et al. (1998). En primer lugar, el estudio piloto de Spassova y Turell (2007) se centra específicamente en los n-gramas de tipo *bigramas* y *trigramas* [33], de manera que las combinaciones de diferente número de formantes quedan excluidas. Esta restricción responde a que es de esperar que solo las secuencias de categorías de más de un integrante y menos de cuatro sean equivalentes a estructuras sintácticas reales y no coincidan con frases completas. En segundo lugar, el método que exponen Spassova y Turell (2007) no se basa en las secuencias de categorías sintácticas, sino en las secuencias de etiquetas categoriales que, previamente, el sistema de anotación ha asignado a cada unidad dentro del texto de análisis, y que representan las secuencias de categorías en un texto anotado. Este recurso es mucho más práctico, ya que, como veremos más adelante, posibilita la extracción de las equivalencias de cada secuencia en un texto o en todo un corpus.

[32] Para un estudio similar al nuestro sobre lengua inglesa, véase Hirst y Feiguina (2007).
[33] De dos o tres integrantes consecutivos, respectivamente.

2.3 Estudio preliminar sobre el potencial discriminatorio de las secuencias de etiquetas morfosintácticas más frecuentes

El experimento llevado a cabo por Spassova (2008) sobre una de las estructuras sintácticas de composición de orden fijo y uso extendido en la lengua castellana, ha confirmado la hipótesis según la cual las estructuras sintácticas se encuentran entre los rasgos idiosincrásicos distintivos que caracterizan el estilo de un autor y que permiten distinguirlo de otro, aunque bajo ciertas condiciones relacionadas con el número y el tamaño de las muestras de comparación (2008: 608)[34]. Sin embargo, éste tal vez no sería el caso de otras estructuras en las que los componentes de la construcción no son predeterminados, en el sentido de que no se establece un patrón de la unidad[35] y se toma como variable cualquier estructura de orden libre[36], dejando el análisis de su función sintáctica para una etapa posterior.

[34] Las perífrasis verbales resultan en alto grado sumisas a estas condiciones que las incapacitan, por así llamarlo, como marcas de identificación en casos reales, por el muy reducido nivel de ocurrencias que se registra de ellas en textos de menos de 3.000 palabras.

[35] Las perífrasis verbales se caracterizan por un esquema de organización estructural relativamente constante, lo que permitió que en el estudio sobre su potencial discriminatorio se emplearan patrones de búsqueda para la extracción de datos.

[36] Cabe explicar aquí que con el término 'estructura de orden libre' se hace referencia a cualquier secuencia de categorías gramaticales y/o funcionales que pueda observarse en el corpus de análisis sin delimitar construcciones concretas de la lengua como, por ejemplo, las locuciones de todos los tipos o las frases adjetivas y adverbiales para dar un ejemplo concreto, aunque a veces las secuencias puedan coincidir con ellas. Como consecuencia de esta flexibilidad de la variable se pueden extraer secuencias de diferente tipo según el número de elementos que la constituyan.

Con esta idea en mente realizamos el estudio piloto que se presenta a continuación y que refleja la fase inicial de nuestra investigación en atribución de autoría, en la que se examinan las unidades combinatorias de categorías lingüísticas o segmentos sintácticos y su aptitud para funcionar como marcas identificativas distintivas.

El objetivo de este estudio fue seleccionar el tipo de secuencia que mejor se adapte al perfil de marca identificativa buscado mediante pruebas estadísticas[37]. Si las secuencias de categoría se revelasen efectivas en agrupar una serie de textos según su autor y en atribuir correctamente una texto pseudoanónimo al autor correcto, esto supondría un rasgo identificativo que no tiene las limitaciones que a menudo descartan métodos ya reconocidos en autoría.

Antes de continuar con la descripción del diseño metodológico de nuestro estudio preliminar es indispensable reseñar las características de la variable de análisis y su representación en el corpus experimental. En un corpus anotado como el utilizado en el estudio, a cada palabra o unidad marcada se le adjunta una etiqueta [*tag*]. Las clases de etiquetas pueden variar conforme el tipo de anotación. Concretamente en el caso del corpus de análisis del estudio se aplicó la anotación morfosintáctica, que permite asignar una categoría a cada palabra del texto y una etiqueta morfosintáctica a cada categoría. (Véase un ejemplo del corpus anotado en el **Cuadro 1**). Por consiguiente, cuando nos referimos a secuencias de categorías en un corpus, realmente lo que se tiene en mente son

[37] Véanse apartados 5 y 6 del capítulo 3.

secuencias de etiquetas. De ahora en adelante se hará referencia a ambas con la abreviatura **SEM** (Secuencia de Etiquetas Morfosintácticas).

Cuadro 1. *Muestra de anotación morfosintáctica del Corpus Técnico de IULA*

##	TAG	<p>		
##	TAG	<s tree="1.1.2.1">		
6	TOK	No	BOS	**no**\D4
7	TOK	perdí		**perder**\VDP1S-
8	TOK	el		**el**\AMS
9	TOK	tiempo		**tiempo**\N5-MS
10	TOK	en		**en**\P
11	TOK	preparar		**preparar**\VI----
12	TOK	la		**el**\AFS
13	TOK	huida		**huida**\N5-FS
---	DLD	;	EOS	=\DELIM
##	TAG	</s>		

La decisión sobre el número de formantes o etiquetas que debería contener una SEM para que tuviera más posibilidades como marca de identificación fue predeterminada por las siguientes especificaciones:

1. Una SEM o n-grama compuesta por un único componente solo designa una categoría. Como solamente nos interesan las construcciones sintácticas, este tipo de *n-gramas* quedó excluido.

2. Las SEM de más de cuatro integrantes tienen mayor probabilidad de coincidencia con frases completas.

Queríamos evitar que esto ocurriera, puesto que no era nuestro objetivo analizar la locución misma, sino sus componentes estructurales.

3. El número de constituyentes de la SEM repercute en el análisis estadístico de los datos. Los datos de n-gramas de más de cuatro integrantes procedentes de textos cortos contendrían muchas variables de valor cero que perjudicarían los resultados del análisis. Por este motivo nuestra decisión fue trabajar con las frecuencias de SEM de tipo *bigrama*, *trigrama* y *cuatrigrama*.

El corpus explotado para los fines de la evaluación del potencial discriminatorio de las SEM contenía dos subcorpus de textos de narrativa literaria (N) y un subcorpus de textos de artículos de opinión (AO) producidos por tres autores hispanohablantes nativos: Camilo José Cela, Mario Vargas Llosa y Eduardo Mendoza. En su selección se procuró que fueran autores que compartieran las mismas características sociolingüísticas y que acostumbraran a escribir en los mismos géneros[38]. La extensión de los textos para los artículos de opinión fue 1.500 palabras y 3.000 para los textos de novela. Para cada autor se escogieron 15 textos por género.

[38] Para más información relativa a los textos y los autores del corpus véase el Anexo I

Tabla 2. *Distribución del corpus según autor, género, extensión y número de muestras*

Autor	**Género Textual**	**Extensión de las muestras**	**Número de muestras**
Cela Mendoza Vargas Llosa	Artículo de opinión	1.500 palabras	15
	Narrativa literaria	3.000 palabras	15

Las ocurrencias de cada tipo de secuencia en cada documento del corpus de análisis fueron extraídas de forma automática con la ayuda de una serie de programas escritos en el lenguaje de programación AWK. Para ejecutar el programa de extracción de secuencias de etiquetas se siguieron los siguientes pasos:

a) Se especificó previamente la ruta del directorio donde se encuentra el archivo con los ficheros de los documentos del corpus de análisis;

b) Se indicó el nombre del fichero que contiene los nombres de los textos que han de ser tratados y las nuevas etiquetas que se deben asignar a los componentes de las secuencias[39];

[39] Se han introducido algunos cambios en las etiquetas originales del corpus de cara al manejo fácil y al procesamiento de los datos. En primer lugar se ha eliminado la segunda parte de las etiquetas que indica, por ejemplo, en el caso de los nombres el numero y el género del sustantivo en cuestión (N5-FS → N5) y en el de los verbos los de persona y número (VDR3P → VDR). En segundo lugar, se ha reducido el número de etiquetas que constituyen la lista de las variables

c) Se procedió a la creación de tres ficheros de texto distintos con los resultados de cada paso del procesamiento: uno que contiene la lista completa de ocurrencias de SEM o tipos de SEM; otro similar al primero, pero en el que aparece el número de unidades de ocurrencias de cada SEM; y un tercer fichero con las frecuencias calculadas de las SEM.

Este proceso se repitió para cada tipo de secuencias: bigramas, trigramas y cuatrigramas por separado. AWK sólo extrae las frecuencias del número de las SEM más frecuentes en todos los textos de todos los autores que se han predeterminado en el programa. Los resultados finales del estudio fueron obtenidos a partir de las primeras 100 secuencias de categorías o SEM más frecuentes en el subcorpus AO y el subcorpus N analizados por separado.

En este estudio preliminar sobre el potencial discriminatorio de determinadas estructuras en castellano recurrimos a un análisis innovador que combina dos técnicas clásicas en atribución de autoría, para obtener resultados más significativos. El *modus operandi* de análisis propuesto estriba en la combinación del análisis de componentes principales (ACP) y el análisis discriminante lineal (ADL): los componentes principales del ACP que explican la variabilidad en los grupos de datos, es decir, las

---

analizadas. Para este fin se ha creado un *script* que extrae las ocurrencias de cada variable de forma automática y las va sumando bajo la misma etiqueta (pj. N5-FS, N5-MP, N5-MS → N5).

dimensiones en las que se observa cierta estructura autorial se utilizan como variables en el ADL. La función del ADL toma la matriz que forman las puntuaciones de los componentes y crea un vector que indica cual de las observaciones existentes corresponde a cada caso en el análisis. En base de esta información se hace posible la clasificación de los documentos de análisis.

La representación gráfica de la clasificación generada a partir del análisis de los datos del corpus, que se ilustra en **Gráficos 1**, **2** y **3** (págs.66-68) para los textos de novela y en **Gráficos 4**, **5** y **6** (págs.68-69) para los textos de artículo de opinión, nos permite observar que tanto en el caso de los bigramas, como en el de los trigramas y cuatrigramas, la técnica discrimina con éxito entre los textos de los tres autores. Es aún más relevante, de cara a la aplicabilidad de la técnica en casos reales de autoría, el hecho de que el potencial discriminatorio de las SEM no parece disminuir en la clasificación del subcorpus AO que está compuesto por textos con un número de palabras reducido en 50% con respecto a los textos incluidos en el corpus N.

Gráfico 1. *Resultados del ADL de los textos de novela - bigramas*

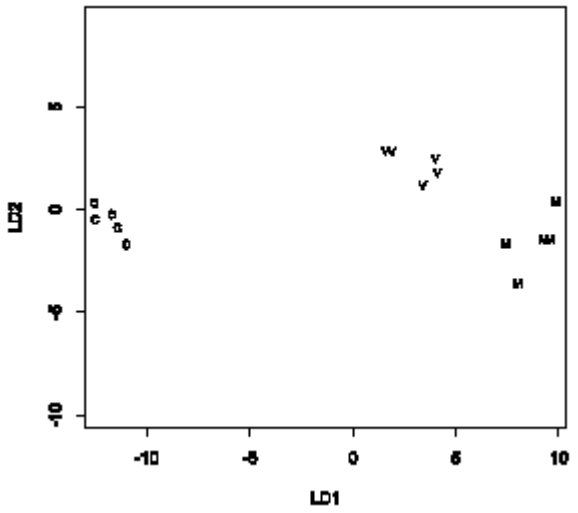


Gráfico 2. *Resultados del ADL de los textos de novela – trigramas*

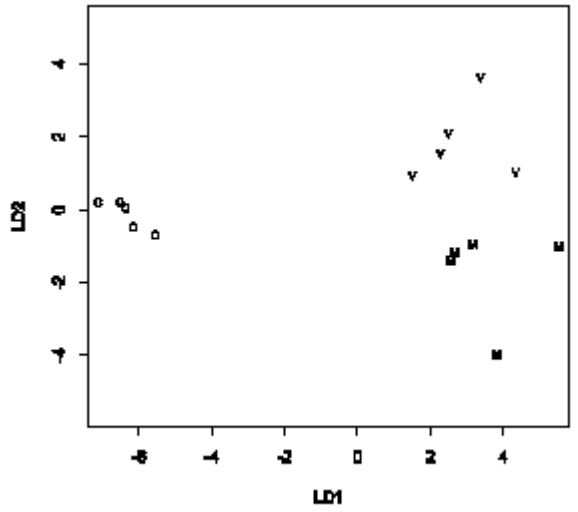


Gráfico 3. *Resultados del ADL de los textos de novela – cuatrigramas*

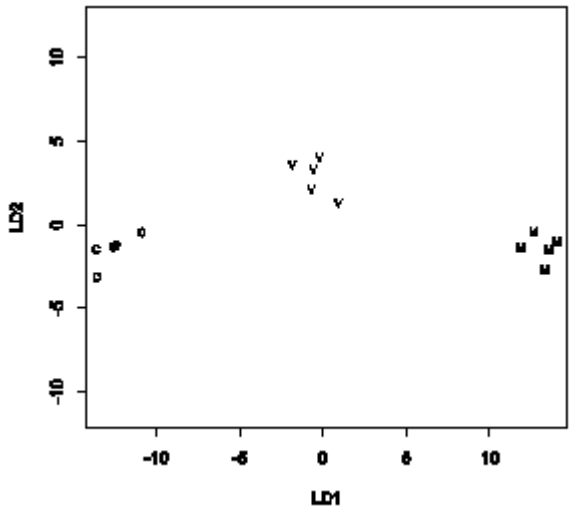


Gráfico 4. *Resultados del ADL de los textos de artículos de opinión – bigramas*

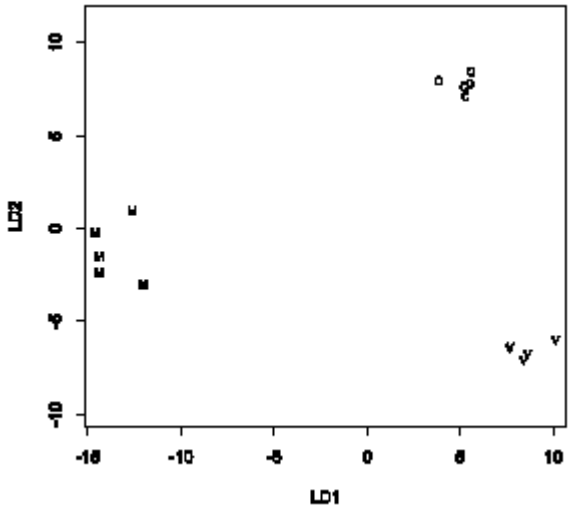


Gráfico 5. *Resultados del ADL de los textos de artículos de opinión – trigramas*

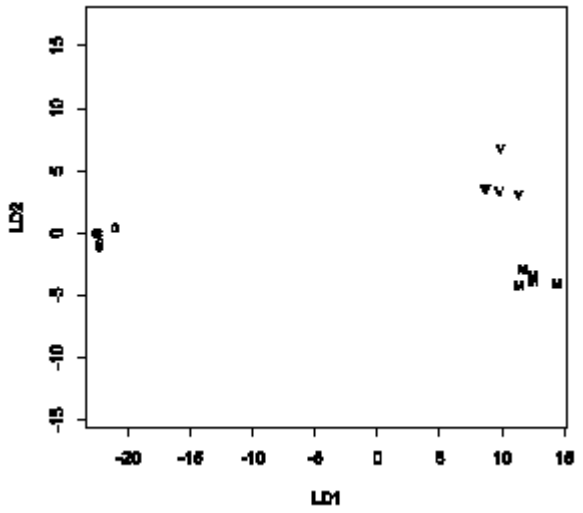
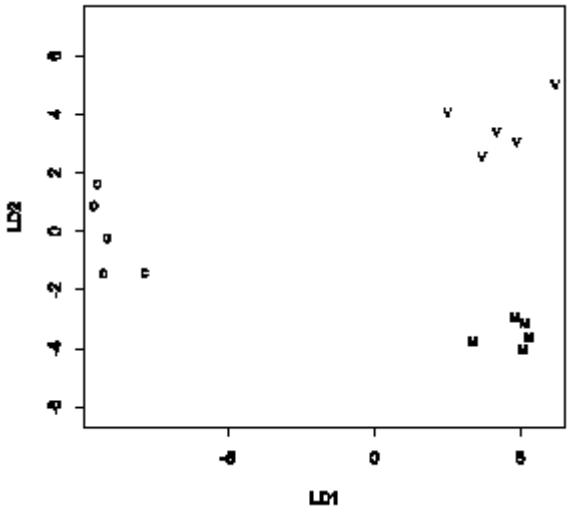


Gráfico 6. *Resultados del ADL de los textos de artículos de opinión – cuatrigramas*



Aparte de ofrecer la base para una primera estimación del valor atributivo de las SEM en el contexto de dos géneros textuales y corpus de diferente extensión, los resultados de los análisis se prestan también a una interpretación relativa a la similitud idiolectal y la homogeneidad estilística de cada escritor en su "uso"[40] de las SEM.

Los marcadores de las muestras de Mendoza y Vargas Llosa son próximos en su situación y ocupan la misma zona de los gráficos de manera recurrente, lo cual indica que estos dos autores probablemente comparten algunos rasgos estilísticos en el modo de construir sus frases y son menos idiosincrásicos en su uso de la sintaxis en comparación uno con otro. No obstante, se observa una determinada distancia entre los dos grupos de marcadores o textos lo que significa que, a pesar de la similitud entre la escritura de Mendoza y la de Vargas Llosa, hay construcciones y/o estructuras que estos escritores emplean en sus escritos de un modo distinto o con mayor o menor frecuencia y de forma relativamente constante.

Partiendo de la representación gráfica de los resultados del análisis del corpus, en cuanto al estilo de Cela podemos concluir que es mucho más homogéneo. Lo demuestra la proximidad en la distribución de los marcadores de sus textos que se da independientemente si se trata de artículos de opinión o fragmentos de novela. En este sentido Mendoza y Vargas Llosa son más

[40] Se puede hablar de uso solo en un sentido figurativo, ya que las SEM no corresponden a unidades lingüisticas empleadas de manera consciente por el autor.

heterogéneos. La distancia dentro del conjunto de marcadores que designan sus textos es mayor y se puede explicar con la mayor variación intra autor[41] que caracteriza estos dos escritores. Podemos decir también que el idiolecto de Cela es notablemente dispar al de los otros escritores cuyos textos hemos analizado.

La información que hemos recabado de la revisión de los resultados del análisis del corpus nos sirve para diseñar y llevar a cabo de manera óptima la evaluación preliminar de la técnica de atribución de autoría que describimos a continuación.

La evaluación de la técnica de identificación de autor mediante las secuencias de categorías de mayor frecuencia de uso implica, antes de la elaboración de un experimento cualificativo que ponga a prueba el valor discriminatorio de las marcas en cuestión, el desarrollo de un test que permita al investigador asegurarse de que los datos de análisis no han sido conformados en el tratamiento automático. Es muy probable que en el trabajo con valores aproximados el programa de procesamiento de datos ajuste los valores de algunas variables, lo cual llevaría a una discrepancia en los resultados finales.

[41] Véase el capítulo 2 para una breve definición del concepto y el capítulo 4 para una discusión más extensa sobre el tema de la variación autorial

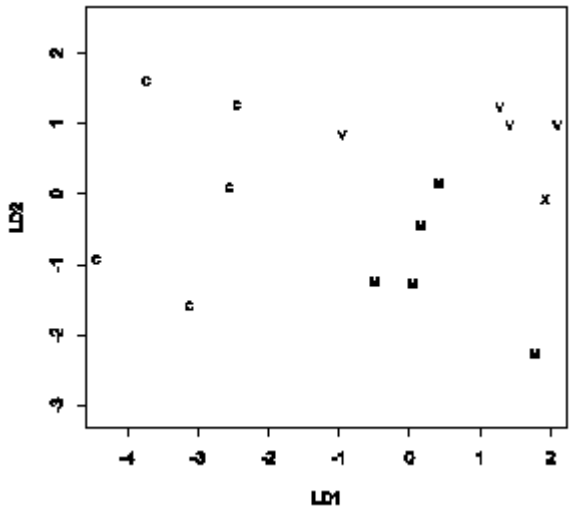
El test que se diseñó con el objetivo de consolidar la validez de los resultados del análisis estadístico realizado con el programa R[42] ha consistido en ocultar la autoría de uno de los textos del subcorpus de artículos de opinión para luego volver a ejecutar el análisis. Los textos de artículos de 1.500 palabras comparten más similitudes estilísticas con el género de los textos delictivos (cartas de amenaza, de extorsión, etc.) que los textos de novela, y de ahí que sean más apropiados para ser empleados como corpus de prueba.

Los parámetros de la réplica del análisis del subcorpus AO han sido los mismos que se han fijado en los análisis previos a la evaluación. La selección de las secuencias de categoría de tipo trigramas como marca identificativa, tanto en este test, como en las pruebas evaluativas del método, se justifica por el hecho de que en el análisis previo del corpus los trigramas no cotizan con valores negativos en las funciones del ADL, es decir, los trigramas más frecuentes que comprenden los componentes principales de la variable presentan menos casos de ocurrencias cero.

En el **Gráfico 7** se visualizan los resultados del análisis discriminante y la clasificación del texto anónimo, (X), según autor.

[42] Free software environment for statistical computing and graphics (http://cran.r-project.org/)

Gráfico 7. *Clasificación del texto X mediante la técnica de las SEM – trigramas*



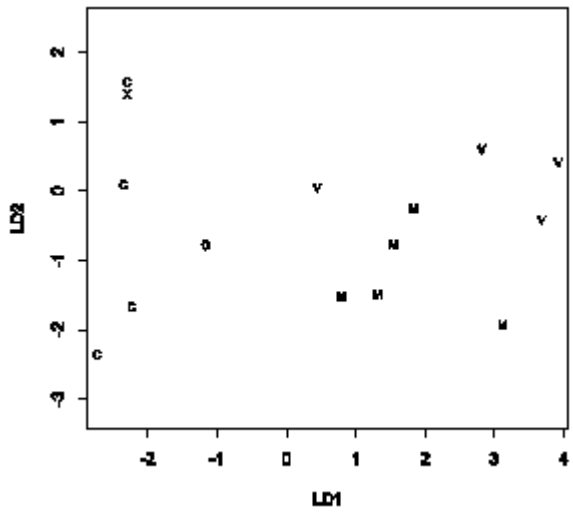
Como se puede observar, la nitidez del agrupamiento entre los marcadores de los textos de cada autor resulta ligeramente distorsionada sobre todo entre los documentos de Mendoza y Vargas Llosa, pero aun así son perceptibles los límites entre los grupos de textos. El texto de autor oculto, que aparece marcado en el gráfico con el marcador X, se sitúa en la zona de los artículos escritos por Vargas Llosa. Dado que Vargas Llosa es el escritor a quien pertenece la autoría del texto, se puede concluir que, aunque los datos hayan sufrido algunos ajustes en el análisis automático, estos han sido mínimos y no afectarán los resultados.

Una vez confirmado que los datos de SEM no experimentan modificaciones de los valores en el análisis automatizado, procedimos a la evaluación de la técnica de identificación de autor mediante las secuencias de categorías tipo trigramas de mayor

frecuencia. La evaluación consistió en un experimento de atribución de autoría de diseño similar al test previo que se acaba de describir. Esta vez el documento que introducimos como anónimo fue un artículo de opinión de Cela que no está incluido en el corpus de análisis.

Como se puede observar en el **Gráfico 8**, el análisis discriminante lineal con los datos añadidos del texto nuevo, determinó correctamente la autoría del texto anónimo clasificándolo en el grupo de los documentos producidos por el escritor Cela, que ocupan la parte izquierda del gráfico.

Gráfico 8. *Atribución de autoría del texto X mediante la técnica de las SEM – trigramas*



Por tanto, las secuencias de categorías morfosintácticas (SEM) han dado resultados muy prometedores en el análisis estadístico y el experimento al que fueron sometidas con el fin de determinar su

valor atributivo. A partir de los resultados finales, pudimos concluir que los n-gramas de dos, tres y cuatro componentes pueden ser empleados para detectar diferencias específicas en el uso de la sintaxis por parte de diferentes autores y, por lo tanto, son marcas válidas de atribución de autoría. Su potencial discriminatorio, además, no parece disminuir a causa del efecto de factores como la reducción del número de palabras[43] y la franja temporal de producción de los textos de análisis.

Con el estudio piloto de atribución de autoría forense mediante n-gramas de secuencias de etiquetas morfosintácticas (Spassova y Turell, 2007) concluimos el panorama general de los estudios más relevantes realizados en autoría que han influido de alguna manera en nuestro trabajo. Las mayores aportaciones provienen de Baayen (1996), en la elaboración del modelo experimental de técnica de atribución de autoría que proponemos; de Hoover (2003), que nos ha dado la idea inicial de centrar el objeto de estudio de la tesis en las marcas de carácter sintáctico; y de Grant y Baker (2001), con los comentarios valiosos sobre las principales cuestiones metodológicas de la disciplina.

[43] En las pruebas estadísticas las SEM juntan los textos por autor en grupos bien separados con la misma precisión, tanto en el análisis de los textos de subcorpus N (3.000 pl.), como en el de subcorpus AO (1.500 pl.).

3. METODOLOGIA

3.1 Objeto de estudio

Una de las asignaturas pendientes en el campo de la comparación forense de textos escritos para la atribución forense de autoría en español estriba en establecer qué elementos lingüísticos (unidades y estructuras) responden a los criterios de selección de marcas identificativas y son aptos de aplicación en la pericia de pruebas lingüísticas.

En esta tesis abordamos este problema con el estudio de un tipo de unidad de carácter sintáctico candidata a ser una marca de identificación, las secuencias de categorías gramaticales. Específicamente tratamos de determinar la aplicabilidad y fiabilidad de las secuencias de categorías gramaticales de dos y tres componentes como marcas de autoría mediante la evaluación estadística de su potencial discriminatorio. Llevamos a cabo esta evaluación en dos contextos diferentes: por un lado, en un corpus compuesto por textos de narrativa y, por otro, en el análisis lingüístico forense de los textos de dos casos reales. Este procedimiento nos permitirá cumplir con un doble propósito: corroborar la viabilidad del método de análisis y confirmar la

independencia de la capacidad discriminatoria de los n-gramas del género textual.

3.2 Variables de análisis

a) Variables dependientes

En los estudios empíricos de la investigación que comporta esta tesis doctoral nos hemos centrado en el análisis de los bigramas (secuencias de dos componentes) y los trigramas (secuencias de tres componentes) y no en la totalidad de variantes de n-gramas que pueden generarse a partir de un texto. La selección de estos dos tipos de SEM no se ha hecho al azar, sino que se basa en los resultados del estudio preliminar sobre el potencial discriminatorio de una serie de n-gramas de más amplio espectro que nos han permitido excluir algunas de las variables[44].

Recordemos del capítulo 2 que los términos SEM y n-grama (bigrama, trigrama, etc.) designan las secuencias de categorías gramaticales[45]. En una frase anotada, a cada unidad lingüística que la constituye se le asigna una categoría representada mediante una

[44] Véase el apartado 3 del capítulo 2 para la descripción del estudio y sus resultados.

[45] Es importante matizar aquí que el término SEM es el que hemos adoptado en esta tesis para referirnos a las secuencias de unidades generalmente definidas como n-gramas, ya que, como veremos más adelante, en el marco de nuestro trabajo el término *n-grama* se usa con otro significado.

etiqueta. Nuestras variables, por lo tanto, son en realidad secuencias de etiquetas. A pesar de esta definición clara, el concepto de SEM puede resultar complejo, sobre todo para el que no está familiarizado o carece de experiencia en el trabajo con corpus anotados. A continuación ofrecemos un ejemplo de las secuencias que se producen en el análisis de una cita de Blaise Pascal, que ayudará a entender mejor la naturaleza de las SEM.

Plus je vois l´home, plus j'aime mon chien.
Cuanto más conozco al hombre, más quiero a mi perro.[46]

Tabla 3. *Ejemplo de una frase anotada con etiquetas de categoría*

Cuanto	más	conozco	a	el	hombre	,	más	quiero	a	mi	perro	.
ER-MS	D	VDR1S-	P	AMS	N5-MS		D	VDR1S-	P	JP116S	N5-MS	

La Tabla 3 muestra las partes del habla de las que está compuesta la frase y las respectivas etiquetas morfosintácticas[47] que le han sido asignadas por el programa de anotación y que aparecen debajo de cada unidad en la fila inferior de la tabla. Posponemos el comentario de las etiquetas y del proceso de anotación para el capítulo metodológico[48].

[46] Traducción libre del francés.
[47] Para la lista completa de etiquetas morfosintácticas consulte el anexo II.
[48] Véase el capítulo 3.

- *Extracción de las variables*

El proceso de extracción consiste de dos etapas consecutivas. La primera etapa consiste en la reducción de las etiquetas y la segmentación de los textos etiquetados en secuencias o SEM de tipo bigrama y trigrama. Denominamos a las SEM variables base, ya que a partir de ellas se forman las variables definitivas, en la segunda etapa de extracción. En esta segunda etapa, mediante un sistema de agrupación de variables automatizado gobernado por reglas específicas, que describiremos más abajo junto con sus criterios de creación, las variables base de mayor frecuencia en el corpus se combinan y se convierten en nuevas variables.

- *Reducción de las etiquetas*

Antes de explicar el método de reducción de las etiquetas cabe mencionar algunas de las características del etiquetario usado para la anotación de los textos para explicar el porque de la modificación en la forma original de las etiquetas que lo constituyen.

El etiquetario del IULA está diseñado de manera que cada etiqueta proporcione información tanto acerca de la categoría gramatical como de las propiedades morfológicas de cada elemento lingüístico que se somete al proceso de anotación. Para los fines de nuestra investigación, no obstante, son importantes sólo las categorías a las que pertenecen las palabras funcionales y léxicas que contiene un

texto. Así que con el propósito de facilitar la extracción automatizada de los n-gramas, hemos eliminado la parte de las etiquetas que no es relevante para nuestro trabajo. Sin embargo, hay que tener en cuenta que el etiquetario del IULA está compuesto por dos tipos de etiquetas : por un lado, aquellas que incorporan una división interna entre las partes denominadoras de función gramatical y morfológica (mediante un guión o dos) y, por otro, etiquetas sin esta división. En los casos en los que no hay división, se ha omitido la reducción automática. Véase la tabla 4 para algunos ejemplos de etiquetas de los dos tipos.

Tabla 4. *Ejemplos de los dos tipos de etiquetas de anotación*

División	**Etiqueta**	**Designador de categoría**	**Designador de morfología**
sí	*N5-MS*	**N5** – nombre sustantivo	**MS** – masculino singular
	JQ -- FP	**JQ** - adjetivo	**FP** – femenino plural
no	*AFS*	**A -** artículo	**FS** – femenino singular
	VDP2S-	**V -** verbo	**DP2S** – indicativo, imperfecto, 2ª p.,sg.

A modo de ejemplo, la tabla 5 muestra como se representaría la cita de Pascal después de que se aplicara la reducción de etiquetas.

Tabla 5. *Ejemplo de una frase anotada después de la reducción de etiquetas*

Cuanto	más	conozco	a	el	hombre	,	más	quiero	a	mi	perro	.
ER	D	VDR1S	P	AMS	N5		D	VDR1S	P	JP116S	N5	

Las combinaciones de las etiquetas reducidas según el método que acabamos de describir, en secuencias de dos o tres, constituyen nuestras variables base en la etapa inicial del proceso de extracción, que culmina con la segmentación de los textos en n-gramas.

- *Segmentación de los textos en SEM*

La segmentación de los textos en SEM sigue determinadas pautas que vienen descritas en el apartado correspondiente del capítulo metodológico, de modo que, sin entrar en mayores detalles, explicaremos el proceso de segmentación con dos ejemplos, uno para cada tipo de n-grama, a partir de la misma frase célebre de Pascal que hemos venido empleando. Como se puede ver en las tablas 6 y 7, la segmentación progresa en escala, siendo el último elemento del n-grama precedente (o los últimos dos en el caso de los trigramas) el primer elemento del n-grama consecutivo en orden lineal.

Tabla 6[49]. *Ejemplo de la segmentación de una frase en bigramas*

1	2	3	4	5	6	7	8	9	10	11	12	13	Bigrama
Cuanto	más												ER.D
	más	conozco											D.VDR1S
		conozco	a										VDR1S.P
			a	el									P.AMS
				el	hombre								AMS.N5
						,							-
							más	quiero					D.VDR1S
								quiero	a				VDR1S.P
									a	mi			P.JP116S
										mi	perro		JP116S.N5
												.	

Tabla 7[50]. *Ejemplo de la segmentación de una frase en trigramas*

1	2	3	4	5	6	7	9	10	11	12	13	14	Trigrama
Cuanto	más	conozco											ER.D.VDR
	más	conozco	a										D.VDR.P
		conozco	a	el									VDR.P.AMS
			a	el	hombre								P.AMS.N5
						,							-
							más	quiero	a				D.VDR.P
								quiero	a	mi			VDR.P.JP116S
									a	mi	perro		P.JP116S.N5
												.	-

El objetivo de aplicar este procedimiento es evitar la pérdida de información (procediendo de otra forma, es decir, realizando una segmentación sin solapamiento, posibles SEM que pueden ser idiosincrásicas y significativas para el análisis forense de autoría podrían ser omitidas en la extracción) y de optimizar el proceso de extracción de variables.

A continuación presentamos las listas de las 10 SEM de tipo bigrama y las 10 SEM de tipo trigrama más frecuentes que han sido

[49] La última columna del lado derecho de la tabla contiene las SEM que se producen tras la segmentación de la frase. Las unidades lingüísticas correspondientes a cada etiqueta formante de la SEM aparecen en las columnas precedentes.

generadas en la segmentación de los textos de nuestro corpus de análisis[51].

Tabla 8. *Lista de las primeras 10 SEM de tipo bigrama más frecuentes en el corpus con sus equivalencias, número de ocurrencias y valor porcentual*

BIGRAMA	EQUIVALENCIA EN EL TEXTO	Nº DE OCURRENCIAS	%
N5.P	[...] *asomo de* [...] [...] *pasión por* [...]	15143	*6,61*
P.N5	[...] *en forma* [...] [...] *sin riesgo* [...]	8324	*3,68*
AMS.N5	[...] *el final* [...] [...] *el alcohol* [...]	8192	*3,62*
AFS.N5	[...] *la vida* [...] [...] *la soledad* [...]	7885	*3,48*
N5.JQ	[...] *aire libre* [...] [...] *mirada ardiente* [...]	6208	*2,74*
P.AMS	[...] *en el* [...] [...] *a-el* [...]	5842	*2,58*
P.AFS	[...] *de la* [...] [...] *con la* [...]	5051	*2,23*
N5.C	[...] *pobreza y* [...] [...] *huevo o* [...]	4392	*1,92*
P.VI	[...] *para alquilar* [...] [...] *a decir* [...]	4238	*1,87*
JQ.N5	[...] *largas explicaciones* [...] [...] *saludable ejercicio* [...]	3413	*1,50*

[51] Véase el anexo III para una lista de las 100 SEM bigramas y las 100 SEM trigramas de mayor frecuencia en el corpus.

Tabla 9. *Lista de las primeras 10 SEM de tipo trigrama más frecuentes en el corpus con sus equivalencias, número de ocurrencias y valor porcentual*

TRIGRAMA	EQUIVALENCIA EN EL TEXTO	Nº DE OCURRENCIAS	%
P.AMS.N5	[...] *en el desierto* [...] [...] *a-el fondo* [...]	4537	2,34
P.AFS.N5	[...] con la cabeza [...] [...] entre la decencia [...]	4256	2,20
N5.P.N5	[...] *martirio de sopa* [...] [...] *cosa de broma* [...]	3994	2,06
AMS.N5.P	[...] *el universo con* [...] [...] *el día en* [...]	2605	1,34
AFS.N5.P	[...] *la piel por* [...] [...] *la magia de* [...]	2575	1,33
N5.P.AMS	[...] *tentación de el* [...] [...] *renta para el* [...]	2189	1,13
P.N5.P	[...] *de vergüenza por* [...] [...] *a pie en* [...]	1908	0,98
N5.P.AFS	[...] *hielo de la* [...] [...] *cena con la* [...]	1900	0,98
N5.JQ.P	[...] *pasiones efímeras por* [...] [...] *fuego invisible de* [...]	1447	0,74
P.AMP.N5	[...] *de los hombres* [...] [...] *por los espacios* [...]	1411	0,72
P.AMS.N5	[...] *en el desierto* [...] [...] *a-el fondo* [...]	4537	2,34

– *Agrupación de las SEM*

La segunda etapa de la extracción de las variables de análisis implica agrupar las variables base de cada tipo, SEM bigrama o trigrama. Ha sido necesario recurrir a la agrupación de SEM a causa del elevado número de variables resultantes de la extracción (véase Tabla 10).

Tabla 10. *Número de trigramas y bigramas extraídas del corpus según el tipo de texto*[52]

Tipo de SEM	**Subcorpus AO**		**Subcorpus N**	
	Nº de palabras	**Nº de realizaciones**	**Nº de palabras**	**Nº de realizaciones**
bigramas	127382	**42000**	266922	**8750**
trigramas		**54000**		**11250**

– *Criterios de agrupación*

El hecho de trabajar con tantas variables puede tener diversas repercusiones negativas en la realización correcta del análisis estadístico de los datos. En primer lugar, la matriz de varianza y covarianza que forman las variables sería demasiado grande y sus

[52] El número de realizaciones de SEM corresponde al recuento hecho tomando las 300 primeras palabras de cada artículo y las primeras 600 de los fragmentos de novela, en los casos en los que su extensión era superior a este máximo.

principios estadísticos[53] que deben seguir para obtener una estimación[54] correcta de los coeficientes, es decir, de las variables que se toman para el análisis. Esto llevaría a que el análisis no fuera significativo y sus resultados fueran trucados.

En segundo lugar, la multitud de variables puede ocasionar problemas de multicolinearidad en el análisis. La multicolinearidad consiste en la correlación o, en otras palabras, en la relación de dependencia entre las variables analizadas. Cuanto más alta es la correlación entre las variables, tanto más aumenta la dificultad de delimitar los efectos de las variables individuales. A consecuencia de ello es posible que los resultados del análisis sean poco fiables y no confirmen ni refuten nuestras hipótesis.

Para evitar esta clase de problemas a la hora de ejecutar el análisis estadístico de nuestros datos hemos preferido la agrupación de las variables a la opción alternativa de excluir parte de ellas y así reducir el conjunto de datos. Los motivos para esta elección son dos. Por un lado, porque eliminar directamente las variables de menor (o poca) frecuencia conllevaría el riesgo de suprimir alguna que podría ser significativa. Por otro lado, porque la técnica estadística de reducción de variables, ACP, que hemos usado en los estudios preliminares sobre SEM para llevar a cabo este proceso ha resultado no ser efectiva, puesto que genera demasiados

[53] Los principios estadísticos a los que debe responder un estimador son consistencia y suficiencia, eficiencia, y se refieren al valor del estimador que además no debe ser sesgado.
[54] Es el valor numérico que toma el estimador sobre la muestra seleccionada.

componentes principales, y su ejecución sobre datos de carácter lingüístico supone algunas carencias metodológicas. Estas carencias radican sobre todo en el método de agrupación que se rige únicamente por los valores cuantitativos de las variables dejando a un lado su significado, es decir, su valor lingüístico. Aunque la combinación arbitraria de SEM está justificada desde el punto de vista de la estadística, donde son decisivos el valor numérico y la distribución de las variables, en la investigación lingüística ésta resulta inaceptable e incluso errónea.

Con el fin de conformar el diseño experimental de la tesis doctoral cumpliendo con los requisitos de ambas disciplinas, agrupamos las SEM de cada tipo según unas reglas específicas, que a su vez establecemos a partir de dos criterios:

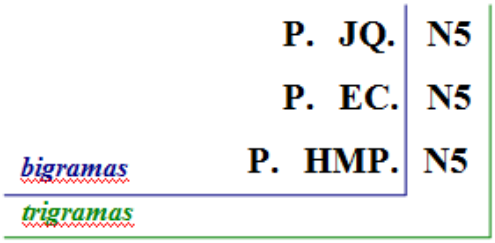
a) la función gramatical de los componentes de la SEM;
b) la distribución de cada SEM en el corpus.

El criterio de agrupación según la función gramatical dicta que se pueden combinar en el mismo grupo exclusivamente las SEM, sean estos bigramas o trigramas, cuyos elementos integrantes pertenecen a la mismas categorías o cumplen funciones gramaticales análogas. De acuerdo con este criterio, las SEM que aparecen en el cuadro 2 se pueden combinar aunque su segundo componente en cada caso sea de diferente categoría. Las etiquetas E, H y J representan un especificador[55], un adjetivo y un adjetivo deverbal, y se pueden

[55] La categoría especificador a la que designa la etiqueta E representa aquellas categorías funcionales que pueden ser tanto adjetivas, como pronominales.

agrupar por ser todas categorías que desempeñan la función de calificar o determinar el sustantivo.

Cuadro 2. *Ejemplo de la aplicación del criterio lingüístico de agrupación de SEM*



El segundo criterio tiene que ver con la frecuencia de los n-gramas en los textos del corpus. Para determinar si es viable la agrupación de dos o más SEM, contrastamos la distribución de cada SEM en los textos de todos los autores del corpus con la distribución de la SEM o las SEM con las que queremos agruparla o agruparlas (véase Tabla 11). Si la distribución revela valores más o menos constantes y próximos en la comparación intra autorial, consideramos que las variables se pueden agrupar. Cómo es lógico, la agrupación sólo se efectúa en caso de que respete el criterio lingüístico presentado anteriormente.

El cuadro de comparación de la Tabla 11 muestra un ejemplo concreto de la aplicación de los criterios de agrupación. Podemos ver que la distribución de los dos bigramas que deseamos combinar (P.AMS y P.AFS) en una nueva variable (PAS), que denominamos n-grama, se expresa en valores aproximados de frecuencia en todos los autores del ejemplo, cumpliendo así con el criterio de la

distribución. Esta agrupación responde también al criterio lingüístico, ya que se combinan SEM cuyos componentes coinciden en la categoría gramatical.

Tabla 11. *Cuadro de comparación de la distribución de los bigramas P.AMS y P.AFS*

SEM	**Autor 1**	**Autor 2**	**Autor 3**	**Autor 4**	***Autor 5***
P.AMS	*69*	*54*	*58*	*83*	*42*
P.AFS	*73*	*68*	*68*	*77*	*54*
Nueva variable *PAS*	**142**	**122**	**126**	**160**	**96**

La aplicación de este criterio nos permite no sólo reducir el número de variables, sino también ver como se distribuyen entre los autores y hacer algunas previsiones sobre el potencial discriminatorio de determinadas SEM. Si la diferencia en la distribución de una SEM en los textos de un autor en comparación con la que se observa en otro autor es muy alta, entonces podemos concluir que la SEM en cuestión es una variable significativa y probablemente discriminante para el análisis.

Por último, mantenemos como variables individuales en su forma original las SEM que no pueden ser agrupados porque no cumplen con el criterio lingüístico, pero sí con el criterio de distribución. No

obstante, si la distribución de la SEM es la misma en todos los autores (es decir, no cumple con el criterio de distribución) y el criterio lingüístico impide su agrupación con otra SEM, la eliminamos de la lista de variables porque no posee valor discriminante.

– *Reglas de agrupación*

Siguiendo los criterios que acabamos de definir, hemos fijado una serie de reglas específicas necesarias para que el proceso de agrupación se pueda producir correctamente y de forma automática, con la ayuda de una herramienta informática que hemos adquirido para la extracción de los n-gramas[56]. La finalidad de estas reglas es ante todo posibilitar la aplicación simultánea de los criterios en la agrupación de las SEM en todo el corpus y minimizar el tiempo de extracción de datos, mediante su incorporación en el mecanismo de búsqueda y extracción de variables.

Establecemos las reglas partiendo de la agrupación de las variables base extraídas del subcorpus de textos literarios. Hemos de notar que el conjunto de datos no incluye las variables que no aportan información para el análisis, ya que como hemos dicho antes pueden perjudicar su realización correcta. Estas son todas las variables con frecuencia menor a 20 ocurrencias en el corpus, que han sido eliminadas antes de proceder a la agrupación.

[56] Véase el apartado 5 del capítulo 3 para la descripción de las características y anexo IV para el funcionamiento de este programa.

Tomamos como modelo base para la formulación de las reglas los datos del subcorpus de fragmentos de novela porque los textos que lo componen son más extensos que los del subcorpus de artículos de opinión[57]. Como hemos explicado en los apartados anteriores, la longitud del texto es proporcional a las ocurrencias de SEM de tipo bigrama y trigrama. Es decir, cuanto más larga sea la muestra textual tanto más alto es el número de n-gramas obtenidos en la fase de extracción. Por lo tanto, el hecho de trabajar con textos más extensos nos brinda la posibilidad de detectar una mayor variedad de n-gramas que podrían ocurrir en el universo de textos escritos en lengua española.

El diseño de las reglas de agrupación es bastante rudimentario, pero ha sido nuestra intención mantenerlo lo más simple posible, conforme a los propósitos metodológicos de esta tesis y las limitaciones que pueden tener la búsqueda y la extracción automática de n-gramas. Con el análisis cuantitativo de los n-gramas nos proponemos en última instancia revelar las estructuras sintácticas del español que representan y que podrían tener un carácter idiosincrásico y discriminar entre las producciones escritas de los usuarios de la lengua española. Para ello, una vez hayamos concluido el análisis de los distintos grupos de n-gramas y hayamos detectado los grupos de mayor potencial discriminatorio será preciso "disolver" dichas agrupaciones de variables para localizar las equivalencias textuales de sus integrantes y estudiar su aspecto

[57] Para la descripción detallada del corpus de análisis véase el capítulo 4.

lingüístico. Esta tarea es más fácil de realizar si las reglas de agrupación que usamos no son demasiado complejas.

En lo que se refiere a la extracción de datos, la simplicidad de las reglas reduce la probabilidad de errores de extracción automatizada, que suele ser mucho mayor cuando se emplean reglas sofisticadas que combinan varios patrones de búsqueda a la vez. Además, dado que el programa de extracción es un prototipo de herramienta con perspectivas de desarrollar más sus prestaciones en el futuro[58], nos conviene que el diseño de las actuales opciones de búsqueda, de las que forman parte las reglas de extracción, se mantenga a un nivel de complejidad que nos permita introducir modificaciones sin tener que alterar todo el sistema de interrogación de la base de datos.

- *Breve descripción de las reglas de agrupación*

Las reglas de agrupación se pueden dividir en 12 grupos según la categoría del elemento inicial de la secuencia, siendo válida esta división para ambos tipos de SEM: los bigramas y los trigramas. Para denominar cada grupo empleamos la primera letra de la etiqueta del elemento que encabeza la SEM, con la excepción de aquellos casos en los que hay más de una etiqueta que comienza por la misma letra, en los que se toma como denominador la etiqueta completa (véase Tabla 12). Esta división en grupos tiene como

[58] Nos referimos a la posibilidad de introducir nuevas prestaciones que permitan la atribución forense de autoría mediante el método de los n-gramas de forma automática.

único objetivo facilitar la tarea de catalogar las reglas en la base de datos.

Tabla 12. *Lista de los grupos de reglas con la correspondencia categorial de cada denominador*

Sigla(s) etiqueta	**Categoría gramatical**
A	***Artículo***
ANS	*Artículo neutro "LO"*
C	*Conjunción*
D	*Adverbio*
E	*Especificador*
J	*Adjetivo*
H	*Adjetivo deverbal*
N	*Sustantivo*
P	*Preposición*
R	*Pronombre*
V	*Verbo*
VI[59]	*Forma deverbal*

Las siglas de Tabla 12, además de designar su tipología, nos sirven para formular las reglas de agrupación. En la estructura de las reglas, las siglas adquieren la función de sustitutos de las etiquetas completas de los componentes de la SEM. Al sustituir las etiquetas

[59] Para distinguir las formas no personales del verbo del resto, en la tabla 12 la etiqueta VI se usa como denominador común para referirse al participio, al gerundio y al infinitivo, aunque en el etiquetario designa solo la categoría gramatical de verbo en infinitivo.

por sus siglas correspondientes, evitamos crear reglas que se repitan para cada variante de la SEM. El programa de extracción precisa de más parámetros, aparte de la letra inicial, para identificar, entre los componentes de las SEM, las etiquetas que no están constituidas por una sola letra, por ejemplo "C" (conjunción), por lo que el resto de siglas de dichas etiquetas vienen representadas en las reglas por asteriscos (*) (por ejemplo R6EZZZZ – R******) (véase también la Tabla 13, ej.4). La única etiqueta cuyas siglas introducimos sin modificaciones es la etiqueta del pronombre neutro "lo" (ANS), por su semejanza con las etiquetas del artículo definido en singular (A*S).

Frente a la necesidad de agrupar las SEM que pertenecen al mismo grupo, pero que los criterios de agrupación determinan como formantes de variables diferentes, recorremos a la sustitución solo de algunas de las letras por asteriscos (*), y mantenemos aquellas que designan propiedades diferenciales de las unidades que representan las etiquetas. En la Tabla 13 podemos ver algunos ejemplos de este procedimiento: en el ejemplo 1 se conserva la letra de la etiqueta que representa el número; en el ejemplo 2, la que denota el género y, en el 3, la del tiempo verbal, etc.

Tabla 13. *Ejemplos de la formulación de algunas reglas de agrupación*

№ ej.	SEMs	Regla de agrupación
1.	AMS.N5 […] *el vicio* […]	A*S + N5
	AFS.N5 […] *la ira* […]	
2.	AMS.N5.D […] *el rasgo más* […]	AM* + N5 + D
	AMP.N5.D […] *los reyes no* […]	
3.	AFS.N5.V8R6S […] *la memoria pierde* […]	A*S + N5 + V*R***
	AMS.N5.VDR3S […] *el problema es* […]	
4.	VI. REE636S.P […] *perdonar-le a* […]	V* + R******* + P
	VI. REEC3FS. P […] *abrazar-la con* […]	

Como conclusión de esta breve descripción de las reglas de agrupación (que ampliaremos en el capítulo metodológico de la tesis), cabe especificar que el total de reglas formuladas para la

agrupación de las SEM de tipo bigrama ha sido 502 y para las de tipo trigrama, 2417[60].

Las reglas de agrupación comprenden opciones combinatorias específicas. Estas opciones dependen de la casuística de categorías consecutivas dentro de una SEM tras la segmentación de los textos del corpus de análisis en SEM, y también del orden de la estructura oracional de la lengua en la que están escritos dichos textos. De ahí que las reglas que empleamos aquí son exclusivamente válidas para la agrupación de variables base generadas a partir de documentos en español.

Por último, las variables definitivas, resultado del proceso de agrupación, que explotamos en los análisis estadísticos de los experimentos que llevamos a cabo en esta tesis aparecen en el anexo 5. En la Tabla 14 contigua se muestran algunos ejemplos de los n-gramas de cada tipo.

Tabla 14. *Modelo de las variables de análisis definitivas*

N-grama	**Bigramas**	**Trigramas**
1.	DVR	PAV
2.	EAP	PAVI
3.	RAN	AECP
4.	VVP	CNA

[60] Las reglas de agrupación se listan en los anexos V.

Los nombres de los n-gramas son combinaciones de las siglas iniciales de las etiquetas que integran las SEM (variables base) mediante cuya agrupación han sido formados. Por ejemplo el bigrama 3 de la tabla 14, EAP, representa la combinación de las SEM que contienen en su estructura las etiquetas de especificador (E) y artículo en plural (A). Cuando la coincidencia de siglas ha hecho imposible diferenciar entre variables, hemos añadido letras para facilitar su distinción. Los trigramas PAV y PAVI ejemplifican la aplicación de este procedimiento en la nomenclatura de los n-gramas. La "I" del nombre del n-grama PAVI que designa los componentes preposición (P), artículo (A) y forma del verbo (V*) , sirve para distinguirlo del PAV (preposición + artículo + verbo conjugado (V****)).

En este apartado hemos descrito las variables dependientes y hemos aclarado los aspectos de su naturaleza que podrían resultar ambiguos por ser estas representaciones no sustantivas de estructuras y construcciones lingüísticas. A continuación, describiremos aquellas variables que pueden restringir el comportamiento de los n-gramas como unidad de análisis y también, por lo tanto, su carácter discriminatorio como marcas idiosincrásicas. Nos referimos a las variables independientes.

b) Variables independientes

En la investigación que refleja esta tesis y, por lo general, en la investigación en atribución forense de autoría, las variables independientes que resultan relevantes y que es preciso tener en cuenta son aquellas cuya presencia conllevaría cambios en el idiolecto escrito de una persona (variables que dependen del individuo), y aquellas que podrían dificultar la ejecución correcta del análisis lingüístico forense o incluso imposibilitarlo (variables que dependen del texto).

A continuación, definimos cada tipo de variable independiente y explicamos su influencia en la práctica de comparación lingüística forense para la atribución de autoría de textos escritos. En la parte empírica de la tesis (capítulos 5 y 6), estudiamos el grado en que el valor idiosincrásico y la capacidad discriminatoria de los n-gramas se ven influidos por cada una de las variables independientes que reseñamos.

– *Variables ligadas al autor*

Las variables ligadas al autor son aquellas que están relacionadas con las características específicas del individuo, y que se considera que podrían influenciar el análisis lingüístico por suponer cambios y diferencias estilísticas. De las variables que comprende este grupo,

en este trabajo exploramos solo la variedad lingüística y su efecto en el estilo escrito.

- *La variedad lingüística del autor*

Variedad lingüística es el término que se ha adoptado en lingüística para denominar las diferencias lingüísticas entre los individuos usuarios de la misma lengua sin incurrir en el uso del término *dialecto* que a menudo puede resultar ambiguo e incluso peyorativo[61]. En el marco de este trabajo nos centramos exclusivamente en las diferencias en el uso del lenguaje que ocurren a causa de la distancia geográfica entre los hablantes, es decir, en la variedad geográfica. Sin embargo, para referirnos a este tipo de variedad utilizaremos el término genérico variedad lingüística. Las diferencias que abarca se contemplan en los principales niveles lingüísticos (fonético, léxico y sintáctico) y pueden ser detectadas tanto en el habla como en la escritura de una persona.

Desde el punto de vista de la lingüística forense y la atribución de autoría, podemos hablar de variedad lingüística aún cuando se trata de diferencias que se detectan en el uso de la lengua de una persona no nativa en comparación con otra nativa. En esta tesis nos interesan solo las diferencias lingüísticas entre los hablantes de la

[61] La ambigüedad en torno a dicho término se crea a raíz del hecho de que no existe un conjunto de criterios universalmente aceptado que permita determinar cuando dos variedades deben ser consideradas como la misma lengua y cuando como dialectos. El término dialecto recibe además en la lengua habitual connotaciones peyorativas que desacreditan la variedad lingüística en cuestión de su importancia social y cultural (Rojo, 1986: 41-43; Wardhaulgh, 1986: 24-25).

misma lengua, concretamente las que se dan entre los usuarios de las dos principales variedades del español: peninsular y latinoamericana[62].

Analizar estas diferencias lingüísticas en lingüística forense, y en atribución de autoría en concreto, nos puede servir para crear un modelo para el trazado de perfiles lingüísticos basado en las marcas idiolectales que son propias de cada variedad. Tratamos este tema en el estudio descrito en el capítulo 5 con el propósito de comprobar que los n-gramas pueden discriminar entre variedades lingüísticas y demostrar que sería factible integrarlos como marca en un modelo para el trazado de perfiles lingüísticos del español.

– *Variables ligadas al texto*

Las variables ligadas al texto tienen que ver con las características textuales de las pruebas lingüísticas. Las que tienen mayor peso a la hora de tomar una decisión sobre la viabilidad del análisis lingüístico forense, y que por consiguiente hemos tomado en consideración en la tesis, son la extensión, el género textual y el tiempo de medición en la producción de los escritos.

[62] Aunque en nuestro trabajo delimitamos únicamente dos variedades del español, somos conscientes de que esta división es muy generalista. Tanto el español peninsular, como el de América Latina integran una amplia gama de variedades lingüísticas habladas en sus territorios, a las que no es nuestra intención menospreciar. No obstante, un estudio piloto como el que llevamos a cabo en esta tesis no puede profundizar en el análisis de las distintas variedades propias de cada comunidad y país, puesto que su objetivo es detectar las diferencias lingüísticas en relación a nuestra variable de análisis a una escala general.

- *La extensión del texto*

La extensión de los documentos que componen el corpus de análisis en un caso real (textos dubitados e indubitados) es la variable independiente decisiva para el peritaje. Es así, porque, pese a la experiencia y a la alta competencia en la lengua y sus variedades y particularidades, el lingüista forense es incapaz de llevar a cabo la pericia si solo tiene en su poder un único texto cuyo contenido se limita a unas pocas palabras o líneas (por ejemplo, "Págame o te mato"). Este es el ejemplo de un caso extremo, pero no inverosímil, pues también es cierto que la mayoría de los textos dubitados que llegan a las manos de los expertos forenses raras veces exceden una centena de palabras, y por lo general suelen ser más cortos. Por ello, la buena práctica en comparación lingüística forense para la atribución de autoría exige que los textos que se peritan tengan una extensión que permita que el lenguaje escrito de los sujetos implicados pueda someterse al análisis cuantitativo y cualitativo[63]. Esta exigencia se refiere tanto a los textos dubitados, como a los indubitados. Una excepción a esta regla son los casos en los que se dispone de un número considerable de los dos tipos de texto y sujeto de forma que el volumen compensa las carencias del corpus en longitud.

[63] Esta extensión depende del número de textos disponibles para la comparación lingüística forense y el método y las herramientas de análisis. En el caso concreto de los n-grama hemos podido comprobar que la extensión mínima de los documentos analizados no debe ser inferior a 200 palabras.

En atribución de autoría otra cuestión en relación a la extensión de las muestras del corpus concierne las marcas identificativas. Esta cuestión tiene que ver con el grado de dependencia que existe entre la aplicabilidad de un rasgo idiosincrásico como marca y el tamaño de las muestras de texto. Es decir, la probabilidad de que una marca de cualquier tipo ocurra en el texto de análisis y su potencial discriminatorio como tal disminuyen a la par que la extensión del texto.

Para estimar el nivel en el que la variable extensión del texto incide en el potencial discriminatorio de los n-gramas, realizamos un estudio en el que aplicamos el análisis lingüístico forense basado en los n-gramas a textos cortos (300 palabras) y a textos largos (600 palabras) y contrastamos los resultados obtenidos (véase el capítulo 5).

- *El tiempo de medición*

Cuando hablamos de *tiempo de medición* en atribución de autoría nos solemos referir al tiempo que ha transcurrido entre la producción de una prueba lingüística oral o escrita y otra de la misma tipología. A diferencia de la variación relacionada con la variable *edad*, cuya influencia también implica el transcurso del tiempo, la variación ligada a la variable *tiempo de medición* es más fácil de medir ya que podemos obtener observaciones de los eventuales cambios idiolectales y estilísticos intra autor a causa del efecto del tiempo de medición, con la recogida de muestras de cada

punto intermedio en el período de estudio. Con la variable *edad* sería muy difícil, si no imposible, hacer lo mismo porque la evolución intelectual y cultural del ser humano no son procesos que se desarrollan de manera uniforme en las fases de su ciclo vital. Esta es la razón por la cual los escasos estudios en estilometría[64] sobre la variación intra autor (Can y Patton, 2004), incluido el que se presenta en el capítulo 4 de esta tesis, prefieren centrarse en el análisis del efecto del tiempo de medición. En particular, nuestro estudio se ocupa del análisis de la variación estilística intra autor en dos contextos: uno en el que la distancia del tiempo de medición entre los escritos de un autor es mayor, y otro en el que es menor.

- *El género textual*

Concluimos la descripción de las variables independientes y el comentario introductorio sobre su repercusión en la metodología en atribución de autoría[65] con la variable más problemática en relación a la evaluación experimental y la aplicación de las técnicas de comparación de pruebas lingüísticas de casos reales: el género textual.

El género textual al que pertenece un escrito se determina en función de sus características específicas y conforme criterios socio-culturales y funcionales. Las pruebas lingüísticas, es decir, los

[64] La estilometría es la disciplina que se dedica al estudio del estilo mediante su análisis estadístico.
[65] Ampliamos la información relativa a cada variable en los capítulos correspondientes a cada estudio de evaluación de su efecto sobre el potencial discriminatorio de los n-gramas.

textos dubitados, son difíciles de clasificar dentro de la tipología textual convencional por su carácter "camaleónico".

> Many texts, [...], which are analysed as part of forensic casework, are not inherently criminal; they may be more mundane including for instance, personal letters and diaries. […]Given the variety of texts subject to forensic analysis there is real danger in attempting to make generalizations about their character.
>
> (Grant, 2008: 216)

Al tiempo que comparten muchas de las propiedades de un género textual, los textos forenses poseen sus propias características peculiares y únicas. Por ejemplo, un mensaje circular de empresa que contiene información infamatoria y ofensiva sobre uno de los empleados puede tener la mayoría o todas las características típicas de una carta formal. Es así porque en realidad un texto, cualquiera que sea su género, se convierte en un documento dubitado cuando se le considere de carácter delictivo y surja la necesidad de su análisis lingüístico.

En la investigación en el campo de la atribución de autoría para el fin de la evaluación de los métodos y las técnicas de análisis lingüístico forense, desde los inicios de la disciplina se ha recurrido casi siempre al uso de corpus constituidos por textos del género literario (Ellegard, 1962; Mosteller y Wallace, 1964; Holmes, 2001

y anteriores; Baayen et al., 1996; Hoover, 2003 y anteriores, entre otros) [66].

Esta práctica, sin embargo, ha suscitado críticas y serios debates en los congresos europeos e internacionales de lingüistas forenses sobre su adecuación para la evaluación de las técnicas de atribución de autoría. La discusión proviene del hecho de que el género literario no es el género prototípico de los textos dubitados (ni de los indubitados) y es muy probable de que los resultados de los experimentos de evaluación basados en ellos no se puedan generalizar a los contextos de trabajo con textos de casos reales.

> If there are some differences in the character of texts in more literary authorship analysis when compared to those of forensic case work, this raises the interesting question of whether methods and assumptions from the more academic field can be transferred to the applied setting […].
>
> Grant (2008: 217)

A pesar de que compartimos estas ideas, creemos que el trabajo experimental con textos literarios es un paso previo substancial por el que empezar en el proceso de evaluación de técnicas de atribución de autoría.

[66] Para un repaso de los estudios citados y otros estudios similares véase capítulo 3.

Además de esta última que acabamos de destacar, existen otras razones por las que las investigaciones en esta rama de la lingüística forense se han basado habitualmente en textos literarios. Una es que el acceso a textos de casos reales suele estar restringido a los órganos de la ley y a los abogados de las partes implicadas, mientras que los textos literarios son de dominio público y se encuentran al alcance del investigador científico. Y otra razón es que la experimentación en atribución de autoría basada en este género permite seleccionar muestras de una extensión significativa, mientras que los textos de casos reales suelen ser en general bastante cortos.

En la actualidad es más importante tratar de resolver otro problema trascendental que tiene que ver con la variable *género textual*, la disparidad en la tipología textual de los documentos dubitados e indubitados habitual en una pericia. Esto ocurre porque muchas veces no se pueden aportar para el análisis lingüístico textos indubitados del sospechoso o los sospechosos que pertenezcan al mismo género que los dubitados. Chaski (2001) pone en duda que en estos casos se pueda realizar el peritaje argumentando que cada género se caracteriza por sus propios esquemas textuales, construcciones argumentativas, estructuras y léxico, por lo que resulta más fácil discriminar entre textos de diferentes géneros que entre textos del mismo género. Debido a estas diferencias la comparación de los textos puede llevar a conclusiones erróneas respecto a su autoría.

Con el propósito de averiguar si los n-gramas son una marca identificativa cuyo potencial discriminatorio depende del género textual de los escritos, en esta tesis analizamos textos de dos géneros diferentes: el literario (fragmentos de novela) y el periodístico (artículos de opinión).

Como conclusión de este apartado sobre las variables independientes queremos enfatizar el hecho de que las que acabamos de describir no agotan la diversidad de posibles variables que pueden influir en la comparación lingüística forense de textos escritos. En el contexto de una pericia, como resultado de las características propias del corpus de análisis y de las personas implicadas, como también de otra índole, pueden surgir muchas más variables que resultan difíciles de prever y de considerar en una sola investigación. Sin embargo, las que hemos destacado y que contemplamos en esta tesis doctoral son las variables que, basándonos en nuestra experiencia en el trabajo con textos forenses, podemos decir que son comunes a la mayoría de casos y por ello debe prevalecer su estudio.

A lo largo de este capítulo hemos definido las variables de la presente tesis doctoral. En primer lugar, hemos descrito el carácter de las variables dependientes, los procesos de agrupación de los que resultan y los motivos por los que se ha recurrido a este procedimiento. En segundo lugar, hemos comentado las variables independientes cuyo efecto puede influir en el análisis lingüístico forense y hemos descrito brevemente los estudios que llevamos a

cabo para comprobar como afectan el comportamiento de los n-gramas como variable candidatas a ser marcas de atribución de autoría. Prestamos especial atención a aquellas para las cuales pretendemos verificar si existe dependencia, y establecer su grado, en relación con el potencial discriminatorio de los n-gramas.

3.3 Objetivos

A la hora de fijar los objetivos de esta tesis han sido determinantes algunas consideraciones respecto al porvenir de nuestro trabajo. Nos interesa que, al tratar cuestiones problemáticas y muy recurrentes en la práctica habitual de atribución forense de autoría, los resultados a los que llegamos sean de utilidad y sirvan de referencia a otras investigaciones sobre lengua española en el campo de la comparación forense de textos escritos. Pero sobre todo nos interesa que nuestra propuesta de método de atribución de autoría llegue a formar parte del conjunto de técnicas lingüísticas forenses aprobadas por la comunidad de lingüistas peritos forenses y que sea acogida en su práctica.

En la literatura sobre autoría forense se hace hincapié en los contextos problemáticos que se presentan a menudo en las condiciones reales de trabajo con textos dubitados, que resultan sumamente restrictivos para el potencial discriminatorio de las marcas de identificación y la atribución correcta de autoría en

general[67]. Por lo tanto, para nuestra investigación ha sido importante, por una parte, evaluar los n-gramas[68] como marcas discriminantes en contextos en los que se presentan factores cuyo efecto podría dificultar el análisis de las pruebas lingüísticas, o incluso impedir la realización del peritaje lingüístico forense; y por otra parte, validar su eficacia y aplicabilidad en contextos similares a las pericias lingüísticas forenses. Con la realización de esta tesis doctoral nos planteamos alcanzar una serie de objetivos conceptuales y metodológicos.

a) Objetivos conceptuales

Los objetivos conceptuales son los que proponemos para corroborar el carácter idiosincrásico y la capacidad de los n-gramas de discriminar entre autores de cara a su aplicación como marcas identificativas. Estos objetivos son los siguientes:

OC1 **Demostrar que la variación estilística idiolectal es mayor a nivel inter autor que a nivel intra autor.**

OC2 **Determinar qué n-gramas o grupos de n-gramas discriminan mejor entre los autores del corpus de análisis y pueden ser usados como marcas de identificación comunes en la**

[67] Para su discusión véase el capitulo 3
[68] A lo largo de esta tesis empleamos n-gramas como término genérico para referirnos a bigramas y trigramas.

comparación lingüística forense para la atribución de autoría de textos escritos en lengua española.

OC3 **Medir el efecto del tiempo de medición en el potencial discriminatorio de los n-gramas.**

OC4 **Establecer si los n-gramas permiten discriminar entre las producciones lingüísticas escritas según la variedad lingüística del autor.**

OC5 **Establecer si los n-gramas pueden discriminar entre las producciones lingüísticas escritas según el género biológico del autor.**

OC6 **Establecer el grado en el que el potencial discriminatorio de los n-gramas se ve restringido por el tamaño de las muestras de análisis y el género textual.**

b) Objetivos metodológicos

En lo que atañe a los objetivos metodológicos, estos deben asegurar la idoneidad del método de atribución de autoría basado en los n-gramas como marcas identificativas que hemos desarrollado, para su aplicación en los casos reales de pericia lingüística forense. Para tal fin nos proponemos los objetivos siguientes:

OM1 **Evaluar el potencial discriminatorio de las secuencias de categorías gramaticales de tipo bigrama y trigrama mediante el análisis estadístico.**

OM2 **Desarrollar una propuesta de técnica de atribución forense de autoría de textos escritos que implemente los n-gramas como marca identificativa.**

OM3 **Evaluar la aplicabilidad de la técnica de atribución de autoría en la experimentación con textos de casos reales.**

OM4 **Crear un protocolo de análisis lingüístico forense mediante n-gramas que pueda incorporarse a una herramienta de atribución de autoría semiautomática.**

3.4 Hipótesis

Esperamos poder alcanzar los objetivos propuestos arriba validando las hipótesis generales y específicas que formulamos a continuación.

a) Hipótesis generales

HG1 **Cada individuo, como usuario de una lengua concreta, dispone de su propia gramática interna que forma parte de su idiolecto y viene modificada por varios factores externos. Estos**

cambios quedan reflejados a todos los niveles lingüísticos, y por lo tanto, podemos considerar únicas las realizaciones fonéticas, léxicas y sintácticas de cada autor.

La Lingüística Forense, en sus dos ramas especializadas en la comparación lingüística forense para la identificación de autor mediante el análisis de sus producciones orales (fonética forense) o escritas (autoría forense), parte del concepto de la unicidad de las realizaciones lingüísticas de los usuarios de una lengua. En dicho concepto se sustenta la hipótesis general que acabamos de exponer y así mismo lo hacen todas las investigaciones de autoría que se realizan dentro de la disciplina de lingüística forense. No obstante, en esta tesis nos limitamos a hipotizar la unicidad de las producciones sintácticas, por ser estas las representaciones de la lengua de principal interés. Para su argumentación teórica, nos apoyamos por un lado en la teoría de la variación (véase capítulo 1.1) y, por otro, en las teorías de adquisición del lenguaje y producción lingüística (véase capítulo 1.4), que conciben la unicidad e idiosincrasia lingüísticas como resultado del efecto de diversos factores que en su conjunto influyen en que los individuos alcancen diferentes grados de competencia y destreza lingüísticas y usen el lenguaje de una manera particular y distintiva.

HG2 Se hipotiza que **la autoría de un texto escrito puede atribuirse a un determinado autor a partir de las estructuras y construcciones sintácticas que emplea con mayor frecuencia.**

Las estructuras sintácticas suelen ser descartadas como posibles marcas de identificación en los estudios de autoría por la dificultad considerable y el tiempo a invertir que suponen su tratamiento para quien desea utilizarlas en el análisis lingüístico forense. Sin embargo, tienen una gran ventaja frente a las marcas de tipo léxico. Las marcas léxicas son relativamente fáciles de localizar y de tratar mediante métodos estadísticos, pero esto las hace también susceptibles a imitaciones, mientras que falsear la sintaxis de una persona requeriría los conocimientos avanzados de un lingüista.

Creemos que el individuo desarrolla ciertas preferencias a determinadas estructuras y construcciones que se encuentran en su lexicón sintáctico y tiende a usarlas con mayor frecuencia que otras que pueden cumplir las mismas funciones dentro de la oración. Según estudios en el campo de la adquisición de la lengua escrita[69], esto puede ocurrir a causa de las muchas influencias externas que recibe una persona en el proceso de aprender a escribir, empezando por su propio modo de hablar, la escolarización, y llegando al contacto con la lectura. El uso repetitivo de estructuras lingüísticas en escritura individual se explica además por el fenómeno de la persistencia sintáctica[70], que se estudia en psicolingüística.

[69] Véase el apartado 4 del capítulo 1 para el comentario más extenso de los estudios a los que nos referimos.

[70] También conocida como facilitación sintáctica, consiste en recurrir a la misma construcción que se ha utilizado con anterioridad al crear oraciones para formar nuevas. Véase el capítulo 1 para más detalles sobre este fenómeno y algunas referencias sobre estudios exhaustivos del tema.

En el capítulo 2.2 hemos visto las características a las que tienen que responder los elementos y unidades lingüísticas candidatas a marcas identificativas. Una de las que volvemos a destacar aquí entre las principales, debido a su relevancia para el análisis cuantitativo de los datos, es la frecuencia. El número de casos por texto es importante, ya que la falta de ocurrencias en el corpus de control (el corpus de textos indubitados) hace imposible no solo el análisis cuantitativo, sino también el análisis cualitativo. De ahí que ha sido fundamental centrarnos en estructuras de uso frecuente. Las secuencias de categorías gramaticales[71] son una variable que cumple con este criterio, pero cuya naturaleza implica *per se* el análisis cuantitativo[72].

HG4 Se hipotiza que **las secuencias de categorías gramaticales pueden ser empleadas como marcas identificativas en la comparación lingüística forense de textos escritos.**

En cuanto a su organización, un texto escrito visto fuera del contexto estrictamente lingüístico representa una secuencia de objetos y signos de puntuación ininterrumpida desde el principio hasta el final. Esta secuencia sería, sin embargo, el resultado de la agrupación de palabras o categorías que forman construcciones simples y complejas entre sí. La lingüística y la sintaxis se han ocupado de dar nombre a cada una de estas construcciones creando una jerarquía de constituyentes sintácticos. Hasta ahora la práctica

[71] Véase el capítulo 4.1
[72] Comentamos este aspecto de la variable en mayor detalle en el capítulo 4.

común en comparación lingüística forense para la atribución de autoría ha consistido en tratar de identificar una o varias construcciones o segmentos de construcciones como marcas fiables, estudiando las características de su comportamiento lingüístico y evaluando su eficacia como tales (Baayen et al., 1996; Van Halteren, 2004; Hoover, 2003). Muchas veces el empleo de una estructura en lugar de otra depende del contexto y, pese a que la idiosincrasia estilística del autor lo incite a hacer la misma elección siempre, o por lo menos en gran parte de los casos, su uso puede quedar restringido por la irrelevancia funcional de la estructura[73]. Por este motivo, en la tesis hemos juzgado más acertado empezar la búsqueda de marca de carácter sintáctico a un nivel inferior al funcional, donde las construcciones sintácticas representan simples secuencias de categorías[74].

b) Hipótesis específicas

HE1 Se espera que **la variación de los n-gramas a nivel intra autor sea inferior a la que se produce a nivel inter autor.**

[73]Esto ocurre sobre todo con las construcciones que poseen un significado fijo. Tomemos un ejemplo de las perífrasis verbales. Una acción futura puede ser expresada de dos maneras: por los tiempos gramaticales de futuro o presente o mediante la perífrasis verbal formada por el verbo ***ir*** en su forma conjugada del presente o el imperfecto de indicativo + la preposición ***a*** + el segundo verbo. Supongamos que el individuo suele usar la perífrasis y no las otras dos variantes. Si no necesitara comunicar una acción que se realiza en el futuro, no usaría la perífrasis.

[74] Establecer la relación entre el n-grama y la construcción y el papel que esta construcción cumple en la oración es una tarea que queda para una etapa posterior a la tesis doctoral.

Se ha demostrado empíricamente que los individuos pueden variar no sólo en su uso del lenguaje en comparación con otros usuarios de la lengua (variación inter autor), sino también al crear sus propios textos (variación intra autor) (Pennebaker y Stone, 2003; Can y Patton, 2004). Los dos tipos de variación autorial ocurren a causa de factores fisiológicos, socioculturales, idiolectales y lingüísticos[75]. Sin embargo, su efecto no es el mismo, incluso en escritores con las mismas características demográficas. La mente humana es una *terra incognita* en la que los mecanismos cognitivos de cada individuo se desarrollan a su manera, distinta a la de los demás. El individuo tampoco se muestra receptivo de forma uniforme a todos los estímulos y influencias lingüísticas externas a las que se ve expuesto. Por consiguiente, es de esperar que cuando haya diferencias en el uso del lenguaje de un texto a otro del mismo autor estas diferencias serán siempre menores que las diferencias entre textos de dos autores distintos, es decir, que el grado de variación intra autor será inferior al de la variación inter autor.

HE2 Se hipotiza que **los n-gramas pueden discriminar entre las producciones lingüísticas escritas según el género biológico de su autor.**

La investigación sobre la diversidad en el uso del lenguaje entre hombres y mujeres es muy amplia y cuenta con un gran volumen de

[75] Véase el capítulo 1 para una descripción de estos factores.

trabajos que aportan pruebas de que realmente existen diferencias en el habla y en la escritura, y de que además se pueden detectar en todos los niveles del sistema lingüístico. Dado que la marca cuyo potencial discriminatorio estamos evaluando representa la estructura sintáctica, podemos hipotizar que el análisis discriminante a partir de datos de SEM permitirá distinguir entre los textos escritos por la mano de mujeres y aquellos escritos por hombres.

HE3 Se hipotiza que **los n-gramas pueden discriminar entre autores en base a la variedades lingüísticas del español que emplean en su producción escrita.**

El español es una de las lenguas más habladas en el mundo gracias a que su uso como lengua nativa se extiende a más de un continente y a una decena de países. A raíz tanto de las diferencias culturales y económicas, como también de los periódicos cambios demográficos que se han experimentado en las dos regiones donde se concentra el mayor número de usuarios de la lengua española, España y América del Sur, el español se ha estructurado en dos grandes variedades lingüísticas[76]. La distinción entre estas dos variedades, según los hispanistas que se dedican a su estudio (Alvar, 1991; 1996), se divisa de forma más prominente en el contraste de la fonética y en el vocabulario, pero se percibe también en la sintaxis.

[76]Aunque hay subvariedades en el español peninsular y en el español latinoamericano donde cada ámbito territorial por países podría considerarse una variedad lingüística específica.

Para esclarecer la idea base de esta hipótesis podemos tomar como ejemplo el caso de los tiempos gramaticales que se utilizan para expresar una acción realizada en el pasado próximo. En la variante americana el usuario de la lengua recurriría al tiempo indefinido (*salí* – verbo –, corresponde a una etiqueta de categoría gramatical en un texto anotado), mientras que el hablante de la península emplearía el pretérito perfecto (*he salido* – verbo + participio –, dos etiquetas). Esto resultaría en una diferencia clara entre las secuencias de etiquetas que se extraen de los textos escritos en las dos variedades lingüísticas.

HE4 Se hipotiza que **el comportamiento lingüístico de los individuos no cambia de forma significativa en el transcurso del tiempo y, más concretamente, que los eventuales cambios idiolectales a causa del tiempo de medición entre la producción escrita realizadas en tiempo aparente y tiempo real no influye en la capacidad discriminatoria de los n-gramas**.

Uno de los aspectos problemáticos de la atribución y determinación de autoría en condiciones reales, que se comenta muy a menudo en las reuniones y los eventos científicos de lingüistas forenses, es el hecho de tener que trabajar sobre textos indubitados que el sospechoso ha redactado años antes que el documento o documentos dubitados.

En la psicolingüística se han llevado a cabo estudios (Kemper, 1990; Pennebaker y Stone, 2003) que sugieren que el estilo de una

persona experimenta ciertos cambios a lo largo de su vida, pero se deconoce si los cambios en cuestión son tan significativos como para que el lenguaje escrito de un autor en un punto del pasado sea completamente diferente al de su lenguaje actual. Teniendo en cuenta la relevancia de esta pregunta sorprende la escasez de trabajos que tratan de averiguar el grado de interdependencia entre el factor tiempo y la variación intra autor. Nos consta que las palabras más frecuentes (Can y Paton, 2004) y la densidad proposicional (Cooper, comunicación personal[77]) están entre los únicos elementos lingüísticos cuyas alteraciones a través del tiempo y su implicación para la lingüística forense han sido examinadas. De acuerdo con los resultados de estos dos experimentos independientes, los ratios de cambio que se observan para ambas marcas no son altos. Este hecho no implica una confirmación inmediata de la aplicabilidad de las palabras más frecuentes y la densidad proposicional como marcas de autoría, pero sí que es un paso crucial en la evaluación de su aplicabilidad con fines forenses. Si logramos demostrar que el uso distintivo de los n-gramas no obedece a las variaciones en el estilo escrito de una serie de autores en el tiempo, nuestra marca habrá superado una de las pruebas de su potencial discriminatorio.

HE5 Se espera que **las frecuencias de uso de los n-gramas no varíen por género textual con lo cual se demostraría la**

[77] Agradezco al Dr. Cooper de la University of Alaska Faribanks que haya compartido conmigo los resultados de su trabajo.

aplicabilidad y el potencial discriminatorio de los n-gramas como marcas identificativas.

Uno de los puntos débiles por el que la metodología de la investigación en atribución de autoría recibe mucha crítica es la homogeneidad del corpus que usa. Con alguna que otra excepción, en las que los experimentos se llevan a cabo sobre un corpus de textos cuya producción ha sido inducida por el investigador, los textos de análisis pertenecen a un solo género textual, el de narrativa literaria. Puesto que cada género tiene sus propiedades estilísticas, que abarcan desde un léxico hasta unas estructuras específicas, haber comprobado la fiabilidad de una marca en el contexto de un género no prueba nada en cuanto al resto de géneros y menos aún sobre su aplicabilidad en el género de los textos de casos reales que posee unas características peculiares. En esta tesis doctoral tratamos este problema mediante la evaluación en dos etapas de la técnica de atribución de autoría basada en los n-gramas: una con textos narrativos y periodísticos, y otra con textos de casos reales tomados del "book" del laboratorio forense del IULA, ForensicLab.

3.5 Corpus

La finalidad del trabajo de investigación que refleja esta tesis doctoral es ante todo ofrecer una propuesta analítica teóricamente justificada y metodológicamente sólida que pueda implementarse

como técnica de atribución de autoría en el conjunto de métodos lingüísticos forenses que se suelen usar en la pericia de textos de casos reales. Nos hemos ocupado de la fundamentación teórica en los capítulos de la parte 0, donde explicamos los conceptos y las premisas que constituyen la base teórica de la lingüística forense en su rama de la comparación lingüística forense para la atribución de autoría y también la de esta tesis. En los capítulos siguientes nos dedicamos a la descripción del diseño metodológico de los estudios que llevamos a cabo en la evaluación del potencial discriminatorio de las secuencias de categorías gramaticales (n-gramas) en la atribución forense de autoría de textos escritos en español.

Para que una metodología sea considerada válida y fiable, sobre todo en el contexto del trabajo forense, que pone en juego no sólo la autenticidad de la técnica y la reputación del perito que la aplica, sino que puede tener consecuencias para el futuro de una persona, es indispensable someterla a una evaluación lo más completa posible. Con el fin de cumplir con este propósito, en la evaluación de los n-gramas como marca de autoría recurrimos a la recogida y la explotación de dos tipos de corpus: uno de análisis y otro de evaluación. El corpus de análisis nos sirve para establecer los límites de la capacidad discriminatoria de los n-gramas. Su diseño y los experimentos que llevamos a cabo han sido ideados teniendo en cuenta los principales factores cuyo efecto se considera que puede afectar (reducir o anular) el potencial discriminatorio de una marca indentificativa. Con la explotación del corpus de análisis testamos la técnica de atribución de autoría mediante n-gramas a nivel de la

lengua general. El corpus de control, en cambio, nos permite llevar a cabo pruebas de evaluación de nuestra propuesta analítica en el contexto de los textos forenses. A continuación, detallamos las características específicas, tanto del corpus de análisis como del de control.

a) Corpus de análisis

El corpus de análisis que usamos es un corpus de textos literarios. A la vista de los comentarios críticos que suscita este planteamiento de trabajo metodológico en los círculos de peritos lingüistas (Grant, comunicación personal), no podemos dejar de empezar la descripción de nuestro corpus sin prestar atención a aquellas críticas que consideramos que son fundadas y exponer los argumentos por los que hemos decidido recurrir a la explotación de un corpus de este tipo.

Lo que sobre todo provoca las críticas respecto la elaboración y desarrollo de nuevas técnicas de análisis lingüístico forense a partir de la experimentación con corpus literarios es la dificultad de generalizar las conclusiones acerca de su fiabilidad y de las marcas identificativas que se comprueban con ellos. En nuestra opinión, las críticas de esta índole son contundentes y difícilmente se podrían tachar de infundadas cuando surgen a propósito de la tendencia de algunos estudios de incitar a creer que una técnica "validada" en un corpus de narrativa daría los mismos resultados en los textos de un caso real (Grant y Baker, 2001). Las pruebas de delitos lingüísticos,

exceptuando los casos de plagio, no son nunca obras literarias, y tratándose de dos géneros distintos con sus propias características estilísticas, producidos para un público diferente y en circunstancias muy distintas, es posible que un método cuya eficacia ha quedado comprobada en los textos de narrativa resulte inaplicable o de rendimiento negativo[78] en el análisis de textos forenses. De ahí que no se puede concluir que una técnica de atribución de autoría es igualmente válida para el peritaje de textos forenses sin que se haya ejecutado previamente su evaluación en un corpus de casos reales. Por este motivo, en nuestro trabajo sobre el potencial discriminatorio de los n-gramas incluimos entre los análisis que llevamos a cabo una serie de experimentos con documentos de dos casos de autoría peritados por el Laboratorio de Lingüística Forense - Forensiclab del Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF) (véase el apartado 3 del capítulo 5).

En cambio, discrepamos con las críticas que cuestionan la conveniencia de la elección de un corpus literario para la investigación en general sobre marcas lingüísticas de atribución forense de autoría. Actualmente, la metodología en autoría está en un estado de desarrollo cuyos avances dependen en mayor grado de la investigación que de la práctica en el campo. Aunque en la actualidad las opiniones y habilidades expertas de los lingüistas forenses se solicitan más a menudo que hace una década, por

[78] Se dice del método que produce resultados erróneos en la clasificación de los textos analizados.

cuestiones de confidencialidad el perito no puede divulgar su trabajo de análisis de casos de crímenes lingüísticos o que implican pruebas lingüísticas, o usar como corpus los documentos de los casos en su "book" profesional. Por lo tanto, a la hora de elaborar métodos de análisis para fines forenses, los lingüistas han tenido que recurrir al uso de corpus más accesibles, como los corpus de textos literarios. Esta vía de progreso metodológico no desfavorece la disciplina cuando la investigación tiene las características de la que se realiza en atribución de autoría forense.

La investigación en el campo de la comparación forense de textos escritos es exploratoria y experimental. Nos permite calificarla de exploratoria el hecho de que la mayoría de propuestas analíticas parten de la búsqueda, en el vasto universo de la lengua, de unidades y fenómenos lingüísticos que pueden manifestarse con usos idiosincrásicos en los idiolectos de los individuos usuarios de esta lengua. Es experimental porque su diseño está concebido de manera que mediante la formulación y evaluación de hipótesis haga posible llegar al mismo tiempo a conclusiones sobre la idiosincrasia y el potencial distintivo de la unidad candidata a marca identificativa y sobre el efecto de determinados factores que podrían influirlos.

Teniendo en cuenta todo lo mencionado anteriormente, creemos que un corpus de narrativa, a pesar de no ser la opción ideal, como sería el caso de un corpus forense, responde a las necesidades de los estudios en esta línea de investigación por varias razones. En primer

lugar, compilar un corpus representativo de la lengua de análisis que permita estudiar la variable o marca en el comportamiento lingüístico de un gran número de hablantes y hacer generalizaciones sobre su carácter idiosincrásico (no sobre su aplicabilidad como marca de identificación) es más factible y, tratándose de una investigación en variación, metodológicamente más correcto. Incluso si dispusiéramos de suficientes textos forenses como para crear un corpus, nos encontraríamos con una serie de problemas de muestreo. Por un lado, los sujetos no serían caracterizados socialmente y no seríamos capaces de controlar los factores que podrían influir en la validez de los resultados del análisis. Por otro lado, muy probablemente si tuviéramos muchos sujetos de estudio no contaríamos con el mismo número de textos por autor. En consecuencia, la distribución de los textos no sería proporcional y nuestro corpus no sería homogéneo[79]. En segundo lugar, hemos de resaltar que el corpus de narrativa brinda al investigador la oportunidad de realizar experimentos a la medida de sus objetivos e hipótesis.

Por todo lo expuesto, podemos concluir que el trabajo con textos de una tipología genérica como los narrativos constituye un buen punto de partida en la evaluación de una técnica de comparación textual forense conducente a la atribución de autoría.

[79] Explicaremos la importancia de la homogeneidad del corpus de análisis en el apartado sobre la selección de muestras.

b) Selección del corpus

Para la evaluación inicial, basada en el corpus de narrativa, de la marca identificativa y de la técnica de análisis que proponemos en esta tesis doctoral, hemos recogido las muestras textuales de 17 escritores contemporáneos, usuarios nativos de la lengua española. Estas muestras proceden de una selección de su obra literaria y periodística y constituyen nuestro corpus de análisis.

– *Criterios de selección*

En la selección del corpus de análisis, para lograr la validez interna de los resultados de los estudios que llevamos a cabo, hemos seguido las pautas recomendadas por Biber et al. (1998) para el diseño de un corpus representativo de la lengua adaptándolas a los objetivos de la investigación y a los interrogantes principales de la disciplina.

• *Géneros textuales*

Uno de los puntos clave del diseño del corpus según Biber (1998: 247) estriba en lograr "incluir el rango de variación que existe en la lengua". Sin lugar a duda, lo mejor para una investigación, que tiene que ver con la identificación y evaluación de marcas de autoría, sería disponer de un corpus que reuniera textos de la lengua de

estudio de la más diversa tipología, que reflejaran la variación en el uso de las unidades lingüísticas que pueden revelarse como marcas. A pesar de que la variación de las secuencias de categorías gramaticales en el lenguaje a nivel inter autor es relevante para nuestro trabajo, en esta tesis predomina el interés por la variación individual. Estudiar la idiosincrasia de una unidad candidata a marca en relación a su variación a nivel intra autor supone recopilar textos de diferentes géneros, en nuestro caso, escritos por el mismo individuo.

Resulta difícil encontrar muestras de la escritura de una persona que pertenezcan a más de tres o cuatro géneros textuales y aún más cuando se trata de hacerlos coincidir en el mismo aspecto con los de otros individuos[80]. Por consiguiente, en la presente tesis optamos por trabajar con textos de dos géneros: el literario y el periodístico. Del género literario hemos escogido la narrativa y del periodístico, el artículo de opinión. La narrativa comprende novelas de diversa temática (psicológica, policíaca, social, picaresca, biográfica)[81], mientras que los artículos de opinión tratan en su mayoría diferentes temas de la vida cotidiana. Hemos seleccionado estos dos tipos de textos en concreto por sus características: Las novelas suelen abarcar una gran variedad de dimensiones textuales, desde descriptivas a argumentativas, lo que nos brinda la posibilidad de aplicar la metodología de análisis que hemos desarrollado a muchos contextos textuales. Además, según Halliday et al. (1964:97), un

[80] Aquí se exceptúan los casos de investigación basada en textos que han sido recogidos mediante experimentos de producción lingüística (escrita) inducida.
[81] Véase el anexo I para el listado completo de novelas utilizadas.

texto escrito en prosa reflejaría con mayor exactitud las características léxicas y estructurales del estilo de un autor, porque estimula la creatividad y no obedece a las reglas estrictas de construcción y uso de vocabulario determinado que rigen otros géneros.

> Literature forms only a small part of written language, but it is the part in which we are more aware of the individual and most interested in the originality of the individual's language. At the same time it is of the essence of creative writing that it calls attention to its own form, in the sense that unlike other language activity, written or spoken, it is meaningful as activity in itself and not merely as part of a larger situation: again, of course, without a clear line of demarcation. This remains true whether or not the writer is consciously aiming at creating an individual variety. Thus the linguistic uniqueness of a work of literature is of much greater significance than the individuality of a variety of language in any other use.

Los artículos de opinión, a su vez, comparten algunos rasgos de los escritos de carácter forense, por un lado, porque son textos cuya producción es relativamente espontánea, aunque a diferencia de los escritos forenses (usualmente destinados a ser leídos por una sola persona), en general están dirigidos a un amplio círculo de lectores y son editados. Por otro lado, se podría decir que existe un paralelismo en la motivación de ambos. Tanto el autor de un

artículo de opinión, como el autor de, pongamos por ejemplo, una carta infamatoria anónima, escriben sobre algo que los involucra emocionalmente. En el caso del columnista puede ser un tema polémico de la sociedad que le afecta o la actuación de algún personaje famoso en un evento puntual que desaprueba; en el del calumniador, puede ser la conducta o cualquier otro aspecto relacionado con el infamado, que causan su descontento o lo irritan. Es probable que el lenguaje del escritor experimentado (el columnista) sea más medido que el del infamador, pero la intención de recriminar y poner en evidencia estarán presentes, aunque de modo mucho más sutil. Por ello consideramos que, dado el enfoque de la tesis, este género es uno de los que mejor se ajustan a las necesidades de la investigación.

- *Autores*

En la selección de los autores han sido determinantes el género biológico, la variedad lingüística de su lengua materna y el género textual de su obra, aunque hemos procurado que compartan, dentro de lo posible otras características sociolingüísticas importantes, como la edad[82] y el nivel de formación.

Según el género biológico, dentro de nuestro corpus se forman dos grupos de autores. El grupo femenino, tal y como se muestra en la

[82] Los autores del corpus representan dos grupos generacionales: de autores relativamente jóvenes y autores mayores con una trayectoria publicística más larga. Esta selección ha sido importante de cara al estudio de la variación intra autor en tiempo aparente y tiempo real.

Tabla 15, está compuesto por: Isabel Allende (IA), Vladi Kociancich (VK), Laura Restrepo (LR), Carmen Posadas (CP), Rosa Montero (RM), Lucía Etxebarría (LE), Alicia Giménez Bartlett (AG) y Carmen Laforet (CL)[83].

Tabla 15. *Leyenda de los autores del grupo femenino*

ID del autor	Nombre del autor	Género	Variedad lingüística	Origen	Nacida en:
IA	Isabel Allende	F	A	Chile	1942
VK	Vladi Kociancich	F	A	Argentina	1941
LR	Laura Restrepo	F	A	Colombia	1950
CP	Carmen Posadas	F	A	Uruguay	1953
RM	Rosa Montero	F	P	España	1951
CL	Carmen Laforet	F	P	España	1921
LE	Lucia Etxebarria	F	P	España	1966
AG	Alicia Gímenez Bartlett	F	P	España	1951

Los escritores Mario Vargas Llosa (MV), Gabriel García Márquez (GG), Mario Benedetti (MB), Antonio Muñoz Molina (AM), Arturo Pérez-Reverte (AP), Eduardo Mendoza (EM), Javier Marías (XM) y Miguel Delibes (MD) integran el grupo masculino. Resumimos la información relativa a los autores de este grupo en la Tabla 16.

[83] De ahora en adelante nos referiremos a los autores del corpus mediante su identificador.

Tabla 16. *Leyenda de los autores del grupo masculino*

ID del autor	Nombre del autor	Género	Variedad lingüística	Origen	Nacido en:
MV	Mario Vargas Llosa	M	A	Perú	1936
GG	Gabriel García Marquez	M	A	Colombia	1927
MB	Mario Benedetti	M	A	Uruguay	1920
AM	Antonio Muñoz Molina	M	P	España	1956
AP	Arturo Pérez-Reverte	M	P	España	1951
EM	Eduardo Mendoza	M	P	España	1943
XM	Javier Marías	M	P	España	1951
MD	Miguel Delibes	M	P	España	1920
JM	Juan José Millás	M	P	España	1946

En lo que respecta a la selección según la variedad lingüística del autor, hemos elegido limitar el marco de trabajo de nuestra investigación a las dos variedades principales de la lengua española: la peninsular (P) y la latinoamericana (A). Somos conscientes de que tanto el español de la Península, como el de América del Sur presentan una amplia gama de variedades lingüísticas que merecen ser estudiadas por la lingüística forense a fin de establecer un perfil de rasgos lingüísticos que permitan determinar el origen, por ámbito lingüístico o por país, del autor de un texto. Sin embargo, en esta tesis, por las limitaciones de tiempo y medios, solo nos proponemos determinar si los n-gramas captan las diferencias lingüísticas entre las dos variedades mayores y si poseen valor discriminatorio para distinguir entre los autores según su variedad lingüística.

En el corpus, la variedad latinomericana del español está representada por escritores de Argentina (VK), Colombia (GG, LR), Chile (IA), Perú (MV) y Uruguay (CP, MB), y la variedad peninsular, por autores de las comunidades autónomas de Andalucía (AM, AP), Castilla y León (MD), Castilla la Mancha (AG), Cataluña (EM, CL), Madrid (JM, RM, XM) y Comunidad Valenciana (LE) (véanse tablas 15 y 16). Esta diversidad en la procedencia de los autores responde a las recomendaciones de Biber et al. (1998:248) para la constitución de un corpus representativo de la población estudiada. El español peninsular comprende la mayor parte de nuestro corpus, puesto que en una etapa inicial de recogida de muestras nos hemos visto obligados a excluir algunos de los autores de la variedad sudamericana para los que no hemos podido recuperar textos de los dos géneros de análisis.

Por último, la selección de los sujetos ha dependido de los géneros textuales que hemos fijado para el corpus. Es decir, en el corpus han sido incluidos solo autores que han publicado tanto prosa como artículos de opinión. Una de las mayores dificultades que hemos encontrado en la elaboración de la tesis doctoral ha sido la localización y recuperación de los textos, sobre todo los de los artículos. Al no ser posible adquirir u obtener por otras vías ambos tipos de documentos de algunos de los sujetos que habíamos elegido, en última instancia hemos decidido excluirlos, puesto que esta reducción no desequilibra el corpus (véase Tabla 17). La reducción se ha hecho en paralelo en los dos subcorpus que podemos delimitar conforme a la tipología de los textos del corpus,

el de novela (N) y el de artículos de opinión (AO). El primer subcorpus consta de los fragmentos de novela de 17 autores y el segundo, de los artículos de 10 de los autores que figuran en el subcorpus N. Estos escritores son: AP, CP, EM, IA, JM, LE, MB, MV, RM, XM.

En los distintos experimentos que llevamos a cabo empleamos el corpus con distribución diferente de acuerdo con los objetivos y las hipótesis del estudio del que forman parte[84].

Tabla 17. *Distribución de los autores en los subcorpus según el género biológico y la variedad lingüística*

Subcorpus	**Total de sujetos**	**Género biológico**		**Variedad lingüística**	♀	♂
N	17	♀	8	***P***	4	6
		♂	9	***A***	4	3
AO	10	♀	4	***P***	1	2
		♂	6	***A***	3	4

– *Muestras*

• *Selección de los documentos de origen*

[84] Describimos en detalle la distribución del corpus de análisis pertinente a cada estudio en los apartados correspondientes de los capítulos 4 a 5).

Los documentos de origen son los textos completos de las novelas y los artículos de opinión de los que extraemos las muestras que constituyen el corpus. En la selección de las novelas hemos usado como criterio el año de su publicación. Para conseguir muestras de diferentes períodos de la actividad escritora de los autores y poder constatar los cambios a lo largo del tiempo que experimenta el lenguaje de cada uno de ellos en relación a nuestra marca de estudio, hemos consultado los fondos bibliográficos en busca de novelas publicadas con un mínimo de dos años de distancia entre sí. Los artículos han sido recuperados de la columna de opinión de la versión electrónica de los periódicos *El País* y *La Vanguardia,* con la excepción de los artículos de IA, adquiridos del periódico venezolano *El nacional*, y los de AP y MB, que han sido publicados en formato de libro y han sido recuperados de la misma forma que las muestras de novelas[85]. Siguiendo un método de selección parecido al usado con las novelas, hemos buscado artículos publicados en diferentes años[86].

- *Selección de muestras*

Visto que los escritos con los que trabajamos no son de carácter forense y nuestro objetivo es determinar si la técnica que estamos evaluando en la tesis puede aplicarse en la comparación forense de textos escritos, las muestras que seleccionamos para el corpus deben tener una serie de características para que los resultados de los

[85] Explicamos el proceso de recuperación de las muestras en la segunda parte de este capítulo.
[86] Véase anexo I.

análisis tengan validez externa. Por una parte, deben guardar cierta similitud con los textos forenses en cuanto a extensión. Estos últimos, por lo general, tienden a ser bastante cortos, de entre 80 y 1000 palabras, por lo que es preferible que la longitud mínima de las muestras esté por encima del mínimo indicado, y la máxima por debajo del máximo. Asimismo, las muestras no deben contener formas de discurso comunes a los géneros del corpus, pero inusuales en los documentos forenses. Así, por ejemplo, en los textos de casos reales no encontraremos diálogos como los que aparecen en muchas novelas, ni tampoco citas de otros textos. Por otra parte, ya que el análisis de los datos es cuantitativo, es importante que todas las muestras tengan aproximadamente la misma extensión y que dispongamos del mismo número de casos para cada autor. De este modo nos aseguramos de que nuestro corpus será homogéneo. La homogeneización de los textos es imprescindible de cara al análisis estadístico, que impone que para que la estimación de los datos sea correcta y los resultados válidos, las muestras han de tener aproximadamente el mismo tamaño.

La extensión de las muestras[87] que recogemos para el subcorpus N es de 600 palabras, y de 300 para el AO. Esta selección nos permite establecer si el valor discriminatorio de los n-gramas y el tamaño del documento analizado son interdependientes mediante el

[87] Se trata de una extensión de valor aproximado, ya que todas las muestras acaban en punto final. Si este no coincide con la sexcentésima palabra, tomamos la frase hasta el punto siguiente.

contraste de los resultados de los experimentos con cada subscorpus.

Al seleccionar las muestras, en las novelas hemos buscado fragmentos que no contengan diálogos y en los artículos de opinión, textos sin citas directas. Cuando parte del cuerpo del fragmento está en estilo directo, la eliminamos solo si dicha parte supera una o dos líneas. Para compensar el texto reducido, tomamos el número de frases necesario para completar la extensión de la muestra del párrafo inmediato. No obstante, este tratamiento, que también hemos aplicado a las citas de los artículos, ha sido empleado en un escaso número de casos, por lo que no creemos que se hayan ocasionado problemas de coherencia textual en el corpus. Los textos originales de los artículos no han precisado ninguna otra modificación, a parte de la eliminación de citas cuando superaban el número predeterminado para este tipo de muestras, ya que la herramienta que utilizamos para extraer los datos de n-gramas nos permite fijar el número de palabras a partir de las que se hace la extracción[88].

Para una investigación como la nuestra, que comprende estudios estilométricos, es crucial que el diseño de su corpus incluya un número de muestras que, por un lado, pueda reflejar el idiolecto de cada escritor analizado y, por otro, represente la variación inter autor que existe entre ellos.

[88] Vease capítulo 3 para la descripción del mecanismo de extracción.

Desde el punto de vista de la estadística, el número de muestras adecuado es siempre aquel que nos proporciona una distribución normal[89] de las variables estudiadas. Según el teorema del límite central, se cuenta con la distribución normal de una variable cuando se ha tomado un promedio muestral elevado. Normalmente, el promedio ideal para los estudios estadísticos suele ser superior a 20 muestras. Por consiguiente, para poder considerar que las muestras recuperadas en el corpus son suficientes para la estimación estadística correcta de los n-gramas, hemos establecido el número de muestras partiendo de esta pauta. En la recopilación del subcorpus N hemos extraído cinco muestras de cada una de las cinco novelas de los 17 escritores de análisis, que suman en total 25 muestras por autor. El número de artículos que hemos recogido de los 10 autores del subcorpus AO ha sido 20.

Tabla 18. *Distribución de los textos de análisis*

SUBCORPUS	Nº DE OBRAS POR AUTOR	Nº DE MUESTRAS POR AUTOR	EXTENSIÓN POR MUESTRA EN PALABRAS
N	**5**	**25**	**~600**
AO	**20**	**20**	**~300**

Tabla 19. *Distribución del corpus de análisis*

[89] Es la distribución de probabilidad de variables continuas.

SUBCORPUS	ID DEL CORPUS	Nº TOTAL DE AUTORES	Nº TOTAL DE OBRAS	Nº TOTAL DE MUESTRAS	Nº TOTAL DE PALABRAS
	N	17	85	425	266922
	AO	10	200	200	127382
CORPUS DE LA TESIS		***17***	***285***	***625***	***394304***

– *Corpus de control*

Con el fin de corroborar, contrastar y, en definitiva, evaluar los resultados obtenidos de la explotación del corpus de análisis llevamos a cabo una serie de experimentos de atribución de autoría de los textos de dos casos reales. Estos textos forman el corpus al que nos referimos como corpus de control, y provienen de dos casos de peritaje lingüístico en los que el Laboratorio de Lingüística Forense – Forensiclab del IULA (UPF) ha actuado como experto independiente. Hemos escogido estos casos en concreto para este corpus, puesto que su extensión permite la aplicación correcta del análisis en la evaluación del método de análisis forense mediante n-gramas, pensado en comparación lingüística forense sobre todo para la atribución y determinación de autoría. A continuación ofrecemos una breve descripción del corpus de control que ampliaremos en el capítulo 0 que recoge el estudio de evaluación de los n-gramas como marca identificativa.

El caso real 1 (CR1) consiste de 12 textos que están distribuidos de la siguiente manera: 4 faxes que corresponden a los textos

indubitados; 4 e-mails que representan los textos dubitados, y 4 textos de documentos de otro caso real no relacionado con CR1. Estos últimos han sido incluidos en el análisis con el fin de llevar a cabo la comparación lingüística forense mediante el análisis discriminante.

El caso real 2 (CR2) del corpus de control a su vez contiene 26 textos. De ellos 16 son textos indubitados y 10 son textos dubitados de carácter anónimo. Cabe aclarar que en el análisis del CR2 no ha sido necesario ampliar el número de grupos con documentos de otro caso ya que los indubitados proporcionados por el cliente para la pericia provienen de dos sujetos sospechosos diferentes.

Tabla 20. *Distribución del corpus de control*

Subcorpus	**No de textos indubitadas**	**No de textos dubitados**	**Extensión del total de textos en palabras**
CR1	8	4	7976
CR2	16	11	10127

c) Procesos de tratamiento computacional del corpus

En el apartado anterior hemos descrito los criterios aplicados en la selección de los corpus que utilizamos en esta tesis doctoral. A continuación explicaremos los procesos de tratamiento computacional de los textos de análisis.

- *Digitalización y codificación de los documentos del corpus*

El requisito básico del procesamiento computacional de textos es que todos los documentos de entrada deben estar en formato electrónico. Por lo tanto, antes de aplicar las herramientas de anotación morfosintáctica a los textos del corpus de análisis de esta tesis, hemos digitalizado las muestras que no hemos podido recuperar directamente como ficheros de texto electrónico.

El preproceso, la etapa inicial del proceso de anotación morfosintáctica, requiere un formato específico de digitalización de los textos a procesar (de extensión *.txt), por lo que hemos convertido las muestras en ficheros de texto.

En cuanto a la codificación de los textos muestrales, hemos usado un sistema alfanumérico en el que los números indican los datos identificativos de las muestras (texto de origen y número de muestra) y las letras, a cual de los dos subcorpus pertenecen y quien es su autor. Los datos identificativos de las muestras que se incluyen en la parte numérica del código de cada fichero son relativos al texto de origen o procedencia (solo en el caso de los

fragmentos de novela) y el número de la muestra. Las letras del código correspondientes al identificador del subcorpus son: "n" para el subcorpus de novela y "a" para el subcorpus de artículos de opinión.

Tabla 21. *Ejemplo de codificación de las mestras del corpus*

Texto de origen	**Nº de muestra**	**Id de subcorpus**	**Id del autor**	**Código**
01	04	n	mb	*0104nmb*
00	17	a	mb	*0017amb*

– *Procesamiento del corpus*

En el tratamiento del corpus ha sido necesario el uso de herramientas informáticas sobre todo debido al carácter de nuestra variable de análisis. Recordemos del capítulo 4 que los n-gramas son combinaciones de secuencias de categorías gramaticales (SEM) agrupadas según criterios y reglas específicos[90]. Por lo tanto, para recuperar sus ocurrencias en el corpus, precisamos en primer lugar de un sistema de anotación que asigne una etiqueta a cada unidad constituyente de los textos originales y, en segundo lugar, de un programa que, siguiendo determinadas reglas, fragmente los textos anotados en SEM, las extraiga y agrupe de forma automática según el tipo de n-grama (bigrama o trigrama).

[90] Para la descripción de los criterios y reglas de agrupación de SEM, véase la sección 2 del capítulo 3.

Para las tareas de etiquetaje de los corpus de esta tesis hemos utilizado las herramientas de procesamiento de textos desarrolladas en el Corpus Técnico del IULA. En cuanto a la extracción de datos, ha sido posible gracias a la aplicación *Legolas 2.0*[91].

- ***Anotación morfosintáctica del corpus. Etapas***[92]

El proceso automático de anotación morfosintáctica o el etiquetaje de los textos transcurre en tres etapas: preproceso, análisis morfológico y desambiguación. Explicamos cada una de estas etapas en los apartados siguientes.

- ***Preproceso***

Se podría decir que la etapa de preproceso es la etapa fundamental en el procesamiento computacional de un corpus, ya que de él depende la ejecución correcta del análisis lingüístico posterior. El preproceso consiste en el marcaje de los documentos del corpus. Se trata de un marcaje estructural, es decir, que detecta y anota las unidades estructurales menores del cuerpo de un texto (divisiones, títulos, párrafos, frases, etc.), y las entidades (como nombres

[91]La licencia indefinida de este programa ha sido adquirida por el Laboratorio de Lingüística Forense del Institut de Lingüística Aplicada, Universitat Pompeu Fabra.

[92] El contenido de los apartados sobre el procesamiento del corpus está basado en la información disponible en la página del proyecto Corpus Técnico del IULA.

propios, números, abreviaturas, elementos no analizables, etc.), que no se encuentran en el diccionario que usa la herramienta de procesamiento. He aquí un ejemplo de un fragmento preprocesado:

Tabla 22. *Resultado del preproceso de un texto*

Texto original	Texto preprocesado
En el noroeste de Bulgaria, el Danubio inundó la mayor parte de la zona industrial en la ciudad de Vidin, donde los niveles del agua alcanzaron los 9.66 metros. Un fuerte viento del sudeste amenazaba las barreras de bolsas de arena.	<s>En el noroeste de <name>Bulgaria</name>, el<name>Danubio</name> inundó la mayor parte de la zona industrial en la ciudad de <name>Vidin</name>, donde los niveles del agua alcanzaron los<num>9.66</num> metros.</s><s>Un fuerte viento del sudeste amenazaba las barreras de bolsas de arena.</s>

- *Análisis morfológico*

En la etapa de análisis morfológico, a cada una de la unidades lingüísticas que componen el texto procesado se le asignan de forma automática las etiquetas morfosintácticas correspondientes a todas las categorías a las que puede pertenecer la unidad. También se marcan con etiquetas los signos de puntuación y el resto de unidades extralingüísticas. El siguiente ejemplo de la tabla 23

muestra el resultado del análisis morfológico de un trabalenguas. Se trata de un texto que contiene una serie de unidades lingüísticas que presentan ambigüedad categorial. Por ejemplo, la palabra "cura" puede cumplir tanto la función de sustantivo, como la de verbo de indicativo o de imperativo, por lo que el análisis morfológico propone tres posibles etiquetas: N5-6S, VDR3S- y VRR2S-, respectivamente.

Tabla 23. *Análisis morfológico de un trabalenguas*

Texto original	Resultado del análisis morfológico			
El amor es una locura que ni el cura lo cura y si el cura lo cura es una locura de cura.	**Forma**	**Etiqueta 1**	**Etiqueta 2**	**Etiqueta 3**
	El	el:AMS	-	-
	amor	amor:N5-MS:	-	-
	es	e: N5-FP	ser: VDR3S-	-
	una	un: E6--FS	-	-
	locura	locura:N5-FS	-	-
	que	que:C	que:D	que:RR---66
	ni	ni:C	-	-
	el	el:AMS	-	-
	cura	**cura:N5-6S**	**curar:VDR3S-**	**curar:VRR2S-**
	lo	lo:ANS	pr: REEC3MS	-
	cura	**cura:N5-6S**	**curar:VDR3S-**	**curar:VRR2S-**
	y	y:C	y: N5-FS	-
	si	si:C	si: N5-MS	pr: REO-366
	el	el:AMS	-	-
	cura	cura:N5-6S	curar:VDR3S-	curar:VRR2S-
	lo	lo:ANS	pr:REEC3MS	-
	cura	cura:N5-6S	curar:VDR3S-	curar:VRR2S-
	es	N5-FP:e	VDR3S-:ser	-
	una	un:E6--FS	-	-
	locura	locura:N5-FS	-	-
	de	de:P	-	-
	cura	cura:N5-6S	curar:VDR3S-	curar:VRR2S-

- *Desambiguación*

La desambiguación es la etapa en la que se resuelven las ambigüedades resultantes de la anotación morfosintáctica. Se eliminan las etiquetas superfluas y se mantienen aquellas que el desambiguador, la aplicación que lleva a cabo este proceso, considera adecuadas conforme a las reglas de desambiguación que incorpora[93]. En la tabla 24, que se muestra a continuación, se puede ver el resultado de la desambiguación del trabalenguas del tabla 23. Los lemas que presentaban ambigüedad en el análisis morfológico están marcados en amarillo.

[93] Omitimos la descripción de las reglas de desambiguación ya que no son relevantes en el marco de este trabajo.

Tabla 24. *Resultado de la desambiguación del trabalenguas*

Resultado de la desambiguación				
Tipo	Forma	BEOS	Lema	Etiqueta
TAG				
TAG	<s>			
TOK	El	BOS	el	AMS
TOK	amor		amor	N5-MS
TOK	es		ser	VDR3S-
TOK	una		uno	E6--FS
TOK	locura		locura	N5-FS
TOK	que		que	RR---66
TOK	ni		ni	C
TOK	el		el	AMS
TOK	cura		cura	N5-6S
TOK	lo		pr	REEC3MS
TOK	cura		curar	VDR3S-
TOK	y		y	C
TOK	si		si	C
TOK	el		el	AMS
TOK	cura		cura	N5-6S
TOK	lo		pr	REEC3MS
TOK	cura		curar	VDR3S-
TOK	es		ser	VDR3S-
TOK	una		uno	E6--FS
TOK	locura		locura	N5-FS
TOK	de		de	P
TOK	cura		cura	N5-6S
DLD	.	EOS	=	DELS
TAG	</s>			
TAG				

La anotación morfosintáctica del corpus finaliza con la etapa de desambiguación, tras la cual el sistema genera un fichero individual que contiene el resultado del etiquetaje de cada texto procesado. Utilizamos estos ficheros para crear la base de datos de la herramienta de extracción de n-gramas. No obstante, antes de proceder a introducir los ficheros en el programa, llevamos a cabo la revisión manual de las etiquetas que contienen.

- *Corrección de etiquetas*

La corrección de etiquetas es una etapa anexa a la anotación morfosintáctica. Cabe aclarar aquí que el protocolo de procesamiento del Corpus Técnico del IULA que hemos seguido no incluye ninguna etapa relativa a la corrección de las etiquetas posterior a la desambiguación. Se trata de una modificación del protocolo que hemos introducido a propósito de la presente tesis doctoral y que ha sido motivada por las razones que expondremos seguidamente.

A pesar de que los sistemas de anotación disponibles en la actualidad son muy avanzados, ninguno es capaz de sustituir completamente al ser humano en las tareas de desambiguación. Incluso el etiquetaje mediante herramientas como las que hemos utilizado, cuya tasa de precisión es relativamente alta, produce errores en la asignación de etiquetas. Dado que las etiquetas son las formantes directas de nuestra variable de análisis, es de suma importancia que todos los textos del corpus estén etiquetados correctamente. Este es uno de los motivos por los que hemos considerado oportuna la revisión manual de las etiquetas. El otro motivo tiene que ver con el hecho de que el corpus fue procesado en dos períodos diferentes. La primera parte de los textos de los subcorpus N y AO fueron etiquetados en el año 2006 y la segunda parte, en el 2008. En el tiempo intermedio el etiquetario del IULA sufrió algunos cambios en su nomenclatura que afectaban a la

sistematicidad de etiquetas en el corpus de análisis (véase Tabla 25).

Tabla 25. *Modificaciones en la nomenclatura de las categorías gramaticales*

Categoría	Etiquetas		
	Corpus 2006	**Corpus 2008**	**Corpus de análisis**
determinante	**A**	**A**	**A**
conjunción	**C**	**C**	**C**
especificador	**E**	**E**	**E**
interjección	**I**	**I**	**I**
locución	**LC/LD/LP**	**LC/LD/LP**	**ETIQUETAJE INDIVIDUAL DE LOS COMPONENTES**
preposición	**P**	**P**	**P**
fecha	**T**	**T**	**T**
verbo	**V**	**V**	**V**[94]
identificador	**B**	**B**	**B**
adverbio	**D4/D6**	**D**	**D4/D6/D**
adjetivo / participio	**H**	**J**	**H**
adjetivo	**J**	**J**	**J**
nombre	**N**	**N**	**N**
pronombre	**R**	**R**	**R**
no analizable	**W**	**W**	**W**

[94] Las etiquetas marcadas en gris son las que se han sometido a corrección, como veremos más adelante.

cifra	**X**	**X**	**X**

Puesto que los textos inicialmente anotados representaban la mayor parte del corpus y que su reprocesamiento hubiera conllevado repetir una fase del trabajo de investigación que requiere mucho tiempo y que ya estaba realizada, hemos decidido corregir solo el corpus procesado en el 2008, de manera que su anotación quede unificada con el del corpus del 2006. Para lograr la unificación de las etiquetas hemos seguido unas pautas, predeterminadas por la corrección ya aplicada al corpus del 2006 y por las modificaciones en el etiquetario. Presentamos dichas pautas a continuación.

- *Pautas de corrección*

Las pautas de corrección conciernen cuatro categorías: adjetivo/participio (H), adverbio (D), locución (L) y verbo (V). Cabe decir que la categoría de locución no ha experimentado ningún cambio de modificación en el marco del etiquetario del IULA, pero la peculiar manera de su anotación ha hecho que la consideremos en la corrección.

a) H – Adjetivo/participio

Originariamente, la etiqueta H fue diseñada para designar los adjetivos deverbales y/o la forma de participio de los verbos (Morel et al., 1997). Visto que existen otras etiquetas que cumplen la misma función (J para adjetivo y VC para participio) y no presentan

problemas para la desambiguación automática, la H ha sido eliminada del etiquetario actual. Como consecuencia de esta modificación en la nomenclatura de las etiquetas, las concordancias[95] de los participios en el corpus del 2006 han sido marcadas como H y los del 2008, como VC. Hemos observado que las concordancias del corpus del 2006, en las que el desambiguador selecciona la etiqueta H frente a VC, el participio está en plural y/o femenino y forma parte de una construcción pasiva como por ejemplo la de la siguiente frase:

Ej.1 *La marcha* ha sido suspendida *por la lluvia.*

Tabla 26. *Resultado del análisis de desambiguación del Ej.1*

[95] La concordancia es el contexto en que aparece una unidad lingüística en el corpus.

Desambiguación 1 (2008)	Desambiguación 2 (2006)
TAG TAG <s> TOK La BOS el\AFS TOK marcha marcha\N5-FS TOK ha haber\VDR3S- TOK sido ser\VC--SM *TOK suspendida* *suspender\VC--SF* TOK por por\P TOK la el\AFS TOK lluvia lluvia\N5-FS DLD . EOS =\DELS TAG </s> TAG	TAG TAG <s> TOK La BOS el\AFS TOK marcha marcha\N5-FS TOK ha haber\VDR3S- TOK sido ser\VC--SM *TOK suspendida suspender\HFS* TOK por por\P TOK la el\AFS TOK lluvia lluvia\N5-FS DLD . EOS =\DELS TAG </s> TAG

En esta concordancia tanto la etiqueta H como la etiqueta VC son correctas, pero aparentemente prevalece la regla que asigna la H[96]. Para tratar este problema en la corrección de etiquetas del corpus 2006 hemos adoptado una pauta de unificación según la cual decidimos etiquetar estas unidades de acuerdo con su concordancia estándar[97]. Establecemos que la concordancia estándar del ejemplo anterior es la etiqueta H porque, es más frecuente que VC en este contexto de coincidencia de las dos. Por lo tanto, en todos los casos de construcción pasiva cambiamos la etiqueta VC del último

[96] Probablemente esto se debe a que la herramienta de anotación morfosintáctica estaba programada para distinguir los participios de los adjetivos deverbales en base al sufijo. De ahí que todos los participios que tienen forma de femenino o de femenino o masculino en plural se etiquetaran como H.

[97] El contexto prototípico en el que a una unidad lingüística se le asigna una determinada etiqueta.

participio por H. Hemos seguido la misma pauta de unificación en la corrección del corpus del 2008.

El cambio en el etiquetario no afecta a los casos de adjetivos que no son deverbales en los textos anotados, ya que sus etiquetas fueron substituidas por J en la corrección del corpus del 2006.

b) D – Adverbio

El etiquetario del 1997, que es el que hemos usado en la anotación morfosintáctica de del corpus del 2006, recoge tres etiquetas para la categoría de adverbio: adverbio común (D4), adverbio derivado documentado (D5) y adverbio derivado virtual (D6). En la actualidad, las tres han sido sustituidas por la etiqueta D. Este cambio no es significativo de cara a la corrección del corpus del 2008, puesto que la información que aporta esta división de tipos de adverbios no es relevante en esta tesis. Sin embargo, es relevante para la agrupación de SEM, porque requiere formular las reglas de modo que cubran todas las secuencias que incluyen una de las cuatro posibles etiquetas de adverbio. Para evitar tener que crear reglas repetitivas para la agrupación de las SEM del corpus 2006 que llevan un adverbio marcado como D4 o D6 en su estructura, empleamos el asterisco (*) para sustituir el número distintivo de cada tipo de adverbio y así formulamos una sola regla (véase Tabla 27).

Tabla 27. *Ejemplo de una regla de agrupación de SEM que contienen adverbios etiquetados como D4 o D6.*

SEM	Regla de agrupación	Equivalencias textuales
C.D4.VDA6S	**C + D* + VD*** **	[...]pero no trabajaba[...] [...]porque aún quedaban[...]
C.D6.VDA6S		[...]*y realmente tenía*[...] [...]*cuando finalmente lograba*[...]

c) L – Locución

Las etiquetas de locución son tres: LC (locución conjuntiva), LD (locución adverbial), LP (locución prepositiva). Sin embargo, estas etiquetas no se aplican en la anotación morfosintáctica de textos. La sigla inicial "L" es un código que figura solo dentro de la nomenclatura del etiquetario. En el etiquetaje, a cada locución se le asigna la etiqueta correspondiente a su rasgo particular[98], de modo que las locuciones adverbiales reciben la etiqueta de adverbio, las conjuntivas de conjunción, y así sucesivamente.

[98] Se denomina rasgo particular la parte del código de una etiqueta que representa los rasgos específicos de la categorías que designa dicha etiqueta (Morel et al., 1997). Por ejemplo, el tiempo verbal es un rasgo particular de las etiquetas de la categoría verbo.

A parte de esta cuestión, el mayor problema que presentan las locuciones es que en realidad constituyen una secuencia de etiquetas, pero en la anotación se consideran una única categoría a la que corresponde una sola etiqueta. El hecho de que nuestro estudio esté basado en el análisis de n-gramas hace que, si conserváramos la anotación de las locuciones que aplica la herramienta de procesamiento del corpus del IULA, estaríamos perdiendo información por el hecho de trabajar con secuencias (n-gramas) que incluyen otras secuencias (locuciones). Por lo tanto, en nuestro corpus las locuciones no están etiquetadas como tales.

De modo que la pauta de corrección que aplicamos en el tratamiento de las locuciones ha consistido en dividir sus componentes en unidades aisladas y asignarles manualmente las etiquetas según su función gramatical. Para una muestra de los dos modos de etiquetaje de una locución, en el corpus de IULA y en el corpus de la tesis, véase el ejemplo 2.

Ej.2 Anotación morfosintáctica de la oración *Sin duda son un ejemplo a seguir, pero no a ciegas.*

Anotación en el corpus del IULA			Anotación en el corpus de la tesis		
TAG			TAG		
TAG	<s>		TAG	<s>	
TAG	**<loc pos='D'>**		**TOK**	**Sin** **BOS**	**sin\P**
TOK	**Sin duda**	**BOS sin duda\D**	**TOK**	**duda**	**duda\N5-FS**
TAG	**</loc>**		TOK	son	ser\VDR3P-
TOK	son	ser\VDR3P-	TOK	un	uno\J6--MS
TOK	un	uno\J6--MS	TOK	ejemplo	ejemplo\N5-MS
TOK	ejemplo	ejemplo\N5-MS	TOK	a	a\P
TOK	a	a\P	TOK	seguir	seguir\VI----
TOK	seguir	seguir\VI----	DLD	,	=\DELIM
DLD	,	=\DELIM	TOK	pero	pero\C
TOK	pero	pero\C	TOK	no	no\D
TOK	no	no\D	**TOK**	**a**	**a\P**
TAG	**<loc pos='D'>**		**TOK**	**ciegas**	**ciego\JQ--FP**
TOK	**a ciegas**	**a ciegas\D**	DLD	. EOS	=\DELS
TAG	**</loc>**		TAG	</s>	
DLD	. EOS	=\DELS	TAG		
TAG	</s>				
TAG					

d) V – Verbo

El etiquetaje automático de los verbos suele presentar dificultades por la coincidencia que existe entre algunas de las formas de los modos imperativo y subjuntivo (por ejemplo: cante – 3ª prersona de singular de imperativo o 3ª persona de singular de subjuntivo). El etiquetario del IULA, en su nomenclatura original, prevé una serie de etiquetas para estos casos en los que el desambiguador no es capaz de distinguir entre una forma u otra, de manera que asigna una tercera etiqueta que representa la posibilidad que sea un verbo en imperativo o en subjuntivo. En 2008, las herramientas de procesamiento del IULA han evolucionado y ya permiten la

desambiguación entre estos modos verbales, de forma que las etiquetas ambivalentes resultan redundantes y han sido eliminadas del etiquetario.

Con respecto a nuestro corpus de análisis, en la parte procesada en el 2006, desambiguamos manualmente todas las etiquetas ambivalentes, sustituyéndolas por las de imperativo o subjuntivo correspondientes. Por lo tanto, no ha sido necesario aplicar ninguna corrección en el corpus procesado en 2008, donde no existen estas etiquetas.

Una vez concluida la etapa de corrección de etiquetas procedemos a la incorporación de las muestras en la base de datos del programa de extracción de secuencias de categorías gramaticales y n-gramas, *Legolas 2.0*.

d) Extracción de datos

La etapa final del tratamiento computacional de los textos de análisis consiste en la extracción de las ocurrencias de n-gramas en nuestro corpus. La hemos llevado a cabo con la ayuda de la aplicación informática del entorno Microsoft, *Legolas 2.0*.

En los apartados siguientes, describiremos los mecanismos de funcionamiento de cada una de las prestaciones del programa.

– *Prestaciones del programa*

El programa *Legolas 2.0* es el prototipo de una herramienta en desarrollo que permite la explotación de corpus anotados para los fines de la comparación interidiolectal lingüística forense de textos escritos. Dicho programa ha sido diseñado específicamente para responder a las necesidades concretas de la investigación que refleja la presente tesis doctoral, por lo que sus prestaciones se limitan a cumplir tres funciones básicas, a saber: la extracción de secuencias de categorías gramaticales o SEM, la conversión de las SEM en n-gramas y la búsqueda de equivalencias textuales de etiquetas y de secuencias de etiquetas.

• *Extracción de SEM. Descripción del proceso de segmentación*

En el diseño conceptual de la herramienta de extracción de SEM, hemos tenido en cuenta dos consideraciones, que luego también han sido implementadas en su programación, para asegurarnos de que los datos de SEM sean correctos, es decir, que cada SEM tenga un equivalente de secuencia de categorías lingüísticas. En primer lugar, es importante que las SEM extraídas no contengan en su estructura ninguna de las etiquetas que designan signos de puntuación y/o

etiquetas de marcaje estructural[99]; y, en segundo lugar, que la segmentación de los textos anotados en SEM se realice observando la estructura oracional.

El proceso de segmentación y extracción de SEM se rige por las siguientes pautas:

1. **Pauta de reconocimiento.** El reconocimiento de las etiquetas implica la reducción de las etiquetas a los componentes denominadores de la categoría gramatical de la unidad lingüística que designan[100].

2. **Pautas de segmentación**
   i. La segmentación en SEM empieza a partir de la primera palabra del primer párrafo del texto anotado, excluyendo de este modo del recuento el texto del título del documento, que no aporta información sustancial.

   ii. El recuento de SEM acaba y se reinicia después de cada punto de final de frase. Para evitar que en la segmentación se produzcan secuencias incompletas (por ejemplo, que en la extracción de trigramas se obtengan SEM de menos de tres componentes), el

[99] Véase la tabla 22 del apartado "Preproceso" de este capítulo para un ejemplo de texto con marcaje estructural.
[100] Véase la sección 2 del capítulo 3 para una descripción más detallada de la estructura de las etiquetas.

programa está configurado de manera que el último elemento de una SEM sea el primer componente de la SEM consecutiva[101].

3. **Pauta de exclusión.** Las etiquetas DELS y DELIM, que designan signos de puntuación, son consideradas con el mismo valor que el punto final en la segmentación, y no se toman como formantes de las SEM. Cuando el programa encuentra una de estas etiquetas finaliza la segmentación y la reinicia después de la etiqueta.

4. **Pauta de extracción.** En la extracción de los datos se excluyen de forma automática del resultado final, es decir, del fichero de datos que genera el programa, las SEM cuyo valor de la suma de ocurrencias:
    - es igual a cero en todos los textos de todos los autores;
    - es inferior a 10 en todos los textos de un autor;
    - es inferior a 5 en un conjunto de textos (muestras de una novela).

    Con la aplicación de esta pauta nos aseguramos de que en los datos extraídos no habrá variables que por su baja frecuencia no aportan ningún tipo de información para el análisis estadístico y podrían tener un efecto negativo en los resultados.

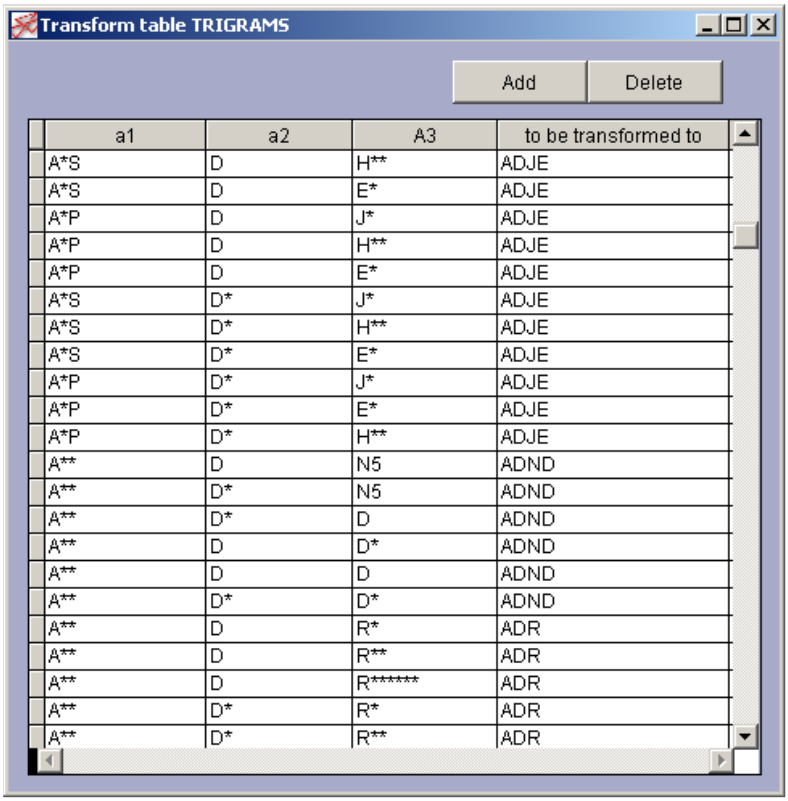
[101] Véanse la tabla 6 y 7 para un ejemplo de segmentación de un texto.

- *Conversión de las SEM en n-gramas*

La conversión de las SEM en n-gramas es la prestación de *Legolas 2.0* que permite la extracción y agrupación simultánea de SEM. Como ya hemos explicado en el capítulo 4.2, la agrupación de las variables se lleva a cabo mediante la aplicación automatizada de una serie de reglas. En el mismo capítulo, hemos descrito los criterios de formulación de dichas reglas de agrupación y su estructura. A continuación, complementamos esta descripción en los aspectos referentes a la implementación de las reglas en el programa.

La implementación de las reglas de agrupación se realiza mediante su incorporación en las *tablas de conversión* de las que dispone *Legolas 2.0* (véase imagen 1).

Imagen 1. *Captura de pantalla de la tabla de conversión de SEM en n-gramas de tipo trigrama*



a1	a2	A3	to be transformed to
A*S	D	H**	ADJE
A*S	D	E*	ADJE
A*P	D	J*	ADJE
A*P	D	H**	ADJE
A*P	D	E*	ADJE
A*S	D*	J*	ADJE
A*S	D*	H**	ADJE
A*S	D*	E*	ADJE
A*P	D*	J*	ADJE
A*P	D*	E*	ADJE
A*P	D*	H**	ADJE
A**	D	N5	ADND
A**	D*	N5	ADND
A**	D*	D	ADND
A**	D	D*	ADND
A**	D	D	ADND
A**	D*	D*	ADND
A**	D	R*	ADR
A**	D	R**	ADR
A**	D	R******	ADR
A**	D*	R*	ADR
A**	D*	R**	ADR

Como podemos ver en la captura de pantalla de la imagen 1, en las dos tablas que determinan la agrupación de las SEM en bigramas y trigramas, la disposición de las reglas es la siguiente: cada regla ocupa una fila y sus integrantes están distribuidos en columnas individuales. La última columna de cada tabla contiene el nombre del n-grama que se forma tras la conversión. El contenido de las tablas se puede modificar directamente desde la pantalla del programa haciendo clic en la fila cuyo contenido se quiere corregir. También es posible introducir nuevas reglas o eliminar reglas existentes utilizando los botones “ADD”, que añade filas en blanco

a la lista de la tabla, o "DELETE", que elimina las filas de una en una.

Poder editar las tablas de conversión en todo momento compensa un aspecto negativo del modo de introducción de la reglas en la herramienta que comentaremos seguidamente.

En relación a la implementación de las reglas de agrupación en la aplicación informática *Legolas 2.0*, cabe recordar que estas son patrones combinatorios en los que cada elemento de la estructura de las SEM está representado por siglas y asteriscos. Los asteriscos cumplen una doble función: por una parte, desempeñan el papel de designadores de las propiedades no diferenciales[102] de la unidad que representa una etiqueta y, por otra, sirven para evitar la reiteración de reglas. A pesar de aminorar de forma efectiva el número de reglas a incorporar en el programa, y así reducir el tiempo de extracción de datos, su uso puede ser la causa del mal funcionamiento de la prestación de conversión, que resulta en graves errores en la agrupación de SEM y en los cálculos de ocurrencias de n-gramas. Esto se debe al hecho de que el orden de ejecución de cada regla depende del orden de entrada de la misma en la tabla de conversión. Si, por ejemplo, en el recuento de trigramas, la regla

A*S + N5 + D

[102] Cada sigla del nombre de una etiqueta designa una propiedad de la unidad lingüística que denota. Consideramos no diferenciales aquellas siglas (propiedades) que no son relevantes para la agrupación conforme a los criterios establecidos.

precede a

ANS + N5 + D

las SEM que están encabezadas por la etiqueta ANS serán contadas dos veces, como ocurrencia de ANS y también de A*S, ya que el asterisco sustituye cualquier carácter. Por lo contrario, si el orden de las reglas es inverso, las ocurrencias de ANS se recogerán en primer lugar y no entrarán en el recuento de las reglas consecutivas que contienen A*S en la misma posición y tienen componentes idénticos. Se trata de una peculiaridad de la configuración de *Legolas 2.0* que impone que todas las reglas sean introducidas de modo que aquellas que son más específicas aparezcan antes que aquellas que son más genéricas.

- *Búsqueda de equivalencias textuales de etiquetas y tipos de n-grama*

La prestación de búsqueda de equivalencias textuales de *Legolas 2.0* ha sido diseñada, por una parte, para facilitar en el marco de este trabajo la ejemplificación de las variables de análisis junto con sus realizaciones lingüísticas y, por otra, para hacer posible para las futuras explotaciones del programa la visualización de los equivalentes de SEM. Esta prestación permite la consulta de tres tipos de equivalencias en todos los textos que se encuentran en la base de datos del programa: entre etiquetas y palabras concretas, y entre SEM bigramas o SEM trigramas y secuencias de palabras.

Como conclusión de esta descripción de las prestaciones de *Legolas 2.0*, cabe decir que, aunque el programa ha sido creado para la extracción de SEM y su conversión en n-gramas a partir de textos en español, este hecho no implica que su empleo se limite únicamente a la lengua en cuestión. Las herramientas de anotación morfosintáctica del IULA permiten procesar además textos en catalán, alemán, inglés y francés, por lo que también pueden ser tratados con *Legolas 2.0*[103] textos en estas lenguas. La posibilidad de aplicar la técnica de análisis lingüístico forense basada en n-gramas a textos en alguno de estos idiomas no ha sido explorada en esta tesis, que se centra de forma exclusiva en la lengua española; pero constituye una futura vía de trabajo.

[103] En relación a esta última afirmación, hemos de mencionar que la versión actual de *Legolas 2.0* no admite la extracción de SEM y n-gramas de textos anotados mediante otros sistemas distintos a los del Proyecto Corpus Técnico del IULA.

e) Análisis de los datos

Antes de entrar en detalles sobre el procedimiento analítico de esta tesis, conviene comentar en breve el planteamiento metodológico común que se adopta en la comparación lingüística forense para los fines de la atribución de autoría.

– *El planteamiento metodológico en comparación lingüística forense*

El análisis que conlleva la comparación lingüística forense de textos escritos en general comporta enfoques cualitativos y/o cuantitativos. La mayoría de los estudios llevados a cabo hasta el momento en el campo de la autoría se han basado en el análisis cuantitativo, ya que no solo facilita y hace más precisa la evaluación del carácter idiosincrásico y del valor discriminante de las unidades o las estructuras lingüísticas candidatas a ser marcas identificativas, sino que también es el instrumento principal en el proceso de establecer métodos de análisis lingüístico forense fiables y de llegar a conclusiones objetivas y válidas sobre su aplicabilidad (McMenamin, 2002: 138). El predominio de los enfoques cuantitativos implica un consabido problema metodológico, es decir, que cuanto menos sorprende la posibilidad de anteponer el análisis cuantitativo al cualitativo, no puede haber cuantificación sin cualificación para poder definir las variables. De hecho, el análisis

cualitativo de los datos es una parte inherente de la fase inicial de una investigación, sin la que resulta imposible delimitar el objeto de estudio y tener un enfoque metodológico correcto, por lo que todo análisis cuantitativo está precedido por un análisis cualitativo.

Cabe decir también que un trabajo de comparación forense de textos puede realizarse exclusivamente mediante un análisis cualitativo. Esto ocurre cuando la variable de análisis no es una variable discreta y el volumen de datos no permite una cuantificación fiable. Por lo tanto, en relación a las opciones de análisis que existen en el contexto forense de estudio de casos reales o de experimentos de evaluación es preciso enfatizar que la selección del tipo de análisis depende de los parámetros del conjunto de muestras textuales (número de las muestras escritas, longitud de los textos, etcétera) y del carácter discreto o continuo de la marca identificativa tomada como variable.

Los n-gramas son variables de componentes múltiples, es decir, cada n-grama resulta de la agrupación de una serie de variables (SEM) que representan diversas combinaciones de unidades lingüísticas[104]. Su análisis meramente cualitativo sería muy costoso en cuanto al tiempo invertido, ya que comportaría estudiar todas las SEM constituyentes de cada uno de los n-gramas. A parte de este problema de viabilidad, analizar los n-gramas desde una perspectiva cualitativa resultaría metodológicamente incoherente porque

[104] Para la descripción detallada de la variable de análisis consúltese la sección 2 del capítulo 3.

implicaría tener observaciones previas sobre la idiosincrasia y el carácter distintivo de todos los n-gramas, propiedades que por la propia naturaleza de las variables sólo pueden establecerse mediante cálculos porcentuales y estadísticos. De ahí que en esta tesis hemos considerado que el análisis de los datos más adecuado, vistas las características de las variables de estudio, es el cuantitativo. Sin embargo, hay que tener en cuenta que los n-gramas son unas variables de categorización cualitativa, puesto que las SEM constituyen representaciones de secuencias de categorías gramaticales y las categorías que designan son de carácter cualitativo.

Otro argumento a favor del enfoque analítico cuantitativo que se adopta en los estudios de autoría forense, incluidos entre ellos los que se llevan a cabo en este trabajo, se encuentra en el principal objetivo del análisis de comparación forense para la atribución de autoría. El análisis lingüístico forense de autoría, tal y como lo define Grant (2007: 3) de forma simplificada es, entre otras, una cuestión de clasificación que consiste en asignar un texto dubitado a un grupo de textos del mismo autor (o definido a partir de otro criterio), o bien excluirlo de dicho grupo. Este problema está siendo tratado de manera muy extensa en el procesamiento del lenguaje natural, ámbito en el cual se han desarrollado diversas técnicas estadísticas para la clasificación de textos, algunas de las cuales han sido adoptadas en la comparación lingüística forense de textos para la atribución de autoría. Entre estas técnicas están el análisis discriminante y el análisis de varianza (ANOVA) que hemos

implementado en los estudios que integran la parte empírica de esta tesis.

Los motivos que han hecho que consideremos estas técnicas apropiadas para el análisis de los n-gramas han sido los siguientes: en primer lugar, porque se trata de dos técnicas de análisis estadístico que permiten analizar al mismo tiempo una gran cantidad de variables. Y, en segundo lugar, porque ambos métodos estadísticos han sido aplicados a preguntas de autoría en otros estudios anteriores a este trabajo (véase Holmes y Forsyth, 1995 y Stamatatos et al., 2001 para algunos casos en los que se emplea el análisis de varianza; y Baayen, 2002, Chaski, 2005, Grant, 2007, Spassova y Turell, 2007 y Stamatatos et al., 2001, entre otros, para los que se usa el análisis discriminante), con resultados alentadores que apuntan a su eficacia y precisión para la evaluación de variables candidatas a marcas de autoría que podrían usarse en la comparación lingüística forense de textos.

Seguidamente ofrecemos una breve descripción del análisis discriminante y del análisis de varianza (ANOVA) a modo de introducción a los fundamentos de cada técnica.

– *Técnicas de análisis estadístico*

A modo de preámbulo a la descripción de las técnicas que hemos aplicado en la parte empírica de la presente tesis, cabe aclarar algunos conceptos básicos de la estadística y explicar cómo se

traducen en el contexto de la comparación lingüística forense para los fines de la atribución de autoría de textos escritos. Estos conceptos tienen que ver, por una parte, con la denominación de los elementos de análisis (población, grupo y caso) y con la interpretación de los resultados (hipótesis nula y p-valor), por otra.

Cuando en atribución de autoría se hace referencia a la agrupación de individuos de los que se sospecha que pueden haber producido un escrito cuya autoría se cuestiona, se habla siempre del universo de probables autores. La estadística emplea el término *población* para designar este universo o cualquier otro conjunto de posibles sujetos de análisis, y *grupo* para denominar los diferentes miembros de la población. En un estudio de autoría hipotético, los autores serían los grupos y los textos que se incluyen en el análisis estadístico para cada autor, incluidos los del autor desconocido o anónimo, representarían los *casos*. Los textos para los que no se dispone de información acerca de su grupo de pertenencia (en el contexto del análisis lingüístico forense estos serían los textos anónimos o dubitados) se definen en estadística como *casos desagrupados*.

Los conceptos de hipótesis nula y p-valor están relacionados con la tarea de establecer si los resultados obtenidos mediante el análisis estadístico son significativos y cuál es su nivel de significación. Para corroborar la validez de los resultados de la comparación lingüística forense en atribución de autoría es crucial demostrar que las diferencias que se observan en el uso de una o varias marcas

identificativas en distintos autores no se dan a consecuencia de la variación en el muestreo de la población de análisis o de otro factor, sino que son a causa de la variación inter autor. Las técnicas estadísticas incluyen pruebas específicas, llamadas pruebas de significación, que se aplican con el fin de determinar cuál es la probabilidad de que la diferencia entre dos o más grupos se deba o no al azar. El grado de probabilidad o el nivel de significación de un análisis está reflejado por el *p-valor*. Las pruebas de significación parten siempre de una hipótesis que asume lo opuesto a lo que se pretende demostrar con la realización del análisis estadístico. Esta hipótesis es conocida con el nombre de *hipótesis nula*. En el caso de un análisis lingüístico forense, la hipótesis nula sería que no existe ninguna diferencia entre los grupos. Si al testar esta hipótesis el p-valor que se obtiene es inferior o igual a 0,05, las diferencias entre los datos se pueden considerar significativas y, por consiguiente, se descarta la hipótesis nula.

- *El análisis discriminante*

El análisis discriminante (AD) es una técnica estadística compleja que comprende una serie de pruebas estadísticas que responden a dos finalidades: por un lado, las que hacen posible identificar las características que marcan las diferencias entre los grupos y permiten distinguir entre ellos y, por otro, las pruebas que tienen como fin la clasificación de nuevos casos en el grupo más probable de pertenencia (Klecka, 1980: 8-9).

Las pruebas que identifican las características diferenciales de los grupos se ejecutan en primer lugar al iniciar el AD. Estas pruebas toman los valores de las variables de análisis y mediante diversos cálculos crean tantas ecuaciones matemáticas como variables se hayan incluido en el AD. Estas ecuaciones se denominan funciones discriminantes y son la combinación lineal de las variables de análisis. De la totalidad de funciones que se generan, el segundo tipo de pruebas aprovecha para la clasificación solo una función discriminante: la que logra separar lo más posible los grupos en el espacio multidimensional que ocupan. Cada grupo tiene un valor en la función discriminante que es igual a la media de las puntuaciones de todos los casos que pertenecen al grupo en particular. La estadística se refiere a este valor promedio como centroide de grupo. Para la clasificación de nuevos casos, en el análisis se ponderan las distancias entre los centroides de los grupos previos y el grupo de casos desagrupados para determinar cuál es el grupo con el que los casos nuevos guardan mayor grado de similitud. Cuanto más cercano esté el centroide del grupo no identificado al centroide de un grupo conocido, tanto más alta es la probabilidad de que los casos desagrupados pertenezcan a este último grupo. En esta tesis hemos utilizado el análisis discriminante exclusivamente en los estudios sobre la variación inter autor y en los de evaluación de los n-gramas en casos reales.

- *El análisis de varianza (ANOVA)*

El análisis de varianza o ANOVA es una técnica de análisis estadístico que sirve para la comparación de grupos en base a la media de cada grupo para cada una de las variables de análisis. El resultado de la comparación puede revelar si existen diferencias o similitudes entre los grupos analizados. El grado de disimilitud en la comparación está reflejado en los resultados por el estadístico F (*F-statistic*), que representa el ratio de variación poblacional estimada entre las medias de cada grupo (la variación inter autor) respecto a la variación poblacional estimada que se observa dentro de cada grupo (la variación intra autor) (Pardo y Ruiz, 2002). Cuando el valor del estadístico F es mayor que 1, la diferencia entre las medias y los grupos comparados se puede considerar estadísticamente significativa, mientras que cuando su valor es inferior o próximo a 1, esto indica que las medias poblacionales son iguales y los grupos muy similares. Por la función que cumple, el ANOVA constituye en el ámbito de atribución de autoría una técnica muy adecuada para las pruebas que se realizan en el análisis lingüístico forense con el propósito de establecer si el uso de una unidad lingüística varía de forma significativa a nivel inter autor y no a nivel intra autor, de modo que podría considerarse marca de autoría.

En el marco de esta tesis, la técnica estadística ANOVA se aplica en el análisis llevado a cabo en los estudios sobre variación intra autor,

sin que forme parte de la propuesta analítica de la presente tesis doctoral, para cumplir un doble objetivo: por un lado, determinar si la distancia intertextual dentro del mismo autor es inferior a la distancia entre los textos de diferentes autores y, por otro, confirmar que, aunque el tiempo de medición puede conllevar algunos cambios idiolectales en la producción lingüística escrita del individuo, no influye en la capacidad discriminatoria de los n-gramas como marcas de autoría.

3.6 Propuesta analítica de la tesis doctoral

A parte de la validación de determinadas hipótesis, la propuesta analítica que se ofrece en esta tesis doctoral representa una técnica de comparación de textos escritos en español para los fines de la atribución de autoría basada en los n-gramas como marcas idiosincrásicas de autoría. La técnica de análisis lingüístico forense que hemos desarrollado implica un procedimiento de tres pasos: en el primer paso se lleva a cabo la selección entre la totalidad de variables obtenidas en la extracción de los n-gramas que serán analizados; en el segundo, se estandardizan los datos de ocurrencia de los n-gramas seleccionados; y finalmente, en el último paso, se realiza el análisis estadístico de los textos de los autores del corpus, de acuerdo con los objetivos e hipótesis de la tesis que se plantean y se testan en el tipo de estudio en particular. A continuación explicamos en mayor detalle en que consiste cada paso.

a) Selección de n-gramas

Una de las principales hipótesis que se espera poder validar en este trabajo es que los n-gramas más frecuentes pueden discriminar entre las producciones lingüísticas escritas de diferentes individuos. El fichero de datos que genera la herramienta de extracción, Legolas, contiene todos los n-gramas presentes en el corpus procesado y este hecho impone la necesidad de hacer una selección de la lista completa de variables obtenidas en función de algún criterio.

De cara al análisis discriminante, las variables seleccionadas deben ser dos menos que el número de casos. Sin embargo, dado que para la mayoría de los estudios de variación inter e intra autor el número de casos (muestras escritas) varía entre 100 y 400, tendríamos que trabajar con 98 o 398 variables, dependiendo del estudio concreto, entre las que se incluirían muchos n-gramas de baja frecuencia. Por lo tanto, hemos optado por fijar un criterio de selección a partir de un número mínimo de ocurrencias. Este mínimo, tanto para los n-gramas de tipo bigrama como de trigrama, es de 100 y es calculado a partir del recuento de n-gramas en cada corpus específico analizado. Siguiendo este criterio hemos podido disponer para el análisis de cada estudio de una cantidad similar de variables: entre 80 y 89 n-gramas[105].

En relación al estudio de evaluación de los n-gramas como marca discriminante en casos reales, cabe aclarar que ha sido una

[105] Para un listado de las variables de análisis de cada estudio véase el anexo VI.

excepción de la aplicación del criterio referido previamente, ya que el volumen de datos es mucho más reducido, se trabaja con bastante menos textos y su extensión es inferior a la de los del subcorpus de fragmentos de novela (subcorpus N) y de artículos de opinión (subcorpus AO), lo que implica una N pequeña de variables de frecuencia mayor. En consecuencia, en este estudio la selección comprende un número de variables que concuerda con el número de documentos de análisis.

b) Estandarización de los datos de n-gramas

Los datos de n-gramas que proporciona la herramienta de extracción son recuentos numéricos de las ocurrencia de cada unidad de análisis en cada muestra y autor. Para ser tratados correctamente en el análisis que aplicamos precisan ser convertidos en valores estandarizados. Estandarizar los datos en el marco de los estudios que llevamos a cabo es importante para evitar que los n-gramas con el mayor rango de variación dominen en los análisis cuando puede haber otros n-gramas menos frecuentes, pero significativos, que poseen valor discriminante.

Existen diversos métodos de estandarización. El que hemos empleado en esta tesis doctoral es el que mejor se ajusta al propósito de la conversión de los datos, ya que elimina las grandes diferencias en la distribución entre las variables de valores más altos y valores más bajos. Dicho método de estandarización consiste en extraer la media del grupo del valor de cada variable y dividir el

valor resultante por la desviación estándar[106]. Los datos así tipificados son los que usamos en el análisis estadístico descrito a continuación.

c) Análisis estadístico de los textos

El análisis de los textos que culmina el proceso de aplicación de la técnica de comparación lingüística forense presentada en esta tesis incluye una serie de pruebas estadísticas que, conforme a la finalidad que cumplen, se pueden dividir en grupos de pruebas de clasificación, de determinación y de evaluación.

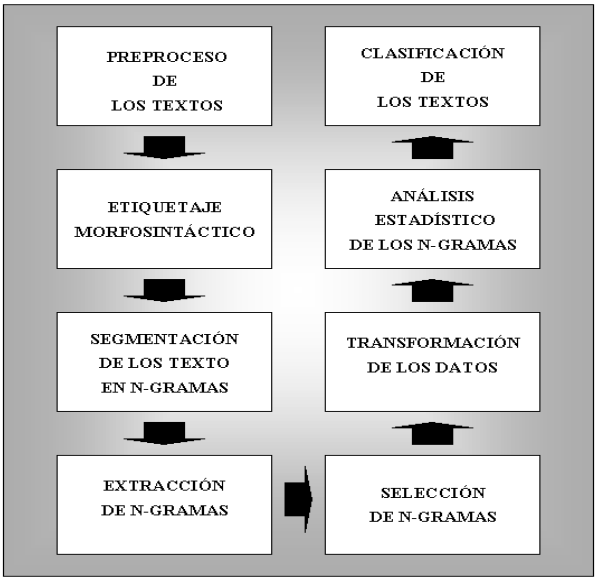
En las *pruebas de clasificación* se analizan los textos del corpus o subcorpus concreto con el propósito de determinar en qué grado los n-gramas más recurrentes son capaces de discriminar entre los distintos escritores y de agrupar correctamente los textos según su autor. Las *pruebas de determinación* pretenden establecer cuales son los n-gramas discriminantes en cada contexto de análisis y, por lo tanto, poseen un potencial discriminatorio más alto. Estas pruebas se realizan mediante el método de inclusión por pasos del análisis discriminante, que determina la significación individual de cada variable en el modelo discriminante y construye una función discriminante basada solo en aquellas variables que son útiles para la clasificación. Por último, las *pruebas de evaluación*, tienen el objetivo de mostrar si la técnica funciona adecuadamente, es decir,

[106] Recuérdese que la desviación estándar es igual a la raíz cuadrada de la varianza de la distribución del grupo.

si rinde resultados válidos y fiables en la clasificación (atribución) de textos de autor(es) anónimo(s) y cuál es su nivel de precisión y margen de error. Las pruebas de este tipo se basan en los datos de los n-gramas que han sido identificados como portadores de rasgos estructurales idiosincrásicos y discriminantes en las pruebas de determinación previas.

Como conclusión de este apartado presentamos un esquema (véase el cuadro 3) a modo de resumen del procedimiento completo de la propuesta de técnica de comparación lingüística forense para los fines de la atribución de autoría.

Cuadro 3. *Esquema de las etapas de aplicación de la técnica de comparación lingüística forense para los fines de la atribución de autoría mediante n-gramas*



En este capítulo hemos abordado las principales cuestiones en torno a la metodología aplicada en esta tesis doctoral: en primer lugar, la descripción del diseño experimental en el que se basa, los procesos de recuperación de datos y el funcionamiento de la herramienta de extracción *Legolas 2.0,* y en segundo lugar la exposición del procedimiento y los método de análisis estadísticos implementados en los estudios sobre variación inter e intra autor. Finalmente, hemos descrito la propuesta analítica de la tesis doctoral.

4. ESTUDIOS SOBRE LA VARIACIÓN INTRA AUTOR

La variación es un rasgo inherente al lenguaje y al uso lingüístico que se observa tanto en la comparación del estilo idiolectal de los diferentes individuos como en el marco de la actividad lingüística oral y escrita de una persona. La variación que se produce entre los distintos usuarios de una lengua es la que se denomina variación inter autor y se ve afectada por distintos factores tanto internos (lingüísticos) como externos (estilísticos, sociales y pragmáticos) [107]. En consecuencia, en el ámbito de la lingüística y la sociolingüística es posible hablar de dialectos, sociolectos e idiolectos entre otros, como variedades lingüísticas de la misma lengua. La premisa de que este tipo de variación tiene sus rasgos particulares en el estilo individual de cada persona y que es posible detectarla en sus producciones lingüísticas constituye la base de la lingüística forense en su rama de la comparación forense que usa los rasgos estilísticos idiosincrásicos detectados en el análisis lingüístico para atribuir o determinar el autor más probable de textos cuya autoría se cuestiona. El otro tipo de variación, la que se percibe en los textos escritos y orales de un individuo, es la que se define como variación intra autor. Esta variación se expresa en los cambios que puede experimentar el estilo de una persona a causa de

[107] Véanse las secciones 1 y 2 del capítulo 1.

factores relacionados con las propiedades del texto como, por ejemplo, el género textual y otros ligados al individuo[108].

A diferencia de la variación inter autor, la variación intra autor puede comportar implicaciones negativas directas para el trabajo práctico que tiene que llevar a cabo el experto en lingüística forense, en este caso, en la atribución forense de autoría. Dichas implicaciones radican, por una parte, en el hecho de que muchas veces las pruebas dubitadas y las indubitadas del sospechoso o los sospechosos disponibles en un caso forense real pertenecen a diferentes géneros textuales. Por otra parte, también se puede dar la circunstancia de que entre los momentos en los que fueron escritos los dos conjuntos de textos haya transcurrido un largo o considerable período de tiempo. Si los autores tendieran a variar en el uso de determinadas variables (parámetros o marcas) lingüísticas de forma significativa de un texto a otro por la influencia del tiempo o del género textual, las pruebas periciales serían muy difíciles de realizar o darían resultados erróneos.

En este capítulo se abordan las dos preguntas de investigación o hipótesis que se formulan en torno a las implicaciones de la variación intra autor en la metodología y la práctica de la comparación de textos escritos mediante n-gramas (bigramas y trigramas) para los fines de la atribución forense de autoría. En primer lugar, en el estudio sobre la variación en tiempo aparente y en tiempo real, se trata de establecer si el tiempo de medición

[108] Para el comentario de estos factores véase la sección 2 del capítulo 4.

produce variación intra autor en relación a la recurrencia de los bigramas y los trigramas y cuál es el nivel de significación de esta variación. En segundo lugar, se concluye si hay variación intra autor estadísticamente significativa entre los textos del mismo autor producidos en dos géneros textuales diferentes: novela y artículo de opinión.

4.1 Estudio sobre la variación en tiempo aparente y tiempo real real

Este estudio tiene como objetivo confirmar una de las principales hipótesis que se formula en esta tesis doctoral, en el sentido de que la variación que se observa en las producciones lingüísticas escritas de un mismo individuo (intra autor) es menor que la variación que se da entre los escritos de este individuo y los de otros usuarios de la misma lengua (inter autor), en este caso el español, utilizando como marca de autoría los n-gramas de tipo bigrama y trigrama.

a) Corpus del estudio

El corpus del estudio sobre la variación en tiempo aparente y en tiempo real está constituido por los textos del subcorpus N de dos grupos de autores[109]. El grupo 1 comprende 5 escritores de cuyos textos se recogen los fragmentos de las novelas que reflejan el estilo

[109] Metodológicamente, la consideración de estos dos grupos tienen que ver con el hecho de que no se disponía de lo tres tiempos de producción y de la misma distancia en la producción para todos los autores.

escrito de cada autor del grupo en tres tiempos de producción: tiempo aparente (TA), tiempo intermedio (TI) y tiempo real (TR). Los autores de este grupo son Eduardo Mendoza (EM), Mario Vargas Llosa (MV), Gabriel García Márquez (GG), Miguel Delibes (MD) y Carmen Laforet (CL). El grupo 2 está formado por el mismo número de autores que el grupo 1 de los cuales se incluyen en el corpus también los textos de novelas representativas de tres tiempos de producción en su carrera como escritores. Los escritores que forman parte del grupo son Alicia Giménez Bartlett (AG), Antonio Muñoz Molina (AM), Arturo Pérez-Reverte (AP), Carmen Posadas (CP) y Rosa Montero (RM). En el caso del grupo 1 la distancia entre los tiempos de producción es mayor (entre 20 y 30 años) y cada período está representado por los fragmentos de una novela (véase tabla 28) mientras que en el caso del grupo 2 la distancia entre los tiempos de producción es menor (entre 5 y 8 años) y para cada período se dispone de los fragmentos de dos novelas (véase tabla 29)[110].

Tabla 28. *Distribución de escritores por mayor distancia entre períodos de producción (Grupo 1)*

Identificadores de los sujetos	Tiempo de producción		
	Tiempo aparente (TA)	Tiempo intermedio (TI)	Tiempo real (TR)
EM	1975	1990	2002
MV	1963	1987	2003
GG	1967	1985	2004
MD	1947	1966	1998
CL	1945	1963	2004

[110] Para consultar los datos bibliográficos de las novelas usadas en este estudio véase el anexo I.

Tabla 29. *Distribución de escritores por menor distancia entre períodos de producción (Grupo 2)*

Identificadores de los sujetos	Tiempo de producción		
	Tiempo aparente (TA)	Tiempo intermedio (TI)	Tiempo real (TR)
AG	1991	1999	2006
AM	1986	1991	2000
AP	1988	1990	2002
CP	1983	1996	2001
RM	1981	1988	1997

b) Variables

Los n-gramas de tipo trigrama y bigrama en los que se basan los análisis del presente estudio han sido seleccionados mediante la prueba de determinación, descrita en el capítulo 4, que permite establecer cuáles son las variables de mayor potencial discriminatorio y hacen posible diferenciar entre los escritos de los distintos autores del corpus. La selección se ha llevado a cabo a partir de los datos de los n-gramas más frecuentes en los textos de análisis, 54 bigramas y 79 trigramas para el grupo 1, y 71 bigramas y 88 trigramas para el grupo 2[111]

c) Análisis

Para los análisis que abarca este estudio se ha usado la técnica estadística ANOVA, ya que ha sido específicamente diseñada para el contraste de la variación que se da a nivel de inter e intra autor.

[111] Se trata de una selección basada en las variables previamente escogida según el criterio descrito en la sección 6 del capítulo 3. Para el listado de estas variables véase el anexo V.

Las particularidades analíticas de esta técnica han sido descritas en detalle en el capítulo 0. Cabe recordar, sin embargo, que la hipótesis de entrada que se testa mediante el ANOVA asume que no existe diferencia entre las medias de los grupos analizados o en el caso concreto de este trabajo, que no se producen cambios estadísticamente significativos en la recurrencia de los n-gramas (bigramas y trigramas) entre las novelas escritas en distintos años por los autores del corpus. Por lo tanto, si los resultados de análisis asignan un p-valor igual o inferior a 0,05 a un n-grama, la interpretación que se debe hacer es que el uso de ese n-grama experimenta cambios y no se mantiene con una frecuencia estable en los textos analizados; en cambio, si el p-valor es superior al de significación, el n-grama puede considerarse una variable que no experimenta cambios y se mantiene con una frecuencia estable en los textos analizados.

- *Análisis de los textos con mayor distancia en el tiempo de producción (Grupo 1)*

El análisis de los n-gramas de los textos de mayor distancia en el tiempo de producción se basa en los 12 bigramas y los 11 trigramas más discriminantes en el corpus del estudio. El listado de ambos tipos de n-gramas se presenta en la tabla 30.

Tabla 30. *Lista de los bigramas y los trigramas más discriminantes en los textos de mayor distancia en el tiempo de producción*

BIGRAMAS	TRIGRAMAS
VC	CREV
VAL	PASN
VRR	NCNP
VG	NPT
PN	RRVDA
DE	PEN
WV	VJEN
NV	VPV
DA	VVPC
DVP	DPT
VY	NEJT
DT	

- *Resultados del análisis de bigramas*

A continuación se ofrecen los resultados individuales del análisis ANOVA de los fragmentos de novela de los autores que forman el grupo de sujetos del corpus del estudio con mayor tiempo de medición en la producción de las obras de las que han sido extraídos los textos de análisis. Las tablas detalladas de resultados para cada autor y variable, generadas por el programa SPSS se pueden consultar en el anexo VI.

Resultados para el autor EM

En la comparación de los textos de EM, el análisis detecta que hay una variación significativa en la recurrencia de los bigramas DVP, VAL y DA entre los conjuntos de textos representativos de los tres tiempos de medición en la producción lingüística del autor

estudiado. La comparación por parejas de los textos en tiempo aparente (TA) con los textos de tiempo de producción intermedio (TI) y real (TR) indica que la diferencia entre las novelas se observa en 7 de los bigramas originales (PN, WV, VG, DVP, VAL, VY Y DA). 5 de estas variables son las que marcan la variación intra autor estadísticamente significativa entre las novelas de los años 1975 y 2002 que representan el estilo del autor EM en TA y TR. Para los textos de novela de TI y TR que están más próximos, la variación significativa se contempla en 3 bigramas. Estos bigramas son VAL, DA y VG en la comparación de datos del 1990 y 2002 y WV, DVP y VY en la de 1975 y 1990. Estos resultados significan que en el caso de EM la variación intra autor podría ser un factor que afecte la clasificación o atribución correcta de textos de este autor.

Resultados para el autor MV

El ANOVA de los textos del autor MV muestra que los bigramas PN y VAL son las únicas variables que varían de forma significativa entre TA, TI y TR. Los fragmentos de la novela en tiempo aparente (1963) y tiempo real (2003) son la que más se distinguen entre sí en las frecuencias de los bigramas más discriminantes. Dado que se trata de una distancia de 40 años entre producciones es de esperar que en el estilo del autor se produzcan variaciones en cuanto al uso de determinadas estructuras y construcciones y de ahí en la recurrencia de los n-gramas. No obstante, se observa una variación intra autor solo en la frecuencia de 3 de los bigramas analizados, en concreto el PN, VG y VAL. Estos resultados conducirían a concluir que cuando se analizan

textos de este autor con fines forenses el factor variación intra autor no va a influir en los resultados si los textos seleccionados son de un período inferior a 30 años.

Resultados para el autor GG

Los bigramas del conjunto de variables más discriminantes que comportan variación intra autor en los textos del autor MD, son dos: VC y VAL. Estos n-gramas son los mismos que junto con DA distancian estilísticamente las novelas que reflejan el estilo del autor en TI (1985) y TR (2004). Los textos de MD en TA (1967), sin embargo, tienen valores de las medias que los distinguen de los textos de la novela en TI (1985) únicamente en la frecuencia de uno de los bigramas que discriminan a nivel intra grupo: el bigrama VAL. Estos resultados significan que la variación intra autor tampoco influye en el estilo del autor cuando no han transcurrido más de 20 años.

Resultados para el autor MD

El análisis de la varianza aplicado a los textos del escritor MD indica que los brigramas que presentan variación intra autor son DE, DT, VG y VRR. Esta variación es más destacada entre los fragmentos recogidos en TI (1966) y TR (1998) y en TA (1947) y TI. Las diferencias estilísticas entre el binomio TI y TR se muestran también en la recurrencia del bigrama PN. Una explicación plausible de este comportamiento sería que el estilo del autor en la novela de TI es diferente a causa de otro factor como por ejemplo el tema o el género literario.

Resultados para la autora CL

Para los textos de novela de la autor CL el análisis registra valores de la variación estadísticamente significativos en las frecuencias de los bigramas WV, VAL, DT y VRR. La variación intra autor más notable se produce en el intervalo temporal en el que CL escribió las novelas del 1945 y 1963. En el caso de esta escritora, nuevamente podemos concluir que su estilo no es susceptible a cambios significativos cuando el tiempo entre las producciones lingüísticas es inferior a 20 años.

– *Conclusiones*

Como conclusión de los resultados expuestos es importante anotar que, del total de variables analizadas, menos de la mitad se muestran susceptibles a la variación intra autor en los escritores de este estudio. Los resultados demuestran, además, que la variación estadísticamente significativa expresada en el cambio de la frecuencia de las estructuras idiosincrásicas y distintivas del estilo del autor, representadas por los n-gramas de tipo bigrama, se manifiesta en el redactado de un escritor experimentado tras períodos de escritura de entre 20 y 30 años. Estos resultados confirman la hipótesis inicial de que los n-gramas no varían en su frecuencia en el transcurso del tiempo.

– *Resultados del análisis de trigramas*

Resultados para el autor EM

Los resultados del análisis de los fragmentos de novela del EM revelan que del total de 11 trigramas de mayor potencial discriminatorio solo dos (PASN y NPT) experimentan una variación intra autor estadísticamente significativa en la escritura del autor. En la comparación múltiple de los textos en los tres tiempos de producción analizados, esta variación se detecta exclusivamente entre los fragmentos de novela que representan el estilo de EM en TA (1975) y TR (2002). El bigrama NPT es también objeto de variación en el tiempo entre la novela del 1990 y del 2002, pero se trata de un caso que no se registra para ninguna otra variable. Estos resultados confirman la hipótesis inicial ya que no revelan ninguna variación intra autor significativa durante un período largo.

Resultados para el autor MV

La variación entre los tres conjuntos de textos del escritor MV según los resultados obtenidos en su análisis alcanza niveles de significación estadística de modo singular en el trigrama PEN. En lo que respecta a la variación intra autor observada en la comparación de los grupos por parejas, los datos del ANOVA atribuyen valores de las medias con diferencias significativas para las frecuencias de los trigramas PASN, PEN, VPV y CREV. Todas estas variables varían en su recurrencia de forma significativa cuando se aplica la comparación concreta de los fragmentos de las novelas del TR (2003) y el TA (1963). Los trigramas PEN y CREV muestran variación entre los textos representativos de la producción escrita de

MV en TI (1987) y TA y entre el TI y el TR, respectivamente. El análisis de este autor produce resultados que confirman la hipótesis planteada en la tesis sobre la variación intra autor.

Resultados para el autor GG

Los trigramas que varían en su recurrencia en los textos del escritor GG son dos: VVPC y VJEN. Entre los tres grupos de textos de análisis se presentan diferencias de variación en las frecuencias de las mismas variables al comparar las novelas en TI (1985) y TA (1967). Los resultado para el autor GG también indican que los fragmentos de las novelas en TI y TR se distinguen significativamente solo en el trigrama VJEN, lo cual indicaría que en un posible análisis de la variación inter autor entre GG y otros autores, la discriminación sería posible.

Resultados para el autor MD

Los resultados del análisis de los trigramas de los textos del escritor MD muestran que en 4 de los n-gramas de este tipo para los que se ha determinado que son más discriminantes se produce una variación intra autor estadísticamente significativa. Las variables en cuestión son PASN, PEN, NCNP y NEJT. Cabe destacar que MD es el autor con el mayor número de trigramas que experimentan cambios en su recurrencia en el transcurso del tiempo. En cuanto a los resultados de la comparación entre los conjuntos muestrales de cada novela, los trigramas arriba enumerados y RRVDA son los que marcan las diferencias, sobre todo entre el conjunto que representa

la producción escrita de MD en TR (1998), lo cual indica que los textos que le corresponden son completamente diferentes a los de los otros años en los valores de recurrencia. Se observa también una variación de p-valor significativo entre la frecuencia del trigrama NEJT de los textos de TA y los textos de TI. Estos resultados permitirían concluir que la producción escrita de este autor no presenta variación hasta un período mínimo de 50 años.

Resultados para la autora CL

Los resultados para la autora CL indican que los trigramas que varían de forma significativa en los textos de las novelas seleccionadas de su obra, son PASN y VPV. La frecuencia de estas variables diferencia los fragmentos de novela de cada conjunto de textos en la comparación por parejas. Según los datos del análisis los textos representativos del TA (1945) y TI (1963) se distinguen significativamente de los textos del TR (2004).

– *Conclusiones*

Los trigramas más discriminantes, cuya frecuencia se ve afecta por el factor tiempo de medición, constituyen en la gran parte de los autores analizados una tercera parte de las variables empleadas en el análisis de la varianza, a diferencia de los bigramas para los cuales las variables no alteradas por el transcurso del tiempo constituían la mitad. Los casos que se han observado de variación intra autor estadísticamente significativos se dan entre los fragmentos que

provienen de novelas que han sido escritas con un mínimo de 10 años de diferencia en el tiempo.

Desde la perspectiva de un enfoque centrado en los autores individuales es posible concluir que los escritores MV, EM y CL son más constantes en la frecuencia de los trigramas que se obtienen en la segmentación de sus textos que en los bigramas. El autor en el que se manifiesta el mayor número de bigramas y trigramas que varían en sus frecuencias entre los grupo de textos es MD. En cambio, los escritores en los que se aprecia muy poca variación intra autor en el caso de los bigramas, en contraste con los otros tres sujetos de análisis, son MV y GG. Por último, el escritor que refleja muy poca variación en el caso de los trigramas, en comparación con los otros cuatro sujetos, es MV.

– *Análisis de los textos con menor distancia en el tiempo de producción (Grupo 2)*

En el análisis basado en n-gramas de los escritores cuyos textos han sido producidos con un tiempo de medición menor se han utilizado los 14 bigramas y los 13 trigramas más discriminantes entre esos autores. La lista de estos n-gramas se presenta en la tabla 31.

Tabla 31. *Lista de los bigramas y los trigramas más discriminantes en los textos de menor distancia en el tiempo de producción*

BIGRAMAS	**TRIGRAMAS**
NC	ANP
CP	ENVAP

NV	VPVI
ASN	JNT
PN	CPAS
DT	ASNE
EVR	CVAC
NVR	VVTI
EN	ECE
VAK	VVPA
EC	ASNV
DE	VEJ
VC	JHPT
CVR	

- *Resultados del análisis de datos de bigramas*

En este apartado se presentan los resultados del análisis de varianza de los sujetos del corpus basado en bigramas de los fragmentos de novela del grupo de autores de los cuales se dispone de textos con menor distancia en el tiempo de producción. La tabla detallada de los resultados del análisis llevado a cabo con el SPSS para cada autor se incluye en el anexo VI.

Resultados para la autora RM

El análisis de los textos de la autora RM ha arrojado resultados que indican que hay dos bigramas en cuya recurrencia se observan diferencias significativas en los fragmentos de novela de cada tiempo de medición. Estas variables son EN y CVR. En la comparación de las novela por parejas, además de los bigramas que se acaban de citar, la variable del grupo analizado para la que se detecta una variación intra autor significativa entre los textos

representativos del TA (1981) y TR (1993) es DT. Entre el TI (1988) y el TR de la escritora RM se observa variación solo en los valores de las medias de la frecuencia de los bigramas VC y VAK. Estos resultados implican que la variación en la manera de escribir de esta autora no es significativa a corto plazo (menos de 10 años aproximadamente).

Resultados para el autor AP

Del total de bigramas analizados en el caso de AP solo la frecuencia de la variable EN muestra una variación intra autor estadísticamente significativa entre todas las novelas seleccionadas de este autor. Al comparar entre sí los textos de cada período, el análisis encuentra variación en un grado significativo entre el TA (1988) y el TI (1991) en las variables ASN y EC. Este resultado llevaría a concluir que este autor se mantiene relativamente constante en su uso de los recursos lingüísticos.

Resultados para la autora AG

Los resultados del análisis de los bigramas de los textos de la autora AG revelan que los bigramas que varían de forma significativa en su recurrencia en el transcurso del tiempo son EN y CP. El bigrama EN distingue sobre todo los fragmentos de la novela de TR (2006) de los fragmentos de novela de TI (1999) y TA (1991), mientras que la variación en la frecuencia de CP implica la diferencia entre los textos de TA y TI. Estos resultados comportan que el estilo de

esta autora no cambia de forma significativa en un espacio temporal de 15 años.

Resultados para el autor AM

Según el ANOVA, AM es el autor para el cual la variación intra autor se da en el mayor número de bigramas. Los bigramas en cuestión son NC, ASN, NVR y NV. Junto a DT tres de estas variables, NC, ASN y NVR, varían significativamente en el tiempo de medición entre el año 1986 y el 2000. Los fragmentos de la novela de TA (1986) y TI (1991), en cambio, se distinguen en los valores de media de la frecuencia de los bigramas EN, ASN y PN. En la comparación de textos del TI y el TR (2000) el único bigrama que manifiesta variación en su recurrencia es el NC. Estos resultados confirmarían la hipótesis según la cual los n-gramas no varían en su frecuencia de forma significativa en el transcurso del tiempo.

Resultados para la autora CP

Los resultados obtenidos para la autora CP indican que la variación en los textos representativos de su producción escrita en el intervalo que abarca los tres tiempos de medición, posee un p-valor de significación en el bigrama del mismo acrónimo (CP). Las diferencias que marca esta variable la distinguen de las demás novelas, en particular, la que la escritora publicó en el 1983. Este resultado llevaría a concluir que esta autora en concreto no varía en

su estilo en el margen de tiempo de su producción lingüística representado en el corpus analizado para este estudio.

- *Resultados del análisis de datos de trigramas*

Resultados para la autora RM

El análisis de trigramas de los textos de RM indica que en esta autora se manifiesta variación intra autor solo en la frecuencia de una variable, CVAC. Se trata del mismo trigrama cuyos valores de las medias se muestran significativamente distintos en la comparación de las novelas de TA (1981) y TR (1993). Este resultado permite concluir que no hay diferencia en la frecuencia con la que los n-gramas de tipo trigramas recurren en sus textos y por lo tanto se confirma la hipótesis de la variación mínima a corto plazo.

Resultados para el autor AP

Los resultados para el autor AP muestran que, de entre todas las variables más discriminantes entre los escritores del grupo 2, en los textos de las novelas que reflejan su estilo escrito únicamente varía a un nivel de significación relevante el trigrama VVPA. En la comparación de los fragmentos de narrativa de TA (1988) y TI (1990) se observa otra variable, ECE, que presenta variación solo en el caso descrito. Estos resultados confirmarían la hipótesis según la cual la variación intra autor en la frecuencia de los n-gramas no

alcanza niveles de significación cuando se trata de un período corto entre las producciones lingüísticas.

Resultados para la autora AG

Los resultados del análisis de la varianza en los textos de la autora AG revelan que la variación intra autor en los textos de AG se detecta a un nivel estadísticamente significativo en los trigramas ASNV y CPAS. Estas son las únicas variables que presentan variación en la producción escrita entre la publicación de las novelas en TA (1991) y TR (2006) y las novelas en TI (1999) y TR. Este resultado permite concluir que esta autora es bastante constante en su estilo.

Resultados para el autor AM

Los resultados de trigramas de mayor potencial discriminante de los textos del escritor AM indican que solo la variable JHPT obtiene puntuaciones de las medias diferentes con un p-valor significativo de variación. El análisis detecta que se produce una variación intra autor significativa entre los textos de novela de TA (1986) y TR (2000) en la frecuencia de los trigramas ANP y VVTI. La variable VVTI es la única en la que se observa una variación de p-valor de significación en la comparación por parejas de los fragmentos de TI (1991) y TR. Estos resultados permiten confirmar la hipótesis planteada de que la frecuencia de los n-gramas no varía de forma significativa en diferentes producciones lingüísticas del mismo individuos.

Resultados para la autora CP

Según el análisis, en los textos de CP no se dan casos de trigramas que varíen de forma significativa a nivel intra autor.

– *Conclusiones*

Los resultados obtenidos en el análisis de n-gramas de tipo bigrama y trigrama de los textos de novela de los autores del grupo 2 conducen a las siguientes conclusiones. En primer lugar, los trigramas han demostrado ser la variable que presenta un menor número de casos en los que se produce variación intra autor. Además, uno de los autores, CP, no presenta variación en la recurrencia de ninguno de los trigramas analizados. En cuanto a los resultados de bigramas para esta autora, el número de variables que varían también es muy limitado. En segundo lugar, se ha podido constatar que la variación intra autor se suele manifestar sobre todo en la comparación de los textos en los tiempos de medición TA y TR. Sin embargo, dentro del conjunto de sujetos analizados se observa en algunos autores una tendencia hacia la variación sólo en el transcurso de períodos más cortos (cinco años aproximadamente), es decir del TA al TI o del TI al TR. Estos autores son AP y AM para los bigramas y AG y AM para los trigramas. Este resultado significa que el estilo escrito de los autores del grupo 2, en lo que viene reflejado por la variable de análisis, no cambia a corto plazo

(cinco años) pero el cambio es patente cuando ha transcurrido un período más largo.

Para terminar, se confirma la hipótesis de esta tesis de que los individuos no tienden a variar en gran medida cuando se trata de un intervalo de tiempo corto respecto a la frecuencia de los n-gramas (bigramas y trigramas). Dado que en este trabajo sólo se ha analizado la variación intra autor en el margen de un período entre 20 y 40 años y entre 5 y 8 años, de cara a una futura investigación en variación intra autor, sería interesante investigar cuál es el comportamiento de la variable en textos que han sido producidos en un tiempo intermedio más limitado del que se ha analizado en este estudio.

4.2 Estudio sobre la variación según el género textual

Con la realización de este estudio se pretende obtener resultados preliminares sobre el potencial clasificatorio y la aplicabilidad de los n-gramas de tipo bigrama y tigrama como marca de autoría en contextos de comparación textual forense que impliquen el uso de textos escritos en diferentes géneros textuales.

a) Corpus del estudio

El corpus del estudio sobre variación intra autor según el género textual consiste en textos de novela y artículos de opinión (en adelante, subcorpus NAO) de 10 de los autores del corpus de la

tesis[112]. Estos autores son: Arturo Pérez Reverte (AP), Carmen Posadas (CP), Eduardo Mendoza (EM), Isabel Allende (IA), Juan José Millás (JM), Lucía Etxebarría (EL), Mario Benedetti (MB), Mario Vargas Llosa (MV), Rosa Montero (RM) y Javier Marías (XM). Con el fin de disponer del mismo número de muestras de cada tipo de texto por sujeto, para este estudio, se han tomado, en el caso de los fragmentos de novela del subcorpus N, los textos correspondiente a las primeras cuatro novelas de cada autor[113] y en el caso de los artículos de opinión, todas las muestras disponibles en el subcorpus A0[114]. De esta manera el subcorpus resultante, NAO, ha quedado proporcionalmente distribuido con 20 muestra por autor y tipo de texto tal y como se muestra en la tabla 32.

Tabla 32. *Distribución del subcorpus NAO*

ID del autor	**Nº de textos N por autor**	**Nº de textos AO por autor**
AP	20	20
CP		
EM		
IA		
JM		
LE		
MB		
MV		
RM		
XM		
Total	***200***	***200***

[112] La razón que explica esta selección se ha presentado en el apartado 5a del capítulo 3.
[113] Los fragmentos de narrativa incluidos en el subcorpus NAO, provienen de las primeras cuatro novelas de cada autor que se listan en el anexo I.
[114] Para el detalle de los artículos recogidos en el subcorpus AO véase anexo I.

b) Variables

Dado que los fragmentos de novela y los artículos de opinión no tienen la misma extensión, para evitar la desproporción en los datos de n-gramas (bigramas y trigramas) obtenidos, *Legolas 2.0* se ha configurado para limitar la extracción a las primeras 300 palabras, equivalentes a la extensión mínima de los artículos. Siguiendo el criterio de selección de variables descrito en el apartado 6 del capítulo 4, del total de n-gramas de cada tipo extraídos se han escogido para los respectivos análisis 75 bigramas y 76 trigramas[115].

c) Análisis

La técnica estadística que se ha empleado en este estudio ha sido el análisis discriminante (AD), porque a diferencia del análisis de varianza, no solo permite establecer si se da una mayor variación en la recurrencia de los n-gramas a nivel intra o a nivel inter autor, sino que también comprueba si la similitud intra autor que ha facilitado la clasificación correcta de los autores en los estudios sobre la variación inter autor[116] es independiente del género textual de las muestras analizadas.

El procedimiento que se ha seguido ha sido el de introducir los textos de los dos géneros a la vez, identificando los conjuntos de textos de los dos géneros textuales según su autor y ejecutando dos análisis consecutivos de bigramas y trigramas.

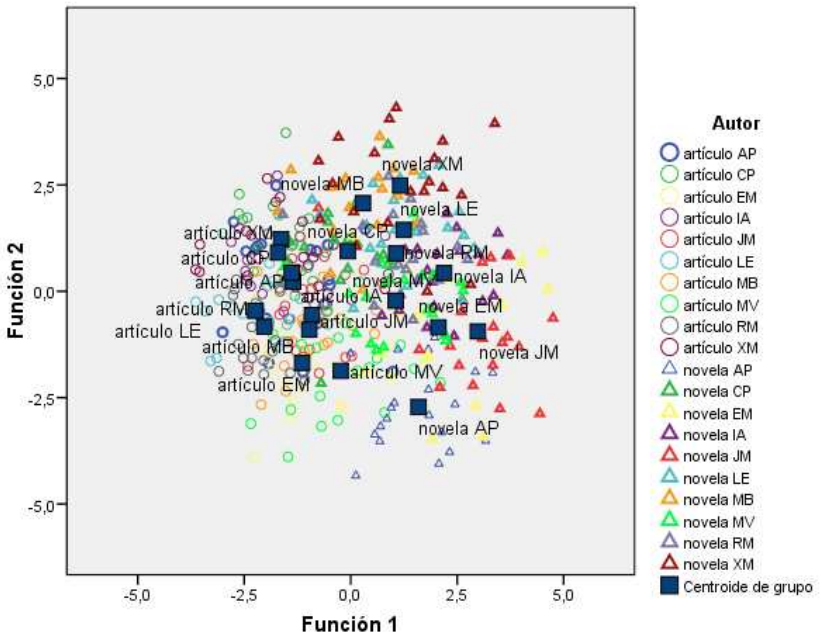
[115] El listado de ambos tipos de n-gramas se puede consultar en el anexo V.
[116] Véase capítulo 5.

d) Resultados

Teniendo en cuenta que en este estudio se están analizando los textos de los mismo sujetos, se espera que, si existen alguna similitud entre el estilo en el que un autor escribe cuando redacta un artículo y cuando escribe una novela, el análisis calcularía puntuaciones similares para los textos de los dos géneros textuales y clasificaría los fragmentos en el grupo de los artículos o en el grupo de las novelas siempre del mismo autor. Para poder aceptar que los resultados son fiables y que el autor no varía de en su estilo escrito entre géneros se considera que la clasificación "incorrecta" de textos de este autor tiene que ser de un mínimo de 50%.

– *Resultados del análisis de bigrama*

Gráfico 9. *Representación gráfica de los resultados de bigramas en subcorpus NAO*



Los resultados del análisis de bigramas que se muestran en el gráfico 9, conducen a dos observaciones inmediatas. La primera observación tiene que ver con la situación de los centroides. En el gráfico se aprecia claramente como los textos de artículo de opinión y sus centroides de grupo ocupan la parte izquierda mientras que los centroides de los fragmentos de novela se ubican en la parte derecha.

Estos resultados parecen indicar que en el caso del género textual no se confirmaría la hipótesis de trabajo según la cual los escritores,

en lo que al uso de secuencias de categorías gramaticales se refiere (en este caso, bigramas), no varían demasiado su estilo idiolectal cuando escriben en diferentes géneros textuales, sino que más bien indicarían la circunstancia opuesta, es decir, que en cuanto a la selección de estas secuencias los escritores parecen tener preferencias diferentes según escriban una novela o un artículo de opinión.

Respecto a la distribución que refleja el gráfico, hay que mencionar dos excepciones: una referente a los centroides y gran parte de los textos del autor JM en el caso de los artículos, y otra en cuanto a los centroides y los textos de la autora CP en el caso de los fragmentos de novela. Es particularmente interesante el hecho de que el centroide de los textos de narrativa de la autora CP sea el único caso en el que los centroides de los textos de los dos géneros textuales del mismo autor están situados cerca uno del otro. La consulta de los datos estadísticos por casos, que se puede realizar en el análisis con el SPSS, revela, sin embargo, que ninguno de los textos de cualquiera de los dos géneros se asigna al grupo opuesto. No obstante, este dato no descarta la posibilidad de que la variación inter género en esta autora sea mucho menor que en el resto de los sujetos, sino que solo indica que no hay una coincidencia completa entre textos concretos. Respecto a la posición del centroide de los textos de novela de la autora CP, hay que hacer notar también que se solapa con el de los artículos del escritor AP. Este resultado probablemente significa que existen ciertas similitudes entre las pautas seguidas por CP en su uso de recursos lingüísticos

(sintácticos) para escribir sus novelas y las pautas de uso que sigue AP al elaborar sus artículos.

La segunda observación importante en relación al gráfico 9 sobre bigramas concierne a la cohesión entre los grupos que representan los textos de los autores que pertenecen al género de artículo de opinión. En comparación con los correspondientes grupos de textos de novela, los artículos de opinión presentan cierta aglomeración que resulta en varios casos del solapamiento y la aglutinación de los centroides de grupo. Esta disposición de los centroides podría ser debida a que que los estilos de los autores analizados en el género del artículo de opinión son muy parecidos. La interpretación más verosímil de este resultado, sin embargo, sería que el análisis discriminante encuentra en el conjunto de los datos de los dos géneros una mayor similitud entre los artículos de opinión de los distintos autores que entre los textos de artículos y de fragmentos de novela del mismo autor, lo cual llevaría a concluir de nuevo que el estilo idiolectal sintáctico (en este caso, en cuanto a los n-gramas de tipo bigrama) varía por género textual más de lo que se había contemplado en la hipótesis inicial de trabajo de esta tesis.

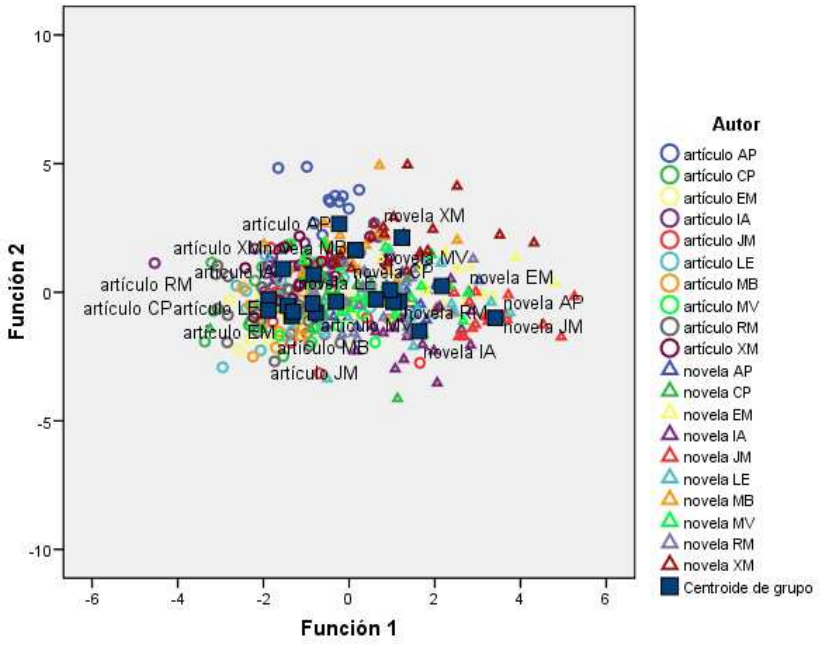
Por último, cabe decir que el equivalente porcentual de los resultados de la clasificación mediante los bigramas más frecuentes de los textos de los dos géneros textuales (artículos de opinión y narrativa) descritos hasta ahora es de un 84,3%. Entre los casos mal clasificados solo se observan dos (uno en los artículos de IA y otro en los textos de narrativa de RM) en los que realmente se trata de

una atribución de muestras del mismo autor, pero que sin duda se debe al azar.

– *Resultados del análisis de trigrama*

A continuación se exponen los resultados del análisis realizado mediante los n-gramas de tipo trigrama más frecuentes en el subcorpus NAO. El resultado de dicho análisis se muestra en el gráfico 10.

Gráfico 10. Representación gráfica de los resultados de trigramas en el subcorpus NAO



Como se puede observar, la clasificación basada en los trigramas genera resultados muy similares a los que se han obtenido en el análisis mediante bigramas, por lo que los comentarios que se han

hecho en este último caso en cuanto a la distribución de los grupos y la posición de sus centroides son válidos para los trigramas y responden a la misma argumentación. No obstante, cabe anotar algunos aspectos diferenciales. En primer lugar, en cuanto a la conglomeración de los centroides del grupo de los textos de artículos de opinión que destacamos en la descripción de los resultados de bigramas, es preciso decir que esta se hace patente en mayor medida en el caso del análisis de trigramas. Además se produce una dispersión notable de los marcadores que representan los textos de ambos géneros y un solapamiento múltiple. Este solapamiento que se da entre los centroides, por una parte, en el conjunto de grupos de artículos entre los autores RM, LE, CP, JM, EM, MB y, por otra, en el conjunto de grupos de textos de novela entre los autores LE, AP y MV, hace que la discriminación entre dichos escritores resulte prácticamente imposible. No debe sorprender, por lo tanto, que la clasificación mediante trigramas según los resultados porcentuales sea de un 53%. Esta clasificación implica solo un caso en el que el análisis asigna un texto de narrativa al grupo de artículos del mismo autor, pero este dato no puede considerarse significativo.

– *Conclusiones*

Los resultados de los análisis mediante los bigramas más frecuentes en el subcorpus NAO llevan a la conclusión de que el alto potencial discriminatorio que se ha observado en el análisis separado de artículos de opinión y de fragmentos de novela no parece extenderse

a los contextos del análisis conjunto de textos de más de un género, por lo cabe decir también que, a través de la realización de este estudio y con el corpus del que se ha partido, no parece confirmarse la hipótesis (que se pretendía extender a las variables de tipo sintáctico) según la cual el estilo idiolectal de un escritor es relativamente similar cuando este escribe en géneros textuales diferentes.

No obstante, esta es una conclusión preliminar ya que la distancia que se ha observado entre los centroides de grupo de algunos autores en los dos géneros no es excesivamente grande y no se excluye la posibilidad de que las idiosincrasias estilísticas estén presentes en los textos, pero que no se puedan detectar debido a alguna limitación del corpus de análisis. En este sentido se plantea la necesidad de seguir realizando más estudios y experimentos en español y en otras lenguas, con tipos de corpus más amplios e incluso con el contraste de un mayor número de géneros textuales.

5. EL POTENCIAL DISCRIMINATORIO DE LOS N-GRAMAS MÁS FRECUENTES. ESTUDIOS SOBRE LA VARIACIÓN INTER AUTOR

En este capítulo se recogen los estudios que reflejan los experimentos llevados a cabo en el corpus de análisis para la confirmación de las hipótesis formuladas respecto al carácter idiosincrásico y la capacidad discriminante de los n-gramas. Son, por lo tanto, estudios que, desde el punto de vista metodológico, forman parte de la evaluación de los n-gramas como unidad candidata a marca de autoría y de la técnica que los implementa, y desde el punto de vista conceptual, son estudios sobre la variación inter autor, ya que comparten como objeto de análisis unidades de las que se espera demostrar que evidencian las diferencias lingüísticas escritas que se dan entre diferentes individuos.

Se incluye también entre estos estudio un trabajo de carácter experimental con el que se espera poder llegar a observaciones preliminares sobre el poder discriminante de los n-gramas como marcas que permiten la categorización de textos según el origen de su autor y podrían implementarse en un modelo de elaboración de perfiles lingüísticos.

5.1 Estudio sobre el potencial discriminatorio de los n-gramas en textos de narrativa

En este estudio se analizan los datos de textos de narrativa con el fin de determinar, por una parte, cuál de los dos tipos de n-gramas, bigramas o trigramas, discrimina mejor entre los autores del corpus y, por otra parte, para establecer qué n-gramas o combinaciones de n-gramas de cada tipo en particular poseen el potencial discriminatorio más alto y podrían ser usados como marcas de autoría en la comparación lingüística forense de textos escritos en español.

a) Corpus del estudio

Para cumplir con estos objetivos, los análisis que comprende el estudio sobre el potencial discriminatorio de los n-gramas en textos de narrativa se basan en los datos extraídos del subcorpus N. Este subcorpus contiene 425 fragmentos de novela distribuidos en partes iguales entre los 17 autores del corpus de la tesis. Los pormenores de la selección y la recogida de las muestras textuales han sido detallados en las sección 5 del capítulo 3, por lo que en este punto nos limitamos a recapitular los datos pertinentes a la distribución del subcorpus N en la tabla que se muestra a continuación.

Tabla 33. *Distribución del subcorpus N*

Subcorpus	N° de sujetos	N° de obras por autor	N° de muestras por autor	N° de palabras por muestra	N° total de palabras
N	17	5	25	~ 600	266922

Esta distribución del corpus se ha mantenido en las pruebas de clasificación y determinación, pero ha sufrido algunas modificaciones relativas a la evaluación de los n-gramas como marca identificativa que cabe mencionar. Dado que la última fase del análisis estadístico en la que se aplican las pruebas de evaluación requiere introducir textos nuevos para testar la fiabilidad y la viabilidad de los n-gramas como marcas de autoría, para su realización se ha ampliado el corpus de estudio con 5 textos adicionales. Los textos en cuestión proceden de la novela del escritor Arturo Pérez-Reverte (AP), *El sol de Breda* publicada en el año 1998[117] y han sido seleccionados siguiendo los mismos criterios que el resto de muestras del subcorpus N[118].

b) Análisis estadístico de los n-gramas del subcorpus N

En este estudio, como en los demás estudios sobre variación inter autor que veremos más adelante[119], se aplica el análisis que se

[117] Para la información relativa a esta novela, véase el anexo I.
[118] Para los criterios de selección del corpus, véase capítulo 3.
[119] El estudio sobre el potencial discriminatorio de los n-gramas en textos de artículos de opinión y el estudio de evaluación en casos reales.

integra en la propuesta de técnica de comparación lingüística forense mediante n-gramas para los fines de la atribución de autoría de textos escritos en español. En el capítulo 4 hemos descrito el procedimiento de aplicación de dicha técnica y las etapas en las que transcurre. La última etapa corresponde al análisis estadístico, que se lleva a cabo después de la estandardización de los datos y consiste en la realización de tres tipos de pruebas: de clasificación, de determinación y de evaluación. El análisis del estudio se desarrolla en el mismo orden, asó como la presentación de los resultados.

A continuación se describen en detalle los tres tipos de pruebas que comprenden el análisis estadístico de los estudios de evaluación del potencial discriminatorio de los n-gramas.

- *Descripción de la prueba de clasificación*

La prueba de clasificación tiene como objetivo determinar el grado de eficacia de los n-gramas en discriminar entre los sujetos y clasificar correctamente los textos del corpus de análisis según su autor. A continuación se presentan los resultados de su realización junto con los comentarios pertinentes a los aspectos específicos de la prueba.

Mediante esta primera prueba se obtiene una estimación general de la capacidad clasificatoria de las funciones discriminantes que se forman a partir del análisis del conjunto de datos de cada variable,

en este caso los bigramas, en los textos de los 17 autores del corpus del estudio.

– *Descripción de la prueba de determinación*

Uno de los objetivos de este estudio pretende establecer cuáles de los n-gramas, mediante los que se lleva a cabo primero la clasificación de los textos de análisis y luego la atribución de textos pseudoanónimos, poseen el mayor potencial discriminatorio. La prueba de determinación esta diseñada para cumplir con este objetivo. Como se ha explicado en el capítulo 4, el análisis discriminante (AD) en que se basa la prueba no emplea todas las variables que se introducen, sino que hace una selección de aquellas cuya combinación capta las diferencias entre los grupos analizados, las variables clasificadoras, y permite discriminar entre ellos. Por lo tanto, la finalidad de la prueba de determinación es permitir obtener información de la significación individual de cada variable en la función discriminante que crea el AD.

Para realizar esta prueba en el análisis de los datos se aplica el método de AD de inclusos por pasos. Este método emplea un algoritmo de selección que consiste en incluir en la función discriminante una por una en orden sucesivo las variables de análisis empezando por la de mayor valor discriminante. El criterio que se usa en el algoritmo para medir el valor discriminante de cada variable en el caso de la prueba de determinación es el del

estadístico *lambda de Wilks*[120]. Se introducen solo las variables que provocan un cambio significativo en el valor de la lambda de Wilks al incorporarlas en la función discriminante. El cambio del valor de la lambda se mide por el así llamado estadístico *F* y se considera significativo cuando es igual o mayor de 3,84 y no significativo cuando es igual o menor de 2,71[121]. El algoritmo concluye la selección cuando ninguna de las variables que quedan es significativa de cara a la clasificación y a la discriminación entre los grupos analizados.

– *Descripción de las pruebas de evaluación*

La evaluación del potencial discriminatorio de los n-gramas se desarrolla en tres pruebas consecutivas a las que nos referimos como pruebas de evaluación.

• *Prueba de evaluación 1*

La primera prueba consiste en introducir en el análisis estadístico 5 textos nuevos que pertenecen a uno de los autores que constituyen el corpus del estudio, pero cuya autoría permanece oculta en la ejecución del análisis. El autor de los textos pseudoanónimos en

[120] El estadístico lambda de Wilks expresa la proporción de variabilidad total no debida a las diferencias entre los grupos; permite contrastar la hipotesis nula de que las medias multivariantes de los grupos [comparados] (los centroides) son iguales. La lambda de Wilks es la cociente entre el determinante de la matriz de la varianza y covarianza intragrupos y la matriz de varianza y covarianza total. (Pardo y Ruiz, 2002: 506).

[121] Valores de F preestablecidos en el programa de análisis estadístico SPSS para Windows.

todas las pruebas de evaluación de este estudio es el sujeto de análisis AP[122].

- *Prueba de evaluación 2*

La segunda prueba de evaluación replica la prueba inicial de evaluación pero se diferencia de ella en que para su realización se han llevado a cabo algunas modificaciones del corpus. Estas modificaciones radican en la reducción del número de textos por autor y se aplican, por un lado, para simular el contexto de análisis de un caso forense real en cuanto a número de muestras y, por otro, para cerciorarse de que la desproporción muestral entre posibles autores y los textos pseudoanónimos a clasificar no afecta los resultados del análisis.

La reducción de textos para la prueba 2 se ha realizado de manera automática mediante la opción de SPSS de selección de casos arbitraria evitando así la necesidad de una decisión justificable e imparcial sobre qué textos excluir y qué textos mantener en el análisis. De las dos opciones que ofrece el SPSS para la selección de submuestras aleatorias se ha optado por la de seleccionar una proporción en vez de eliminar un número determinado de muestras porque así se evita obtener un número igual o muy similar de textos para cada autor. El motivo para querer evitar que esto ocurra tiene que ver con el propósito de evaluar la técnica en un contexto

[122] Para mayores detalles sobre los textos pseudoanónimos véanse la descripción del corpus de análisis en la sección 1 de este capítulo y el anexo I.

parecido al de un caso real de atribución forense de autoría donde no es habitual disponer del mismo número de textos de todos los sospechosos. De ahí que la proporción que se ha fijado en este estudio para la reducción del número de muestras por autor ha sido del 20%.

- *Prueba de evaluación 3*

La tercera y última prueba de evaluación que se realiza para establecer el potencial discriminatorio de los n-gramas emplea el mismo corpus y procedimiento que las dos pruebas anteriores. Lo particular de esta prueba, sin embargo, estriba en que además de reducir el número de muestras como se ha hecho en la prueba 2, aquí la reducción se efectúa también en el número de sujetos de análisis. De este modo la prueba 3 pretende imitar incluso más un caso típico de atribución de autoría en el que los posibles sospechosos son muy pocos.

La selección de autores que conservar en el análisis que conlleva la prueba 3 se ha hecho en base a los resultados de la prueba previa de clasificación. Se han tomado los textos de aquellos autores entre cuyos estilos existe un grado de similitud con el autor pseudoanónimo. Este criterio de selección pone a prueba la técnica de comparación para los fines de atribución forense de autoría en un contexto complejo donde se dan *a priori* las condiciones para una atribución errónea y en el que la atribución correcta demostraría el

valor idiosincrásico de los n-gramas (bigramas o trigramas, o ambos) y su potencial discriminatorio.

c) Resultados y discusión del análisis del subcorpus N

– *Resultados del análisis basado en los datos de bigramas*

En este estudio se exponen en primer lugar los resultados de análisis mediante n-gramas de tipo bigrama. Los datos en los que se basa dicho análisis corresponden a los valores estandarizados de los primeros 75 bigramas más frecuentes seleccionados según el criterio establecido[123]. El listado de estos bigramas se puede consultar en el anexo VI.

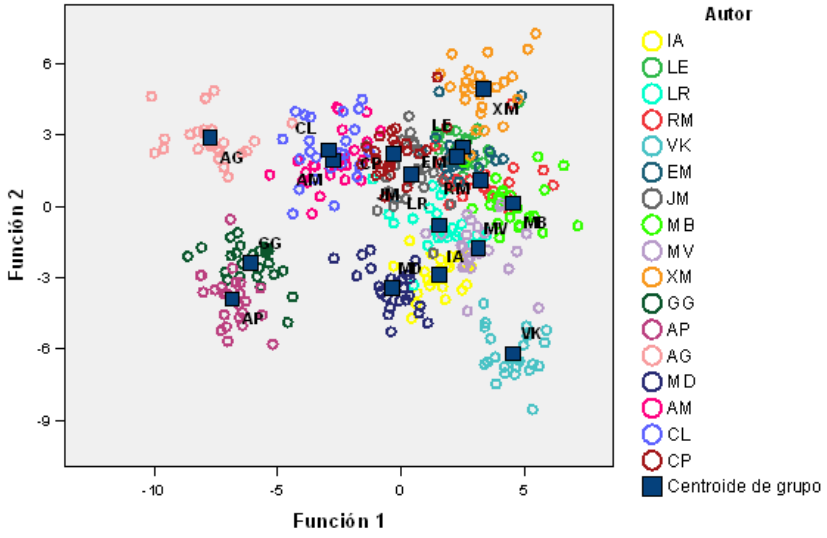
– *Resultados de la prueba de clasificación para bigramas*

Los resultados de la prueba de clasificación de los fragmentos de novela basada en los datos de bigrama se muestra en el gráfico 11. Como se puede ver en la representación gráfica de las funciones que mejor discriminan entre los sujetos de análisis y sus textos, aunque hay una clara separación entre la mayoría de los grupos (autores), también se observa cierto solapamiento entre los casos (textos) de algunos de los escritores. Las diferencias en la frecuencia con la que

[123] Véase el capítulo 4.

los bigramas más frecuentes ocurren en los textos de los autores AG, VK, XM, AP, MD, IA, CP y GG son más distintivas, ya que la distancia entre los centroides de grupo de sus textos es mayor en comparación con la del resto de los autores, lo que los sitúa en los extremos del gráfico. La proximidad que se observa en la disposición de los textos de las parejas de autores AP y GG, MD y IA, IA y MV, y MV y MB, indica que existe algún grado de similitud entre estos escritores en cuanto a la recurrencia de los n-gramas de tipo bigrama, pero que no es lo suficientemente alto como para impedir la discriminación entre ellos. Por lo que respecta a los grupos cuyos centroides están en parte solapados (CL y AM, LE y EM), se puede considerar que los estilos escritos guardan un mayor parecido entre sí que probablemente puede afectar la atribución correcta de más textos de los mismos autores.

Gráfico 11. Representación gráfica de las funciones discriminantes de clasificación de los textos de los 17 autores del subcorpus N mediante bigramas



La clasificación del gráfico 11 permite hacer también algunas observaciones acerca de la homogeneidad o heterogeneidad del estilo escrito de los autores analizados. Se aprecia muy poca dispersión de los textos de los autores con la excepción de RM. Esta congestión en el modo en el que están dispuestos los marcadores de los casos respecto el centroide de su grupo de pertinencia se puede interpretar como una homogeneidad en el uso de los recursos lingüísticos por parte de los autores. La autora RM, en cambio, resulta ser una escritora más heterogénea en su manera de escribir ya que los marcadores de sus muestras escritas están más dispersos.

En conclusión, la prueba de clasificación de los textos del subcorpus N basada en bigramas en general ha dado un resultado positivo. Los casos analizados han sido agrupados según su autor en su totalidad, por lo que la capacidad discriminatoria de los bigramas queda confirmada y además podemos calificarla como alta.

– *Resultados de la prueba de determinación*

A continuación se presentan los resultados de la prueba de determinación realizada con datos de n-gramas de tipo bigrama extraídos del corpus del estudio.

La tabla 34 contiene el listado de los 30 bigramas que según la prueba de determinación poseen el mayor potencial discriminatorio entre el total de 75 variables introducidas en el análisis.

Tabla 34. *Lista de los bigramas de mayor valor discriminante*

NV	NC
DE	NJ
CP	EN
VRR	DT
ASN	CD
NVR	VAL
RVR	RT
WV	VR
PAS	CA
PC	VRE
PV	ND
EVR	AMR
PAN	EC
CVR	RAP
AFR	ED

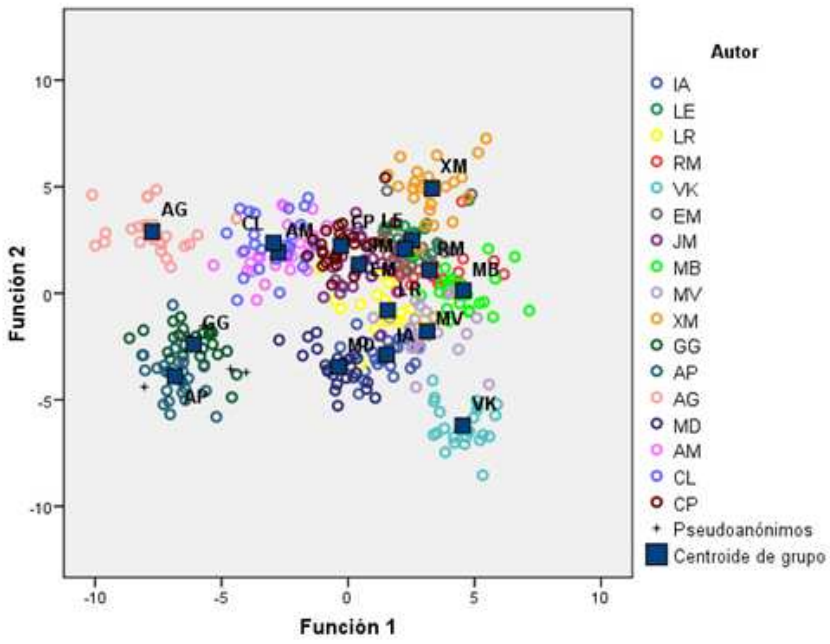
- *Resultados de las pruebas de evaluación para bigramas*

En la prueba de clasificación se ha podido comprobar que los bigramas poseen una capacidad clasificatoria relativamente alta. Sin embargo, como acertadamente apuntan Pardo y Ruiz (2002), aunque en el análisis quede demostrado el elevado potencial clasificatorio de una o varias variables, de este hecho no se desprende que la capacidad discriminante de dichas variables sea también alta o ni tan solo existente. Esta última afirmación se puede constatar solo mediante la clasificación de nuevos casos. De modo que una vez se ha establecido el potencial clasificatorio de los n-gramas (bigramas y trigramas) y se ha determinado cuáles en concreto reflejan en mayor grado las diferencias entre los autores

del corpus de análisis, es necesario evaluar su capacidad discriminante en la atribución de textos pseudoanónimos.

- *Resultados de la prueba de evaluación 1*

Gráfico 12. *Representación gráfica del resultado de la prueba de evaluación 1 de los bigramas en textos de narrativa*



Como se puede ver, los textos pseudoanónimos que aparecen marcados en el gráfico 12 con cruces negras se sitúan en el espacio del centroide del grupo que representa el escritor AP, quien es su autor verdadero, pero en proximidad inmediata a los textos del autor GG con el que AP ha mostrado un nivel relativamente alto de similitud. Los datos estadísticos revelan que el análisis

discriminante atribuye correctamente los casos nuevos o pseudoanónimos en un 90%, es decir, 3 de los textos a AP, y los 2 textos restantes (10%) a GG. Este resultado es muy alentador vista la similitud estilística que se ha detectado previamente entre los dos autores y reafirma el potencial discriminatorio de los bigramas.

- *Resultados de la prueba de evaluacion 2*

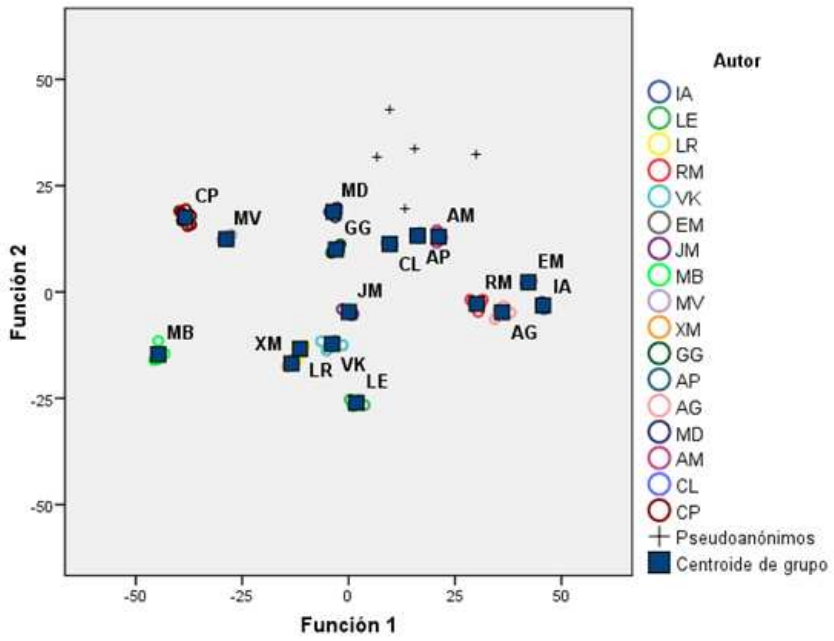
Al aplicar la opción de "selección por casos", el corpus del estudio ha quedado reducido a 85 textos distribuidos entre los autores de manera que el número de muestras correspondientes a cada uno de ellos varía entre 3 y 7 muestras por autor (véase tabla 35).

Tabla 35. *Número de textos por autor en la prueba 2 con bigramas (N)*

ID del autor	Nº de textos en el análisis
IA	**2**
LE	**4**
LR	**7**
RM	**6**
VK	**8**
EM	**4**
JM	**4**
MB	**5**
MV	**4**
XM	**5**
GG	**4**
AP	**5**
AG	**5**
MD	**6**
AM	**6**
CL	**2**
CP	**8**

Seguidamente se muestra el gráfico 13 de los resultados obtenidos en el análisis conjunto del corpus modificado y los textos pseudoanónimos objeto de atribución.

Gráfico 13. *Representación gráfica del resultado de la prueba de evaluación 2 de los bigramas en textos de narrativa*



En la prueba 2, los textos de la submuestra del corpus del estudio se clasifican en su totalidad según su grupo de pertenencia real. El gráfico muestra como los centroides de los grupos de análisis están claramente separados y la distancia de los textos al centroide es mínima, lo que indica una alta cohesión del grupo. Los centroides de algunos autores se sitúan a corta distancia entre sí, por lo que se puede concluir que sus valores en la función discriminante que se

crea en la prueba de evaluación son similares, pero no a un nivel que impida la discriminación. Los marcadores de textos pseudoanónimos se encuentran en proximidad a los centroides de los escritores MD y AP. Los cálculos estadísticos, no obstante, indican que su grupo más probable de pertenencia es AP, con un porcentaje de probabilidad del 5% de que el autor de uno de los textos sea MD.

- *Resultados de la prueba de evaluación 3*

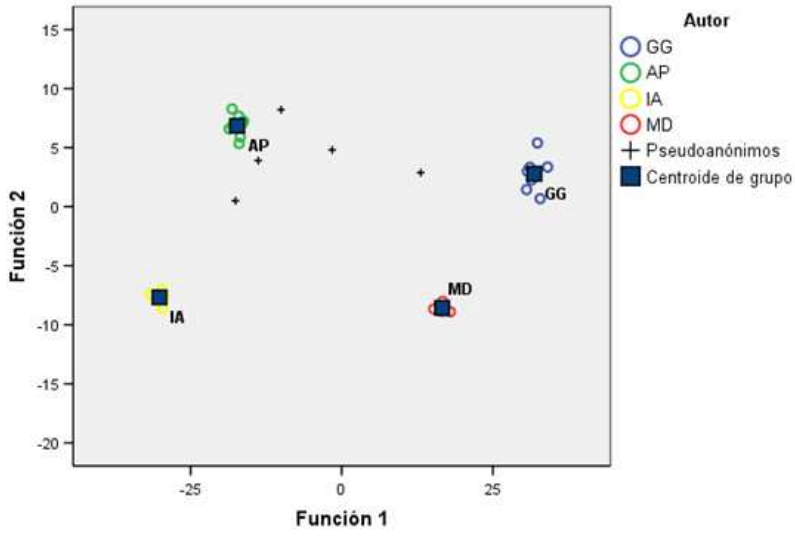
Antes de realizar la reducción de casos para poder llevar a cabo la prueba 3 de la forma correcta, se han seleccionado los autores cuyos textos se mantendrían en el análisis y se ha eliminado el resto. La prueba de clasificación mediante bigramas ha revelado que la similitud estilística con el autor pseudoanónimo (AP), criterio que hemos establecido para la selección de los autores a incluir en la prueba de evaluación final, se da con los sujetos de análisis GG, IA y MD. Por lo tanto, estos han sido los autores elegidos para la prueba 3. Una vez completada la selección de autores, se ha hecho la reducción del número de textos en el corpus utilizando la opción de SPSS descrita arriba. Después de la reducción, el corpus del estudio estaba constituido por 26 textos. El número de muestras por autor variaba entre 5 y 9 (véase tabla 36).

Tabla 36. *Número de textos por autor en la prueba 3 con bigramas (N)*

ID del autor	Nº de textos en el análisis
GG	**7**
AP	**9**
AG	**5**
MD	**5**

A continuación se puede ver la representación gráfica del resultado del análisis de la submuestra del corpus.

Gráfico 14. *Representación gráfica del resultado de la prueba de evaluación 3 de bigramas en textos de narrativa*



En el gráfico que genera el análisis basado en los bigramas más frecuentes en la submuestra del corpus del estudio se puede ver que

los centroides de los cuatro autores están muy separados y los textos forman grupos compactos en torno a ellos. Los textos pseudoanónimos ocupan la parte superior izquierda del gráfico cerca del centroide del autor AP. El análisis atribuye todos estos textos a AP con un 100% de seguridad, porcentaje de aciertos que también se obtiene para el resto de textos. Que el análisis clasifica correctamente todos los textos de control no debe sorprender, ya que dado el reducido número de muestras y autores, el AD construye una función discriminante optimizada, es decir que capta al máximo las diferencias entre los grupos. Sin embargo, el hecho de que atribuya los textos pseudoanónimos a su autor verdadero y no al autor con mayor número de textos en el análisis por ser este del que se tiene más información representa un resultado significativo.

- *Resultados del análisis basado en los datos de trigramas*

En este estudio sobre el potencial discriminatorio de los n-gramas en textos de novela se presentan en segundo lugar los resultados para los trigramas. Los datos que han sido usados para el análisis comprenden los primeros 76 trigramas más frecuentes en el subcorpus N. La lista de estos trigramas se puede ver en el anexo V.

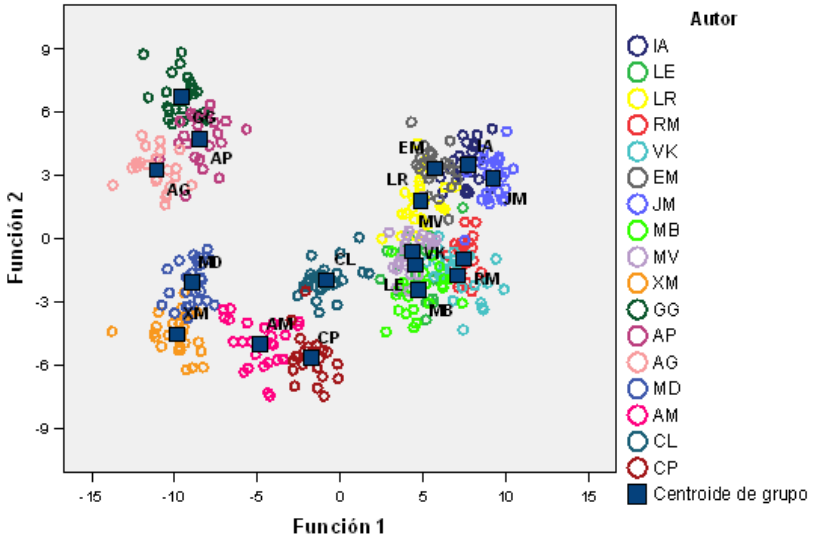
- *Resultados de la prueba de clasificación para trigramas*

Los resultados de la prueba de clasificación se muestran en el gráfico 15. La clasificación mediante trigramas replica en cierto grado la obtenida en base a los bigramas. Por una parte, los casos analizados también se aglomeran en grupos bien delimitados, aunque debe precisarse que en comparación con los resultados de la prueba de clasificación mediante bigramas, dentro del grupo de cada autor se observa una mayor conglomeración de los textos. Este hecho puede interpretarse como una mayor similitud estilística a nivel intra autor y mayor idiosincrasia a nivel inter autor en el uso de las combinaciones de unidades lingüísticas representadas por los trigramas más recurrentes. Por otra parte, la situación en el plano de los centroides de grupo es parecida. El análisis vuelve a posicionar relativamente cerca los centroides de grupo de los autores parecidos en cuanto a la recurrencia de los bigramas más frecuentes (AP, GG, CL, AM) (véase el gráfico 11).

Aún así, el resultado de la prueba basada en trigramas muestra algunas particularidades que es preciso comentar. Mientras que el análisis de bigramas permitía discriminar entre una serie de autores en los que las ocurrencias de la unidad candidata a marca de autoría era muy distintiva (AG, XM, VK), en el análisis de trigramas ninguno de los autores revela el mismo nivel de unicidad. En cambio, en la representación gráfica de las funciones discriminantes

de clasificación formadas a partir del análisis de trigramas (gráfico 15), se pueden observar una separación de los autores en tres grupos. Por un lado, un grupo grande, en el que los centroides de los autores son inmediatamente próximos y por lo tanto se puede considerar que la similitud entre ellos es bastante alta. Por lo tanto, es lógico que en él se observen más casos de solapamiento entre los textos de análisis. Este grupo está compuesto por los escritores EM, LR, IA, JM, MV, VK, RM, LE, CL y MB. Por otro lado, dos grupos más pequeños cuyos centroides no están tan aglutinados, es decir, similares en sus puntaciones en la función discriminante y en la recurrencia de los trigramas más frecuentes en sus textos. La similitud se hace patente de forma más notable en los casos en los que hay solapamiento entre los textos analizados. El primer grupo está formado por los autores MD, CP, XM y AM y, el segundo, por los escritores GG, AP, AG (véase gráfico 15). Entre las parejas de autores MD y XM, y AP y GG, como se ha comentado antes, es donde se percibe un mayor parecido y se puede prever que en un análisis de atribución de nuevos textos es posible obtener un resultado erróneo.

Gráfico 15. *Representación gráfica de las funciones discriminantes de clasificación de los textos de los 17 autores del subcorpus N mediante trigramas*



Por último, teniendo en cuenta que en la prueba de clasificación mediante trigramas el resultado es satisfactorio, ya que todos los casos están correctamente clasificados, se puede concluir que los trigramas poseen una capacidad clasificatoria alta, que indica que la aplicación de estas unidades como marcas de autoría con fines forenses puede ser viable.

- *Resultados de la prueba de determinación para trigramas*

A continuación se muestra en la tabla 37 la lista de trigramas para las cuales la prueba de determinación ha designado un valor discriminante alto y estadísticamente significativo en la función discriminante.

Tabla 37. *Lista de los trigramas de mayor potencial discriminatorio*

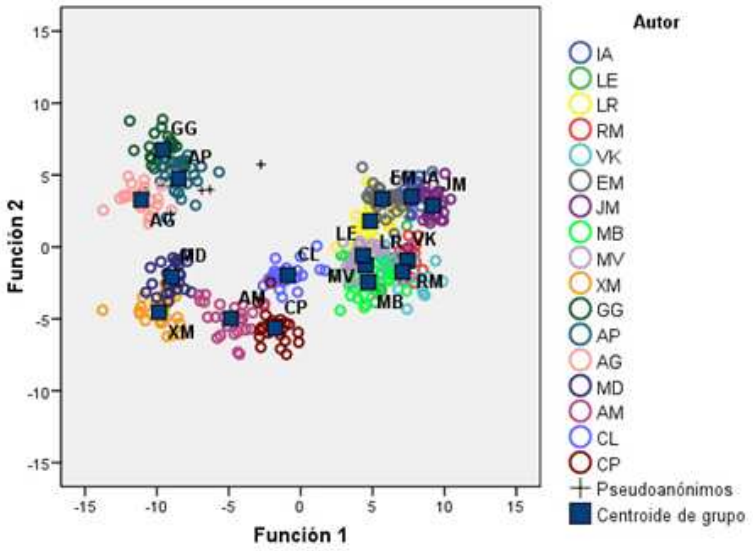
VPV
DPT
ANP
CVAC
ASNT
NCNP
APNE
VAPEJ
NPT
PASN
PEN
ECE
NEJT
RVAV
JHPT
ENVAP
ASNE
NDHJ
ENCP
VAN

Los 20 trigramas que contiene son las variables que se seleccionan para el análisis que se lleva a cabo en las pruebas de evaluación subsiguientes.

- *Resultados de las pruebas de evaluación para trigramas*

  - *Resultados de la prueba de evaluación 1*

Gráfico 16. *Representación gráfica del resultado de la prueba de evaluación 1 de los trigramas en textos de narrativa*



El gráfico 16 muestra el resultado de la prueba de evaluación 1, que se basa en los datos de trigramas en el subcorpus N. Como se puede ver, el análisis clasifica los textos de los autores conocidos con el

mismo resultado obtenido en la prueba de clasificación, por lo que es posible considerar que los valores para las variables de análisis de los nuevos casos introducidos no difieren de forma significativa del resto del corpus. Los marcadores de los textos pseudoanónimos son posicionados cerca de los centroides de los autores AP y GG para los que previamente ya se ha observado que comparten un cierto grado de similitud. El análisis asigna 3 de los 5 textos nuevos a su autor real, AP (90%) y 2 a los escritores GG (5%) y AG (5%), respectivamente. A pesar de que no todos los fragmentos de novela pseudoanónimos son atribuidos a su grupo de pertenencia real, este resultado es muy alentador vista la similitud entre el autor verdadero AP y los otros dos autores en los que se agrupan 2 de los textos de forma errónea. Partiendo de la atribución obtenida mediante los trigramas en el contexto descrito, se podría decir que este tipo de n-gramas son portadores de idiosincrasias lingüísticas y que poseen un potencial discriminatorio considerable.

- *Resultados de la prueba de evaluación 2*

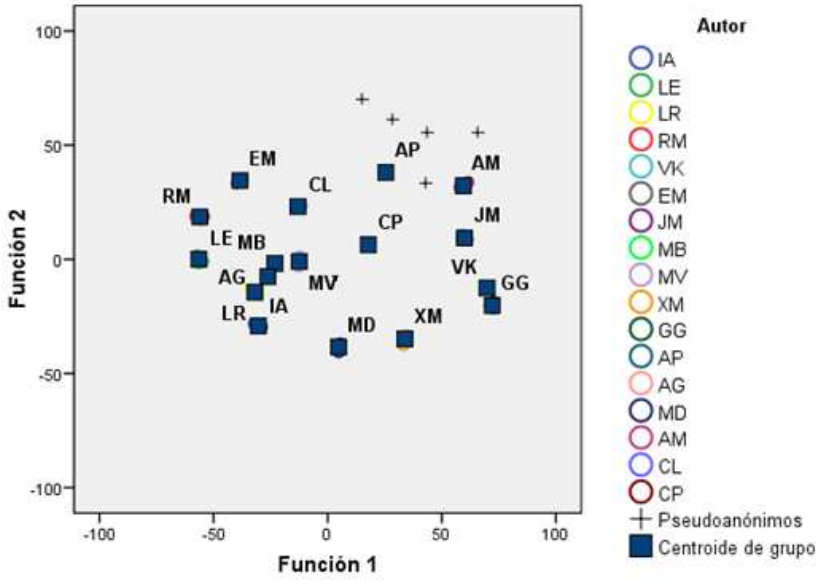
La submuestra del corpus del estudio con la que se ha llevado a cabo la prueba 2 comprende 83 textos a los que se añaden los 5 textos pseudoanónimos. La distribución de los textos que se incluyen en el análisis como textos indubitados varía de 3 a 9 textos por autor (véase tabla 38).

Tabla 38. *Número de textos por autor en la prueba 2 con trigramas (N)*

ID del autor	Nº de textos en el análisis
IA	**5**
LE	**9**
LR	**8**
RM	**5**
VK	**4**
EM	**4**
JM	**5**
MB	**3**
MV	**7**
XM	**6**
GG	**4**
AP	**5**
AG	**3**
MD	**3**
AM	**6**
CL	**3**
CP	**3**

El resultado de la prueba de evaluación 2 está reflejado en el gráfico 17 que se muestra a continuación.

Gráfico 17. *Representación gráfica del resultados de la prueba de evaluación de los trigramas en textos de narrativa*



En el gráfico 17 se percibe claramente como la reducción de los textos de análisis ha afectado la nueva clasificación según autor. No hay solapamiento entre los centroides de los grupos, que en su mayoría están bien distanciados entre sí. Se exceptúan las parejas de autores MB y AG y VK y GG, que resultan de valores próximos en la función discriminante que se crea en el análisis de la prueba 2 y ocupan posiciones de cercanía en el gráfico.

En cuanto a la clasificación de los nuevos casos, el análisis posiciona sus marcadores en la zona del centroide del autor AP. A diferencia del resultado de la prueba 1 los textos pseudoanónimos

presentan menos dispersión respecto el centroide del grupo. Conforme a la estadística del análisis, el 95% de los textos pseudodubitados tiene como grupo más probable de pertenencia el grupo del escritor AP, mientras que el 5% (que equivale a uno de los fragmentos de novela) corresponde al autor AM. Este resultado confirma en mayor medida el potencial discriminante de los trigramas en el contexto de aplicación de la técnica actual.

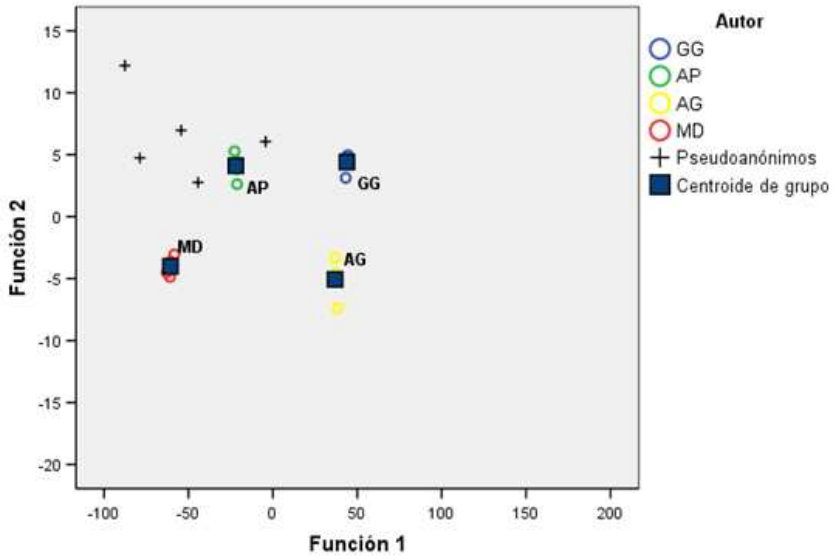
- *Resultados de la prueba 3*

Los textos a los que ha quedado reducido el corpus del estudio tras la selección aleatoria de muestras y que se han utilizado para la realización de la evaluación en la prueba 3 han sido 27. El total de textos del que se excluyen los 5 textos objeto de atribución se dividen entre los cuatro autores previamente seleccionados en base al criterio establecido en fracciones de 4 a 7 textos por sujeto (véase tabla 39). Los autores en cuestión son AG, GG, MD y AP, siendo el último el autor al que pertenece la autoría real de los textos pseudoanónimos.

Tabla 39. *Número de textos por autor en la prueba 3 con trigramas (N)*

ID del autor	Nº de textos en el análisis
GG	**6**
AP	**4**
AG	**4**
MD	**7**

Gráfico 18. *Representación gráfica del resultado de la prueba de evaluación 3 de trigramas en textos de narrativa*



Como se puede ver en el gráfico 18, la clasificación que devuelve el análisis discriminante de la prueba 3 distingue claramente entre los cuatro autores y sus textos. Las muestras de cada autor se sitúan cerca del centroide del grupo con la excepción del escritor AG, para el cual se observa una ligera dispersión de los marcadores. Cabe anotar que los centroides de los escritores AP y GG estan más próximos que los de los otros dos autores en la prueba, probablemente debido a la similitud que se ha podido detectar en los análisis anteriores. En el gráfico, los textos pseudoanónimos se posicionan en el ángulo superior izquierdo. Parte de ellos están a una distancia relativamente corta del centroide del autor AP, mientras que los demás se sitúan más lejos sin llegar a acercarse al

espacio de los escritores más próximos. La estadística de casos señala que todos los casos originales han sido clasificados correctamente y todos los pseudoanónimos también han sido atribuidos al sujeto de análisis AP.

d) Conclusiones del análisis del subcorpus N

A partir de los resultados obtenidos en los análisis de los bigramas y los trigramas de los textos del subcorpus N descrito en este apartado, se puede llegar a dos conclusiones principales respecto a su potencial discriminatorio. En primer lugar, las pruebas de clasificación y evaluación han demostrado que tanto los bigramas como los trigramas discriminan entre los autores de forma efectiva independientemente del número de textos y autores empleados en el análisis discriminante. En comparación con los bigramas se observa que los trigramas se muestran como marca que capta en mayor grado las idiosincrasias que permiten discriminar entre los autores analizados. Sin embargo, la diferencia en los datos porcentuales de la clasificación obtenida mediante unos u otros, no es muy alta. En el caso de los bigramas es de 88% en la validación cruzada y en el de los trigramas de 92%. En segundo lugar, cabe decir que ambos tipos de n-gramas empleados como marcas indentificativas atribuyen un porcentaje alto (del 80% al 95%) de los textos pseudoanónimos a su autor real, lo que confirma los resultados obtenidos en el estudio preliminar sobre el potencial discriminatorio

de las variables de análisis[124] y su validez en el contexto de los textos de carácter no forense.

5.2 Estudio sobre el potencial discriminatorio de los n-gramas en textos de artículos de opinión

a) Corpus del estudio

Como corpus de este estudio se han utilizado los textos contenidos en el subcorpus de artículos de opinión (subcorpus AO). Como hemos explicado en el capítulo 4, el subcorpus AO está constituido por los artículos de solo 10 del total de 17 escritores analizados anteriormente, debido a la falta de documentos disponibles de este género textual de todos los autores que integran el corpus de análisis de la tesis. Estos autores son: Arturo Pérez Reverte (AP), Carmen Posadas (CP), Eduardo Mendoza (EM), Isabel Allende (IA), Juan José Millás (JM), Lucía Etxebarria (LE), Mario Benedetti (MB), Mario Vargas Llosa (MV), Rosa Montero (RM) y Javier Marías (XM). A continuación se ofrece una tabla recordatoria de la distribución del subcorpus AO que ha sido comentado en el capítulo 4.

[124] Véase apartado 4 del capítulo 2.

Tabla 40. *Distribución del subcorpus AO*

Subcorpus	N° de sujetos	N° de muestras por autor	N° de palabras por muestra	N° total de palabras
AO	10	20	~ 300	127382

Del mismo modo que en el estudio anterior, para poder llevar a cabo las pruebas de evaluación ha sido necesario disponer de textos que no figuran en el subcorpus AO. En este caso como textos suplementarios en función de muestras de autor anónimo se han usado 7 artículos de la escritora Rosa Montero (RM) procedentes del mismo fondo que las demás muestras de esta autora que se recogen en el subcorpus AO[125]. Los datos relativos a estas muestras adicionales se pueden consultar en el anexo I donde aparecen marcados en *cursiva*.

b) Análisis estadístico de los n-gramas del subcorpus AO

Este estudio informa de los resultados de los análisis llevados a cabo en la tesis como parte de la investigación sobre la variación inter autor. Los objetivos de dicho estudio son dos: en primer lugar, evaluar el potencial discriminatorio de los dos tipos de n-grama objeto de estudio de esta tesis y determinar cuales son los bigramas y trigramas de potencial discriminatorio más alto; y, en segundo

[125] Véase el anexo I para la información sobre la fuente y las fechas de publicación.

lugar, comprobar si la aplicabilidad y la eficacia de la técnica basada en los n-gramas como marca de autoría depende de la extensión de las muestras de análisis.

Para alcanzar estos objetivos aplicamos la técnica de análisis lingüístico forense que se está evaluando en este trabajo a textos de artículos de opinión que tienen un tamaño inferior a los fragmentos de novela del estudio anterior y que se asemejan más en su extensión a los documentos de los que por lo general se dispone en un caso forense real.

c) Resultados

– *Resultados del análisis basado en los datos de bigramas*

Los resultados que se presentan a continuación han sido obtenidos mediante el análisis estadístico de los datos de los primeros 58 bigramas más frecuentes en el subcorpus AO. Para el listado de estos bigramas, consúltese el anexo V. Las distintas pruebas que constituyen la técnica de análisis han sido descritas en el capítulo 3 y en la sección 1 de este capítulo.

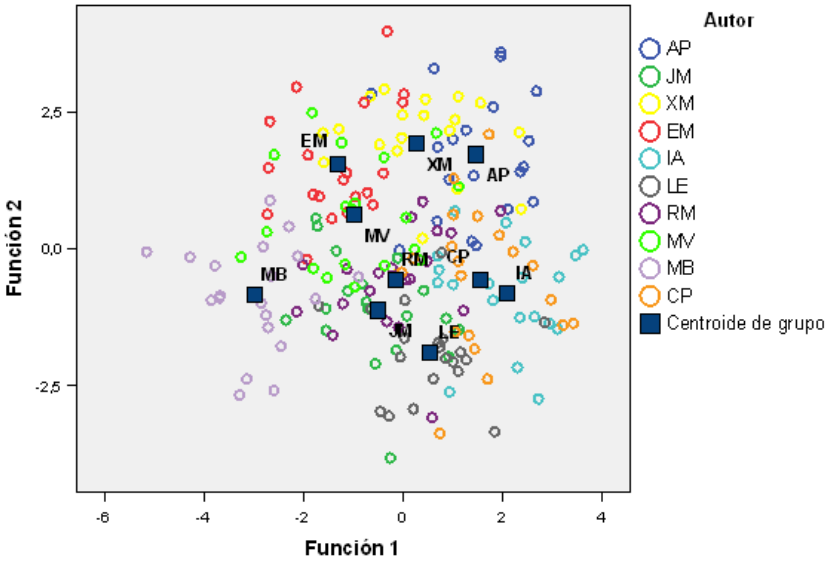
- *Resultados de la prueba de clasificación para bigramas*

El resultado de la prueba de clasificación con datos de los bigramas de mayor frecuencia en el corpus de análisis se muestra en el gráfico 19. En esta representación gráfica se aprecia una clara separación entre los grupos, aunque hay casos de solapamiento entre los textos de algunos autores. El mayor número de solapamientos ocurren entre los autores LE, CP y JM y entre RM y MV. No obstante, visto que el solapamiento es parcial, es decir, solo de algunos textos individuales, y no afecta la clasificación es posible considerar que se trata de casos únicos en los que los diferentes escritores emplean los recursos lingüísticos de un modo muy parecido, probablemente por el hecho de escribir en el mismo género.

Así mismo, se hace patente la mayor dispersión de los marcadores que representan los artículos de opinión en contraste con la que observamos en la clasificación de los fragmentos de novela del estudio anterior. Esta dispersión podría explicarse por el hecho de que los fragmentos de novela están en cierto modo unidos por el argumento que se narra, lo que puede incitar el uso repetido de los mismos recursos lingüísticos del repertorio del autor y de ahí resultar en una mayor cohesión interna de los textos, mientras que los artículos que suelen tratar distintos temas de la actualidad que no están relacionados entre sí carecen de dicha cohesión y por su

extensión más corta no dejan lugar a una reiteración notable en el lenguaje empleado.

Gráfico 19. *Representación gráfica de las funciones discriminantes de clasificación de los textos de los 10 autores del subcorpus AO mediante bigramas*



Las funciones discriminantes que captan en mayor grado la diferencia entre los grupos consiguen clasificar un 93% de los textos analizados. Dentro de esta clasificación se asigna correctamente el total de documentos de análisis (20 textos) que corresponden a los artículos de MB, MV y RM. Para el resto de los sujetos el resultado de la clasificación también marca un número alto de aciertos: para los escritores AP, IA y LE el número de textos clasificados correctamente por autor según su grupo de pertenencia es de 19 (95%) y para XM y CP, de 18 (90%). Los autores en los que se observa el mayor número de muestras textuales de

clasificación errónea son EM y JM. En el caso de EM los textos mal clasificados (3 textos) se atribuyen al escritor MV y en el caso de JM (4 textos), a EM y MV. Este resultado indica que existe cierta semejanza en el uso de la lengua en cuanto a las variables objeto de estudio (bigramas) en los artículos en cuestión, pero que dado el número limitado de casos que se atribuyen erróneamente es más probable que esta fuera debida al efecto de factores externos que no forman parte de los objetivos que se propone esta tesis y no a la similitud estilística de los tres autores.

Estos resultados permiten constatar que la capacidad clasificatoria de los n-gramas de tipo bigrama es elevada. Visto que en comparación con los resultados del análisis obtenidos utilizando la misma variable en la clasificación de los textos del subcorpus N, el poder clasificatorio de los bigramas disminuye en un grado relativamente bajo (7%) en el análisis de textos más cortos, se puede concluir que los bigramas son unos buenos candidatos a marcas de autoría.

- *Resultados de la prueba de determinación para bigramas*

En la prueba de determinación basada en los 58 bigramas más frecuentes en el corpus del estudio se ha establecido que las variables de mayor potencial discriminatorio en la función discriminante responsable de la clasificación correcta de los casos

de análisis son 21. La lista de estos bigramas se da en la tabla 41 que se presenta a continuación. Las variables de la lista son las que se utilizan en las pruebas subsiguientes de evaluación.

Tabla 41. *Lista de los bigramas de mayor potencial discriminatorio en los textos de artículos de opinión*

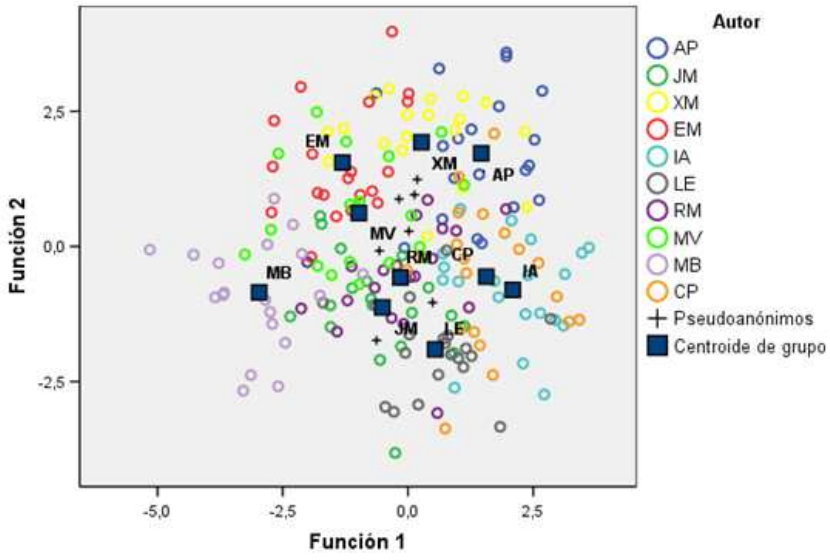
NJ
DT
ASN
NC
EC
RR
VRE
DE
PR
PN
VR
WV
CE
CD
AMR
VG
APE
ASE
VAK
CP
VC

- *Resultados de las pruebas de evaluación para bigramas*

Resultados de la prueba de evaluación 1

La primera prueba de evaluación en textos de artículos de opinión se ha llevado a cabo con el subcorpus AO sin ningún tipo de modificación. Con el fin de testar la técnica de comparación de textos con fines forenses en el análisis del corpus del estudio se han introducido también 7 textos adicionales de la autora RM. Estos textos pseudoanónimos constituyen el corpus de textos dubitados cuya autoría se pretende atribuir a uno de los posibles autores del corpus del estudio, cuyos textos representan el corpus indubitado. Los resultados de la prueba 1 se muestran en el gráfico 20.

Gráfico 20. *Representación gráfica del resultado de la prueba de evaluación 1 de bigramas en textos de artículos de opnión*



En el gráfico 20 no se puede apreciar ningún tipo de cambio en la distribución de los centroides de los autores del corpus como consecuencia de la introducción de los textos pseudoanónimos. Se produce la misma dispersión de los marcadores de los textos de autor conocido y el solapamiento parcial entre algunos de los grupos, como por ejemplo, es el caso de los escritores RM, CP y MV. No obstante, la prueba 1 clasifica con éxito 5 de los textos pseudoanónimos como textos producidos por su autora real, RM. La autoría de los 2 textos restantes se asigna a CP y XM.

Resultados de la prueba de evaluación 2

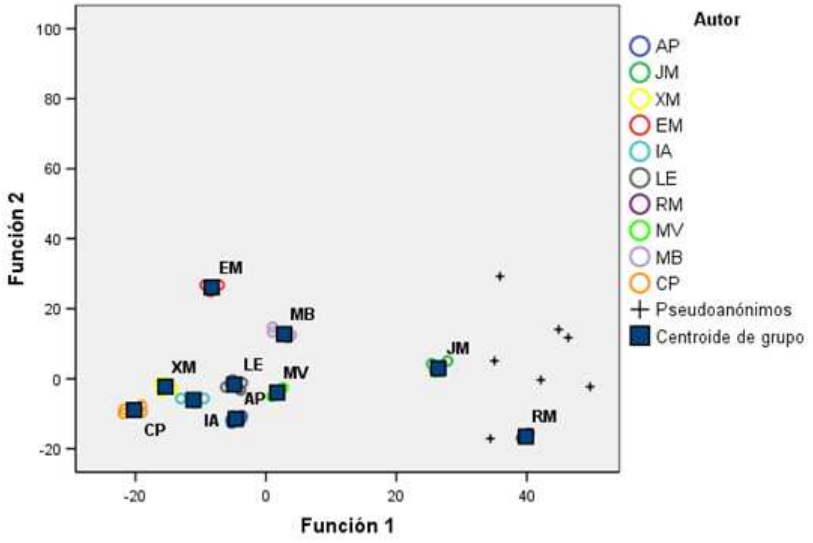
La proporción de textos que se ha fijado para obtener una submuestra de artículos de opinión de los 10 autores que constituyen el corpus del presente estudio para esta clase de prueba ha sido del 30%. Esta proporción corresponde a un mínimo de 3 y a un máximo de 10 muestras textuales por autor (véase tabla 42). El número total concreto que se ha usado en la prueba 2, cuyos resultados se exponen a continuación, suma 52 artículos.

Tabla 42. *Número de textos por autor en la prueba 2 con bigramas (AO)*

ID del autor	Nº de textos en el análisis
AP	**5**
JM	**6**
XM	**5**
EM	**3**
IA	**4**
LE	**5**
RM	**3**

MV	**5**
MB	**10**
CP	**6**

Gráfico 21. *Representación gráfica del resultado de la prueba de evaluación 2 de bigramas en textos de artículos de opinión*



Los resultados de la prueba 2 se muestran en el gráfico 21. Como se puede ver, la reducción del número de textos por autor en el análisis lleva a la discriminación de dos grupos de autores. El primer grupo, que es también el grupo más numeroso, se sitúa en el ángulo inferior izquierdo del gráfico y está constituido por los centroides que representan los autores XM, CP, IA, AP, MV, LE, MB y EM. Dentro de este grupo se observa cierta aglomeración de centroides que probablemente se debe a la similitud en la frecuencia de los bigramas en los textos de estos autores. En cambio, en el segundo grupo, que solo está formado por dos escritores, JM y RM, la

distancia entre los centroides de grupo deja clara las diferencias estilísticas, reflejadas por los bigramas, que existen entre ellos. Este resultado significa que la función discriminante que se crea por los bigramas identificados en la prueba de determinación posee un potencial discriminante significativo en cuanto a la distinción entre los autores del segundo grupo, al tiempo que en lo que se refiere al primero, su valor es unitario. Por lo tanto, los bigramas que diferencian una serie de autores podrían ser las mismas variables en el uso de las cuales dichos autores se asemejan.

En cuanto a los textos desagrupados, el análisis posiciona sus marcadores en la zona cercana a los centroides de los escritores JM y RM. Del total de 7 pseudoanónimos solo atribuye 6 de los textos pseudoanónimos a su autora, RM. Para el otro texto el análisis determina que la probabilidad de que haya sido escrito por JM es mayor, de modo que lo asigna dentro de su grupo.

Resultados de la prueba de evaluación 3

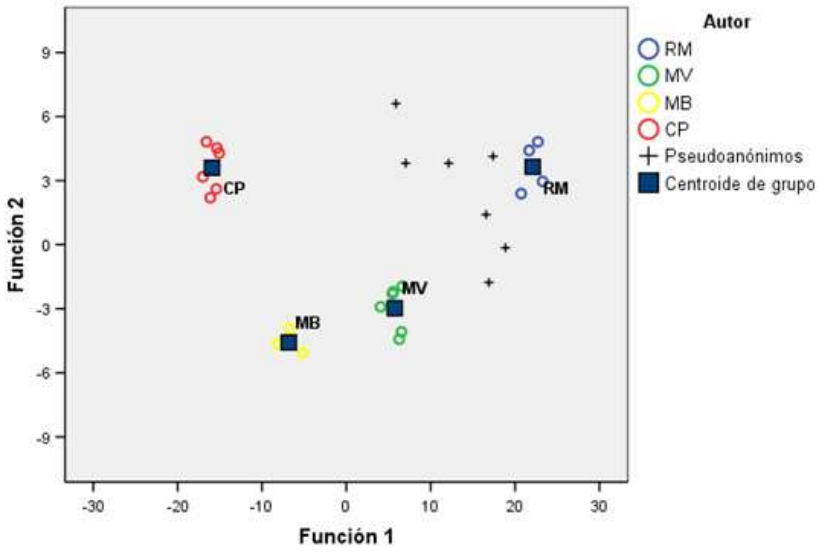
Tal y como se ha especificado en el capítulo anterior, la prueba de evaluación 3 implica limitar el análisis a los textos de solo aquellos autores que en la prueba previa de clasificación tienen puntuaciones similares, por lo que se considera que comparten similitudes estilísticas. En el caso de la clasificación mediante bigramas el parecido con el autor real de los textos pseudoanónimos (RM), se observa en los escritores MB, CP y MV. La selección aleatoria de casos en el programa SPSS en el caso de la prueba 3 ha reducido los

textos de los autores seleccionados a 20, repartidos en conjuntos muestrales de 4 y 6 textos (véase tabla 43).

Tabla 43. *Número de textos por autor en la prueba 3 con bigramas (AO)*

ID del autor	Nº de textos en el análisis
RM	**4**
MV	**6**
MB	**4**
CP	**6**

Gráfico 22. *Representación gráfica del resultado de la prueba de evaluación 3 de bigramas en textos de artículos de opinión*



El gráfico 22 muestra la clasificación de los cuatro autores de análisis y sus textos en la prueba 3. Como se puede ver, la función discriminante capta las diferencias en la frecuencia de los bigramas en los textos de los distintos grupos y separa sus centroides a una

distancia considerable. Las diferencias son más notables en los escritores CP y RM, que están más apartados uno del otro en los dos extremos del gráfico. Los marcadores de los pseudoanónimos ocupan la zona intermedia, relativamente cerca del centroide de RM. Los resultados indican que 6 de los textos desclasificados pertenecen, con una probabilidad del 95%, a la escritora RM y el restante 5%, que corresponde a uno de los textos, al autor MV.

– Resultados del análisis basado en los datos de trigramas

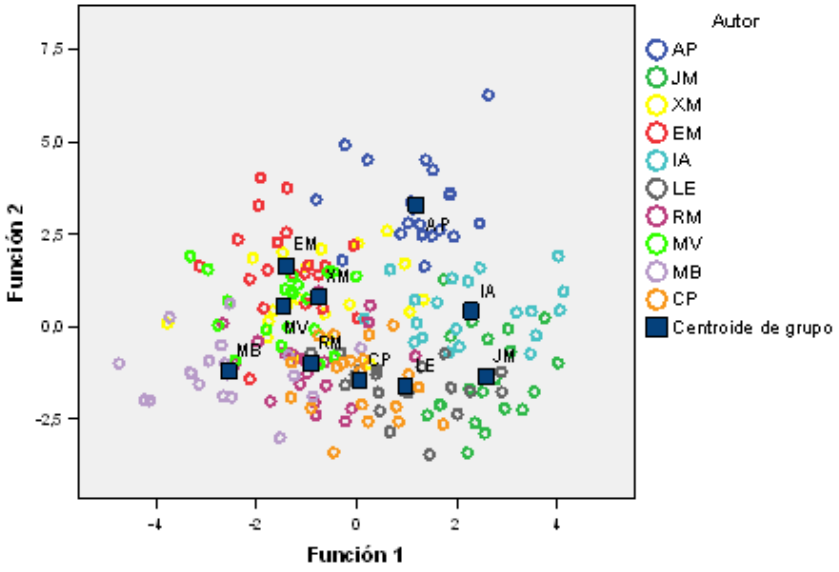
Los resultados que se exponen seguidamente proceden del análisis de los primeros 78 trigramas más frecuentes en el subcorpus AO. La lista de estos trigramas puede verse en el anexo V.

• *Resultados de la prueba de clasificación para trigramas*

Los resultados de la prueba de clasificación aplicada a los textos de artículos de opinión se muestran en el gráfico 23. Como se puede observar, las funciones discriminantes que se generan en la prueba logran delimitar los textos de análisis en grupos separados entre cuyos centroides no se observa ninguna clase de solapamiento, que en caso contrario sería indicativo de un muy alto grado de similitud entre los autores que representan estos grupos. Sin embargo, del mismo modo que el análisis basado en los datos de bigramas presentaba algunos casos en los que parte de los textos de

determinados autores se solapaban o se encadenaban en el mismo espacio, en el de trigramas esto ocurre también, pero entre un número limitado de autores, entre los que se encuentran RM, MV, CP y LE. A diferencia de los bigramas, los trigramas como parámetro de agrupación producen una clasificación con menos dispersión entre los textos del mismo grupo. Esto significa que los escritores del corpus son más idiosincrásicos en el uso de estructuras de mayor complejidad o de mayor número de componentes constructivos, como los trigramas, que en el uso de construcciones de menos integrantes composicionales, como los bigramas. Esta idiosincrasia es más destacada en los autores AP (marcador en azul marino), MB (marcador en lila), EM (marcador en rojo) y JM (marcador en verde). En la clasificación de los trigramas se aprecia claramente como los centroides de los grupos de los autores en cuestión se distancian del resto y destacan como grupos individuales más compactos.

Gráfico 23. *Representación gráfica de las funciones discriminantes de clasificación de los textos de los 10 autores del subcorpus AO mediante trigramas*



La clasificación basada en los trigramas, según los resultados del análisis estadístico, es correcta en un 94%. Dentro de esta clasificación, el porcentaje de casos que se agrupan según su pertenencia por autor en cada grupo es del 100% para los sujetos de análisis AP, XM, RM, MV, MB, JM y EM, y del 80% para los sujetos IA, LE y CP. El análisis discriminante detecta en el autor LE un 20% de similitud con el estilo escrito de los escritores JM y CP cuyos centroides se sitúan próximos uno del otro en el gráfico 23. Lo mismo ocurre con IA y CP, que según el análisis guardan el mismo grado de similitud con el estilo de LE y entre sí mismos.

Estos resultados muestran que los trigramas tienen un potencial clasificatorio bastante alto. Cabe decir también que a pesar de que la eficacia de la técnica no difiere de forma significativa en la comparación de los dos tipos de n-gramas (93% para los bigramas y 94% para los trigramas), parece que los trigramas captan mejor las idiosincrasias estilísticas que caracterizan a los autores del corpus y podrían ser mejores candidatas a marcas de autoría.

- *Resultados de la prueba de determinación para trigramas*

La prueba de determinación ha establecido que entre los 78 bigramas que han sido usados para su realización solo 20 poseen un potencial discriminatorio alto y son estadísticamente significativos en la función discriminante que crea el análisis discriminante para la clasificación de los textos del corpus de análisis. Las variables se listan en la tabla 44.

Tabla 44. *Lista de los trigramas de mayor potencial discriminatorio en los textos de artículos de opinión*

NCNP
ASNE
VDAN
ECE
AJN
VDEJN
PASN
DJT
VGIRP
PAR
VDPN
APNE
NPT
PEJT
EHJPA
RVDR
VRVE
RRVDR
PAPN
CPJE

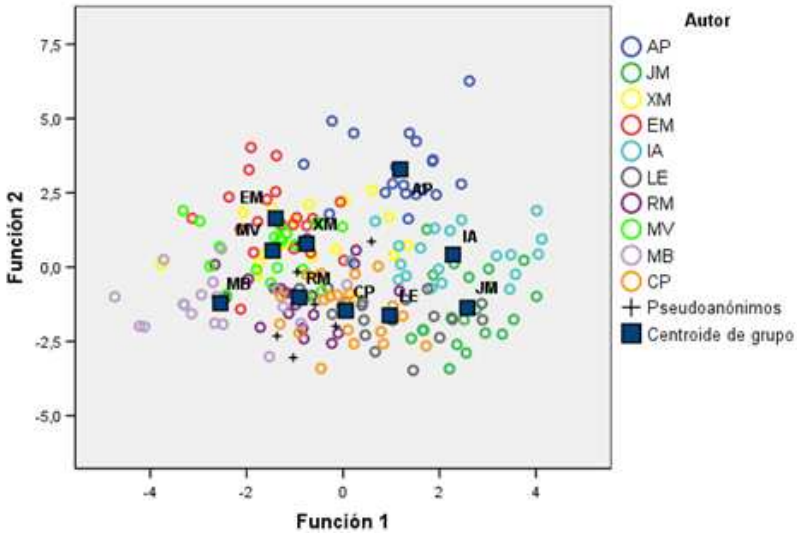
El resultado de esta prueba se aplica a la hora de realizar las pruebas de evaluación de la técnica, puesto que permite reducir el número de variables a aquellas que son de valor discriminante más alto.

- *Resultados de las pruebas de evaluación para trigramas*

Resultados de la prueba de evaluación 1

Para la realización de la prueba 1 se emplean los bigramas que acuerdo con la prueba de determinación ha permitido concluir que poseen el valor discriminante más alto entre el conjunto de variables de este tipo. El gráfico 24 muestra el resultado de su análisis de evaluación.

Gráfico 24. *Representación gráfica del resultado de la prueba de evaluación 1 de trigramas en textos de artículos de opinión*



En la representación gráfica de los resultados se puede ver cómo los marcadores de los textos pseudoanónimos (marcados con +) están ubicados en el centro del gráfico, donde también se encuentra el centroide de su autor verdadero, RM. La alta dispersión de los textos de los autores que se da en esta zona del gráfico los sitúa en proximidad inmediata con los marcadores de otros autores como CP, IA y MV. Sin embargo, de los 7 casos de textos de autor desconocido, solo 2 son atribuidos erróneamente. El sujeto que según la estadística de casos es el autor más probable de los dos casos desclasificados es CP.

Resultados de la prueba de evaluación 2

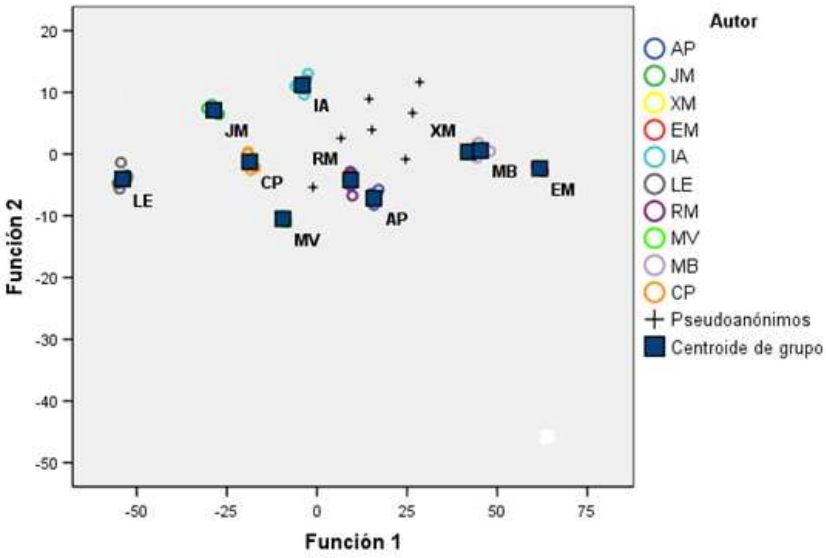
La prueba de evaluación 2 usa una submuestra del corpus del estudio, generada según ha sido explicado en la sección 5.1, que contiene 58 de los textos originales. Estos textos se reparten entre los sujetos de análisis en grupos de 3 a 8 (véase tabla 45).

Tabla 45. *Número de textos por autor en la prueba 2 con trigramas (AO)*

ID del autor	Nº de textos en el análisis
AP	**5**
JM	**5**
XM	**6**
EM	**3**
IA	**10**
LE	**9**
RM	**6**
MV	**4**
MB	**5**
CP	**5**

El gráfico 25 muestra los resultados de la prueba que incluye también los 7 textos pseudoanónimos de la autora RM.

Gráfico 25. *Representación gráfica del resultado de la prueba de evaluación 2 de trigramas en textos de artículos de opinión*



El resultado de la clasificación que se obtiene en la prueba 2 muestra que los datos de trigramas procedentes de los textos de artículos de opinión son capaces de discriminar bien entre los autores, incluso cuando las muestras de escritura disponibles son de número limitado. Los centroides de grupo y los casos de cada grupo no presentan ningún tipo de solapamiento. Entre la totalidad de grupos se observa una particularidad en los autores JM y RM cuyos centroides se sitúan a cierta distancia de los del resto de escritores.

Esta particularidad puede ser debida a la diferencia significativa en la frecuencia de los trigramas en los textos de estos dos autores y a una mayor idiosincrasia lingüística en comparación con los demás. En cuanto a los textos desagrupados o pseudoanónimos, el análisis atribuye 6 de ellos a RM, su autora real, y un texto al autor JM.

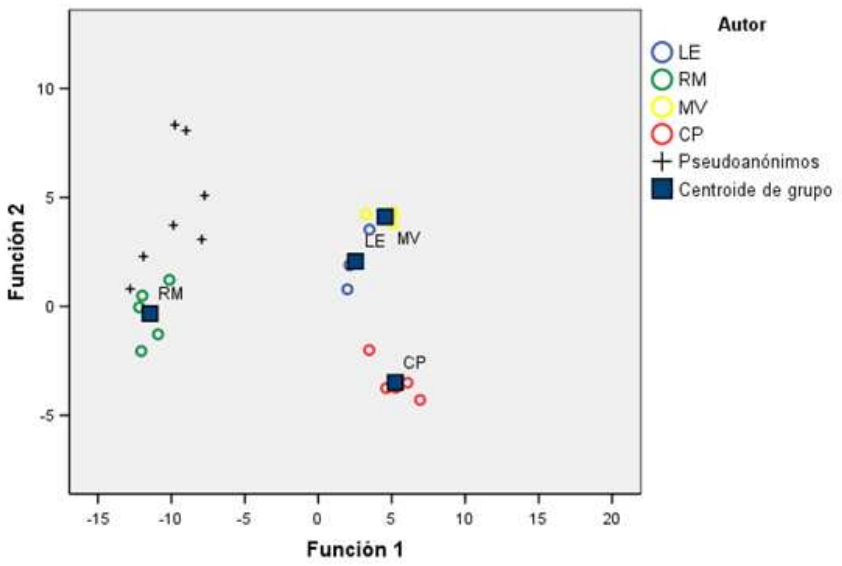
Resultados de la prueba 3

En el análisis que implica la última de la serie de pruebas de evaluación se han usado 18 de los textos que componen el corpus del estudio. Esta submuestra está distribuida en grupos de textos entre los 4 sujetos que han sido elegidos de los 10 escritores periodistas del corpus original. El resultado de la prueba de clasificación que se ha llevado a cabo en primer lugar ha permitido determinar que por la proximidad en los valores de recurrencia de los trigramas en sus escritos y los de RM, autora de los textos pseudoanónimos en el análisis, la selección comprenda los autores CP, LE y MV. Las 18 muestras finales están distribuidas entre los escritores que se acaban de mencionar en partes desiguales de 3 a 7 textos (véase tabla 46). El resultado de la prueba se muestra en el gráfico 26.

Tabla 46. *Número de textos por autor en la prueba 3 con trgramas (AO)*

ID del autor	Nº de textos en el análisis
LE	**3**
RM	**5**
MV	**4**
CP	**6**

Gráfico 26. *Representación gráfica del resultado de la prueba de evaluación 3 de trigramas en textos de artículos de opinión*



El gráfico 26 muestra el resultado de la prueba 3 con la que se evalúa el potencial discriminatorio de los n-gramas de tipo trigrama en un contexto similar al de los casos forenses. El análisis discrimina de forma efectiva entre los autores, aunque se observa cierta proximidad entre los centroides de los grupos de los escritores LE y MV. En comparación con el resultado de la prueba 3 basada en textos de narrativa, los casos analizados muestran una mayor dispersión. Los casos desagrupados están dispersos en la parte izquierda del gráfico. Es allí donde se encuentra también el centroide de RM, a quien el análisis atribuye la autoría de los textos sin excepciones. El porcentaje de clasificación correcta en esta prueba es del 95%.

d) Conclusiones

En base a los análisis realizados y los resultados obtenidos de los datos de bigramas y trigramas en el subcorpus AO, se pueden sacar las siguientes conclusiones. Los n-gramas de ambos tipos muestran un potencial discriminatorio alto en el género periodístico que por sus características se distingue mucho del género de narrativa al que pertenecen los textos en los que hemos evaluado en primer lugar con éxito la técnica de comparación lingüística forense para los fines de la atribución de autoría. La extensión de los textos que por lo general comporta problemas para la clasificación correcta no parece disminuir de forma significativa el potencial discriminatorio de los bigramas, ni tampoco el de los trigramas. La validación cruzada de las pruebas de clasificación indican un 83% de precisión en la aplicación de los bigramas como marca discriminante y 89% en la de los trigramas. No obstante, queda por comprobar en el capítulo siguiente, si estos resultados positivos implican una independencia de la técnica del género textual de los textos de análisis.

5.3 Estudio de evaluación de la técnica en casos forenses reales

En la primera sección de este capítulo se ha explicado como se ha llevado a cabo la primera parte de la evaluación de la técnica de

comparación lingüística de textos escritos en español para los fines de la atribución de autoría forense mediante los n-gramas de tipo bigrama y trigrama en los subcorpus N y AO del corpus de análisis de la tesis y se han presentado los resultados de esta evaluación. La segunda parte de la evaluación de la técnica, que comprende el análisis de textos de dos casos forenses reales, está descrita en los apartados siguientes.

El presente estudio tiene como objetivo la evaluación de la técnica de análisis de comparación lingüística forense mediante n-gramas para los fines de la atribución de autoría en el contexto de trabajo con textos forenses reales. Se espera, por un lado, poder confirmar el potencial discriminatorio significativo de los n-gramas, ya registrado en la evaluación de la técnica en el corpus general y, por otro, poder corroborar la posibilidad de una aplicación válida y fiable de la técnica en casos forenses reales.

Es importante subrayar el hecho de que en esta tesis doctoral se habla estrictamente de evaluación y no de validación, término cuyo uso erróneo se ha propagado en los estudios del área de atribución forense de autoría, probablemente por influencia de la estadística usada como herramienta base en los análisis. Resulta ingenuo pensar que, por impecable que fuera el diseño del corpus de análisis y los experimentos utilizados, estos bastarían para validar un método o una técnica de análisis lingüístico forense. En el sentido estricto de la palabra, en el marco de las ciencias forenses tampoco sería admisible usar el término en estos casos por lo que significa:

validar una técnica conlleva someterla a la revisión por pares, determinar las tasas de error y establecer un protocolo estandarizado de trabajo basándose en los resultados de un número ilimitado de casos; evaluar, en cambio, implica hacer mediante diversas pruebas una estimación del potencial discriminatorio de la marca identificativa y de la eficacia del método o la técnica de análisis lingüístico forense que la incorpora. Dicho de otro modo, la evaluación que nos ocupa en este trabajo representa la fase inicial de la validación.

Para llevar a cabo la evaluación de la técnica de comparación lingüística de textos escritos en español para los fines de atribución de autoría forense mediante n-gramas desarrollada como parte de esta tesis doctoral, se ha empleado un corpus de control compuesto por los textos de dos casos forenses reales cuyas características se han descrito brevemente en el capítulo 4. Seguidamente se completará la información relativa a ambos casos, pero antes cabe recordar las preguntas de investigación sobre atribución de autoría que se plantean en el análisis de sus textos.

En el caso forense real 1 (en adelante CR1) se pretende establecer cual es la probabilidad de que el autor de una serie de mensajes enviados por fax sea también el autor de una serie de mensajes anónimos de carácter difamatorio enviados por correo electrónico. Y en el caso forense real 2 (en adelante CR2), la pregunta de investigación tiene que ver con la posibilidad de determinar quién

de los dos posibles sospechosos del caso es el autor de varios mensajes anónimos enviados por vía electrónica.

a) Caso forense real 1 (CR 1)[126]

– *Corpus del CR 1*

El corpus original del CR1 comprende los textos de una serie de 8 cartas, 4 de las cuales fueron enviadas por fax y las otras 4, vía correo electrónico. El primer grupo de textos representa las muestras lingüísticas indubitadas y el segundo, las muestras dubitadas. Dado que la atribución de autoría usa el análisis discriminante y esto implica la presencia de más de dos grupos (autores) cuando en un caso forense como el que nos ocupa solo hay uno, en el análisis se ha ampliado el grupo de muestras indubitadas con 4 textos de un segundo posible autor que proceden de otro caso forense real. Por lo tanto, el número total de textos que incluye el corpus del estudio es de 12, distribuidos según se muestra en la tabla 47.

Tabla 47. *Distribución del corpus del CR1*

[126] Agradezco a la Dra.Turell que me haya cedido los datos relacionados con este caso forense real para su uso de control y comparación con los experimentos llevados a cabo en esta tesis

Tipo de texto	Número de textos
Faxes indubitados	4
Anónimos de otro caso	4
E-mails dubitados	4
Total	**12**

La extensión de las muestras varía entre 428 y 935 palabras, tal como muestra la tabla 48. Para evitar que la desproporción en los datos comprometa los resultados del análisis estadístico, a partir del uso del programa *Legolas 2.0*, se ha fijado un máximo de palabras al que se limita el proceso de extracción, equivalente a la extensión mínima posible de los textos, es decir, a 400 palabras.

Tabla 48. *Extensión de las muestras del corpus del CR1 según el tipo de texto*

Tipo de texto / Muestra	Faxes indubitados	E-mail dubitados	Anónimos de otro caso
T1	753	428	434
T2	480	931	787
T3	935	680	512
T4	903	678	455
Total	***3071***	***2717***	***2188***

– *Variables*

A partir de los textos del CR1 la herramienta de extracción de *Legolas 2.0* ha generado 95 bigramas y 186 trigramas. El número de

n-gramas (bigramas y trigramas) a incluir en los análisis se ha seleccionado a partir del número total de textos analizados (12), del cual se restan 2; es decir, que esta selección reduce el total de n-gramas de cada tipo a los primeros 10 n-gramas más frecuentes en el corpus del estudio. La lista de estas variables se presenta a continuación en la tabla 49[127].

Tabla 49. *Lista de los 10 n-gramas de tipo bigrama y trigrama en el corpus del CR1*

BIGRAMAS	**TRIGRAMAS**
NC	PASN
EN	PEN
ASN	ANP
RVR	NPT
PAS	ENCP
PJ	NPAS
PN	PNCP
NJ	NJCP
WV	RRVDR
EC	ENJ

[127] El número de ocurrencias por texto y tipo de texto se puede consultar en el anexo VII.

– *Resultados del análisis del CR1*

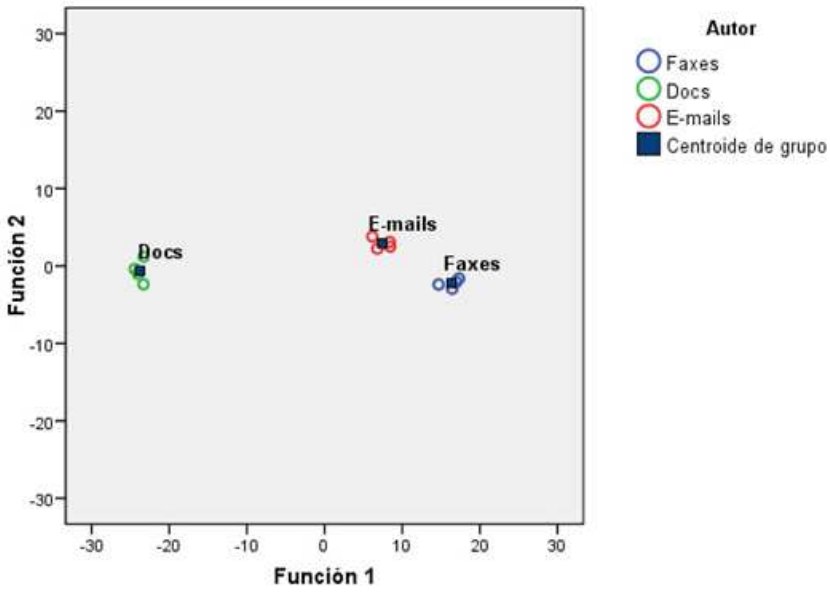
Antes de presentar los resultados cabe mencionar algunas cuestiones del análisis estadístico de los datos del CR1 que resultan problemáticas y la manera como han sido resueltas. Estas cuestiones tienen que ver con el número de sujetos y el método de análisis estadístico.

En la clasificación de nuevos casos en el análisis discriminante, los casos a clasificar se introducen como textos sin identificar, es decir, sin información acerca de su pertinencia, por lo que dichos textos aparecen en los resultados como casos desagrupados. En el caso del CR1, no obstante, para que el programa estadístico pueda generar un gráfico combinado de los resultados y no gráficos individuales para cada grupo de textos, que es lo que ocurre cuando el número de grupos identificados es inferior a 3, como sucede en el análisis del primer caso forense real, los tres conjuntos de textos han sido identificados como grupos independientes. Como resultado de este procedimiento, la probabilidad de pertenencia de los textos dubitados (e-mails) se establecerá no a partir de la distancia entre cada texto dubitado y el centroide de grupo de los dos tipos de textos, sino a partir de la distancia entre los centroides del grupo de anónimos del otro caso y cada uno de los otros dos grupos, es decir, los faxes indubitados y los e-mails dubitados. Cuanto menor sea la distancia, tanto más alta será la probabilidad de pertenencia al grupo.

- *Resultados del análisis mediante bigramas*

Los resultados del análisis discriminante basado en los bigramas más frecuentes en corpus del CR1 se muestran en el gráfico 27.

Gráfico 27. *Representación gráfica de los resultados del análisis de atribución de los e-mails anónimos en el CR1 mediante bigramas*



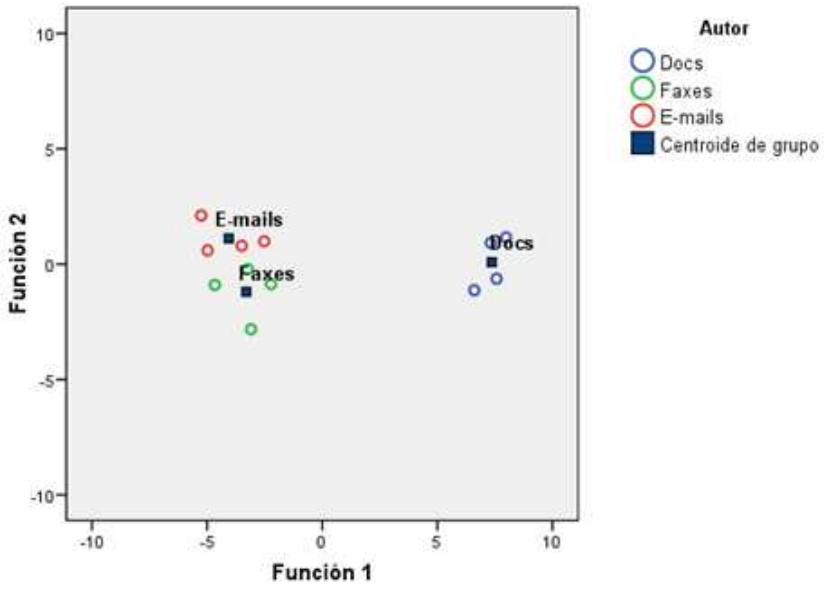
Como se puede ver en el gráfico 27, el análisis sitúa el centroide de grupo de los e-mails dubitados a cierta distancia de los centroides de los anónimos del otro caso y los faxes indubitados enviados por el también presunto autor de los e-mails dubitados. Sin embargo, la

distancia entre los anónimos (Docs en el gráfico) y los e-mails dubitados es mucho mayor e indica unas diferencias estilísticas mucho más significativas que la distancia entre los e-mails dubitados y los faxes indubitados.

Para obtener los resultados porcentuales de esta clasificación, en el análisis se aplica el método de validación cruzada. Se trata de un método que permite comprobar la capacidad predictiva de la función discriminante, y así la fiabilidad de los resultados obtenidos en la clasificación. En la validación cruzada se genera una función discriminante para cada caso en el análisis sin tener en cuenta uno de los casos que luego clasifica mediante la función de la que no forma parte. Según los resultados de la validación cruzada, 2 de los e-mails dubitados pertenecen al grupo de los textos de faxes indubitados, mientras que los dos textos restantes se mantienen en su grupo original, sin ser atribuidos al grupo de los anónimos del otro caso, cuyos textos se clasifican en su totalidad dentro de su grupo real. En resumen, la validación del análisis muestra que este es correcto en un 80% y que existe una probabilidad bastante alta de que el autor de los e-mails dubitados sea el mismo autor que el de los faxes indubitados.

- *Resultados del análisis mediante trigramas*

Gráfico 28. *Representación gráfica de los resultados del análisis de atribución de los e-mails anónimos en el CR1 mediante trigramas*



En el gráfico 28 generado por el análisis de atribución del CR1 basado en trigramas se observan unos resultados similares a los obtenidos en el análisis que emplea los bigramas. Los centroides de los grupos de los e-mails dubitados y de los faxes indubitados vuelven a posicionarse en la misma zona del gráfico, mientras que el centroide de los anónimos del otro caso ocupan el lado contrario y se sitúan a una distancia relativamente grande en comparación con la que se da entre los centroides de los e-mails dubitados y los faxes indubitados. Esta distancia es inferior a la que se aprecia en el

análisis basado en los bigramas, lo que significa que existe una mayor similitud en el uso de las construcciones representadas por los trigramas que el de los bigramas.

Los resultados porcentuales de la clasificación obtenida mediante el análisis de los trigramas indica que el 70% (es decir, 3) de los e-mails dubitados son producidos por el mismo autor de los faxes indubitados. La gran similitud entre los anónimos y los faxes se confirma en la validación cruzada, que a su vez atribuye 2 de los faxes al grupo de los e-mails y 2 de los e-mails al grupo de los faxes. Pese a que esta clasificación es correcta solo en el 42%, estos resultados confirmarían con una probabilidad moderada que el autor de los faxes indubitados es también el autor de los e-mails dubitados.

Por tanto, a modo de conclusión se puede afirmar que, según los resultados obtenidos en el análisis de los dos tipos de n-gramas, tanto los bigramas como los trigramas exhibirían un potencial discriminatorio bastante alto. Sin embargo, los trigramas se muestran como una marca de autoría que capta en mayor grado las similitudes a nivel intra autor y las diferencias a nivel inter autor.

b) Caso forense real 2 (CR 2)

– *Corpus del CR 2*

El corpus del CR2 a partir del cual se ha llevado a cabo la evaluación de la técnica de atribución forense de autoría propuesta en esta tesis doctoral comprende 27 textos. De ellos, 11 corresponden a los anónimos enviados vía correo electrónico que constituyen las pruebas lingüísticas dubitadas del corpus de análisis del estudio y 16, a las muestras indubitadas producidas por los dos sujetos, a los que en adelante nos referiremos como Autor A (8 textos) y Autor B (8 textos), uno de los cuales se sabe con certeza que ha escrito y enviado los e-mails anónimos. Los textos indubitados de los autores A y B son de dos tipos: mensajes de correo electrónico (e-mails) e informes (docs). Su distribución en el corpus se puede ver en la tabla 50.

Tabla 50. *Distribución de los textos en el corpus del CR2*

Autor / **Tipo de textos**	**Autor A**	**Autor B**	**Autor X**
E-mails	4	4	
Docs	4	4	
Anónimos			10
Total	***8***	***8***	***10***

La extensión de los textos dubitados varía entre 218 y 597 palabras y la de los textos indubitados entre 225 y 285, en el caso de los mensajes electrónicos, y entre 367 y 960, en el caso de los informes (véase la tabla 51 para la extensión exacta de cada muestra). Por lo tanto, en el programa de extracción de datos el límite en la extensión de los textos en el que finaliza la extracción de n-gramas se ha fijado en 200 palabras.

Tabla 51. *Extensión de las muestras del corpus del CR2*

Texto	**E-mails del autor A**	**Emails del autor B**	**Docs del autor A**	**Docs del autor B**	**Dubitados**
T1	240	233	960	434	462
T2	285	260	337	787	597
T3	203	225	631	512	276
T4	193	285	367	455	462
T5					214
T6					431
T7					455
T8					337
T9					218
T10					268

– *Variables*

La selección de variables en el CR2 se ha realizado siguiendo el mismo criterio que en el CR 1 y se ha limitado el análisis a los primeros 25 n-gramas más frecuentes del total de 92 bigramas y 469 trigramas. Estos n-gramas se muestran en la tabla 52[128].

Tabla 52. *Lista de los 10 n-gramas de tipo bigrama y trigrama en el corpus del CR2*

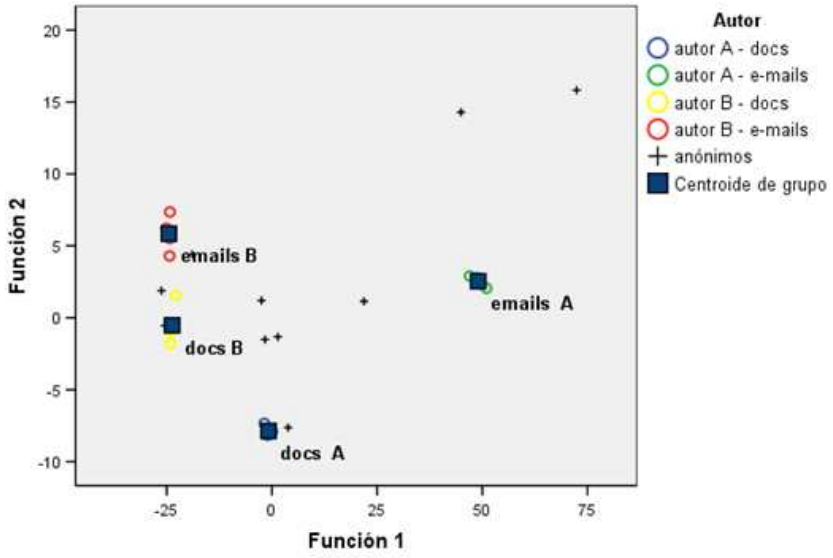
BIGRAMAS	TRIGRAMAS
NC	ANP
ASN	PASN
EN	NPT
PAS	NPAS
PN	PEN
RVR	PNCP
NJ	ENCP
EC	NJCP
PJ	PAPN
APN	EHJPA
WV	NPAP
VG	RVRV
PAP	VVPC
VB	VDAN
VC	NCNP
PV	ENJ
DT	VAN
VRE	PNJH
NR	ASNT
CD	ASNE
EH	AJN
CVR	VDEJN
DVR	RRVDR
RR	NPE
RT	RVDR

[128] El número de ocurrencias de cada variable por autor y texto puede consultarse en el anexo VII.

– *Resultados del análisis del CR2*

- *Resultados del análisis mediante bigramas*

Gráfico 29. *Representación gráfica de los resultados del análisis de atribución de los e-mails anónimos en el CR2 mediante bigramas*



El gráfico 29 muestra los resultados obtenidos mediante el análisis basado en los bigramas más frecuentes en el corpus del CR2. Los cuatro grupos de indubitados están bien separados y los marcadores de sus textos están situados a una distancia mínima de los centroides de grupo. Los centroides de los dos tipos de textos disponibles del autor B están posicionados cerca uno del otro en la parte izquierda del gráfico y presentan cierta dispersión en comparación con los casos del resto de grupos. Esta proximidad es

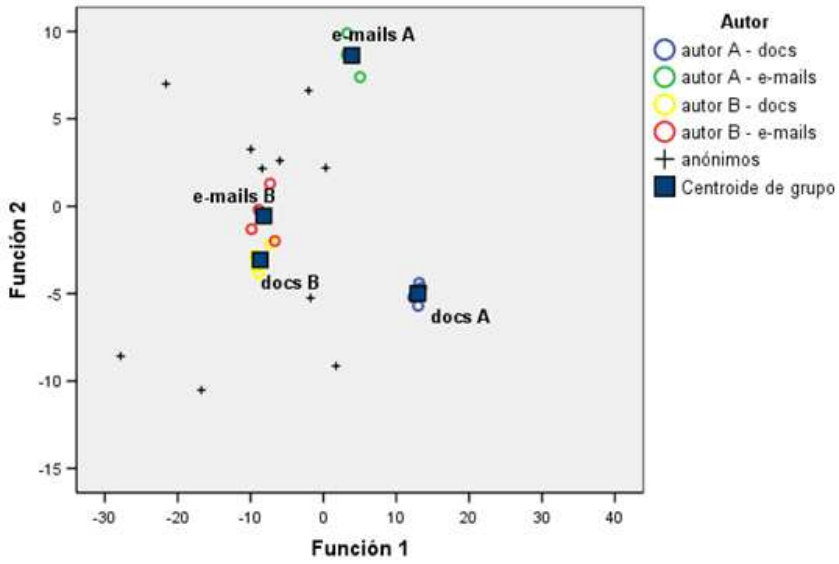
un indicio de la idiosincrasia en la manera de escribir del Autor B, que presenta menos variaciones con independencia del género textual que el Autor A, cuyos centroides de los textos de e-mails y los de docs se encuentran aislados en el otro extremo del gráfico.

El análisis atribuye 7 de los 10 textos dubitados, que aparecen marcados en el gráfico con cruces negras, al Autor B. En el gráfico 27 se puede ver cómo sus marcadores se posicionan en la zona ocupada por los textos de este autor. Del resto de casos, uno se asigna al grupo de los textos de docs del Autor A y dos, al grupo de e-mails del mismo autor. Según los cálculos estadísticos que realiza el análisis discriminante, la probabilidad de que el Autor B sea el autor de los anónimos es moderada (65%) mientras que la probabilidad de que sea el Autor A es baja (35%).[129]

[129] En el CR2 no se ha usado el método de validación cruzada, ya que por defecto solo se aplica a los casos de análisis de los grupos identificados y excluye los casos objeto de atribución.

- *Resultados del análisis mediante trigramas*

Gráfico 30. *Representación gráfica de los resultados del análisis de atribución de los e-mails anónimos en el CR2 mediante trigramas*



Como se puede observar en el gráfico 30, los resultados de la clasificación mediante trigramas no difieren de los anteriores en cuanto a la situación de los centroides de los cuatro grupos de análisis. Los grupos de e-mails y docs del Autor B vuelven a posicionarse en proximidad inmediata uno del otro con la única diferencia de que en este caso se observa también cierto solapamiento entre los dos tipos de textos. Los centroides de los grupos del Autor A se mantienen alejados, ya que el análisis tampoco encuentra ningún tipo de similitud entre ellos a nivel de los trigramas. En cuanto a los textos anónimos, el análisis determina

que con la excepción de uno que es asignado al grupo de e-mails del Autor A, todos pertenecen al Autor B, atribuyendo 5 al grupo de los e-mails y 4, al de los docs. En términos porcentuales este resultado se traduce en un 85% de probabilidad de que el autor de los anónimos sea el Autor B mientras que la probabilidad de que sea el Autor A es baja (15%).

– *Conclusiones*

Los resultados expuestos más arriba permiten llegar a dos tipos de conclusiones. Por un lado, estos resultados reafirman las conclusiones de los análisis realizados a partir del corpus de análisis y del corpus de control del CR 1, según los cuales los n-gramas de tipo bigrama y trigrama son marcas de autoría con un potencial discriminatorio relativamente alto, pero superior en el caso de los trigramas, mediante los que se obtienen resultados más precisos.

5.4 Estudio sobre la capacidad de los n-gramas de discriminar entre las dos principales variedades lingüísticas del español

Este estudio presenta los resultados de los experimentos con n-gramas que tienen como objetivo validar la hipótesis formulada en esta tesis doctoral acerca de la capacidad de los n-gramas de discriminar entre autores según la variedad lingüística del español que usan en sus escritos. Las dos variedades que nos conciernen aquí son la peninsular y la latinoamericana, sin tomar en cuenta la diversidad dialectal, es decir, las sub-variedades que ha identificado y descrito la dialectología hispánica en el marco de esas dos macro-variedades. Son de interés, por lo tanto, las diferencias que se observan entre ambas a una escala general, siendo específicamente importantes en el estudio, por la tipología de la marca que empleamos en el análisis, los rasgos divergentes que se manifiestan en la estructura y la composición oracional. Cabe aclarar que, al tratarse de un estudio exploratorio de las futuras vías de trabajo y aplicación forense de los n-gramas, este estudio no se centra en definir las unidades sintácticas divergentes y establecer su correspondencia con las SEM como formantes de la variable de análisis, sino que tiene como objetivo determinar si los n-gramas captan las diferencias sintácticas inherentes a las dos variedades lingüísticas para poder distinguir entres sus usuarios a nivel de variación inter autor.

a) El marco socio-cultural de la diferenciación diacrónico lingüística del español

Es común en las lenguas cuyas fronteras lingüísticas han traspasado los límites de su país de origen y se han asentado de forma oficial en varios continentes - y el español es un ejemplo claro de este tipo de lenguas - que propicien las variedades lingüísticas entre las cuales las diferencias se hacen patentes. La diferenciación suele ser debida a hechos históricos y socioculturales, que en el caso de la lengua española han sido varios. Ofrecer una descripción exhaustiva de los factores y los acontecimientos que crearon las condiciones para que el español de la península y el de América se desarrollaran de distinto modo y con marcadas diferencias no es el eje central de esta tesis ni tampoco de este estudio. Sin embargo, es preciso mencionar aquí aquellos factores que la mayoría de los autores consultados sobre el tema señalan como primarios, a saber: el contacto del español traído de la Península Ibérica con las lenguas autóctonas del nuevo continente y con algunas lenguas africanas durante el proceso de hispanización, la situación sociocultural en cada región del nuevo territorio dominado y la situación lingüística en España en el momento de la conquista.

En la época del descubrimiento del Nuevo Mundo y la llegada de los primeros colonizadores, el territorio de la actual América Latina ya estaba habitado por diversos pueblos indígenas que tenían su propia lengua y cultura. Este hecho impuso que desde un principio

el español conviviera, incluso a veces en competencia, con las lenguas autóctonas e inevitablemente se viera influido por ellas en cierto modo. El desarrollo histórico y social durante el período colonial y los inicios de la independencia, a la par que la estructura sociocultural de cada zona dentro de la amplia geografía hispanoamericana, conllevó que las influencias lingüísticas de las lenguas autóctonas fueran más significativas en algunos países de Latinoamérica que en otros. Este hecho es debido a que el contexto social y cultural de las tribus indígenas en cada región donde se asentaba una colonia española era muy diverso. Algunas de estas tribus estaban conformadas por una población pobre, socialmente aislada y sin prestigio étnico, que adoptaba rápidamente la cultura española y se diluía en la nueva sociedad impuesta por el colonizador sin oponer gran resistencia, mientras que otras tribus vivían en estados bien estructurados y contaban con un rico patrimonio cultural y lingüístico de los que no estaban dispuestas a desprenderse.

Durante su expansión en el Nuevo Mundo, el español estuvo en contacto también con otras lenguas. A pesar de no haber influido en la misma medida que las lenguas indígenas y el lenguaje de los colonizadores en la diferenciación del español latinoamericano de su homónimo peninsular, no se puede menospreciar la marca que dejaron las lenguas propias de los africanos traídos como mano de obra esclava en la cultura y la lengua de toda América. Su influencia en Sudamérica se percibe sobre todo en el lenguaje de los

hispanohablantes de la costa del Caribe, antaño punto de entrada de los barcos de esclavos (Malmberg, 1974).

Desde la otra punta del océano, el español medieval tuvo también un impacto importante. La conquista de América coincide con el período en el que los Reyes Católicos inician la unificación lingüística de España, donde coexisten muchos dialectos del latín vulgar, entre ellos el castellano, la variedad geográfica que se convertiría en la lengua nacional de la península, y que en la época todavía no había concluido su evolución hacia cambios lingüísticos que acabarían dando al español peninsular su fisonomía actual (Frago García y Franco Figueroa, 2001). De ahí que los colonizadores, originarios de distintos puntos de España en los que el castellano se encontraba en contacto con otros dialectos y había desarrollado modalidades distintas, jugaron un papel significativo en la formación del español latinoamericano como una variedad diferenciada (Ramírez Luengo, 2007: 11-12). Hablantes de distintas variedades de un español medieval, que además en aquel momento se caracterizaba por su heterogeneidad lingüística, "transfirieron" sus particularidades dialectales al español del nuevo continente. Ejemplos claros de dicha influencia son algunos fenómenos lingüísticos como, por ejemplo, el voseo, hoy en día presente en Latinoamérica, pero desaparecido en la península.

A consecuencia del efecto de los factores que se acaban de reseñar y de otros que omitimos por no ser tan relevantes, las divergencias

entre las dos variedades del español peninsular y latinoamericana, se extienden a todos los niveles lingüísticos: fonético, léxico, sintáctico, e incluso semántico (véase Alvar, 1996).

El hecho de que las diferencias entre estas dos variedades se den también a nivel sintáctico, nos permite formular la hipótesis de que los n-gramas, como combinaciones de representaciones de construcciones y/o estructuras sintácticas, pueden reflejar estas diferencias y pueden constituir marcas de origen geográfico-lingüístico.

b) Corpus del estudio

El corpus de este estudio sobre el potencial discriminante de los n-gramas entre las dos principales lingüísticas del español está constituido por los textos de narrativa de 14 de los 17 autores originales del subcorpus de fragmentos de novela (subcorpus N). Cada variedad lingüística, peninsular (P) o latinoamericana (A), está representada por 7 autores. Como se puede ver en la tabla 53, el grupo de autores usuarios de la variedad peninsular incluye los fragmentos de narrativa de: Alicia Giménez Bartlett (AG), Antonio Muñoz Molina (AM), Carmen Laforet (CL), Eduardo Mendoza (EM), Juan José Millás (JM), Lucía Etxebarría (LE) y Rosa Montero (RM), mientras que las muestras escritas de la variedad latinoamericana pertenecen a: Carmen Posadas (CP), Gabriel García Márquez (GG), Isabel Allende (IA), Laura Restrepo (LR), Mario

Benedetti (MB), Mario Vargas Llosa (MV) y Vlady Kociancich (VK). Para los análisis del estudio se han tomado la totalidad de textos de cada escritor disponibles en el subcorpus N.

Tabla 53. *Distribución del corpus del estudio según la variedad lingüística del autor*

Identificadores de los sujetos	N° de sujetos	Variedad lingüística	N° de novelas por autor	N° de muestras por autor
AG	7	español peninsular (P)	5	25
AM				
CL				
EM				
JM				
LE				
RM				
CP	7	español latinoamericano (A)	5	25
GG				
IA				
LR				
MB				
MV				
VK				

c) Análisis de clasificación de textos escritos mediante n-gramas según la variedad lingüística del autor

En el análisis discriminante de los datos de n-gramas (bigramas y trigramas), para determinar la variedad lingüística más probable en el que está escrito cada texto, se ha empleado como medida estadística de clasificación la distancia Mahalanobis, que mide la

distancia de cada caso (texto) respecto del centroide de los grupos para establecer su pertenencia a una de las dos variedades lingúísticas. Cuanto más cercanas son las distancias calculadas para un texto del centroide de un grupo, tanto más alta es la probabilidad de que el texto pertenezca a dicho grupo. Como conclusión del análisis, para confirmar la validez de los resultados obtenidos, se ha aplicado el método de la validación cruzada, descrito en el apartado 3 de este capítulo.

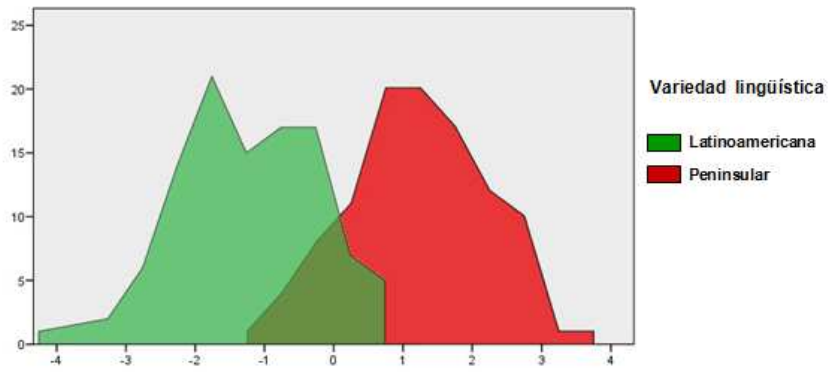
d) Resultados y discusión del análisis

Los resultados que se presentan a continuación se basan en los 75 bigramas y los 77 trigramas[130] más frecuentes en el corpus del estudio. Estos n-gramas han sido seleccionados mediante el mismo criterio descrito en la sección 6 del capítulo 4 y empleado también en la selección de variables del resto de estudios que recoge esta tesis.

[130] Para el listado de estos n-gramas, véase el anexo V.

- *Resultados del análisis basado en los datos de bigramas*

Gráfico 31. *Categorización de los textos del corpus del estudio según la variedad lingüística del autor mediante los bigramas más frecuentes*



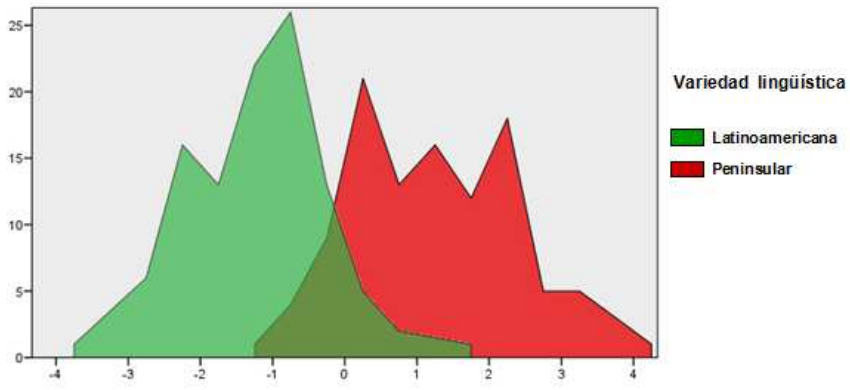
Los resultados del análisis mediante bigramas se muestran en el gráfico 31. En el eje horizontal del gráfico aparece la escala de distancias de los textos respecto los centroides de grupo. Como se puede observar en este gráfico, en el análisis se han calculado valores de distancia desde -4,5 a 0,8 para el grupo de los textos en español latinoamericano (que están representados por la figura en color verde en el gráfico) y desde -1 a 3,8 para el grupo de los textos en español peninsular (que se corresponde con la figura en color rojo). La zona de solapamiento entre las dos figuras, con valores desde -1 a 0,8, representa los casos de textos que han sido asignados incorrectamente a uno de los dos grupos por el análisis.

La equivalencia de estos datos en porcentajes indica que en el caso de los autores usuarios de la variedad latinoamericana un 88,6% (o 93 textos en total) han sido correctamente clasificados y, en el caso de los autores peninsulares, se han clasificado con éxito un 87,6% (o 92 textos). Los textos atribuidos erróneamente a un grupo constituyen un 12,4% (13 textos) de la clasificación para el grupo latinoamericano y un 11,4% (12 textos) para el grupo peninsular.

Según la validación cruzada, los resultados de la clasificación inicial son correctos en el 66,7%. Este porcentaje significa que los bigramas logran discriminar de forma efectiva por encima del umbral crítico del 50%, por lo que es posible considerar que los bigramas poseen un potencial discriminatorio considerable en el contexto de discriminación entre las dos variedades lingüísticas del español y merecen ser estudiados en mayor detalle.

- *Resultados del análisis basado en los datos de trigramas*

Gráfico 32. *Representación gráfica de los resultados de la categorización de los textos del corpus del estudio según la variedad lingüística del autor mediante los trigramas más frecuentes*



Como se puede ver en el gráfico 31, que muestra los resultados de la categorización mediante los trigramas más frecuentes, el análisis consigue discriminar entre los dos grupos de autores que escriben en estas dos macro-variedades: el español latinoamericano y el peninsular. Los valores de distancia del centroide del grupo de textos de autores latinoamericanos oscilan entre 1,8 y -3,8 en el eje horizontal y los valores del grupo de textos de escritores peninsulares, entre -1,2 y 4,2. Como se ha podido observar en el análisis de bigramas, entre las figuras representativas de los dos grupos hay un solapamiento similar. No obstante, en este caso cabe decir que, a diferencia de los resultados de los bigramas, la

intersección de solapamiento que se produce en el análisis de trigramas entre el grupo peninsular y el grupo latinoamericano es mayor (de -1 a 1,8) que la que se da entre el grupo latinoamericano y el grupo peninsular (de -1 a 0). Es decir, el número de textos incorrectamente clasificados en el caso de los textos de escritores peninsulares (14, que equivale a 13,3%) es mayor que el número de textos de escritores latinoamericanos asignados erróneamente (8, que representa un 7,6%). Por otro lado, el porcentaje de clasificación correcta del total de textos en el análisis es de un 89%. El hecho de que los trigramas se muestren como mejores discriminadores en el caso de los escritores originarios de América Latina resulta difícil de explicar sin un previo examen exhaustivo de todos los textos para establecer las secuencias de categorías exactas responsables de la clasificación errónea. No obstante, se hace patente que los trigramas a diferencia de los bigramas captan estructuras lingüísticas que según el análisis estadístico son propios más bien de la variedad latinoamericana de la lengua española.

La validación cruzada devuelve un resultado porcentual muy similiar al obtenido en la ejecución del método en el análisis de los bigramas: en este caso, un 66,2%, resultado que también confirma el alto potencial discriminatorio de los trigramas como posible marca de origen del autor.

– *Conclusiones*

Los resultados obtenidos mediante el análisis de los dos tipos de n-gramas han llevado a la conclusión de que tanto los bigramas como los trigramas poseen un potencial discriminatorio alto, hecho que apunta a su posible aplicabilidad como marca de origen, sobre todo en la creación de perfiles lingüísticos forenses de identificación. Sin embargo, lo realmente interesante desde el punto de vista de la atribución forense de autoría, en cuyo marco hay que apuntar hacia las marcas más idiosincrásicas de estilo idiolectal de un escritor que sin duda se entrelazan con sus rasgos dialectales, será testar el potencial discriminante de los n-gramas de ambos tipos mediante experimentos más específicos y en un corpus en el que todas las sub-variedades lingüísticas, tanto del español peninsular como del español latinoamericano, estén bien representadas.

6. CONCLUSIONES GENERALES

En este último capitulo de la tesis doctoral se exponen, por un lado, las conclusiones finales del trabajo teniendo en cuenta los objetivos y las hipótesis que hemos planteado en su inicio. Por otro lado, se comentan las contribuciones y las posibles vías de trabajo para el futuro.

6.1 Conclusiones generales y específicas

El eje central de esta tesis gira en torno a la evaluación de las secuencias de categorías gramaticales o n-gramas de tipo bigrama y trigrama como marcas de autoría y su aplicabilidad en la comparación lingüística de textos escritos en español con fines forenses. Esta evaluación está vinculada a la confirmación de diversas hipótesis que se formulan a propósito de las principales cuestiones metodológicas pendientes de solución, que preocupan en la actualidad a los lingüistas que se dedican a la pericia lingüística forense y a la investigación en el campo de la autoría.

La principal hipótesis que se ha podido confirmar ha sido acerca el potencial discriminatorio de los n-gramas. Mediante el análisis que se ha llevado a cabo en diversos corpus cuyas características han sido concordadas con los factores el efecto de los que se ha querido estimar en el valor discriminante de los n-gramas, como, por ejemplo, la extensión de los textos, el género textual de los

documentos de análisis y el tiempo de medición en su producción, se ha llegado también a las conclusiones que se exponen en los sub-apartados que siguen.

a) Conclusiones sobre los estudios inter autor

– *Conclusiones sobre el potencial discriminatorio de los n-gramas en textos de narrativa y en textos de artículos de opinión*

Para testar la hipótesis de que los n-gramas (los bigramas y los trigramas, en concreto) poseen potencial discriminatorio, los sometimos a diversas pruebas basadas en el análisis discriminante, en primer lugar, en textos de narrativa y, en segundo lugar en textos de artículos de opinión.

Los resultados de las pruebas realizadas con los textos de novela han demostrado que tanto los bigramas como los trigramas discriminan con un alto porcentaje de clasificación correcta de los casos por autor (de entre 80 y 95%) y de todos los casos (88% para los bigramas y 92% para los trigramas). El potencial discriminatorio además no acusa ninguna relación de dependencia del número de textos y autores empleados en el análisis. Al contrastar los resultados de los dos tipos de variables pudimos observar que los trigramas captan las idiosincrasias que permiten discriminar entre los autores analizados en mayor grado que los bigramas.

Por lo que respecta a las conclusiones del análisis de los textos de artículos de opinión los resultados son también alentadoramente positivos. Los n-gramas de ambos tipos muestran un potencial discriminatorio alto en el análisis del género periodístico que es bastante más diferente de la narrativa. El hecho de que los textos no superasen las 300 palabras, una extensión que suele comportar dificultades de cara a la clasificación correcta, no resultó tener ningún efecto a causa del que el potencial discriminatorio disminuyera ni en los bigramas ni en los trigramas. Los datos porcentuales de la clasificación coinciden en ser altos (89% para los bigramas y 93% para los trigramas). La aplicación de los trigramas como unidad idiosincrásica discriminante vuelve a producir una clasificación más precisa.

En definitiva, los resultados de los dos estudios afirman la validez de la hipótesis de que los n-gramas son buenos candidatas a marcas de autoría ya que poseen un alto potencial discriminatorio evaluado en diferentes contextos de trabajo.

– *Conclusiones sobre casos forenses reales*

La evaluación de los n-gramas de tipo bigrama y trigrama en dos casos forenses reales (CR1 y CR2) descritos en el capítulo 0, ha llevado a resultados que confirman su potencial discriminatorio en el contexto en el que muchas marcas de autoría previamente testadas en corpus generales suelen fallar. Se ha podido establecer que aunque ambos tipos de variables muestran un alto poder

discriminante tanto en el análisis de las pruebas lingüísticas de CR1 y CR2, los trigramas rinden un nivel de precisión más alto que los hace más adecuados para una aplicación con fines forenses.

- *Conclusiones sobre las variedades lingüísticas del español*

Los alentadores resultados obtenidos en el análisis de comparación lingüística mediante bigramas y trigrama nos instigaron a explorar en mayor detalle hasta qué limites de la variación inter autor se extiende su potencial discriminatorio. Concretamente, era de interés saber si los n-gramas iban a ser capaces de discriminar entre las producciones lingüística de los usuarios de la dos variedades principales de la lengua española, la latinoamericana y la peninsular, demostrando así su posible aplicabilidad como marca de origen en la creación de perfiles lingüísticos forenses de identificación. Aunque el análisis reveló que los bigramas y los trigramas discriminan con éxito entre las dos variedades para poder considerar si los resultados son concluyentes es indispensable testar estas variables mediante experimentos más específicos y en un corpus en el que todas las variedades lingüísticas del español estén bien representadas.

b) Conclusiones sobre los estudios intra autor

- *Conclusiones sobre la variación intra autor en tiempo aparente y en tiempo real*

Las posibilidades de que una unidad lingüística de carácter idiosincrásico se convierta en marca de autoría no dependen únicamente de la variación que presenta a nivel inter autor, y de que esta exhiba una significación estadística alta, sino también, e incluso se podría decir, en mayor grado, de que esa variación se manifieste a nivel intra autor.

En el estudio sobre la variación intra autor de los n-gramas en tiempo real y en tiempo aparente, analizando dos grupos de textos que comprenden diversos tiempos de medición, de 10 a 40 años (grupo 1) y de 5 a 8 (grupo 2), pudimos constatar que muy pocos de los n-gramas de mayor potencial discriminatorio varían de forma estadísticamente significativa. Esta variación se produce además sobre todo entre los textos de las novelas escritas con un período intermedio mínimo de 10 años y es más notable en el conjunto de n-gramas de tipo bigramas. Por lo tanto, se ha podido corroborar la hipótesis de que los individuos no tienden a variar mucho respecto a la recurrencia de los n-gramas y esto además ocurre con un número limitado de variables.

- *Conclusiones sobre la variación intra autor según el género textual*

Para terminar, las conclusiones pertinentes del estudio final sobre la variación intra autor, que se centra en el género textual como factor de la variación, se pueden resumir de la siguiente manera. Los n-gramas de mayor frecuencia muestran un alto potencial discriminatorio tanto en los textos de tamaño reducido como en los textos más extensos independientemente del género textual (narrativa o artículos de opinión) al que se aplica. Sin embargo en lo que se refiere al análisis conjunto de textos de géneros diferentes su capacidad clasificatoria y potencial discriminante disminuyen en gran medida. Los trigramas poseen mayor potencial discriminatorio que los bigrama y son menos susceptibles a la variación intra autor.

6.2 Aportaciones de esta tesis

A través de la investigación que refleja esta tesis doctoral esperamos haber contribuido con una propuesta metodológica de análisis basado en los n-gramas de tipo trigrama y bigrama y a la necesidad de poder contar con una marca de autoría fiable y válida en la comparación forense de textos escritos. Los n-gramas no representan una unidad de análisis nueva en el desarrollo científico de la lingüística general, pero su explotación en la comparación de textos escritos a efectos de sustentar de forma más fiable la atribución forense de autoría es bastante reciente (Spassova, 2006; Spassova y Turell, 2007) e innovadora. Los buenos resultados de su aplicación en la comparación lingüística de textos indican que podría constituir una marca útil por su posible implementación en herramientas informáticas que faciliten el proceso de comparación

de textos, como es el caso del programa *Legolas 2.0* que se ha usado en la parte metodológica de esta tesis.

Referencias bibliográficas

Abecassis, M. (2002).«Saliency and frequency in a corpus of 1930's French fims». *Californian Linguistic Notes*, Vol. 17(2). págs. 1-19[Disponible en: http://hss.fullerton.edu/linguistics/cln/fal02/abecassis-saliency.pdf]

Argamon, S. et al. (2003). «Gender, genre, and writing style in formal written texts». *Text*, 23(3), págs.321-346

Argamon, S. y Levitan, S. (2005). «Measuring the usefulness of function words for authorship attribution». En Proceedings of the 2005 ACH/ALLC Conference. Victoria BC (Canada).

Baayen R, H. (2006). *Exploratory data analysis: An introduction to R for the language sciences*. Nijmegen: Interfaculty research unit for language and speech & Max Planck Institute for Psycholinguistics

Baayen, R.H., van Halteren, H. et al. (1996). «Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution». *Literary and Linguistic Computing,* vol.11(3), págs.121-131

Bailey, R.J. (1979). «Authorship attribution in forensic setting». *Advances in Computer-aided Literary and Linguistic Research*, págs.1-20

Bernstein Ratner, N. (1987)«The phonology of parent-child speech». In Nelson, K. y van Kleck, A. (eds.) *Children's language*, Vol.6. Hillsdale, NJ: Erlbaum. págs. 159-174

Bernstein Ratner, N. (1993) «Maternal input and unusual phonological behavior in a child: A case study and its implications». *Journal of Child Language*, 20. págs.191-197.

Biber, D. (1990). «Methodological issues regarding corpus-based analyses of linguistic variation». *Literary and Linguistic Computing*, 5(4), págs.257-269

Biber, D. (1991). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). «Using register-diversified corpora for general language studies». *Computational Linguistics*, 19(2), págs.221-241

Biber, D. (1995). Dimensions of register variation. A cross-linguistic comparison. Cambridge: Cambridge University Press

Biber, D. et al. (1998). *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press

Blanche-Benveniste, C. (1998) *Estudios lingüísticos sobre la relación entre oralidad y escritura*. Barcelona: Gedisa Editorial

Bock, K. J. (1982) «Towards a cognitive psychology of syntax: information processing contributions to sentence formulation». *Psychological review*, 89, págs.1-47.

Bock, K.J. (1986). «Syntactic persistence in language production». *Cognitive Psychology*, 18. págs. 355-387

Bock, K.J. y Loebell, H. (1990) «Framing sentences». *Cognition*, 35(1). págs.1-39

Bock, K.J. y Levelt, W.J. (1994). «Language production. Grammatical encoding». En Gernsbacher, M.A. (ed.) *Handbook of Psycholinguistics*. San Diego: Academic Press. págs. 945-984

Bock, J.K. y Griffin, Z.M. (2000). «The persistance of structural priming : Transient activation or implicit learning*?».* *Journal of Experimental Psychology: General*, 129(2). págs. 177-192

Branigan, H.P. et al. (1999).«Syntactic priming in written production: Evidence for rapid decay». *Psychonomic Bulletin & Review*, 6(4). págs. 635-640

Broeders, P.A. (1999). «Some observations on the use of probability scales in forensic identification». *Forensic Linguistics*, 6(2). págs. 228-241

Bruner, J. (1985) Child's talk : Learning to use language. New York: Norton.

Burrows, J.F. (1987). *Computation into Criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press

Burrows, J. F. (1989). «'An Ocean Where Each Kind...': Statistical Analysis and Some Major Determinants of Literary Style». *Computers and the Humanities,* 23. págs. 309 -321.

Burrows, J.F. (2002). «'Delta': a measure of stylistic difference and a guide to likely authorship». *Literary and Linguistic Computing*, 17(3). págs. 267-428

Burrows, J.F. (2003). «Questions of authorship: Attribution and beyond». *Computers and the Humanities*, 37. págs. 5-32

Butler, C. (1985). Statistics in Linguistics. Basic Blackwell: Oxford

Cabré, M.T., Vivaldi, J. et al. (1998). El corpus de l'IULA: Etiquetaris. *Papers de l'IULA*. Barcelona: IULA, Universitat Pompeu Fabra

Can, F., Patton, J.M. (2004). «Change of writing style with time». *Computers and the Humanities*, 38. págs. 61-82

Castellà Lidon, J.M. (2004). «La llengua oral i la llengua escrita». *Oralitat i escriptura. Dues cares de la complexitat del llenguatge*. Barcelona: Publicacions de l'Abadia de Montserrat. págs. 15-48

Cicres i Bosch, J. (2007). Aplicación de l'anàlisi de l'entonació i de l'alineació tonal a la identificació de parlants en fonètica forense. Barcelona, Institut Universitari de Lingüística Aplicada.[Tesis doctoral dirigida por: Maria Teresa Turell Julià]

Chaski, C.E. (1997). «Who wrote it? Steps towards a science of Authorship identification». *National Institute of Justice Journal*, September'97, págs.15-21

Chaski, C.E. (2001). «Empirical evaluation of language-based author identification techniques». Forensic Linguistics, 8(2). Págs. 1-65

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: The Mit Press

Cohen, M. (1973). *Manual para una sociología del lenguaje*. Madrid: Editorial Fundamentos. [Tit. orig.: *Matériaux pour une sociologie du langage*, 1956; traducción de José Martín Arancibia]

Cooper, B. (2007). «Stabilty of certain stylistic features over time». [Comunicación presentada en el congreso *Language and the Law,* Seattle]

Coulthard, M. (1993). «On beginning the study of forensic texts: Corpus concordance collocation». En Hoey, M. (ed.) *Data Description and Discourse.* London: Harper Collins.

Coulthard, M. (1994). «On the use of corpora in the analysis of forensic texts». *Forensic Linguistics*, 1(1), págs.26-43

Coulthard, M. (2004). «Author identification, idiolect, and linguistic uniqueness». *Applied Linguistics*, 25(4), págs.431-447

Coulthard, M. (2005a). «The linguist as expert witness». *Linguistics and the Human Sciencies*, 1(1). Págs. 39-58

Coulthard, M. (2005b). «Algunas aplicaciones forenses de la lingüística descriptiva». En Turell, M.T. (ed.). *Lingüística forense, lengua y derecho. Conceptos, métodos y aplicaciones*. Barcelona. DOCUMENTA UNIVERSITARIA. págs. 249-274

Coulthard, M. (2007a). «Idiolect and uniqueness of encoding». En Coulthard, M. y Johnson, A. (2007) *An introduction to forensic linguistics: Language in evidence*. Oxford: Routledge. págs. 161-

180.

Coulthard, M. (2007b).«In my opinion». En Turell, M.T.; Cicres, J.; Spassova, M. S. (ed.) (2007). *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006*. Barcelona: DOCUMENTA UNIVERSITARIA. Págs. 43-56

Diab, M. et al. (1998). «A preliminary statistical investigation into the impact of an n.gram analysis based on word syntactic categories toward text author classification». En Proceedings of 6th International Conference on Artificial Intelligence Applications. Cairo, Egipto

Ellegard, A. (1962). *A statistical method for determining authorship: The Junius Letters, 1769-1772*. Gothenburg: University of Gothenburg.

Forsyth, R. y Holmes, D.I. (1996) «Feature-finding for text classification». *Literary and Linguistic Computing*, 11(4). Págs. 163-174.

French, J.P. y Harrison, P. (2007). «Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases». *International Journal of Speech, Language and the Law,* 14(1). págs.137-144

Garside, R. et al. (1997). *Corpus Anotation. Linguistic information from computer text corpora*. New York: Longman

Garrett, M.F. (1975) «The analysis of sentence production». En Bower, G.H. (ed.) *The psychology of learning and motivation*, Vol.9. New York: Academic Press. págs.133-177.

Garton, A. y Pratt, C. (1989) *Learning to be literate: The development of spoken and written language*. Basil Blackwell: London

Gleason, J.B. y Berstein Ratner, N. (2001). *Psicolingüística*. 2ª ed. Madrid: McGraw-Hill [Título original.: *Psycholinguistics*, 1998; traducción de Ana María Esquinas]

Grant, T. y Baker, K. (2001) «Identifying reliable, valid markers of authorship: a response to Chaski». *Forensic Linguistics*, 8(1), págs.66-79

Grant, T. (2004). Authorship attribution ina a forensic context. Birmingham, University of Birmingham. [Tesis doctoral dirigida por: Malcolm Coulthard]

Grant, T. (2007). «Quantifying evidence in forensic authorship analysis». *International Journal of Speech, Language and the Law,* 14(1), págs. 1-25

Grup Corpus (2002). Procediment per a l'adquisició de textos amb l'escànner i posterior etiquetatge estructural. *Protocols IULA*. Barcelona: IULA, Universitat Pompeu Fabra

Gundlach, R.A. (1981).«On the nature and development of children's writing». En Frederiksen, C.H y Dominic, J.F. (eds.) *Writing: The nature, development and teaching of written communication*, Vol.2. *Writing: Process, development and communication*. New Jersey: Hillsdale. págs. 133–151

Halliday, M.A.K. et al. (1964). *The linguistic sciences and language teaching*. London: Longman

Halliday, M.A.K. (1979)«Differences between written and spoken language: Some implications for literacy teaching». En Glenda, P. et al. (eds.). *Communication through reading. Proceedings of the 4th Australian Reading Conference*. Adelaide, Australian Reading Association. págs. 37-52

Halliday, M.A.K. (1985). *An introduction to Functional grammar*. London: Edward Arnold

Hartsuiker, R.J. et al. (1999). «Priming word order in sentence production». The Quarterly Journal of Experimental Psychology, 52A(1). págs. 129-147

Hjelmslev, L. (1974). *Prolegómenos a una teoría del lenguaje*. 2ª ed., Madrid: Editiorial Gredos

Hernández, C.A. (1986). *Gramática funcional del español*. Madrid: Gredos

Hirsh-Pasek, K. et al. (1987). «Clauses are perceptual units for young infants». *Cognition*, 26(3), págs. 269-286

Hirst, G. y Feiguina, O. (2007). «Bigrams of syntactic labels for authorship discrimination of short texts». *Literary and Linguistic Computing*, 22(4). págs. 405-417

Holmes, D.I. (1998). «The evolution of stylometry in humanities scholarship». Literary and Linguistic Computing, 13(3). págs.111-117

Holmes, D.I. (1992).«A stylometric analysis of Morton scripture and related texts». *Journal of the Royal Statistical Society*, 155, part 1. págs.91-120

Holmes, D. I.(1994). «Authorship Attribution». *Computers and the Humanities,* 28(2), págs.87-106

Holmes, D.I. y Forsyth, R.S. (1995). «The Federalist revisited: New directions in authorship attribution». *Literary and Linguistic Computing*, 10(2). Págs. 111-127

Holmes, D.I. et al. (2001). «Stephen Crane and the New York Tribune : A case study in traditional and non-traditional authorship attribution». Computers and the Humanities, 37. págs. 315-331

Hoover, D.L. (2001). «Statistical analysis and authorship attribution: an empirical investigation». *Literary and Linguistic Computing*, 16 (4). págs. 421-444

Hoover, D.L. (2002). «Frequent words frecuencies and statistical stylistics». Literary and Linguistic Computing, 17(2). págs. 157-180

Hoover, D.L. (2003). «Frequent collocations and authorial style». Literary and Linguistic Computing, 18(3). págs. 261-286

Hudson, R. (1980). *Sociolinguistics*. Cambridge: Cambridge University Press

Hyams, N.M. (1986). *Language acquisition and the theory of parameters*. Dordrecht, The Netherlands: D.Reidel Publishing Company.

Hymes, D. (1972). «Models of the interaction of language and social life». En Gumperz, J. y Hymes, D. (eds.). *Directions in*

sociolinguistics: The ethnography of communication. New York: Holt, Rinehart & Winston.

Kantor, K. L. y Rubin, D.L. (1981) «Between speaking and writing: processes of differentiation». En Kroll y Vann R. J. (eds.) *Exploring speaking-writing relationships: connections and contrasts*. Urbana, ILL.: National Council of teachers of English. págs.55-81

Kemper, S. (1990)**.** «Adults' diaries: Change made to written narratives across the life span. Discourse processes, 13. pág.207-223.

Keulen, F.(1986). The Dutch Computer Corpus Pilot Project. En Aarts, J. y Meijs, W. (eds.) *Corpus Linguistics. II. New studies in the analysis and exploitation of computer corpora*. Rodopi: Amsterdam

Khmelev, D. (2000). «Disputed authorship resolution through using relative entropy for markov chains of letters in human language texts». *Journal of Quantitative Linguistics*, 7. págs. 115-126

Khmelev, D. y Tweedie, F.J. (2001). «Using Markov chains for identification of writers». *Literary and Linguistic Computing*, 16(3). págs. 299-307

Kjell, B. (1994). «Authorship determination using letter pair frequency features with neural network classifiers». *Literary and Linguistic Computing*, 9(2). págs. 119-124

Kjell, B. et al. (2003). «N-gram based author profiles for authorship attribution». En *Proceedings of Pacific Association for Computational Linguistics,* Halifax, Canadá. págs. 254-264

Koppel, M. et al. (2002). «Automatically categorizing written texts by author gender». *Literary & Linguistic Computing*, 17(4), págs.401-412

Kress, G. (1994). *Learning to write*. 2ª ed. London: Routledge

Kroll, B.M. (1981) «Developmental relationships between speaking and writing». In Kroll y Vann R. J. (eds.) (1981) *Exploring speaking-writing relationships: connections and contrasts*. Urbana, ILL.: National Council of teachers of English. págs. 32-54

Labov, W. (1963). «The social motivation of a sound change». *Word*, 19, págs.273-309

Labov, W. (1972). *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania Press

Lentin, L. (1996).«La dependencia de lo escrito respecto a lo oral: parámetro fundamental de la primera adquisición del lenguaje». En Catach, N. (1996) *Hacia una teoría de la lengua escrita*. Barcelona: Gedisa Editorial, págs.145-156.

Le Page, R.B. (1968). «Problems of Description in Multilingual Communities». *Transactions of the Philological Society*, 67(1). págs. 189-212.

Levelt, W.J.M. (1993). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Love, H. (2002). *Attributing authorship. An introduction*. Cambridge: Cabridge University Press.

McMenamin, G. (1993). *Forensic Stylistics*. Amsterdam: Elsevier Science Publishers B.V.

McMenamin, G. (2001). «Style markers in authorship studies». *Forensic Linguistics*, 8(2), págs.93-97

Mosteller, F. y Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. New York: Springer-Verlag.

Pardo, A. y Ruiz, M.Á. (2002). *SPSS 11. Guía para el análisis de datos*. Madrid: McGraw-Hill

Pennebaker, J.W. y Stone, L.D. (2003). «Words of wisdom: Language use over the life span». Jornal of personality and social psychology, 85(2). pág. 291-301

Perera, K. (1986) «Language acquisition and writing». En Fletcher, P. y Garman, M. Eds. (1986) *Language acquisition. Studies in first language development*. Cambridge: Cambridge University Press, págs. 494-533

Rose, P. y Morrison, G.F. (2008). «A response to the UK Position Statement" [Disponible en: http://forensic-voice-comparison.net/].

Rudman, J. (1998). «The state of authorship attribution studies: Some problems and solutions». *Computers and the Humanities*, 31, págs.351-365

Sánchez Pol, M. (2006). Proposta d'un mètode d'estilística per a la verificació d'autoria: els límits de l'estil idiolectal. Barcelona, Institut Universitari de Lingüística Aplicada. [Proyecto de tesis doctoral dirigido por: Maria Teresa Turell Julià]

Saussure, F. (1990) *Curso de lingüística general.* Madrid: Alianza

Seco, M. (1989). *Gramática esencial del español. Introducción al estudio de la lengua*. Madrid: Espasa Calpe

Scinto, L.F.M. (1986). *Written Language and Psychological Development*. London: Academic Press Inc.

Smith, M. W. A. (1983). «Recent experience and new developments of methods for the determination of authorship». *Literary and Linguistic Computing,* 11. págs.73-82.

Smith, M. W. A. (1987). «Hapax Legomena in Prescribed Positions: An Investigation of Recent Proposals to Resolve Problems of Authorship». *Literary and Linguistic Computing,* 2. págs.145 - 152.

Spassova, M.S. (2006). Las marcas sintácticas de atribución forense de autoría de textos escritos en español. Barcelona, Institut Universitari de Lingüística Aplicada. [Proyecto de tesis doctoral dirigido por: Maria Teresa Turell Julià]

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.; Spassova, M. S. (ed.) (2007). *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006*. Barcelona: DOCUMENTA UNIVERSITARIA. págs. 229-237

Spassova, M.S. (2008). «Las perífrasis verbales del español en la atribución forense de autoría». En Monroy,R. y Sánchez, A. (ed.). *25 años de lingüística en España: hitos y retos. Actas del XXVI*

Congreso de AESLA. Murcia: Universidad de Murcia, Servicio de Publicaciones. págs. 605-614.

Sridhar, S.N. (1988). *Cognition and sentence production. A cross-linguistic study*. New York: Springer-Verlag

Stamatatos, E. et al. (2001).«Computer-based authorship attribution without lexcial measures». *Computers and the Humanities,* 35. págs. 193 – 214

Stotsky, S. (1983) Research on reading/writing relationships: a synthesis and suggested directions. *Language arts,* 60(5). págs. 627-642

Svartvik, J. (1968). *The Evans Statements. A case for forensic linguistics*. Gothenburg University publication: Gothenburg

Totty, R.N. et al. (1987). «Forensic linguistics : the determination of authorship from habits of style. Journal of the Forensic Science Society, 27. págs. 13-28

Turell, M.T. (1995) *La sociolingüística de la variació*. Barcelona: Promociones y Publicaciones Universitarias, S.A.

Turell, M.T. (2003). «El temps apparent i el temps real en estudis de variació i canvi lingüístic».[en línea] *Noves SL. Revista de Sociolingüística*. [Disponible en: http://www6.gencat.net/llengcat/noves/hm03tardor/turell1_4.htm]

Turell, M.T. (2004). «Textual kidnapping revisited: the case of plagiarism in literary translation», *Forensic Linguistics*, 11(1), págs.1-26

Turell, M.T. (2004). «The disputed authorship of electronic mail: linguistic, stylistic and pragmatic markers in short texts». [Comunicación presentada en el congreso *Language and the Law,* Cardiff]

Turell, M.T. (2010). «Los retos de la lingüística forense en el siglo XXI. In Memoriam Enrique Alcaraz Varó. Homenaje a Enrique Alcaraz Varó». Revista Alicantina de Estudios Ingleses. Alicante [en prensa].

Woolls, D. y Coulthard, M. (1998). «Tools for the trade». *Forensic Linguistics.* 5(1). págs. 33-57

Zhao, Y. y Zobel, J. (2005). «Effective and scalable authorsip attribution using function words». En Lee, G.G. et al. (eds.) *Information Retrieval Technology, Second Asia Information Retrieval Symposium, AIRS 2005. Proceedings. Lecture Notes in Computer Science 3689.* Heidelberg: Springer-Verlag. págs. 174-189