Estudio piloto para la evaluación de evidencias lingüísticas en la comparación forense de textos mediante distribuciones poblacionales y relaciones de verosimilitudes

Sheila Queralt Estevez

TESI DOCTORAL UPF / 2015

DIRECTORES DE LA TESIS

Dra. Núria Bel Rafecas

(Institut Universitari de Lingüística Aplicada – Dept. de Traducció i Ciències del Llenguatge, UPF)

Dr. Lawrence M. Solan

(Brooklyn Law School)


upf.
Universitat
Pompeu Fabra
Barcelona

A mis padres Jaume y Paqui,
a mi pareja Kasper y a nuestro hijo Lance.

A Maite, quien tanto me enseñó.

AGRADECIMIENTOS

En primer lugar, quiero dar las gracias a las tres personas que me han hecho quien soy. Gracias a mis padres por inculcarme valores fundamentales como la apreciación de las cosas importantes, la ilusión por lograr retos nuevos y el placer de compartir el éxito. Siempre estaré agradecida a Maite, Dra. M.Teresa Turell, que creyó en mí desde el primer momento en que abrí enérgicamente la puerta de su oficina en la Rambla 32 como estudiante de segundo año y compartimos la pasión por la lingüística forense. Siempre le estaré agradecida porque fue la persona que me dio la oportunidad de cumplir mis sueños profesionales. Tengo mucho que agradecer a Maite y me hubiera gustado haber hecho esto y muchos otros proyectos juntas. Ha sido un honor ser su alumna y colega. Gracias de nuevo a todos vosotros.

En segundo lugar, me gustaría dar las gracias a mis dos directores, Dra. Núria Bel y Dr. Lawrence M. Solan, por aceptar un proyecto empezado con los brazos abiertos y ayudarme a llevar este trabajo hasta el final.

Estoy en deuda con Laura Barrios, Belén Garzón y el Dr. Carlos Delgado por su invaluable orientación a lo largo de la investigación, seguro que sin su colaboración no podría haber llevado adelante este proyecto. Debo agradecer al Instituto de Lingüística Aplicada y la Universitat Pompeu Fabra en especial, a sus verdaderos

profesionales que hacen que mi vida mucho más fácil de diferentes maneras y siempre están dispuestos a ayudar.

Un ambiente de trabajo lleno de apoyo es importante para sobrevivir y mantenerse cuerdo no sólo durante un doctorado, sino también para superar los momentos difíciles y los desafíos inesperados. Quiero dar las gracias afectuosamente al equipo del ForensicLab por su apoyo profesional y personal, así como al resto de compañeras becarias del IULA.

Por último, pero no menos importante, quiero agradecer a mi pareja, Kasper Birk, por su apoyo y amor, contigo he creado la mejor familia.

Este trabajo ha sido subvencionado por el Laboratorio de Lingüística Forense (ForensicLab) y la beca del Instituto de Lingüística Aplicada ambos en la Universitat Pompeu Fabra.

Estudio piloto para la evaluación de evidencias lingüísticas en la comparación forense de textos mediante distribuciones poblacionales y relaciones de verosimilitudes

RESUMEN

La presente tesis propone la implementación de técnicas estadísticas en el análisis de variables lingüísticas con el fin de crear un modelo de distribución poblacional útil en el área de la comparación forense de textos escrito. Finalmente, en una última fase se pretende aplicar el marco teórico y metodológico de la razón de verosimilitud. El objetivo es poder mejorar los resultados en la tarea de atribuir/determinar la autoría con el fin de asesorar de manera más objetiva a los diferentes agentes judiciales y poder proteger a aquellas personas involucradas en procesos judiciales de un posible error de la justicia.

Palabras clave

Comparación forense de textos escritos, atribución de autoría, estilo idiolectal, distribución poblacional, razón de verosimilitud.

Estudi pilot per l'avaluació d'evidències lingüístiques en la comparació forense de textos mitjançant distribucions poblacionals i relacions de verosimilituts

RESUM

La present tesi proposa la implementació de tècniques estadístiques en l'anàlisi de variables lingüístiques per tal de crear un model de distribució poblacional útil en l'àrea de la comparació forense de textos escrits. Finalment, en una última fase es pretén aplicar el marc teòric i metodològic de la raó de verosimilitut. L'objectiu és poder millorar els resultats en la tasca d'atribuir/determinar l'autoria per tal d'assessorar d'una manera més objectiva els diversos agents judicials i poder protegir a totes aquelles persones involucrades en processos judicials d'un possible error de la justícia.

Paraules clau

Comparació forense de textos escrits, atribució d'autoria, estil idiolectal, distribució poblacional, raó de verosimilitut.

Pilot study for the evaluation of linguistic evidences in forensic text comparison by the creation of a Base Rate Knowledge and the implementation of Likelihood Ratios

ABSTRACT

This PhD dissertation proposes the implementation of statistical techniques for the analysis of linguistic variables in order to create a useful Base Rate Knowledge to the area of forensic text comparison. Finally, at a late stage it is intended to introduce the theoretical and methodological Likelihood ratio framework. This new methodology will endeavor to improve the results in the task of authorship attribution in order to assist judicial agents in a more objective way and also to protect lay people involved in judicial processes from possible miscarriages of justice.

Keywords

Forensic text comparison, authorship attribution, idiolectal style, Base Rate Knowledge, Likelihood Ratio.

ÍNDICE

AGRADECIMIENTOS ..... **III**

RESUMEN ..... **VII**

RESUM ..... **VIII**

ABSTRACT ..... **IX**

INTRODUCCIÓN **3**

CAPÍTULO 1: **11**

1. MARCO TEÓRICO Y ESTADO DE LA CUESTIÓN **15**

1.1. LA VARIACIÓN Y EL INDIVIDUO 17

1.2. LA COMPARACIÓN FORENSE DE TEXTOS ESCRITOS 24

1.3. EL MARCO DE LA RAZÓN DE VEROSIMILITUD EN LAS CIENCIAS FORENSES 35

CAPÍTULO 2: **45**

2. EL ESTUDIO **49**

2.1. OBJETIVOS 52

2.1.1. Diseño de una propuesta metodológica *53*

2.1.2. Creación de una distribución poblacional *54*

2.1.3. Implementación de la razón de verosimilitud *54*

2.2. PREMISAS E HIPÓTESIS 55

2.3. CORPUS.....57

2.3.1. Informantes 57

2.3.2. Selección y representatividad 60

2.3.3. Recopilación del corpus 61

2.3.4. Distribución del corpus 67

2.4. VARIABLES LINGÜÍSTICAS.....69

2.5. METODOLOGÍA DE ANÁLISIS 77

2.5.1. Análisis cuantitativo y cualitativo.....77

2.5.2. Diseño experimental del análisis estadístico 80

2.5.3. Métodos estadísticos aplicados.....93

- *Análisis discriminante.....95*

- *Análisis de proximidades intra- e inter- escritor.....98*

  - *Aproximación binaria 99*

  - *Aproximación numérica discreta.....100*

  - *Análisis de las matrices de distancias.....102*

- *Análisis de la razón de verosimilitud 103*

CAPÍTULO 3:..... 107

3. RESULTADOS SOBRE LA DISTRIBUCIÓN POBLACIONAL 111

3.1. DISTRIBUCIÓN POBLACIONAL DE LAS VARIABLES 111

3.1.1. Variables de complejidad.....112

3.1.2. Variables léxicas 115

3.1.3. Variables pragmáticas.....121

3.1.4. Variables sintácticas.....128

3.2. VARIABLES CON UNA DISTRIBUCIÓN POBLACIONAL DIFERENCIADA POR EL SEXO DEL INDIVIDUO.....131

3.2.1. Variables de complejidad.....131

3.2.2. Variables léxicas 135

3.2.3. Variables pragmáticas.....143

3.2.4. Variables sintácticas.....150

3.2.5. Resumen 157

3.3. RESUMEN Y DISCUSIÓN.....158

CAPÍTULO 4:..... 159

4. RESULTADOS SOBRE LOS MODELOS DE CLASIFICACIÓN..... 163

4.1. RESULTADOS SOBRE EL POTENCIAL DISCRIMINANTE DE LAS VARIABLES 164

4.1.1. Variables de complejidad.....164

4.1.2. Variables léxicas 169

4.1.3. Variables sintácticas 173

4.1.4. Variables pragmáticas.....177

4.1.5. Resumen 179

4.2. IDENTIFICACIÓN DE PATRONES.....180

4.2.1. Clasificación variables binarias.....185

4.2.2. Clasificación variables continuas 188

4.3. PROBABILIDADES DE CLASIFICACIÓN A POSTERIORI.....195

4.3.1. Resultados discriminante directo.....197

4.3.2. Resultados discriminante a partir de distancias 202

4.4. RESUMEN Y DISCUSIÓN 205

CAPÍTULO 5:..... 209

5. CONCLUSIONES..... 213

BIBLIOGRAFÍA **219**

CASOS CITADOS **241**

ANEXOS **243**

I. ANEXO 1: LISTA DE PALABRAS MALSONANTES **245**

II. ANEXO 2: ESTADÍSTICOS DESCRIPTIVOS DE LA DISTRIBUCIÓN POBLACIONAL **247**

LISTA DE FIGURAS

Figura 1. Esquema de las variables sociolingüísticas que reflejan la variación de la lengua. 20

Figura 2. Proceso de decisión del tribunal basado en el teorema de Bayes para actualizar la incertidumbre. 41

Figura 3. Esquema de la metodología sociolingüística variacionista. Fuente: Campoy & Almeida, 2005: 295. 51

Figura 4. Ejemplo Matriz de distancias simples. 181

Figura 5. Ejemplo Matriz de distancias euclídeas medias. 181

Figura 6. Ejemplo matriz de grupo pronosticado y probabilidad a posteriori asociada. 183

LISTA DE FÓRMULAS

Fórmula 1. Estadístico F de cambio. 97

Fórmula 2. Matriz de similaridades. 100

Fórmula 3. Conversión de matriz de concordancias a matriz de distancias 100

Fórmula 4. Cálculo distancia euclídea. 101

Fórmula 5. Cálculo razón de verosimilitud. 104

Fórmula 6. Razón de verosimilitud positiva. 105

Fórmula 7. Razón de verosimilitud negativa. 105

LISTA DE TABLAS

Tabla 1. Variables sociolingüísticas controladas en la compilación del corpus. 59

Tabla 2. Distribución del corpus 1. 68

Tabla 3. Distribución corpus 2. 68

Tabla 4. Variables de complejidad analizadas. 71

Tabla 5. Variables léxicas analizadas. 72

Tabla 6. Variables pragmáticas en el análisis. 73

Tabla 7. Variables sintácticas analizadas. 75

Tabla 8. Comparación entre el modelo positivista y hermenéutico. 78

Tabla 9. Medias de las variables de complejidad con una distribución poblacional diferenciada por el sexo del individuo. . 132

Tabla 10. Medias de las variables léxicas con una distribución poblacional diferenciada por el sexo del individuo. 135

Tabla 11. Medias de las variables pragmáticas con una distribución poblacional diferenciada por el sexo del individuo. 143

Tabla 12. Medias de las variables sintácticas con una distribución poblacional diferenciada por el sexo del individuo. 150

Tabla 13. Resumen del potencial discriminatorio por sexos de las variables. 157

Tabla 14. Análisis discriminante sobre variables de complejidad. 164

Tabla 15. Análisis discriminante sobre variables léxicas. 169

Tabla 16. Análisis discriminante sobre variables sintácticas. 173

Tabla 17. Análisis discriminante sobre variables pragmáticas. 177

Tabla 18. Resumen variables discriminantes seleccionadas por bloques. 179

Tabla 19. Variables binarias discriminante directo. 186

Tabla 20. Variables binarias discriminante a partir de proximidades. 187

Tabla 21. Variables continuas discriminante directo..... 189

Tabla 22. Variables continuas discriminante a partir de proximidades. 190

Tabla 23. Evaluación de las clasificaciones. 191

Tabla 24. Intervalos obtenidos a partir de las probabilidades a posteriori mediante análisis discriminante directo. 198

Tabla 25. Resumen intervalos discriminante directo CLAS A. ... 200

Tabla 26. Resumen intervalos discriminante directo CLAS B..... 201

Tabla 27. Intervalos obtenidos a partir de las probabilidades a posteriori mediante análisis discriminante a partir de distancias. 202

Tabla 28. Resumen intervalos discriminante distancias CLAS A y CLAS B. 204

Tabla 29. Resumen resultados sobre los modelos de clasificación. 205

Tabla 30. Variables de complejidad. 247

Tabla 31. Variables léxicas..... 247

Tabla 32. Variables pragmáticas. 248

Tabla 33. Variables sintácticas. 251

LISTA DE GRÁFICOS

Gráfico 1. Distribución poblacional de las variables de complejidad. 113

Gráfico 2. Distribución poblacional de la variable léxica número de errores. 116

Gráfico 3. Distribución individual de la variable léxica errores... 117

Gráfico 4. Distribución poblacional de la variable léxica expresión de la obligación. 118

Gráfico 5. Distribución individual de la variable léxica expresión de la obligación. 119

Gráfico 6. Distribución poblacional de la variable léxica uso de abreviaturas. 120

Gráfico 7. Distribución poblacional de la variable pragmática expresión del énfasis. 122

Gráfico 8. Distribución individual de la variable pragmática número de preguntas. 123

Gráfico 9. Distribución poblacional de la variable pragmática fórmula de salutación 124

Gráfico 10. Distribución poblacional de la variable pragmática fórmula de despedida. 125

Gráfico 11. Distribución poblacional de la variable pragmática marcador discursivo..... 127

Gráfico 12. Distribución poblacional de la variable sintáctica tipo de oración compleja..... 128

Gráfico 13. Distribución poblacional de la variable sintáctica tipo de oración yuxtapuesta. 129

Gráfico 14. Distribución poblacional de la variable sintáctica tipo de oración coordinada. 130

Gráfico 15.Resultado de la distribución poblacional por sexos de la variable de complejidad número de párrafos..... 133

Gráfico 16. Resultado de la distribución poblacional por sexos de la variable de complejidad número de palabras por frase. 134

Gráfico 17. Resultado de la distribución poblacional por sexos de la variable léxica errores de ortografía. 137

Gráfico 18. Resultado de la distribución poblacional por sexos de la variable léxica errores diacríticos. 138

Gráfico 19. Resultado de la distribución poblacional por sexos de la variable léxica errores gramaticales. 139

Gráfico 20. Resultado de la distribución poblacional por sexos de la variable léxica errores de puntuación. 140

Gráfico 21. Resultado de la distribución poblacional por sexos de la variable léxica errores por contacto de lenguas. 141

Gráfico 22. Resultado de la distribución poblacional por sexos de la variable léxica errores por contacto de lenguas mediante barras apiladas. 142

Gráfico 23. Resultado de la distribución poblacional por sexos de la variable pragmática número de preguntas. 145

Gráfico 24. Resultado de la distribución poblacional por sexos de la variable pragmática trato. 146

Gráfico 25. Resultado de la distribución poblacional por sexos de la variable pragmática trato formal. 147

Gráfico 26. Resultado de la distribución poblacional por sexos de la variable pragmática trato informal. 148

Gráfico 27. Resultado de la distribución poblacional por sexos de la variable pragmática signos de puntuación. 149

Gráfico 28. Resultado de la distribución poblacional por sexos de la variable sintáctica oraciones complejas. 152

Gráfico 29. Resultado de la distribución poblacional por sexos de la variable sintáctica coordinadas ‘ni’. 153

Gráfico 30. Resultado de la distribución poblacional por sexos de la variable tipo de oración subordinada. 154

Gráfico 31. Resultado de la distribución poblacional por sexos de la variable sintáctica relativas ‘que’. 155

Gráfico 32 . Resultado de la distribución poblacional por sexos de la variable sintáctica relativas ‘cual’. 156

Gráfico 33. Variación intra- e inter-escritor variable número de párrafos. 165

Gráfico 34. Variación intra- e inter-escritor variable número de tokens. 166

Gráfico 35. Variación intra- e inter-escritor variable número de frases. 167

Gráfico 36. Variación intra- e inter-escritor variable palabras por párrafo. 168

Gráfico 37. Variación intra- e inter-escritor variable errores de ortografía. 170

Gráfico 38. Variación intra- e inter-escritor variable palabras malsonantes. 171

Gráfico 39. Variación intra- e inter-escritor variable errores por contacto de lenguas. 172

Gráfico 40. Variación intra- e inter-escritor variable oraciones subordinadas. 174

Gráfico 41. Variación intra- e inter-escritor variable oraciones simples. 175

Gráfico 42. Variación intra- e inter-escritor variable oraciones complejas. 176

Gráfico 43. Variación intra- e inter-escritor variable ausencia del sujeto yo. 178

Gráfico 44. Clasificación con VP y FN. 193

Gráfico 45. Clasificación con VP y FP. 194

Introducción

Introducción

Durante los últimos 20 años los tribunales de varios países han solicitado de manera creciente expertos en lingüística forense. Los casos en los que los expertos lingüistas ofrecen su testimonio pueden ser diversos, desde disputas sobre plagio, sobre marcas o casos de atribución de la autoría. Los casos más frecuentes en lingüística forense implican la comparación de un texto dubitado (texto anónimo cuya autoría se cuestiona) y un conjunto de textos indubitados (textos cuya autoría no se cuestiona) de un sospechoso o varios sospechosos. La estimación de la similitud o diferencia entre los dos conjuntos de textos ha sido llevada a cabo tradicionalmente por los lingüistas mediante una escala verbal que puede estar basada en estimaciones de probabilidades o en la opinión del experto. Dicho enfoque tradicional ha sido concebido por muchos, hasta cierto punto, subjetivo teniendo en cuenta que el resultado de la pericia se basa en la experiencia del experto lingüista y puede variar de experto a experto.

Este enfoque tradicional ha sido desestimado en otras ciencias forenses que trabajan con pruebas como el ADN, las huellas dactilares o la escritura a mano. Durante las dos últimas décadas, el volumen de pruebas forenses y métodos forenses sofisticados ha aumentado a pasos agigantados y, en consecuencia, se han implementado métodos probabilísticos y multivariantes en un

intento de poder evaluar la fuerza de la comparación de las propiedades cuantificables de las muestras dubitadas e indubitadas.

El método probabilístico más conocido en las ciencias forenses es el marco de la razón de verosimilitud, en inglés conocido como *Likelihood-Ratio framework (LR).* En la última década, la investigación ha demostrado la validez de los modelos basados en la razón de verosimilitud para ayudar a los expertos en ciencias forenses a interpretar la evidencia (Aitken y Taroni, 2004; Evett, 1998) y en palabras de Fenton y Neil (2012: 2) para expresar el "proper use of probabilistic reasoning has the potential to improve dramatically the efficiency and quality of the entire criminal justice system". Además, la metodología de la razón de similitud cumple con las nuevas necesidades en la presentación de pruebas forenses, aplicando procedimientos transparentes y comprobables.

A la luz de las consideraciones antes mencionadas, esta tesis doctoral se propone testar una nueva metodología basada en métodos estadísticos multivariantes y en el marco de la razón de la verosimilitud para la comparación forense de textos a través del análisis de variables lingüísticas. Esta metodología se aplicará a una muestra de textos escritos en español peninsular. A pesar de que

dicha muestra no es probabilística[1], sí que se pretende exportar los resultados del estudio al español peninsular, así como, a otras lenguas y sistemas judiciales.

En esta tesis doctoral se ha reunido un corpus de cartas de amenazas de jóvenes españoles con estudios universitarios lo más semejante posible a la realidad forense. El estudio de este corpus ha permitido cuantificar la frecuencia media de aparición de distintas variables lingüísticas y, por tanto, ver el comportamiento usual de dicha variable en la población. Un comportamiento distinto a la media poblacional podría indicar una peculiaridad del autor y, por tanto, podría ser una marca personal que identificaría sus escritos.

Este corpus también ha permitido establecer diferencias entre hombres y mujeres. De este modo, mediante el análisis de las variables lingüísticas se puede establecer el sexo del autor más probable en el momento de llevar a cabo un perfil lingüístico de una carta de amenazas anónima.

Finalmente, se ha establecido un umbral numérico para poder discernir si dos muestras pertenecen o no a la misma persona. Este umbral numérico se ha llevado a cabo mediante un riguroso

[1] Por limitaciones metodológicas, en esta tesis no se ha realizado una selección aleatoria de la población sino que los sujetos han sido seleccionados en función de la accesibilidad del investigador.

procedimiento metodológico y se ha testado contrastando la eficacia de atribuir los textos correctamente al autor original y de rechazar los textos que no son originales del autor.

En esta tesis se pretende demostrar que a pesar de la casuística de los textos forenses, textos breves y un número reducido de muestras, es posible implementar los marcos metodológicos forenses actuales con resultados satisfactorios.

El estudio propuesto en esta tesis es parte de un proyecto de investigación más amplio titulado 'Hacia la consolidación de una índice de similitud/distancia idiolectal (IS/DI) en idiolectometria forense' (FFI2012-34601), financiado por el Ministerio de Economía y Competitividad. La tesis se ha llevado a cabo en el contexto del Laboratorio de Lingüística Forense (ForensicLab). ForensicLab es parte del grupo de investigación consolidado de la Unitat de Variació Lingüística (2014 SGR 1317) en el Institut Universitari de Lingüística Aplicada (IULA), un centro de investigación de la Universitat Pompeu Fabra (UPF).

Algunas partes de esta tesis doctoral han sido presentadas en comunicaciones nacionales e internacionales. En 2013 en *Eleventh Biennial Conference on Forensic Linguistics/Language and Law* organizada por the International Association of Forensic Linguists (IAFL) y auspiciada por el Grupo de Ingeniería Lingüística (GIL) de la Universidad Nacional Autónoma de México; y, en el curso de doctorado *The Application of Linguistic Methods in Legal and*

Socio-Legal Research organizado por el programa de doctorado de la Juridiske Fakultet y Retslingvistisk Netværk RELINE de la Københavns Universitet de Dinamarca. En el año 2014 se han presentado fragmentos en la tercera edición de las *Jornadas (In)formativas de Lingüística Forense* organizadas por la Facultad de Filosofía y Letras de la Universidad Autónoma de Madrid (UAM). En el año 2015 se han presentado los resultados en *12th Biennial Conference on Forensic Linguistics (Language and Law)* organizada por the International Association of Forensic Linguists (IAFL) y auspiciada por Guangdong University of Foreign Studies (GDUFS) de China.

Esta tesis se divide en cinco capítulos. El capítulo 1 presenta un breve marco teórico y el estado de la cuestión. El marco teórico se divide en varias secciones: la variación y el individuo, la comparación forense de textos escritos y el marco de la razón de verosimilitud en las ciencias forenses, haciendo especial hincapié en la evaluación de la evidencia mediante la razón de verosimilitud.

En el capítulo 2 se plantea el estudio de esta tesis. Con este fin se introducen los objetivos y las hipótesis de investigación, el corpus de estudio y su proceso de compilación, las variables lingüísticas analizadas en la presente tesis y la propuesta metodológica.

En los capítulos 3 y 4 se presentan los resultados sobre la frecuencia de uso de las distintas variables lingüísticas analizadas en esta tesis

con el fin de poder establecer una media poblacional de cada variable. También se muestran aquellas variables que permiten diferenciar el sexo del individuo y, de este modo, poder realizar perfiles lingüísticos. Seguidamente, se presentan los resultados obtenidos tras la aplicación de la razón de verosimilitud. Finalmente, se determina el método que aporta mejores resultados en la comparación forense de textos escritos para la atribución de autoría y se realiza una propuesta metodológica.

En el capítulo 5 se presentan las conclusiones de la investigación. Se evalúa el grado de alcance de los objetivos y la validación o refutación de las hipótesis de partida. Además, se presentan las principales aportaciones y los posibles estudios futuros.

Por último, se muestra la bibliografía utilizada en esta tesis y los anexos, en los cuales se encuentran datos adicionales.

Capítulo 1:

Marco teórico y estado de la cuestión

En este capítulo se presenta un breve estado de la cuestión actual en las ciencias forenses y se motiva la necesidad de realizar un estudio como el de esta tesis. Asimismo, se introducen las tres disciplinas en que se enmarca esta tesis: la sociolingüística variacionista, la idiolectometria y, finalmente, la lingüística forense. En el punto 1.1 se presenta un resumen del estado de la cuestión en sociolingüística y se exponen las principales premisas de dicha disciplina, las cuales son claves para el estudio de la comparación forense de textos escritos. En el siguiente apartado 1.2, se realiza una breve descripción del marco de la comparación forense de textos escritos en que se inscriben las variables de estudio de esta tesis. Finalmente, en el apartado 1.3, se muestra el estado de la cuestión del marco de la razón de verosimilitud en las ciencias forenses en el que se basa la última etapa de la propuesta analítica.

1. Marco teórico y estado de la cuestión

La Comisión para la Identificación de las Necesidades de la Comunidad en Ciencias Forenses del Consejo Nacional de Investigación de EE.UU. (Committee on Identifying the Needs of the Forensic Sciences Community at the National Research Council of U.S.) publicó un documento titulado 'Strengthening Forensic Science in the United States: A Path Forward'(2009:26) en el que declaró:

> For decades, forensic sciences have produced valuable evidence that has contributed to the successful prosecution and conviction of criminals as well as to the examination of innocent people. Over the last two decades, advances in some forensic science disciplines, especially the use of DNA technology, have demonstrated that some areas of forensic science have great additional potential to help law enforcement identify criminals. Many crimes that may have gone unsolved are now being solved because forensic science is helping to identify the perpetrators.

Esta declaración debería hacer que la comunidad científica forense se diera cuenta de la importancia de su papel en la sociedad y, por lo tanto, que fuera plenamente consciente de las consecuencias –positivas y negativas– de sus peritajes. Debido a la importancia de la tarea del experto forense, la comunidad debe establecer una metodología fiable con estándares comunes y "should establish a

professional body that not only promotes these goals but also certifies experts and, where applicable, accredits training programs and laboratories" (Koehler, 2013: 537).

La presente tesis doctoral se propone mejorar los resultados obtenidos en atribución de autoría a fin de asesorar a los agentes judiciales de una manera más objetiva. De este modo sería posible proteger de posibles errores judiciales a las personas implicadas en procesos judiciales. Se persigue alcanzar dicho objetivo mediante la aplicación de técnicas estadísticas avanzadas para la selección y el análisis de variables lingüísticas en el área de la comparación forense de textos.

Este estudio se basa en tres disciplinas interrelacionadas entre sí: la sociolingüística, en particular la teoría de la variación y el cambio lingüísticos; por otro lado, la idiolectometría que será el objeto de estudio de esta tesis; y, finalmente, la lingüística forense, específicamente la comparación forense de textos para la atribución de autoría.

1.1. La variación y el individuo

La Sociolingüística es una ciencia que estudia la interfaz entre el lenguaje y la sociedad. Trudgill (1975: 28) define esta ciencia también como el estudio de las investigaciones de las lenguas hechas en su contexto social. La sociolingüística puede caracterizarse según Campoy y Almeida (2005: 1) con cinco rasgos definitorios e inherentes:

> i) es una ciencia, ii) es una rama de la Lingüística, si bien, como apunta Labov, es una forma distinta de hacer lingüística; iii) mira el lenguaje como fenómeno social y cultural; iv) estudia el lenguaje en su contexto social, en situaciones de la vida real, por medio de la investigación empírica; y v) está relacionada con la metodología y contenidos de las ciencias sociales, principalmente la Antropología Social y la Sociología.

En los inicios de la sociolingüística, los lingüistas se centraban en la microlingüística o lingüística interna, centrada en una lengua y en la competencia sistemática en el hablante. No se tenía presente la macrolingüística o lingüística externa con un habla heterogénea, variable y teniendo en cuenta la actuación del hablante. Gumperz y Hymes (1986: 13) explican la falta de estudio del lenguaje en su contexto social en los inicios:

> In the realm of theory, speech community studies have shown that the question of structural uniformity of languages is

> largely a matter of the linguist's basic assumptions; the extent to which his analysis is abstracted from everyday behaviour and above all of the field elicitation procedures he employs. When studied in sufficient detail, with field methods designed to elicit speech in significant contexts, all speech communities are linguistically diverse and it can be shown that this diversity serves important communicative functions in signalling interspeaker attitudes and in providing information about speakers' social identities. Speech communities vary in the degree and in the nature of the linguistic relationship among intracommunity variables and it is this relationship which is most responsive to social change and most revealing of social information.

De este modo el desarrollo de la investigación sociolingüística se inició en el momento en que se asumió que el lenguaje era variable y que esa variabilidad podía estar relacionada tanto con la sociedad como con el lenguaje (P. Trudgill, 1983: 32). Milroy (1983: 83) destaca que la variabilidad estructural y regular es una propiedad del uso normal de la lengua y es fundamental para comprender los mecanismos del cambio lingüístico.

A lo largo de la historia se han realizado grandes esfuerzos para sistematizar el estudio del cambio lingüístico y constatar la variabilidad del lenguaje.

En los inicios, los estudios de sociolingüística se centraban más en el sistema lingüístico que en el individuo. De hecho, en muchos casos la variación individual se concebía como una desviación del hablante respecto al modelo establecido o como un rasgo de inmadurez (p.ej. Trudgill, 1986).

Más tarde, se advirtió que el cambio en el comportamiento del grupo debía necesariamente empezar con la innovación de un individuo. Varios autores remarcaron la importancia del individuo en el cambio y destacaban "every linguistic creation is always the work of one single individual only. Several no doubt may create similar products, but neither the act of creation nor the product is affected by that" (Paul, 1889: xliii).

Actualmente, los modelos variacionistas intentan proveer métodos para entender mejor la variación de cada individuo, ya que la variabilidad se percibe como un rasgo característico de la expresión de cada individuo. Algunos hablantes utilizan la lengua de una manera próxima a la lengua convencional, pero otros pueden utilizar la lengua de una forma más creativa y, por tanto, más idiosincrática. Ejemplo de algunos de los sociolingüistas que estudian más al límite la idiosincrasia del individuo son LePage and Tabouret-Keller, los cuales identifican al individuo como "the locus of his language" (1985: 116).

Labov (2003: 234) apunta que no existen hablantes exclusivos de un único estilo puesto que todos, sin excepción, manifiestan algún tipo

de variación según las condiciones socio-contextuales en las que se encuentren. De este modo, se puede postular que existe una correlación entre las variables sociolingüísticas y los rasgos socio-demográficos adscritos (edad, sexo, etc.) como adquiridos (nivel socio-económico, nivel educativo, etc.) y con los rasgos contextuales (situaciones y estilos). En la Figura 1 se puede ver de forma gráfica la correlación de las variables sociolingüísticas.

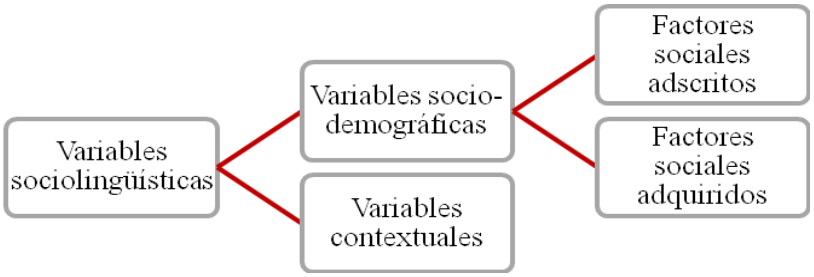

Figura 1. Esquema de las variables sociolingüísticas que reflejan la variación de la lengua.

Para los estudios en lingüística forense es fundamental prestar atención en esa expresión o idiosincrasia del individuo. Por este motivo, los estudios en este campo se centran en la teoría de la variación y el cambio lingüísticos. Dicha teoría postula que la variación es inherente a todas las lenguas y a todos los niveles: fonético y fonológico, morfológico, sintáctico, semántico, pragmático y discursivo. Sin embargo, esta variación no es casual, sino que está subordinada a factores lingüísticos –internos– tales como el tipo de oración, la expresión de la obligación, de la condición o de acciones futuras y a factores sociales –externos–

tales como el sexo, el nivel educativo o el origen geográfico de los individuos.

En lingüística forense se habla de la variación inter-escritor para la variabilidad entre los escritores y de la variación intra-escritor para la variabilidad dentro de un mismo escritor. La variación intra-escritor, es decir, la variación presente en los textos escritos por un mismo autor, es otra de las características intrínsecas de los datos lingüísticos.

> As far as we can see, there are no single-style speakers. Some informants show a much wider range of style shifting than others, but every speaker we have encountered shows a shift of some linguistic variables as the social context and topic change.
>
> Labov (1972: 208)

La variación intra-escritor se da en la selección de palabras, estructuras sintácticas, patrones gramaticales o en otros niveles lingüísticos. Dicha variación puede deberse al género textual, a la distancia temporal entre las producciones, al contexto social, al estilo, al registro o a otros factores externos.

> In the Saussurean view, there are two ways of handling individual variation. One is to treat idiosyncrasy as deviance. The other is to see a linguistic individual as constituted by the set of strategic adaptations he or she makes from a closed set

> of conventional possibilities, in the inter-actions in which he or she takes part.
>
> Johnstone (1996: 14)

Finalmente, la última disciplina en la que se basa esta tesis doctoral es la idiolectometría. Esta disciplina emergente mide la distancia o similitud idiolectal entre hablantes y ha sido capaz de establecer numéricamente el umbral que separaría a una persona del resto de su comunidad. El principal objetivo de esta disciplina es el estudio del estilo idiolectal. En la práctica, las tres disciplinas en las que se basa esta tesis –sociolingüística, lingüística forense e idiolectometría– trabajan bajo la hipótesis de la existencia de un 'estilo idiolectal' único para cada individuo que se define como:

> [The] concept 'idiolectal style', following the use of the term 'style' in pragmatics, is proposed as a notion which could be more relevant to forensic authorship contexts. 'Idiolectal style' would have to do primarily, not with what system of language/dialect an individual has, but with a) how this system, shared by lots of people, is used in a distinctive way by a particular individual; b) the speaker/writer's production, which appears to be 'individual' and 'unique' (Coulthard, 2004) and also c) Halliday's (1989) proposal of 'options' and 'selections' from these options.
>
> Turell (2010: 217)

El concepto de estilo idiolectal ha sido el centro de algunos estudios de variación sociolingüística como Abercrombie, 1969; Biber, Conrad, & Reppen, 1998; Biber, 1988, 1995; Guy, 1980 y también de estudios en lingüística forense como, por ejemplo, Bel, Queralt, Spassova, & Turell, 2012; Cicres Bosch, 2007; Gavaldà Ferré, 2011; Queralt, Spassova, & Turell, 2011; Spassova & Turell, 2007; Spassova & Grant, 2008; Turell, 2004a, 2004b, 2010.

La hipótesis del 'estilo idiolectal' hace posible el establecimiento de una medida de similitud idiolectal para poder afirmar si dos muestras lingüísticas han sido producidas por el mismo autor o no y, por tanto, hace posible que los lingüistas forenses puedan aportar pruebas en peritajes lingüísticos ante los tribunales. Esta premisa teórica es ampliamente aceptada por la comunidad de lingüistas forenses de todo el mundo como el principio para lidiar con el problema de atribución de autoría.

No obstante, la existencia del estilo idiolectal es un axioma que por su naturaleza es indemostrable. Es posible que cada individuo posea un estilo idiolectal único, pero sea ese el caso o no, es totalmente seguro que cada autor desarrolla un estilo y que el estilo de cada escritor es distinguible del estilo de otros escritores. De este modo, el método más exitoso en atribución de autoría será aquel que mida más satisfactoriamente la distancia entre los estilos de los individuos incluso con un perfil sociolingüístico similar.

1.2. La comparación forense de textos escritos

En el ámbito de la lingüística aplicada, la rama de la lingüística forense se puede definir como la interfaz entre el lenguaje y el derecho. Esta disciplina se dedica a la aplicación de los conocimientos lingüísticos en contextos judiciales y se compone por tres grandes áreas según la clasificación que proponen Gibbons y Turell (2005): lenguaje jurídico, el lenguaje judicial y el lenguaje probatorio o evidencial.

En primer lugar, el ámbito del lenguaje jurídico implica el estudio de la comprensión de los documentos legales, la interpretación del significado expresado en las leyes y otros textos jurídicos, las características del discurso jurídico, entre otros.

En segundo lugar, el lenguaje judicial implica el análisis del discurso en contextos jurídicos como el uso de la lengua por los participantes en los tribunales, por ejemplo, el discurso del juez en el momento de instruir al jurado, el lenguaje de las víctimas de delitos de violación o el lenguaje utilizado por los testigos. También abarca el estudio de la lengua utilizada por los agentes policiales en una entrevista policial así como cuestiones de multilingüismo en el contexto judicial.

Por último, el ámbito del lenguaje evidencial o probatorio se refiere al análisis lingüístico presentado ante los tribunales como prueba. La evidencia lingüística se puede proporcionar en casos, por

ejemplo, de comparación forense de habla (también conocido como identificación de hablantes), la comparación forense de textos (que incluiría casos de atribución de autoría y casos de detección de plagio), litigios de marcas registradas, perfiles lingüísticos orales y escritos, entre otros. Sin embargo, la tarea de un lingüista forense no se limita exclusivamente a proporcionar pruebas en casos judiciales, también puede ofrecer asesoramiento a efectos de investigación o contribuir a la obtención de pruebas (Coulthard, Grant, & Kredens, 2011).

Dentro de este subconjunto de campos, esta tesis se sitúa en el ámbito del lenguaje probatorio y, más concretamente, en la comparación forense de textos escritos con la finalidad de atribuir la autoría. En general, la comparación forense de textos escritos conducente a la atribución de autoría se ha definido como aquel proceso mediante el cual se analizan las características lingüísticas de un texto para extraer conclusiones sobre su autoría (Zheng et al., 2003). Además, también supone la tarea de clasificar las características lingüísticas para ser atribuidas a un autor u otro (Chaski, 2001; Grant & Baker, 2001; Kredens, 2001; Love, 2002).

En la comparación forense de textos, como en la comparación forense de habla, el análisis de la evidencia lingüística no consiste únicamente en abordar las características lingüísticas del texto dubitado, también implica cuantificar el grado de similitud entre las características dependientes del escritor obtenidas de la muestra

dubitada y las características obtenidas de las muestras indubitadas (Gonzalez-Rodriguez, Drygajlo, Ramos-Castro, Garcia-Gomar, & Ortega-Garcia, 2006: 332). No obstante, esta comparación "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 del autor, es decir, marcas identificativas (Spassova, 2009: 32)."

La tarea de analizar las características lingüísticas de un texto dubitado para esclarecer su origen ha sido nombrada de distintos modos a lo largo de los años. Sin duda por la influencia de la comparación forense de habla aplicada a la identificación de locutores, se ha extendido el término que se utiliza en esta tesis de *comparación forense de textos escrito*s con finalidad de atribución de autoría (Turell, 2010). No obstante, también se utilizan otros términos como *identificación de autoría* (Solan & Tiersma, 2005), *reconocimiento de autoría* (Hänlein, 1999) o *análisis de autoría* (Grant, 2007; 2008).

Dependiendo de la tarea a realizar en la comparación forense de textos escritos se han realizado diferentes categorizaciones. A continuación, se realiza un resumen de las categorizaciones más recientes. Turell (2011) diferencia entre determinación de autoría para cuando existen diversos candidatos de haber producido el texto dubitado y atribución de autoría para cuando se debe atribuir o no el texto dubitado a un único sospechoso. Koppel, Schler y Argamon (2009) realizan cuatro categorizaciones: en primer lugar, la tarea de

identificación cuando se debe atribuir el texto dubitado entre un conjunto cerrado de candidatos; en segundo lugar, la tarea de realizar el perfil lingüístico del texto anónimo; en tercer lugar, el problema de buscar una aguja en un pajar (*needle-in-a-haystack problem*) en tanto que se debe atribuir el texto dubitado entre miles de posibles candidatos; y, finalmente, la tarea de verificación en la que se debe concluir si el sospechoso ha producido o no el texto dubitado. Abbasi y Chen (2008) diferencian entre identificación cuando se debe atribuir un texto dubitado a un conjunto cerrado de sospechosos y detección de la similitud (*similarity detection*) cuando se compara un texto dubitado con otros textos dubitados. Juola (2008) lleva a cabo una división de acuerdo a tres tareas principales: en primer lugar, clase cerrada (*closed-class problem*) cuando se debe atribuir el texto dubitado a un conjunto cerrado de sospechosos en el cual se encuentra el autor real; en segundo lugar, el problema de clase-abierta (*open-class problem*) que incluye la atribución o no del texto dubitado entre un conjunto de sospechosos en el cual puede o no estar el sospechoso y, también, la identificación de un autor sin posibles sospechosos; finalmente, el tercer problema corresponde a la tarea de realizar un perfil lingüístico.

En el caso de esta tesis doctoral, se emplea el término comparación forense de textos escritos o atribución de autoría refiriendo a la tarea en que tenemos un anónimo y debemos atribuir ese anónimo a

uno de los autores de un conjunto cerrado de posibles autores entre los cuales se encuentra el autor real del texto dubitado.

En el campo de la comparación forense de textos escritos, tanto desde una perspectiva más lingüística como desde una perspectiva más computacional, se ha intentado abordar el problema de cuantificar el grado de similitud entre las muestras mediante una gran variedad de variables. Rudman (1998: 360) declara "approximately 1,000 style markers have already been isolated" después de haber revisado más de 300 publicaciones sobre comparación forense de textos escritos. A continuación, se presenta un resumen de las variables más comunes en este campo.

En primer lugar, las variables de complejidad o estructurales las cuales pueden incluir longitud de palabras, de oraciones, número de palabras o de párrafos por documento (p.ej. Mannion & Dixon, 1997). En este grupo también se pueden incluir otras medidas que calculan la distribución o información estadística de un documento[2] como puede ser la ratio type-token (p.ej. Grieve, 2007; Holmes, 1994; Juola, 2008; Hoover, 2003a). Cabe destacar que estas variables se han utilizado en muchos estudios dada su fácil extracción pero se debe tener en cuenta que los resultados han indicado que se ven influidas significativamente por la longitud del texto.

[2] Dichas medidas derivan del estudio de Zipf (1932).

En segundo lugar, las variables léxicas han sido objeto de números estudios, entre ellas destaca el cálculo de la frecuencia relativa de palabras funcionales o gramaticales (p.ej. Ellegard, 1962; Burrows, 1987, 2003; Grieve, 2007; Stamatatos, 2009a); las distribuciones de frecuencias de palabras (p.e.j. Holmes, 2003) estrechamente relacionadas con los fenómenos de hápax legómena y hápax dislegómena (p.ej. Smith, 1987); el análisis del vocabulario (p.ej. Coulthard, 2004; Woolls & Coulthard, 2007) y el análisis de errores (p.ej. McMenamin, 1993; Stamatatos, 2009; Koppel & Schler, 2003). La alta frecuencia de estas variables permite que se puedan realizar análisis estadísticos para cuantificar su potencial discriminante. Los estudios realizados hasta el momento apuntan que estas marcas poseen un alto potencial discriminatorio pero se debe notar que los resultados también pueden verse afectados por la longitud de los textos.

En tercer lugar, las variables de tipo sintáctico como las reglas de reescritura o construcciones sintácticas (p.ej. Baayen, van Halteren, & Tweedie, 1996); las etiquetas sintácticas en inglés, n-gramas de categorías gramaticales (*Part of Speech n-grams* en inglés) (p.ej. Diab, Schuster, & Bock, 1998; Bel et al., 2012; Queralt et al., 2011; Spassova & Turell, 2007; Turell, 2004a, 2004b); análisis sobre el tipo de oración (p.ej. Svartvik, 1968); los signos de puntuación (p.ej. Chaski, 1996); análisis de frases constituyentes (p.ej. Stamatatos, Fakotakis, & Kokkinakis, 2001); los pronombres utilizados para introducir la oración subordinada (p.ej. Turell,

2011). Gracias a la etiquetación automática, es más fácil estudiar las variables de tipo morfosintáctico puesto que se puede trabajar con un gran volumen de textos en un espacio de tiempo reducido. Los resultados apuntan que las variables morfosintácticas resultan más estables que las léxicas y que pueden diferenciar e identificar el autor de un texto con un alto grado de fiabilidad.

Finalmente, se encuentran las variables pragmáticas que pueden englobar variables como el uso de los diferentes signos de puntuación (p.ej. Chaski, 1996), las fórmulas de salutación y despedida (p.ej. Gains, 1999; Wright, 2013) o los diferentes marcadores discursivos, entre otros. Los resultados de dichos estudios destacan el alto poder discriminante gracias a la baja variación intra-autor y la alta variación inter-autor de estos marcadores. No obstante, su extracción puede resultar más costosa y, es por este motivo, que no son fuente de tantos estudios.

En los inicios de la comparación forense de textos, algunas de estas variables se analizaban en enfoques univariantes y se creía que dichas variables eran medidas capaces de reflejar el estilo idiolectal de un autor (Juola, 2008; Koppel et al., 2009). Durante las últimas dos décadas, se ha destacado la necesidad de utilizar métodos multivariantes en que se analizan varias variables e incluso varios grupos de variables. De hecho, un gran número de estudios ha corroborado que se han obtenido mejores resultados al combinar los distintos grupos de variables. Las combinaciones más frecuentes en estudios de comparación forense de textos escritos son variables

sintácticas, estructurales y léxicas (p.ej. Abbasi & Chen, 2005, 2008; Stamatatos, Fakotakis, & Kokkinakis, 2001b; Zheng, Li, Chen, & Huang, 2006); léxicas y estructurales (p. ej. Orebaugh & Allnutt, 2009; Zheng et al., 2003), léxicas y sintácticas (p.ej. Juola & Baayen, 2005; Koppel et al., 2009; Raghvan, Kovashka, & Mooney, 2010); y, sintácticas y estructurales (p.ej. Gamon, 2004; Hirst & Feiguina, 2007; Rico-Sulayes, 2011).

De este modo, se puede concluir que no existe una variable única que ofrezca buenos resultados sino que es la combinación de diferentes variables la que puede ofrecer una imagen más completa del estilo idiolectal de un autor y, por tanto, optimizar los resultados en la tarea de atribuir la autoría de un texto dubitado. Además, como apunta Picornell (2014: 88):

> La ventaja de analizar todos los rasgos hallados en un texto es que es menos probable que el lingüista forense excluya un diferenciador productivo. A pesar de que siempre habrá rasgos comunes compartidos por varios autores, ciertos rasgos pueden tener un mayor poder discriminatorio para determinados autores, aunque pueden no ser discriminatorios para otros (Rudman, 1998). Por consiguiente, nunca se deberían ignorar rasgos simplemente porque no han resultado ser productivos en análisis previos.

A pesar de todos estos estudios, el marco metodológico en la comparación forense de textos se encuentra todavía en desarrollo.

En comparación con otras ciencias forenses, la lingüística forense no es una ciencia exacta, por ese motivo algunos profesionales ante el presente estado de la cuestión del campo declaran:

> 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'.
>
> M. Coulthard (1994: 31)

Hasta la actualidad, los peritos han basado sus peritajes en análisis puramente descriptivos –puramente cualitativos– o han intentado con gran esfuerzo realizar una combinación de métodos cuantitativos y cualitativos. Solo unos pocos análisis en comparación forense de textos se han beneficiado de una combinación de los métodos cualitativo y cuantitativo y, por tanto, se han podido beneficiar de la naturaleza complementaria de la evidencia lingüística.

El enfoque cualitativo se arriesga a ser visto como una metodología poco objetiva y podría parecer que ofrece conclusiones subjetivas basadas en la opinión del experto. Así, las últimas objeciones a los enfoques cualitativos en la tarea de comparación forense de textos han sido planteadas sobre la base de que no pueden ser considerados como poseedores de una base científica u objetiva, ya que no son comprobables y no proporcionan una tasa de error (Turell & Gavaldà Ferré, 2013: 498).

Por otro lado, el enfoque cuantitativo llevado a cabo por algunos lingüistas forenses pioneros[3] y algunos científicos y lingüistas computacionales[4] ha tratado de identificar las variables discriminatorias y los modelos estadísticos del uso del lenguaje que podrían ofrecer datos de correspondencia entre las diferentes muestras textuales. Sin embargo, no ha sido posible generar evaluaciones de probabilidad debido a la falta de grandes bases de datos con poblaciones de referencia sobre distintas variables lingüísticas y, por tanto, existe un riesgo significativo de confundir o falsear los hechos en el momento testificar como expertos en un caso (Koehler, 2013: 516).

De este modo, es evidente que cuantificar el grado de similitud no es suficiente. Se debe considerar también la rareza o la probabilidad de aparición de esas características en comparación con la población relevante. Coulthard y Johnson en el libro *An Introduction of Forensic Linguistics: language in evidence* preguntan "how can one measure the 'rarity' and therefore the evidential value of individual expressions" (2007: 6). Con el fin de calcular el grado de similitud entre las muestras escritas y su rareza se debe estimar la

[3] Véase p.ej. Grant, 2007; Spassova & Turell, 2007.

[4] Véase p.ej. Juola, 2006; Koppel, Schler, & Argamon, 2009; Stamatatos, 2009.

distribución de esas variables en una población relevante – Distribución Poblacional[5].

> This Base Rate Knowledge implies the collection of data regarding the general usage of the linguistic parameters being considered by a relevant population, or group of language users from the same linguistic community, with which the specific behaviour of the speakers or writers under comparison can be compared.
>
> Turell & Gavaldà Ferré (2013: 499 nota 13)

Estas se pueden abordar con el uso de métodos probabilísticos, como la razón de verosimilitud, que llevaría a cabo análisis empíricos rigurosos. A diferencia de otro tipo de pruebas como el ADN, los lingüistas forenses tratan con datos continuos y datos variables.

> Thus, when considering the assessment of continuous data at least two sources of variability have to be considered: the variability within the source (e.g., window) from which the measurements were made and the variability between the different possible sources (e.g., windows).
>
> Aitken & Taroni (2004: 322)

---

[5] *Base Rate Knowledge*, en inglés.

Para obtener una mayor aceptación sobre la práctica de la comparación forense de textos, la comunidad de lingüistas forenses debe establecer con urgencia un marco metodológico común que sea capaz de proporcionar unos resultados en atribución de autoría fiables mediante la combinación de los métodos cualitativo y cuantitativo. Este marco metodológico debe definir el comienzo del proceso, por ejemplo la naturaleza, el número y el tamaño de las muestras, cuáles son las variables más discriminatorias, y las técnicas estadísticas capaces de medir las tasas de error.

Una de las principales razones por las no se ha materializado un marco común antes en atribución de autoría, como en otras ciencias como el análisis de ADN, es el hecho de que, por ejemplo, el marco de la razón de verosimilitud –el cual se propone en este tesis– requiere una distribución poblacional de las variables lingüísticas para poder proporcionar un valor de la 'rareza' o 'tipicidad' de ese rasgo lingüístico en ese contexto particular, y así ser capaz de proporcionar la fuerza de la evidencia (*strength of the evidence*, en inglés).

1.3. El marco de la razón de verosimilitud en las ciencias forenses

El marco de la razón de verosimilitud es un paradigma presente en las ciencias forenses de comparación que ha sido ampliamente adoptado en las tres últimas décadas y que se inició en los años 90 dentro de la evaluación de la comparación de perfiles de ADN con

fines forenses. La aplicación de este marco implica la sustitución de suposiciones anticuadas sobre la singularidad y la perfección de las pruebas en las ciencias forenses tradicionales por una ciencia más empírica y basada en la probabilística. Esta transición del enfoque anterior al nuevo en las ciencias forenses se ha llamado 'cambio de paradigma'[6] en Saks y Koehler (2005).

El marco de la razón de verosimilitud fue introducido después del fallo judicial de 1993 en el caso Daubert v. Merrell Dow Pharmaceuticals, Inc. mediante el cual el Tribunal Supremo de EE.UU. aclaró los estándares que los jueces deben utilizar para determinar si los métodos científicos de los peritos son fiables y, por tanto, decidir si la prueba pericial es admitida.

En esta sentencia el juez tenía como tarea principal excluir aquellas disciplinas científicas que no fueran válidas para presentar pruebas periciales.

Dicho tribunal se centró en la admisibilidad del testimonio de los peritos. Destacó que dicho testimonio es admisible sólo si es relevante y fiable. Además, sostuvo que las Reglas Federales de Evidencia[7] indican que el juez tiene la obligación de garantizar que el testimonio de un experto responde efectivamente a una base

---

[6] *Paradigm shift*, en inglés.

[7] Las Reglas Federales de Evidencia Se pueden consultar en: http://federalevidence.com/node/638

fiable y relevante para la tarea en cuestión[8]. Asimismo, el tribunal analizó factores más específicos, como las pruebas, la revisión por pares, las tasas de error y la aceptabilidad en la comunidad científica pertinente. Todos o algunos de estos hechos podrían ser de gran utilidad para determinar la fiabilidad de la teoría o la técnica científica[9].

Concretamente, dichos estándares son los siguientes:

1. La metodología debe haber sido probada y debe ser replicable[10]. Según el Tribunal, la prueba empírica es el criterio principal de la ciencia.
2. Debe existir una tasa de error real o probable sobre la técnica aplicada[11].
3. Deben existir y se deben mantener los estándares para el control de la aplicación de la técnica[12].
4. La aceptación general de la metodología aplicada puede tener un papel determinante para el Tribunal. Como el Tribunal señaló "[w]idespread acceptance can be an important factor in ruling particular evidence admissible, and 'a known technique which has been able to attract only

[8] Op. Cit. 597.
[9] Op. Cit. 593-594.
[10] Op. Cit. 593.
[11] Op. Cit. 594.
[12] Op. Cit. 594

minimal support within the community' may properly be viewed with skepticism".

5. La metodología debe haber sido sometida a revisión y publicación. [13] El Tribunal declaró que "submission to the scrutiny of the scientific community is a component of 'good science' ".

En 1994, en el caso Kumho Tire v. Carmichael el tribunal señaló que la Corte Suprema en el caso Daubert limitó explícitamente que los estándares solo cubrían el 'contexto científico' y añadió que a los estándares de Daubert se aplican "only where an expert relies on the application of scientific principles, rather than "on skill- or experience-based observation"[14].

No obstante, pocos años más tarde, en 1999, en el juicio de Kumho Tire Co. v Carmichael se discutió cómo se aplicaban los estándares Daubert a los testimonios de los peritos ingenieros y otros expertos que no son científicos. Se llegó a la conclusión de que 'ciertos límites' no solo se aplican al testimonio basado en el conocimiento 'científico', sino también al testimonio basado en el conocimiento técnico o especializado[15].

---

[13] Op. Cit. 593.

[14] Op. Cit. 1435-1436.

[15] Véase la regla 702 de las Reglas Federales de Evidencia.

Asimismo, el tribunal consideró que la prueba de la fiabilidad es 'flexible', y que la lista de estándares de Daubert ni necesariamente ni exclusivamente se aplica a todos los expertos o a todos los casos.

A pesar de que estos estándares se aplican solo en EE.UU., su aparición no solo ha tenido implicaciones obvias para la ley, sino que también ha tenido un gran impacto en la comunidad científica internacional en general y en la comunidad de lingüistas forenses en particular, ya que establecen un importante cambio conceptual en muchas áreas. Estos estándares determinan cuándo es admisible una evidencia científica y, a su vez, dificultan, como hemos visto en cierto grado, la admisibilidad de metodologías basadas en la experiencia del experto y no en la objetividad.

Por lo tanto, después de la decisión Daubert y de sus consecuencias, muchos científicos forenses de todo el mundo se preguntaban si sus técnicas de evaluación de las pruebas forenses, hasta ese momento, cumplían la objetividad y la transparencia exigida por los criterios Daubert.

Como resultado, Saks y Koehler (2005) propusieron el cambio de paradigma basado en el marco de la razón de verosimilitud afirmando que las ciencias forenses de identificación deben orientarse hacia el modelo de identificación utilizado en las pruebas de ADN. El modelo de las pruebas de ADN se utilizó por primera vez en los tribunales durante los años 80. Después de un período de desarrollo y de ajuste, en la actualidad los protocolos de la razón de

verosimilitud utilizados en el ADN son considerados como el modelo metodológico que debe implementarse en el resto de disciplinas forenses.

No pasó mucho tiempo antes de que los distintos institutos oficiales en ciencias forenses de todo el mundo adoptaran los estándares que emplean el marco de la razón de verosimilitud, por ejemplo, la Asociación de Proveedores en Ciencias Forenses en la República de Irlanda (ASFP, 2009), el Servicio de Ciencias Forenses en Reino Unido (Cook, Evett, Jackson, Jones, & Lambert, 1998), y el Instituto Forense de Holanda (CEH Berger, 2010). Además, Morrison (2012: 5) declara:

> In Evett et al (2011) 31 leading experts in the interpretation of forensic evidence signed a statement to the effect that the likelihood-ratio framework is the logically most appropriate framework for the evaluation of forensic evidence. This statement was also endorsed by the Board of the European Network of Forensic Science Institutes (ENFSI), representing 58 laboratories in 33 countries. […] The statement has also been followed up and expanded upon by a number of articles in refereed forensic-science, general-science, and law journals, including Berger et al. (2011) , Robertson et al. (2011), Redmayne et al., (2011) Fenton (2011), Morrison (2012) , Nordgaard and Rasmusson (2012), and Thompson (2012).

De este modo, el uso de este nuevo paradigma, es decir, el marco de la razón de verosimilitud, se recomienda ampliamente, por una parte, para la evaluación de las pruebas de comparación forense y, por la otra, para evaluar con precisión el valor de las pruebas. El marco de la razón de verosimilitud parece apropiado al contexto legal, ya que permite a los expertos evaluar la fuerza de la prueba, sin especificar cuál es su creencia previa, tanto por parte de la acusación como por la hipótesis de la defensa, consideradas en el teorema de Bayes del cual derivan las relaciones de verosimilitudes. Por lo tanto, el papel del científico y el del tribunal queda bien diferenciado. El científico exclusivamente analiza e interpreta los datos a fin de proporcionar la fuerza de la evidencia mediante el uso de recursos lingüísticos. Es el tribunal el que tiene una opinión subjetiva ante la evidencia y actualiza su opinión a la luz de las pruebas y construye un veredicto final objetivo. La representación del proceso de la decisión judicial basada en el teorema de Bayes se encuentra reflejada en la Figura 2. .

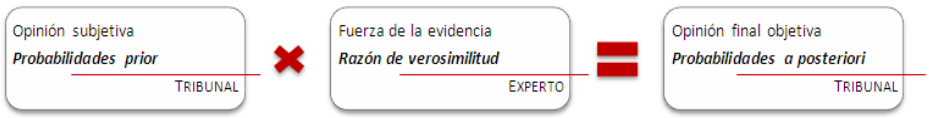

Figura 2. Proceso de decisión del tribunal basado en el teorema de Bayes para actualizar la incertidumbre.

Mortera (2006: 1) declara que "evidence presented in a case at law can be regarded as data and the issue to be decided by the court as a

hypothesis under test". La tarea del científico forense debe ser única y exclusivamente proporcionar una declaración sobre el valor de la evidencia para ayudar al tribunal a responder a la siguiente pregunta:

> How much more likely are the observed differences/similarities between the known and questioned samples to arise under the hypothesis that they have the same origin than under the hypothesis that they have different origins?
>
> Morrison (2009: 299)

Con el fin de dar respuesta a esta pregunta, el experto puede usar las relaciones de verosimilitudes (C. G. G. Aitken & Taroni, 2004) las cuales pueden definirse como la relación de dos probabilidades del mismo evento bajo dos hipótesis distintas, véase la Ecuación 1:

$$LR = \frac{p(E|H_0)}{p(E|H_1)} \quad (1)$$

En la ecuación (1) LR es la razón de verosimilitud, E es la evidencia, H_0 o hipótesis nula es la hipótesis del mismo origen o autor y, H_1 o hipótesis alternativa es la hipótesis de diferente origen o autor. Valores de LR superiores a 1 indican que la evidencia es más probable que ocurra bajo la hipótesis del mismo autor que bajo H_1; valores de LR inferiores a 1 indican que la evidencia es más probable que ocurra bajo la hipótesis de diferente autor que bajo H_0.

La interpretación del valor de las relaciones de verosimilitudes es la frecuencia –más o menos– con la que se obtienen los resultados en las condiciones de una hipótesis y en las condiciones de la otra hipótesis.

Nordgaard y Rasmusson (2012: 12) señalan que la información que se transmite a través de las relaciones de verosimilitudes es a menudo de gran valor para los tribunales, ya que puede apuntar hacia la hipótesis que explica mejor los hallazgos forenses observados. Por lo tanto, las relaciones de verosimilitudes implican ya sea la amplificación o la atenuación de la opinión del tribunal –probabilidades a priori– y como tal representa la medición de la fuerza de la evidencia.

Actualmente, no queda duda de que en los casos de ADN hay una base estadística suficiente para proporcionar resultados con LR y, por ese motivo, son utilizadas sin reservas en los tribunales de muchos países, como Australia[16], Países Bajos, Nueva Zelanda[17], Eslovenia, España, Suiza o Reino Unido[18]. Además, el marco de la razón de verosimilitud no solo se ha aplicado en casos de ADN, sino también en otras disciplinas como la comparación de huellas

[16] Véase p.ej. Regina v GK 2001(NSWCCA 413); R v Berry y Wenitong 2007 (VSCA 202).

17 Véase p.ej. Lapper v R 2005 (NZCA 259).

18 Véase p.ej. R v Doheny y Adams 1997 (1 Cr App R 369, CA).

dactilares (por ejemplo, Neumann, Evett, & Skerrett, 2012), huellas de calzado (por ejemplo, Skerrett, Neumann, & Mateos-Garcia, 2011), armas de fuego y marcas de herramientas (por ejemplo, Champod, Baldwin, Taroni, & Buckleton, 2003), vidrio (por ejemplo, Curran, Champod, & Buckleton, 2000), pintura (por ejemplo, McDermott, Willis, & McCullough, 1999), caligrafía (por ejemplo, Hepler, Saunders, Davis, & Buscaglia, 2012) o comparación forense de la voz (por ejemplo, Morrison, 2009; Delgado, 2001; Gonzalez-Rodriguez et al., 2006). Esta última disciplina está relacionada estrechamente con la temática de esta tesis, la lingüística forense, y ha sido la guía para implementar las relaciones de verosimilitudes en la comparación forense de textos ya que ambas áreas de estudio comparten muchas características debido a la naturaleza de su objeto de análisis.

El marco de la razón de verosimilitud se ha aplicado con éxito para el reconocimiento forense de hablantes por medio de sistemas de reconocimiento automático de hablantes (Meuwly & Drygajlo, 2001), o mediante el uso de técnicas clásicas tales como la fonética acústica (Rose, 2002).

Esta tesis doctoral demuestra que, a pesar de las restricciones, es posible utilizar herramientas probabilísticas con una fiabilidad razonable siguiendo los paradigmas científicos actuales. De este modo, sería posible cumplir con los requisitos de los nuevos estándares aceptados por la mayoría de asociaciones internacionales en ciencias forenses.

Capítulo 2:

El estudio

Después de un primer capítulo en el que se describe el marco teórico y el estado de la cuestión de esta tesis, en el capítulo dos se detalla su metodología. En primer lugar, se explicitan los objetivos (2.1) y las premisas e hipótesis (2.2). A continuación, se procede con la descripción de la obtención de análisis y se describen los corpus de análisis seleccionadas (2.3). Seguidamente, se definen las variables de análisis (2.4). Finalmente, en el punto 2.5 se presenta la metodología de análisis cualitativa y cuantitativa implementada en esta tesis doctoral.

2. El estudio

Esta tesis doctoral sigue el patrón común de cualquier estudio empírico experimental en que se comienza con el planteamiento del problema, la fijación de unos objetivos, la delimitación de la metodología y los procedimientos que se van a aplicar, la obtención de los datos, el análisis de los datos, su interpretación y, finalmente, las conclusiones.

En este estudio, además de implementar el patrón común de cualquier estudio empírico experimental, también se ha aplicado la metodología sociolingüística variacionista basada en una planificación inicial y los métodos de campo.

Con el diseño experimental se ha pretendido establecer un protocolo definido, propio de la metodología cuantitativa, explicitando la teoría de la que se parte, la verificación empírica de las hipótesis formuladas y la determinación de las etapas para verificar o refutar las hipótesis.

Por una parte, en esta investigación se realiza una planificación inicial, partiendo de la definición de las hipótesis de trabajo (véase Apartados 2.1 y 2.2); seguidamente, se estipula el tratamiento que se va a hacer del tiempo, en este caso un estudio transversal en tiempo aparente; a continuación, se determinan las variables lingüísticas (véase Apartado 2.4), se identifican y seleccionan, se fijan los tipos de variables que se van a analizar y se fija el tamaño

de la muestra (véase Apartado 2.3.2). Finalmente, se limitan las variables extralingüísticas (véase Apartado 2.3.1), socio-demográficas y contextuales.

Por otra parte, se precisan los métodos de campo. En primer lugar, se decide el proceso de obtención de los datos y la ética en el trabajo de campo (véase Apartado 2.3.3). Una vez recogidos los datos, se procede a la simplificación de datos mediante la codificación y la tabulación de los datos. Seguidamente, se procede al análisis estadístico compuesto por la catalogación de las variables y la aplicación de las técnicas estadísticas adecuadas (véase Capítulo 4). Para ofrecer un apoyo visual a los resultados de los análisis se realiza la visualización gráfica de los datos y se realiza la interpretación de los resultados. Finalmente, se desarrollan las conclusiones (véase Capítulo 5) considerando las implicaciones teóricas y/o metodológicas.

En la Figura 3 se muestra el esquema de la metodología sociolingüística variacionista seguida en esta tesis.

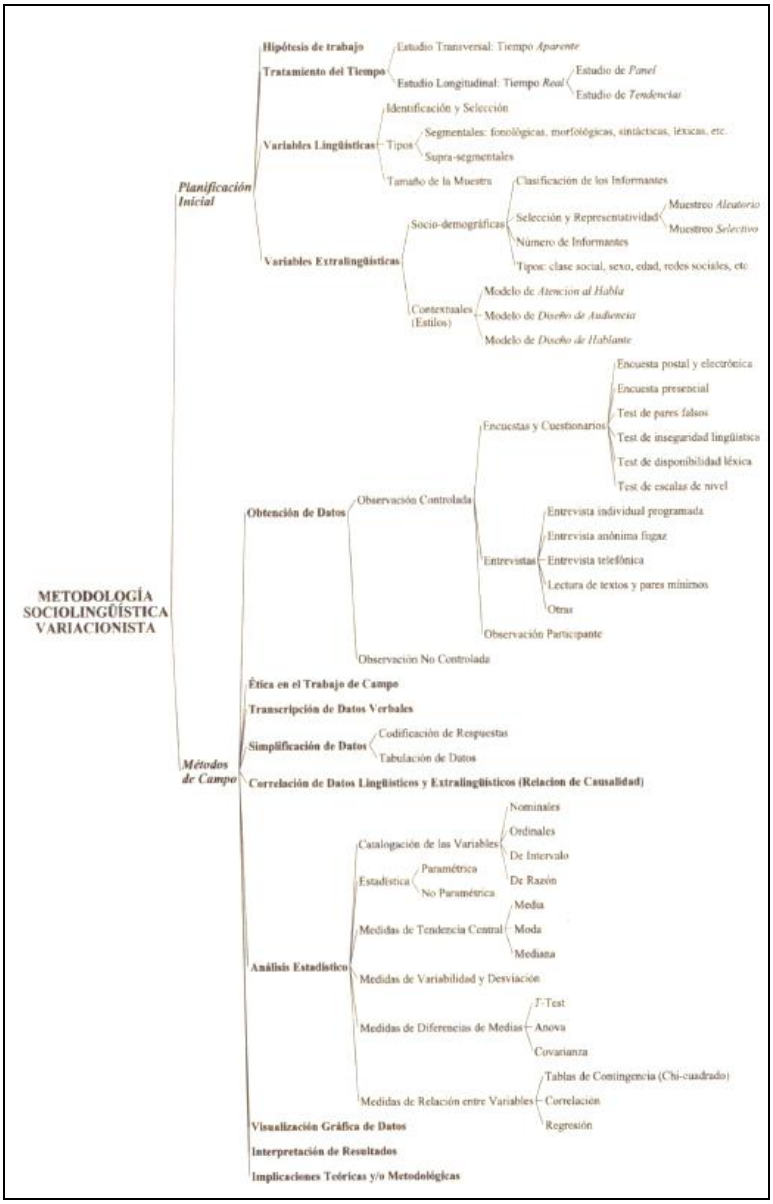

Figura 3. Esquema de la metodología sociolingüística variacionista. Fuente: Campoy & Almeida, 2005: 295.

2.1. Objetivos

El mundo de las ciencias forenses se encuentra en un cambio continuo debido a la evolución de las nuevas tecnologías y la creación de normas más rigurosas. Por consiguiente, con el fin de seguir siendo eficientes y fiables, las ciencias forenses –en este caso en particular la lingüística forense– necesitan adaptarse a los nuevos estándares. Esta tesis doctoral aborda la necesidad de presentar unos resultados más fiables y contundentes en los informes forenses presentados ante los tribunales. El objetivo general que se pretende alcanzar en esta tesis es metodológico: proponer la implementación de métodos estadísticos avanzados para el análisis de variables lingüísticas en la comparación forense de textos. En este sentido, la metodología incluye un análisis cualitativo y un análisis cuantitativo ambos fundamentados en la estadística multivariante clásica y en métodos probabilísticos como el marco de la relación de verosimilitud.

Este objetivo se puede lograr mediante la recogida de una distribución poblacional para algunas de las variables lingüísticas más relevantes en los textos españoles peninsulares y la ulterior implementación de la metodología de la razón de verosimilitud.

Esta propuesta representa un paso adelante para las necesidades y los desafíos de investigación a los que la lingüística forense se ha enfrentado en el siglo XX –obtener pruebas lingüísticas forenses con la misma fiabilidad que otras disciplinas forenses. Ciertamente

en los últimos años, se ha abierto una nueva vía de investigación en la comparación forense de textos. Estudios anteriores sobre la aplicación de métodos multivariantes y probabilísticos como el marco de la razón de verosimilitud en otras ciencias forenses, por ejemplo, el análisis de ADN forense, criminalística, dactiloscopia forense, análisis forense digital o la comparación forense de habla han mostrado que la metodología de la razón de verosimilitud ha resultado ser razonablemente adecuada cuando se utiliza para evaluar la relevancia de la prueba científica en los tribunales.

En particular, los objetivos que se pretenden lograr con esta tesis doctoral son diseñar una propuesta metodológica para seleccionar variables lingüísticas con potencial discriminante; capturar una distribución poblacional de variables lingüísticas en español peninsular y establecer cuáles permiten diferenciar el sexo del autor; y, finalmente, situar un umbral numérico que permita discernir si una muestra anónima se puede atribuir o no a un autor mediante la implementación de la razón de verosimilitud.

2.1.1. Diseño de una propuesta metodológica

Como se ha expuesto anteriormente, el objetivo principal de la tesis es proponer y testar una nueva metodología para la selección y el análisis de las variables lingüísticas fiables en comparación forense de textos. Mediante el uso de la estadística clásica se puede identificar qué variables discriminan entre los autores y mediante

los métodos probabilísticos se puede evaluar su eficacia en una clasificación a posteriori.

2.1.2. Creación de una distribución poblacional

Esta tesis recoge una distribución de datos lingüísticos de textos en español peninsular con rasgos relevantes para la comparación forense de textos escritos de una población potencial de delincuentes. Esta distribución se desarrolla con el fin de poder evaluar el nivel de expectativa o de rareza de las variables en las muestras dubitadas e indubitadas. Uno puede pensar que crear grandes bases de datos puede ser una tarea agotadora, pero otras ciencias como el ADN o la comparación forense de habla ya han superado dicha tarea (Morrison, 2009: 306).

2.1.3. Implementación de la razón de verosimilitud

Esta tesis realiza una puesta en práctica del marco de la relación de verosimilitud para la comparación forense de textos que mejore la fiabilidad de las pruebas lingüísticas ante los tribunales proporcionando unos resultados probabilísticos no solo al juez, sino también al experto lingüista con el fin de llevar a cabo pruebas más rigurosas y análisis de rendimiento de los datos más extensos.

2.2. Premisas e hipótesis

Las premisas y los métodos relevantes en los que se basa esta tesis doctoral son los establecidos por las disciplinas de la lingüística forense, la sociolingüística variacionista y, la estadística multivariante clásica y bayesiana. Por lo tanto, el estudio se lleva a cabo bajo las siguientes hipótesis generales:

1) El uso idiosincrásico y único de la lengua que realiza un individuo permite la caracterización de sus rasgos socio-individuales y socio-colectivos.

2) El uso idiosincrásico y único de la lengua que realiza un individuo permite diferenciarlo de otros hablantes de la misma comunidad lingüística.

La hipótesis particular de esta investigación es que el marco de la razón de verosimilitud es capaz de capturar las diferencias lingüísticas de los hablantes y, de este modo, clasificar correctamente las muestras a su autor original y rechazar aquellas que pertenecen a otro autor.

2.3. Corpus

Una inquietud importante en el momento de diseñar el corpus de estudio fue la obtención de un corpus con prestigio en el ámbito forense. De este modo, el corpus recogido para la tesis doctoral es una simulación de la realidad forense: cartas de contenido amenazante de un número de autores y de muestras por autor, relativamente, reducido. A continuación, se detalla la metodología para la elaboración del corpus.

2.3.1. Informantes

En la fase de confección del corpus de este estudio ha sido esencial reproducir, lo más fielmente posible, las características de un sector de la población. Es por ello que se ha tenido en cuenta, como es habitual en los estudios de sociolingüística, que la muestra de esta tesis esté constituida por individuos de edad, formación, origen y residencia lo más similares posible.

En primer lugar, las variables demográficas. La clasificación de informantes atendiendo al factor edad se debe tener en cuenta dentro del contexto del ciclo de vida en que se encuentre el individuo tal y como propone Eckert (1997). En el caso que nos ocupa se han analizado individuos jóvenes, es decir, entre 18 y 30 años. Son numerosas las investigaciones sociolingüísticas entre el lenguaje y el sexo y, en concreto, las investigaciones llevadas a cabo para la identificación del sexo del individuo en perfiles

lingüísticos[19]. En esta tesis hemos tenido en cuenta el término sexo como un parámetro socio-demográfico independiente y, por tanto, hemos diferenciado entre hombres y mujeres.

Finalmente, es indispensable en todo estudio delimitar el origen geográfico de los informantes, en este caso se han escogido hablantes nacidos en la comunidad autónoma de Catalunya y cuyos progenitores también hayan nacido en Catalunya. El origen de los autores de los correos analizados y el de sus progenitores son variables controladas, ya que en los últimos años Catalunya ha sido destino de oleadas migratorias procedentes de diversos territorios. Según Miret, Salvador, Serracant y Soler (2008: 121) "en l'actualitat més d'una quarta part dels joves són nats a l'estranger i més d'un 30% són nats fora de Catalunya". Consecuentemente, para este estudio se han seleccionado solo los usuarios nacidos en Catalunya, con progenitores también nacidos en Catalunya y residentes en zona metropolitana y rural.

La clase social, aunque resulta siempre un categoría controvertida, se ha delimitado teniendo en cuenta el nivel educativo del individuo y de sus progenitores. El ámbito territorial se ha delimitado teniendo en cuenta el lugar de residencia (urbano/rural).

---

[19] Véase p.ej. Argamon, Koppel, Fine, & Shimoni, 2003; Argamon, Koppel, Pennebaker, & Schler, 2009; Rangel & Rosso, 2013a, 2013b.

Así pues, las variables sociolingüísticas que se han controlado para obtener una muestra lo más homogénea posible y representativa de los jóvenes catalanes son el sexo, la edad, la información demográfica, si están en una situación de contacto de lenguas, el historial lingüístico, el idioma utilizado en el contexto familiar, si han vivido en el extranjero y el nivel educativo del individuo y de sus progenitores. Las variables controladas se especifican concretamente en la Tabla 1.

Tabla 1. Variables sociolingüísticas controladas en la compilación del corpus.

Tipo de variable	Variable	Categorías de la variable
Variables demográficas	Grupo de edad	De 18 a 30 años
	Sexo	Hombres/Mujeres
	Lugar de nacimiento	Catalunya
Variables de origen social	Lugar de nacimiento del padre y la madre	Ambos en Catalunya
	Nivel de estudios más altos de los progenitores	Estudios obligatorios/Estudio universitarios
	Lengua inicial	Tanto catalán como español
	Nivel de estudios	Estudios universitarios
	Tipo de estudios	Traducción e Interpretación
	Lengua de escolarización	Catalán
Variables territoriales	Ámbito territorial	Ámbito urbano/Ámbito rural

En cuanto al historial lingüístico del individuo, se ha tenido en cuenta que la lengua inicial de los usuarios seleccionados fuera tanto el catalán como el español. Así, todos los candidatos se comunican en catalán y en castellano con la familia y los amigos. Además, se ha tenido en cuenta que los individuos fueran escolarizados en catalán.

2.3.2. Selección y representatividad

Por lo que concierne a la selección y representatividad se ha implementado el método más utilizado en las últimas décadas en los estudios sociolingüísticos, el método selectivo cualificado. Este método es reconocido por ser un muestreo de calidad y representativo de la comunidad investigada, ya que consigue anular el riesgo de selección absolutamente casual[20].

De este modo se ha podido satisfacer la necesidad de controlar las variables sociolingüísticas anteriores, ya que se escogen prototipos que se ajustan al perfil del subgrupo socio-demográfico de esta tesis.

Sobre el tamaño de la muestra, los estudios sociolingüísticos no muestran un consenso. Labov (1966) proponía trabajar con un porcentaje mínimo de garantía del 0,025 del universo de muestreo.

[20] Véanse Chambers, 1995: 38-41; Davis, 1990: 11; L. Milroy & Gordon, 2003: 30-33; L. Milroy, 1987: 28.

Gregory Guy (1980) establecía que cinco informantes representativos de la categoría socio-demográfica eran suficientes. En esta tesis, se han seguido las decisiones de Sankoff (1980): a) delimitar el grupo objeto de análisis para definir el universo de muestreo; b) considerar los factores de incidencia que puedan ser influyentes en la variación; y, finalmente, c) determinar el tamaño de la muestra.

Por lo que concierne a la cantidad de datos lingüísticos para poder asegurar la fiabilidad de los resultados, se ha asegurado el mínimo de 10 instancias (suficientes según Labov (1966), Exteberria, Joaristi, & Lizasoaín (1990: 278) y Camacho Rosales (2002: 285)) con 22 hombres y 25 mujeres. Desafortunadamente, no se ha podido conseguir el mínimo de 30 representativo según las leyes estadísticas generales (Guy, 1980a, 1993). No obstante, en todo momento se ha tenido presente el principio de responsabilidad (*principle of accountability*, en inglés) al que aludía Labov (1982: 30) y que define la honestidad del sociolingüista al mantenerse constante en la detección y recuento de ocurrencias en una muestra sin tener en cuenta si pueden confirmar o contradecir la hipótesis de partida.

2.3.3. Recopilación del corpus

Siguiendo el empirismo de la sociolingüística variacionista, la cual basa su teoría en los hechos lingüísticos, en este estudio se ha tenido como objetivo obtener datos científicos reales que permitan una

descripción completa y representativa de la comunidad objeto de estudio.

Para evitar la paradoja del observador, es decir, que el informante sienta que está siendo observado, nos hemos servido de las técnicas desarrolladas por la sociolingüística. En este caso la participación del investigador ha estado de mero observador de acuerdo con los procedimientos más habituales en la sociolingüística variacionista y la distinción que realiza Sevigny (1981): a) mero participante; b) participante como observador oculto; d) observador como participante; y, d) mero observador.

El cuestionario ha sido indirecto, ya que no ha sido presencial, y se ha recurrido un cuestionario a través de internet. De este modo, el procedimiento de recogida del corpus se ha llevado a cabo mediante un formulario durante 2013-2014 y siempre que se ha dudado de la representatividad del candidato en la muestra, éste se ha descartado. Además, para no influir en la actividad del usuario, no se informó al usuario sobre las características lingüísticas objeto de estudio. En todos los casos, sí se informó a los candidatos sobre los fines del estudio y se les pidió permiso para utilizar sus datos. Se anonimizaron los textos para garantizaran la privacidad de los usuarios.

Con el objetivo de reunir un corpus comparable con la realidad forense, se presentaron seis situaciones diferentes a todos los participantes –una cada semana– y debían escribir 4 mensajes de

amenaza con un nivel medio-alto de violencia (supuestos 1 a 4) y 2 correspondencias ordinarias (supuestos 5 y 6). Respecto a la longitud del texto, se pidió que los escritos tuvieran alrededor de 600 palabras a pesar de que todavía no se ha establecido un mínimo de palabras ni cuántos textos son necesarios (Stamatatos, 2009). Los estudios cualitativos pueden trabajar con textos más breves con un mínimo de unas 150-200 palabras mientras que los estudios cuantitativos que emplean programas informáticos pueden trabajar con textos más largos. Por ejemplo, Hirst y Feiguina (2007) trabajaban con escritos de 200, 500 y 1.000 y señalaban que la precisión de sus resultados decrecía significativamente cuando la longitud del texto era inferior a las 1.000 palabras. A pesar de no existir una longitud mínima probada, resulta evidente que difícilmente se podrán determinar rasgos del autor con un texto anónimo de dos líneas (unas 40 palabras aproximadamente).

Este procedimiento de recopilación del corpus tan exhaustivo permite realizar un proceso de homogeneización del corpus. A continuación se muestran las situaciones que se presentaron de forma literal:

a) PRIMER SUPUESTO

Hace un tiempo alquilaste una habitación individual por la cual tuviste que entregar una fianza de 3.000 euros. En este momento te cambias de piso y el antiguo propietario no quiere devolverte ni un euro de la fianza porque alega que has dejado las paredes de la

habitación en malas condiciones y, por tanto, tendrán que ser pintadas. Le has explicado de 100 maneras que eso es mentira, incluso, le enseñas fotos anteriores y posteriores para comparar el estado, pero él insiste en su mal estado y en no devolverte la fianza. Le propones diferentes opciones, como por ejemplo pintar tú las paredes, pero no hay solución. Finalmente, renuncias a una parte importante de la fianza correspondiente al importe de la pintura y de la mano de obra a pesar de que el piso está en perfectas condiciones porque crees que de este modo podrás recuperar una parte de la fianza. No hay resultado, el propietario sigue sin querer devolver ni un céntimo de la fianza. Finalmente, tu única opción es escribirle una carta en la cual le amenazas con el fin de obtener tu fianza.

b) SEGUNDO SUPUESTO

Hace tiempo que sospechas que tu pareja mantiene una relación con otra persona pero no tenías ninguna prueba hasta el día de hoy. Pero acabas de descubrir a través de los mensajes de móvil y de su red social que hoy van a encontrarse para mantener relaciones sexuales. Tú estás extremadamente enamorado/a y quieres mantener la relación con esta persona, así que decides no decirle nada a él/ella para que no piense que le aceptarías una infidelidad y decides enviarle una carta amenazante a la otra persona para que le deje sin decirle a tu pareja que tú se lo has dicho.

c) TERCER SUPUESTO

Vives en un bloque de pisos y tienes un vecino en el piso de al lado que pertenece a un grupo de heavy metal. Este vecino ha decido ensayar hasta altas horas de la mañana con su batería y a todo volumen. Se intentó negociar con él personalmente para que ensayara en otro lugar, con cascos o incluso a otras horas menos molestas puesto que el descanso es imposible, pero no hubo ninguna negociación. Se convocaron reuniones de vecinos para hablar de la situación e intentar resolver el tema pero no hubo éxito. Se ha llamado a la policía repetidas veces e incluso se han interpuesto demandas por el escándalo, pero este vecino sigue ensayando hasta las 4 de la mañana a todo volumen. Todo el bloque está sin poder dormir e incluso han habido propietarios que han decido alquilar un piso en otro sitio porque no podían descansar y estaban al borde de una depresión. La única solución que ves posible es escribirle una carta amenazándolo.

d) CUARTO SUPUESTO

Tu hermano/sobrino/primo pequeño te comenta que hay un hombre muy extraño que le sigue al salir de la escuela hasta llegar a casa cada día y que lleva dos días que se acerca a él ofreciéndole un caramelo pero que él lo rechaza porque siempre le habéis dicho que no se vaya con extraños. El mismo día que te comenta esto esperas al niño al salir de la escuela y el niño te lo señala disimuladamente y sin que el posible agresor se dé cuenta. Rápidamente lo reconoces y,

sabes que es el hijo de tu jefe. Dado que temes la reacción del agresor o incluso de tu jefe si se lo comentas y temes perder tu trabajo decides escribirle una carta amenazándolo para que deje de seguir al menor.

e) QUINTO SUPUESTO

En este último supuesto no se debe redactar una carta de amenaza sino de correspondencia. Queremos que retomes el primer supuesto (propietario que no quiere devolverte la fianza de 3.000 euros porque las paredes deben de ser pintadas) y que le escribas la carta anterior a la carta de amenaza en la cual negocias tu fianza para poder llegar a un buen acuerdo para los dos.

f) SEXTO SUPUESTO

Últimamente con la crisis se están llevando a cabo muchos recortes en la plantilla y los despidos están a la orden del día. En este sexto supuesto, queremos que escribas una carta de correspondencia a tu jefe (no amenazante) para que te tenga en cuenta en su plantilla y no te despida.

Después de la recogida de muestras del primer supuesto, se hizo una observación para comprobar la fiabilidad y la representatividad de la muestra tal y como aconseja López (1994: 128).

Finalmente, con el fin de organizar, clasificar y contextualizar el material obtenido cada informante debía rellenar unos datos personales que formarían una ficha personal. También se registraron datos de carácter técnico como la fecha, la marca de tiempo, nombre y apellidos, identificación y el correo electrónico del individuo. Estos datos son conocidos exclusivamente por la investigadora principal y los ficheros de datos para su posterior tratamiento estarán siempre anonimizados.

2.3.4. Distribución del corpus

Finalmente, se han obtenido dos corpus de textos escritos formados por 47 jóvenes. El corpus 1 está formado por individuos de ambos sexos y con dos muestras (supuesto 1 y 4) con el que se lleva a cabo la distribución poblacional de las variables lingüísticas. El corpus 2 está formado por individuos femeninos con 6 muestras y se utilizará en el paso final para la implementación de las relaciones de verosimilitudes. La distribución de los corpus por sexo y número de muestras se ilustra en la Tabla 2 y la Tabla 3.

Tabla 2. Distribución del corpus 1.

Sexo	N de individuos	N muestras por individuo	N total muestras grupo
Hombre	22	2	44
Mujer	25	2	50

Como queda reflejado en la Tabla 2, para alcanzar los objetivos del primer estudio, el corpus se divide en un 46,8% de individuos (22) de sexo masculino y un 53,5% restante (25) de sexo femenino. De cada escritor, se han analizado 2 cartas de contenido amenazante, por lo tanto, la muestra total es de 94 cartas.

Tabla 3. Distribución corpus 2.

Sexo	N de individuos	N Muestras por individuo	N Total muestras grupo
Mujer	18	6	108

Tal y como se puede observar en la Tabla 3, el corpus 2 para la obtención de las relaciones de verosimilitudes está compuesto por un 100% de mujeres (18) y con 6 cartas de cada una. De este modo, se dispone de una muestra total de 108 cartas.

2.4. Variables lingüísticas

Una variable lingüística es la representación de una característica lingüística que se puede expresar de diferentes maneras con el mismo significado. Cada autor tiene sus propias preferencias cuando expresa una idea, de acuerdo con el marco teórico explicado en el Capítulo 1, el conjunto de opciones o preferencias de un individuo constituye su estilo idiolectal.

Para llevar a cabo la identificación y selección de variables que reflejen más fielmente la posible distribución sociolingüística, las condiciones de variabilidad de sus distintas variantes y su potencial discriminante se ha recurrido a estudios previos en lingüística forense (véase Apartado 1.2 referente a los estudios previos en comparación forense de textos escritos). Además, también se ha tenido en cuenta el grado de conocimiento de la investigadora como nativa de la zona[21] como los resultados obtenidos en estudios y peritajes de casos forenses reales en comparación forense de textos escritos desarrollados por la investigadora en el Laboratorio de Lingüística Forense (ForensicLab) del Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra.

[21] P. Trudgill (1983a: 41) destaca que si los investigadores son nativos del área o gente familiarizada con el dialecto local es menos probable que las preconcepciones sean erróneas, y en el caso de serlas, probablemente se ajustarían más a la realidad.

Las variables lingüísticas de esta tesis deben poseer las siguientes características fundamentales: la variable debe ser muy frecuente y estratificada (Labov, 1972); debe mostrar una alta variabilidad inter-escritor y una baja variabilidad intra-escritor y también debe ser relativamente fácil de extraer y calcular (Nolan, 1983: 11); las variantes de la variable deben ser intercambiables en algunos contextos (Tagliamonte, 2006: 73); cada variable debe ser lo más independiente posible de otras variables (Rose, 2002: 52); la variable debe poseer una independencia relativa del control consciente, es decir, la variable idónea es aquella de la cual el escritor es menos consciente de su elección (Labov, 1963); las variables deben estar definidas en términos lingüísticos de manera clara y precisa (Britain, 1991: 56-63); y, finalmente, otro factor que se tiene en cuenta es la inestabilidad del rasgo, es decir, cuantas más variantes presente la variable más probable es que su uso indique la preferencia estilística del autor (Koppel, Akiva, & Dagan, 2007).

Dada la multitud de variables propuestas en los diferentes estudios en el campo de la comparación forense de textos escritos, muchos autores han propuesto diferentes agrupaciones, la más destacada es la división entre rasgos léxicos, sintácticos y estructurales (Abbasi & Chen, 2005; Juola, 2007, 2008; Koppel et al., 2009; McMenamin, 2001; Zheng et al., 2006). No obstante, no existe un consenso claro sobre qué variables contempla cada subdivisión y, es por este motivo, que en esta tesis se propone agrupar los rasgos en variables de complejidad, variables léxicas, pragmáticas y sintácticas. En

cuanto a las medidas de complejidad (Tabla 4), se analizan el número de palabras por documento, la riqueza de vocabulario calculando el número de palabras distintas, el número de oraciones y párrafos. Los promedios longitud de oraciones y párrafos y, finalmente, se tiene en cuenta la ratio type-token.

Tabla 4. Variables de complejidad analizadas.

Variables de complejidad
Número de palabras
Número de palabras distintas
Número de oraciones
Número de párrafos
Longitud media de palabras por oración
Longitud media de palabras por párrafo
Ratio type-token de palabras

En el análisis de léxico, se calculan frecuencias de palabras malsonantes y errores por muestra. También se consideran características como el uso de *ir a* + infinitivo o el tiempo futuro para expresar futuro, *deber* + infinitivo o *tener que* + infinitivo para expresar obligación o si el autor utiliza *como* o *si* para expresar condición. Las variables analizadas se encuentran en la Tabla 5.

Tabla 5. Variables léxicas analizadas.

Variables léxicas	
Número de errores	Ortografía
	Diacríticos
	Mayúsculas
	Puntuación
	Erratas
	Pleonasmos
	Gramática
	Contacto de lenguas
Expresión del futuro	Tiempo verbal de futuro
	Perífrasis verbal *ir a*
Expresión de la obligación	*Tener que* + infinitivo
	Deber + infinitivo
	Haber de + infinitivo
Expresión de la condición	*Si*
	Como
Expresión del pretérito imperfecto del subjuntivo	Terminación *-ra*
	Terminación -se
Abreviaturas	euros
	€
	EUR
Acortar palabras	
Número de palabras malsonantes	

En cuanto a la pragmática (Tabla 6), se calcula la distribución –la presencia o ausencia– de la primera persona del pronombre personal, por ejemplo, se tienen en cuenta las veces que aparece el *yo* con función de sujeto con el fin de identificar su intensificación. También se analizan las diferentes formas de expresar énfasis, como puede ser el uso de mayúsculas, la repetición o la puntuación.

También se calcula el número de exclamaciones e interrogaciones. La formalidad o informalidad de los pronombres personales constituye otra variable a evaluar. Se tienen en cuenta también los tipos de saludos y despedidas, ya que en estudios anteriores aparecen como posibles marcadores de autoría. Por último, también se evalúa el uso de paréntesis para interrumpir el discurso.

Tabla 6. Variables pragmáticas en el análisis.

Variables pragmáticas	
Intensificación del sujeto 1º persona del singular	Ausencia del *yo*
	Presencia del yo
Intensificación del sujeto 1º persona del plural	Ausencia
	Presencia
Expresión del énfasis	Uso de mayúsculas
	Uso de signos de puntuación
	Uso de la repetición
Trato	Formal
	Informal
Salutación	Buenos días/tardes
	Nombre:
	Hola
	Querido
	Estimado
	Señor
	Sr
	Apreciado
	BuenosXseñorX
	A…

Despedida	Un saludo
	Saludos cordiales
	Att
	Hasta nunca
	Atentamente
	Esperando pronto una respuesta
	Un cordial saludo
	Espero
	Saludos
	Gracias
	Muchas gracias
	Atentamente lo saluda
	Cordialmente
	Muchas gracias por su atención
	Muy cordialmente
	Reciba un cordial saludo
	Estamos en contacto
	Me despido atentamente
	Con todos mis respetos
	Que tenga un buen día
	Salutaciones
	Sin otro cometido, se despide atentamente
	Hasta luego
Marcadores discursivos	¿verdad?
	¿no?
	¿entiendes? ¿entendido?
	¿eh?
	¿sabes?
	¿no es cierto?
	¿no te parece?
	¿no es así?
	¿no crees?
	¿me sigue?

	¿sabes qué?
Signos de puntuación	Aparece apertura y cierre de los signos de exclamación
	Solo aparece cierre de los signos de exclamación
Uso de guiones	
Palabras entre paréntesis	
Número de preguntas	
Número de exclamaciones	

La sintaxis (Tabla 7) es analizada a través de la evaluación de los tipos de cláusulas que el autor utiliza, es decir, oraciones compuestas o simples, el tipo de oración compuesta –coordinadas, yuxtapuestas o subordinadas– y, del mismo modo, también se analiza el subtipo de cada tipo de oración compuesta yuxtapuesta, coordinada o subordinada.

Tabla 7. Variables sintácticas analizadas.

Variables sintácticas	
Tipo de oración	Simple
	Compleja
Tipo de oración compleja	Yuxtapuesta
	Coordinada
	Subordinada
Tipo de oración yuxtapuesta	Coma
	Dos puntos
	Punto y coma
Tipo de oración coordinada	*y/e*
	o/u
	ni
	pero

Tipo de oración subordinada	Relativas introducidas por *que*
	Relativas introducidas por *cual*

Finalmente, el grado de significación de los datos ha sido indicado por el análisis estadístico (Moreno Fernández, 1990: 69-71). A pesar de que las técnicas estadísticas solo son métodos para reflejar modelos de variación y que se acepta que no conseguir significación estadística no implica una irrelevancia sociolingüística (Milroy & Gordon, 2003: 168), en esta tesis se demuestra que es posible conseguir significación estadística mediante un buen tratamiento de los datos.

2.5. Metodología de análisis

2.5.1. Análisis cuantitativo y cualitativo

Una vez obtenidos los datos se procede al análisis de los mismos. Se pueden diferenciar dos tipos de análisis de datos principales: el análisis cualitativo (modelo hermenéutico) y el análisis cuantitativo (modelo positivista).

A grandes rasgos, se puede caracterizar el análisis cualitativo como aquél que ofrece una interpretación rica y detallada de los datos lingüísticos. Durante el análisis cualitativo, las características lingüísticas se identifican, pero no se cuentan frecuencias a las variables lingüísticas. En esta etapa es posible que se reconozcan las ambigüedades inherentes a la lengua, por ejemplo, la palabra *que* se puede encontrar en un corpus como un pronombre relativo o como una conjunción.

El análisis cuantitativo se puede caracterizar por ofrecer resultados de mayor fiabilidad ya que permite diferenciar los rasgos propios de la lengua de los que se deben solo al azar. Durante el análisis cuantitativo se clasifican y se cuentan las variables lingüísticas. Este análisis permite descubrir las características –su distribución y correlación– que identifiquen con una probabilidad alta al escritor y que, por tanto, representen su estilo idiolectal y reflejen su comportamiento de la lengua.

En el caso particular en el que una variable lingüística puede ser difícil de clasificar –por ejemplo, alguna palabra puede ser considerada palabra malsonante para algunas personas pero no para otras– se establecen estándares: se ofrece una lista con las palabras y expresiones consideradas malsonantes en este estudio en el Anexo I.

Ambos tipos de análisis han venido confrontándose durante años en el área de las ciencias sociales. Gummesson (1991: 178) resume las diferencias más significativas entre ambos paradigmas en la.

Tabla 8.

Tabla 8. Comparación entre el modelo positivista y hermenéutico. Fuente: Gummesson, 1991: 178.

Positivistic Paradigm	Hermeneutic Paradigm
Research concentrates on description and explanation	Research concentrates on understanding and interpretation
Vantage point is primarily deductive	Vantage point is primarily inductive
Research concentrates on generalization and abstraction	Research concentrates on the specific and concrete
Search for objectivity	Recognition of subjectivity
Statistical and mathematical techniques for quantitative data processing	Primarily non-quantitative data
Researchers take on the role of an external observer	Researchers want to experience what they are studying from the inside
Clear distinction between reason and feeling	Both feelings and reason govern actions

Actualmente, son muchos los autores que consideran que ambas metodologías no son excluyentes e insisten en que pueden llegar a ser complementarias[22]. Schmied (1993) señala que el análisis cualitativo es a menudo un precursor para el análisis cuantitativo, ya que antes de clasificar y cuantificar las variables lingüísticas, las categorías de clasificación deben ser primero identificadas.

En esta tesis doctoral se ha combinado, siendo el procedimiento más habitual, el protocolo de la metodología cuantitativa, popularizada en sociolingüística por William Labov, y la metodología cualitativa. En este estudio, se entiende que es imprescindible comenzar con un análisis cualitativo en una muestra pequeña de datos para establecer las categorías de clasificación, y después proceder con el análisis cuantitativo para poder extender el análisis a una muestra representativa y llevar a cabo análisis estadísticos.

Tal y como apunta Lavid (2005: 325):

> Es muy importante destacar el hecho de que el objetivo final de todo análisis cuantitativo no es la presentación de cifras y recuentos de rasgos lingüísticos. Éstos sólo son de utilidad cuando sirven de base para llevar a cabo interpretaciones funcionales de tipo cualitativo. En una palabra, los análisis

[22] Véanse Flick, 1998; Ortí, 1999 y Ruiz Olabuénaga, 1999.

> cuantitativos permiten explorar y llevar a cabo descubrimientos sobre los patrones de uso de la lengua de forma rigurosa y fiable, ya que permiten comprobar empíricamente las hipótesis sobre el uso de la lengua.

En resumen, el análisis cualitativo aporta mayor riqueza y precisión, mientras que el análisis cuantitativo permite obtener resultados estadísticamente fiables y generalizables (McEnery & Wilson, 2001: 77).

2.5.2. Diseño experimental del análisis estadístico

El diseño experimental de esta tesis doctoral se lleva a cabo de acuerdo a sus objetivos principales. En primer lugar, se lleva a cabo una prueba piloto con el fin de: a) validar el diseño del cuestionario y si éste debe ser modificado; b) observar el comportamiento de las variables definidas a priori y, por ejemplo, decidir si las variables son adecuadas y si se deben incluir más o menos variables; c) probar los métodos estadísticos por si deben ser ajustados para el análisis con el corpus completo; finalmente, d) determinar el tamaño del corpus necesario y las muestras de los participantes necesarios para poder extrapolar los resultados al conjunto de la población.

En segundo lugar, la fase de experimentos de validación se lleva a cabo en dos iteraciones con el fin de sobreponerse a las posibles limitaciones y ser capaz de implementar la metodología correctamente en la segunda iteración.

Seguidamente, se procede al análisis estadístico divido en dos etapas de acuerdo con los objetivos establecidos: una primera etapa de técnicas estadísticas multivariantes para la creación de la distribución poblacional y una segunda etapa de aplicación de la razón de verosimilitud. En la primera etapa, se aplican técnicas estadísticas multivariantes que se pueden definir como:

> Multivariate statistical analysis is the simultaneous statistical analysis of a collection of random variables. It is partly a straightforward extension of the analysis of a single variable, where we would calculate, for example, measures of variation, check violations of a particular distributional assumption, and detect possible outliers in the data. Multivariate analysis improves upon separate univariate analyses of each variable in a study because it incorporates information into the statistical analysis about the relationships between all the variables.
>
> Izenman (2008: 1)

Las técnicas multivariantes tienen como objetivo:

1. Describir las muestras con el fin de, por ejemplo, filtrar los datos, identificar valores atípicos, explorar y caracterizar las diferencias entre los grupos de casos, etc.
2. Evitar la dependencia o la redundancia en las variables, es decir, las variables pueden estar correlacionadas unas con otras, y ser la medición del mismo constructo. Por lo tanto, se trata de reducir el conjunto de variables y representar la mayor parte de la varianza en las variables observadas.
3. Ordenar los textos en grupos de manera que el grado de similitud entre el texto del mismo grupo sea máxima si pertenecen al mismo grupo y mínima en caso contrario.
4. Predecir la pertenencia a grupos en base a las similitudes de las características lingüísticas del texto.

Pero, cuantificar el grado de similitud no es suficiente. Como se ha expuesto anteriormente, también se debe tener en cuenta la rareza o la expectativa de la distribución y la correlación de esas características similares en comparación con la población pertinente. Esta comparación puede abordarse mediante la utilización de métodos probabilísticos como el marco de la razón de verosimilitud el cual lleva a cabo análisis empíricos rigurosos.

Por lo tanto, la segunda etapa de esta tesis consiste en la aplicación del marco de la razón de verosimilitud. Muchos investigadores y profesionales afirman que el marco de la razón de verosimilitud es muy adecuado para presentar pruebas en los tribunales, porque solo tiene en consideración el impacto de las pruebas analizadas por el experto y no tiene en consideración las creencias previas o posteriores del tribunal. Aitken et al. (2011: 1) manifiestan:

> To form an evaluative opinion from a set of observations, it is necessary for the forensic scientist to consider those observations in the light of propositions that represent the positions of the different participants in the legal process. The ratio of the probability of the observations given the prosecution proposition to the probability of the observations given the defence proposition, which is known as the likelihood ratio, provides the most appropriate foundation for assisting the court in establishing the weight that should be assigned to those observations.

2.5.2.1. Etapas del diseño experimental

El diseño experimental del análisis estadístico de esta tesis doctoral consta de 4 etapas principales –etapas A-D– resumidas a continuación.

Etapa A: Estudio piloto

A partir de los resultados obtenidos en el estudio piloto fue posible validar la metodología de obtención de datos, establecer el tamaño de la muestra y determinar las variables lingüísticas a analizar en esta tesis doctoral.

Etapa B: Experimentos de validación

a) Definición y medición de variables

En esta etapa se codifican las variables y sus variantes. Este proceso se lleva a cabo de forma manual excepto en el caso de las variables de complejidad que se extraen mediante un sistema automático diseñado *ad hoc* por el Dr. Jorge Vivaldi y se revisan de forma manual posteriormente.

b) Transformación y adaptación de las variables para su uso computacional

Durante esta fase, se debe cambiar el formato de algunas variables para su análisis computacional. Las transformaciones a que pueden ser sometidas pueden ser:

- Cambio a escala binaria tomando como referencia la mediana del parámetro lingüístico en la muestra (ausencia=0, aquellos valores menores o iguales a la mediana).

- Tipificación de variables continuas para la homogenización de la escala.
- Estimación de datos perdidos imputándolos por el valor del promedio del autor en la variable a estimar.

Etapa C: Creación de una distribución poblacional

Partimos de una situación en que no existen mediciones de referencia sobre una población definida y acotada. Por tanto, se presenta el trabajo desde una perspectiva del diseño de experimentos frecuentista. El procedimiento consiste en considerar una muestra dada como población de referencia y contrastar varios ensayos de cada individuo de la muestra contra observaciones propias y frente a otros individuos.

El primer paso después de la etapa de los experimentos de validación consiste en un análisis univariante de un conjunto de variables que muestre una Distribución Poblacional de las variables lingüísticas seleccionadas a partir de estadísticos básicos centrales y de dispersión. También es interesante contrastar cuales de ellas presentan mayor poder discriminante entre individuos. De forma complementaria pueden identificarse factores de variabilidad interesantes para posteriores aplicaciones de la técnica como hábitos y capacidades socio-lingüísticas y, otras de tipo de variables sociodemográficas como el sexo, que es el caso de esta tesis.

Etapa D: Aplicación de las técnicas bayesianas

El último paso del análisis consiste en el desarrollo de un modelo de cálculo de probabilidades que permita aproximar una medida de verosimilitud semejante a las utilizadas en los métodos de clasificación basados en la razón de verosimilitud sobre poblaciones conocidas.

La etapa D se desarrolla en seis pasos:

a) Análisis discriminante inicial para identificar las variables clasificadoras

Una vez que se completan los procesos anteriores, se lleva a cabo un análisis multivariante para establecer posibles marcadores de autoría y seleccionar las variables lingüísticas más discriminatorias. Este análisis discriminante se denomina "por bloques de variables", ya que se realiza un análisis por cada bloque de variables. En un segundo procedimiento estos análisis se validan con un análisis ANOVA[23] de las variables discriminantes sobre los 18 individuos que han realizado el ensayo completo produciendo seis textos comparables cada uno. Estos análisis se denominan "directos", ya

[23] El análisis de la varianza (ANOVA, *ANalysis Of VAriance*, según terminología inglesa) se utiliza para probar si las diferencias entre las medias de los grupos y sus procedimientos asociados (tales como la "variación" inter-grupos y intra-grupos) indican si la media de los grupos son o no iguales.

que se realizan sobre las variables discriminatorias seleccionadas en los análisis por bloques.

b) Análisis de proximidades intra- e inter- escritor

El objetivo de esta fase se centra en comprobar la distancia en que se encuentran cada una de las muestras, tanto las que son del mismo autor (variación intra-escritor) como la distancia a la que se encuentran las muestras propias de los demás autores (variación inter-escritor). Es de esperar que en la variación intra-escritor el número total de textos de un autor se clasifique en una sola agrupación. De manera similar, en la variación inter-escritor es de esperar que el conjunto de textos de cada autor constituya un grupo diferente.

El primer paso es calcular las matrices de distancias entre todos los elementos. Para la matriz de datos de expresión binaria, se utiliza el índice de concordancia simple, transformado a continuación en distancia (1-semejanza) y para la matriz de datos de respuesta discreta pero expresión numérica, se utiliza directamente la distancia euclídea.

A partir de estas medidas de distancia se crean dos nuevas bases de datos donde cada observación es representada por la media y desviación de las distancias de esa muestra a las cinco muestras restantes de su propio autor y a cada una de las 6 muestras de los 17 autores diferentes a él. Estas matrices de datos tienen 18x6 filas

correspondientes a las muestras y 18 variables de medias y 18 variables de desviaciones típicas.

c) *Cálculo de probabilidades a posteriori para clasificación con análisis discriminante*

Se han realizado los cuatro algoritmos de cálculo siguientes:

c.1 Análisis discriminante directo con variables binarias: realizado sobre la recodificación binaria de las variables seleccionadas en los análisis "discriminantes por bloques".

c.2 Análisis discriminante directo con variables continuas: realizado sobre las variables continuas seleccionadas en los análisis "discriminantes por bloques".

c.3 Análisis discriminante sobre medias y desviaciones de distancias calculadas sobre variables binarias: revisión de las clasificaciones c.1.

c.4 Análisis discriminante sobre medias y desviaciones de distancias calculadas sobre variables discretas: revisión de las clasificaciones c.2.

En los cuatro escenarios planteados se obtienen dos tipos de variables en las que se centran las evaluaciones finales: la probabilidad a posteriori de cada observación de pertenecer a cada uno de los autores y basada en esta probabilidad, una identificación del autor al que pertenece cada muestra.

Los análisis clasifican una muestra en aquel autor en el que la probabilidad calculada a posteriori sea mayor, es decir, que la probabilidad de clasificarse en ese autor es mayor que la probabilidad de clasificarse en cualquier otro. Por tanto, hipotéticamente, las probabilidades de clasificación en el propio autor deberían ser mayores que las probabilidades de clasificación en el resto de los autores.

Toda la información recogida en estos dos tipos de variables permite calcular las probabilidades de acierto y fallo en las clasificaciones, calculando y valorando la probabilidad exacta de buenas y malas clasificaciones e identificando los individuos que son indubitados de la muestra, fundamentalmente, porque no presentan zonas de solapamiento en el espacio de probabilidad de verdaderos positivos y falsos positivos.

d) Clasificación de las muestras

Para hacer un análisis de las probabilidades y clasificaciones obtenidas se tienen en cuenta los siguientes estadísticos:

- **Verdaderos positivos (VP)**: Recuento de muestras que se clasifican en el autor al que pertenecen. Casos posibles, 6.
- **Falso negativo (FN)**: Recuento de muestras que no se clasifican en el autor al que pertenecen. Casos posibles, 6.

- **Falso positivo (FP)**: Recuento de muestras que se clasifican en un autor dado sin pertenecer a él (siendo de otro autor). Casos posibles, 102.
- **Verdaderos negativos (VN)**: Recuento de muestras que no se clasifican en un autor dado no perteneciendo a él (siendo de otro autor). Casos posibles, 102.

A partir de estos recuentos se debe determinar la validez de las clasificaciones, es decir, en qué medida las clasificaciones obtenidas se ajustarían a procesos más complejos y rigurosos, para ello se tienen en cuenta los conceptos de:

- **Sensibilidad**: Es la probabilidad de detectar las muestras propias, es decir, de que una muestra se clasifique en el autor que le corresponde. Se mide como el cociente entre VP y la suma (VP+FN).
- **Especificidad**: Es la probabilidad de detectar muestras que no son propias, es decir, de que una muestra que no pertenece a un autor no se le adjudique como propia. Se mide como el cociente entre VN y la suma (VN+FP).

e) Razón de verosimilitud

Finalmente, se debe medir cuánto de probables son estas clasificaciones, es decir, la razón de verosimilitud o LR. En esta tesis doctoral se manejan dos conceptos:

- **Razón de verosimilitud positiva (LR+)**: es el cociente entre sensibilidad y la diferencia (1-especificidad), es decir, la probabilidad de que una muestra propia se le asigne a su autor en comparación a que una muestra que no es propia también se le asigne a dicho autor.
- **Razón de verosimilitud negativa (LR-)**: es el cociente entre la diferencia (1-sensibilidad) y la especificidad, es decir, la probabilidad de que una muestra propia no se le asigne a su autor en comparación a que las muestras ajenas se asignen al resto de los autores.

f) Solapamiento del espacio de probabilidad

La bondad de las clasificaciones se mide, no solo por el espacio de probabilidad con que se obtienen las clasificaciones de los verdaderos positivos, sino considerando el espacio de probabilidad en que se sitúan los falsos positivos, o también aquellas muestras que sin ser falsos positivos se acercan en probabilidad. Para validar conjuntamente las clasificaciones y las probabilidades se lleva a cabo una segunda comprobación basada en el solapamiento del espacio de probabilidad.

El solapamiento se mide en términos de intervalo, es decir, se considera como solapamiento el grado o nivel en que el intervalo de probabilidad del propio autor contiene al intervalo de probabilidad del resto de autores. De este modo, se calculan los intervalos de las probabilidades de clasificación:

- **Intervalo de probabilidad propio:** valores máximos y mínimos de las probabilidades del propio autor. Es decir, el mínimo y el máximo con el que una muestra propia se le atribuye a su autor.
- **Intervalo de probabilidad de cualquier otro grupo:** valores mínimos y máximos de las probabilidades de clasificar la muestra en cualquier otro autor.

Un autor queda mejor identificado cuanto mayor número de muestras correctamente identificadas tenga y cuanto menor sea el grado de solapamiento de sus intervalos de probabilidad.

2.5.3. Métodos estadísticos aplicados

A continuación, se presentan los métodos estadísticos implementados para la consecución de los distintos objetivos de esta tesis. En primer lugar, se presentan los métodos para el estudio de la distribución poblacional de las variables (objetivo 2.1.2) consistente en la creación de una distribución poblacional de las variables lingüísticas que permita implementar, en segundo lugar, el marco de la razón de la verosimilitud (objetivo 2.1.3) en un caso real de comparación forense de textos escritos. Dicha implementación permitirá la evaluación de la similitud y la tipicidad de las muestras dubitadas e indubitadas.

Dentro de la estadística se pueden diferenciar dos grandes ramas: la estadística descriptiva y la inferencial. La estadística descriptiva proporciona las características del grupo de individuos que se está analizando, es decir, da cuenta de la forma que adopta la distribución de determinados rasgos –variables lingüísticas– de la muestra estudiada. Este tipo de análisis se ha llevado a cabo en la construcción de la distribución poblacional de las variables (Apartado 3.1). La estadística inferencial permite hacer predicciones sobre los datos y permite proyectar características de una muestra sobre una población. Este tipo de estadística se ha utilizado en el resto de análisis de esta tesis doctoral (Apartados 3.2 y Capítulo 4).

2.5.3.1. Métodos para la distribución poblacional

Las primeras pruebas inferenciales se han realizado para determinar, en un espacio univariante, el efecto de algunas características poblacionales, como el sexo. Todas las variables cuantitativas utilizadas son o bien de carácter binario, ordinal o continuas de intervalo.

En primer lugar se ha constatado la falta de normalidad de las variables con sucesivos tests de Kolmogorov y de Shapiro y Wilks. Hay una gran presencia de datos extremos en las muestras. Para resolver el problema de la falta de normalidad en los siguientes análisis que se utilizan en este estudio, se han calculado las variables denominadas de estimación donde se controla el peso de los extremos y también variables recodificadas en binarias (ausencia/presencia de un suceso). Es decir cada variable puede tener su expresión observada, estimada y recodificada.

Para el contraste del factor sexo con las categóricas (originales o recodificadas) se ha utilizado el test de Chi-Cuadrado de Pearson[24]. Para las variables cuantitativas se ha hecho un estudio previo de

[24] Esta técnica se ha utilizado en múltiples casos de comparación forense de textos escritos como, por ejemplo, Svartvik (1968), Dreher & Young (1969) o Smith (1994).

normalidad, utilizado el test de contraste bilateral de medias cuando ha sido posible (T-test) y test no paramétricos para dos muestras independientes (Test Mann-Whitney). Las tablas de contrastes se muestran en medias, incluso cuando la variable es binaria, pero siempre se ha aplicado el test de contraste adecuado para cada caso. Es de señalar que los resultados para el T-test y los no paramétricos han sido prácticamente siempre coincidentes o de probabilidad muy cercana, con lo que la constatada falta de distribución normal no ha supuesto diferencias en los resultados.

La significación considerada ha sido del 95% en todos los casos.

2.5.3.2. Métodos para los modelos de clasificación

- **Análisis discriminante**

El análisis discriminante (LDA, *Linear Discriminant Analysis*, en inglés) es una técnica estadística multivariante con tres objetivos principales:

1) Describir si las variables introducidas en el análisis muestran diferencias entre individuos estadísticamente significativas.

2) Determinar cuáles son las variables con un potencial discriminatorio mayor entre los individuos.

3) Predecir el grupo de una muestra dubitada. Por ejemplo, en el caso de un texto anónimo determinar cuál es el autor más probable de haber producido ese texto.

En esta tesis estos tipos de análisis se aplican en dos ocasiones comentadas ya en la etapa D del apartado diseño de experimentos y que se detallan a continuación.

El apartado "a" de dicha etapa D, tiene por objetivo encontrar las variables con mayor poder discriminante de cada uno de los bloques que forman parte de nuestro estudio. Estos análisis llamados *discriminantes por bloque* deben perseguir el objetivo 2, es decir, para cada bloque de variables, se ejecuta sobre todos los casos un análisis discriminante para determinar qué variables de cada bloque muestran un potencial discriminatorio mayor para diferenciar individuos. Son análisis discriminantes de inclusión por pasos, es decir, la variables se incorporan a la función discriminante tras evaluar en qué grado contribuyen de manera individual a la discriminación de los grupos.

El estadístico utilizado para la selección de las variables es "Lambda de Wilks", este estadístico selecciona, para incorporar al modelo, aquellas variables que al incorporarse producen un mayor cambio en el valor del estadístico Lambda. Toma valores entre 0 y 1 y cuanto más se aproxime a 0 mayor es el poder discriminante de la variable. Este cambio se evalúa a través del estadístico F de cambio, que se formula en la Fórmula 1:

$$F_{cambio} = \left(\frac{n-g-p}{g-1}\right)\left(\frac{1-\lambda_{p+1}/\lambda_p}{\lambda_{p+1}}\right)$$

Fórmula 1. Estadístico F de cambio.

Donde *n* es el número de casos válidos, *g* es el número de grupos, λp es la lambda de Wilks que corresponde al modelo antes de incluir la variable que se está evaluando y $\lambda p+1$ es la lambda de Wilks que corresponde al modelo después de incluir esa variable.

Las variables que muestran el mayor potencial son las variables candidatas para la aplicación *a posteriori* de las relaciones de verosimilitudes.

En el apartado c de la etapa D del diseño, se mencionan otros análisis discriminantes –los que dan lugar a los 4 abordajes del estudio–, estos análisis discriminantes no son por pasos, sino que se introducen todas las variables independientes juntas y con ellos se pretende lograr el objetivo 3. En un segundo grupo de análisis discriminantes, realizados esta vez sobre casos completos, se obtienen dos análisis llamados "directos", en ellos las variables independientes son las variables discriminantes seleccionadas en cada uno de los bloques. Finalmente, se llevan a cabo otros dos análisis llamados *a partir de distancias* ya que, se utilizaran como variables independientes las medias y las desviaciones típicas

obtenidas de las matrices de distancias medias de los análisis cluster que se especifican en el punto 0.

En ambos tipos de análisis discriminante, el objetivo es medir las proximidades intra- e inter- escritor para determinar si las distancias son diferentes entre las muestras de un autor y entre los autores. Para ello se utilizan las variables de clasificación y probabilidad a posteriori que ofrecen los análisis.

- **Análisis de proximidades intra- e inter- escritor**

El objetivo del análisis de proximidades intra- e inter-escritor se centra en comprobar la distancia en que se encuentran cada una de las muestras tanto de ellas mismas, es decir, de las que pertenecen al mismo individuo, como de las demás y, de este modo, poder llevar a cabo un ensayo de predicción posterior.

Para alcanzar este objetivo es necesario obtener una matriz de similitudes o de distancias a través de un análisis cluster jerárquico de las variables discriminantes. A dicho análisis se le solicita la matriz de similaridades de concordancia simple que se trata de la razón de coincidencias respecto al número total de valores. Se ofrece una ponderación igual a las coincidencias y a las no coincidencias. A continuación, se seleccionan solo los casos completos, es decir, aquellos autores que tienen 6 muestras. Se calculan matrices de distancias entre las 108x108 observaciones,

provenientes de 18 autores y 6 muestras por autor. Este análisis se realiza mediante una aproximación binaria y otra discreta.

- **Aproximación binaria**

Para este análisis se homogeneíza la muestra mediante variables 0-1. Por tanto, es necesario recodificar las muestras para obtener variables 0/1 a partir de la mediana de las variables estimadas. De tal forma que aquellos valores menores o iguales a la mediana toman el valor 0 y aquellos valores superiores a la mediana toman valor 1.

En el análisis de conglomerados o (en inglés, *cluster*) en que se analizan las variables binarias, la matriz que se solicita al análisis es una *matriz de similaridades de concordancia simple*. Esta matriz permite evaluar el grado de parecido o proximidad existente entre dos elementos concediendo mayor importancia a las casillas de concordancia, es decir, las casillas donde coinciden la presencia (*a*, frecuencia de unos) y ausencia (*b*, frecuencia de ceros) de las variables analizadas.

La concordancia o emparejamiento simple se trata de la razón de coincidencias (en 0 y en 1) respecto al número total de valores. Como se observa en la Fórmula 2, se ofrece una ponderación igual a las coincidencias y a las no coincidencias.

$$SM(x,y) = \frac{a + b}{n}$$

Fórmula 2. Matriz de similaridades.

Se considera que dos elementos son más similares entre sí cuanto mayor número de presencias o ausencias comparten, por tanto, los valores más altos indican mayor parecido o proximidad entre los elementos comparados. El valor de *SM(x,y)* es máximo cuando dos elementos se encuentran juntos.

En esta tesis doctoral se obtiene una matriz de concordancias simples de tamaño 108x108 que se debe transformar en una matriz de distancias aplicando la Fórmula 3:

MATRIZ DE DISTANCIAS= 1-MATRIZ CONCORDANCIAS SIMPLES

Fórmula 3. Conversión de matriz de concordancias a matriz de distancias

○ Aproximación numérica discreta

Para confirmar el funcionamiento del método se repite el proceso pero considerando las variables como numéricas (sin recodificar por la mediana), dado que todas las variables excepto ‘cordialmente’ son continuas.

El objetivo es el mismo, comprobar la distancia a la que se encuentra cada una de las muestras tanto de ellas mismas, es decir de las que pertenecen al mismo individuo, como de las demás. En este caso al ser variables a través de un análisis cluster jerárquico se debe obtener una matriz de distancias euclídeas (longitud del segmento lineal que une dos elementos) al cuadrado sobre las variables estandarizadas.

El cálculo de proximidades –distancias euclídeas– se lleva a cabo sobre variables continuas. La distancia euclídea es, quizá, la medida de disimilaridad más conocida y que se utiliza por defecto para datos de intervalo. Su cálculo consiste, como se muestra en la Fórmula 4, en la raíz cuadrada de la suma de los cuadrados de las diferencias entre los valores de las variables.

$$EUCLID(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Fórmula 4. Cálculo distancia euclídea.

En este caso se obtiene de nuevo una matriz de tamaños 108x108, que es directamente una matriz de distancias y, por tanto, no se debe realizar ninguna transformación.

o Análisis de las matrices de distancias

Para ambas aproximaciones –binaria y numérica– se realizan distintas pruebas. Para las medias se realiza un contraste de medias y un análisis ANOVA de un factor para cada una de las variables Med con el fin de determinar si estas distancias medias son diferentes entre sí y entre los individuos. Además se realiza también un análisis discriminante sobre las variables Med y SD, con el objetivo de ver si los individuos se clasifican correctamente (es decir si Med1 y SD1 clasifican a las 6 muestras de C101 dentro del grupo 1) y la probabilidad con la que lo hacen.

En concreto, a partir las matrices obtenidas en ambas aproximaciones –binaria y continua– se calculan las distancias medias y sus desviaciones típicas asociadas para cada una de las muestras analizadas, dando lugar a matrices de distancias medias, de dimensión 108x36 (18 variables de distancias medias y 18 variables de desviaciones típicas asociadas). Sobre estas matrices se realizan los análisis discriminantes a partir de distancias. Las variables de distancias medias se etiquetan como MED (de 1 a 18) y miden la distancia media de las muestras de cada autor a las que le son propias y a las de los demás autores. Las variables de desviaciones típicas se etiquetan como SD (de 1 a 18) y recogen las desviaciones típicas asociadas a las variables MED.

Para cada una de las variables MED se realiza un contraste de medias y un análisis ANOVA de un factor para establecer si estas distancias medias son diferentes entre sí y entre los individuos.

Además, como ya se ha comentado, se realiza también un análisis discriminante sobre cada par de variables MED y SD. El objetivo es observar si todas las muestras de un mismo autor se clasifican correctamente, es decir, si Med1 y SD1 clasifican a las 6 muestras de C101 dentro del grupo 1 y la probabilidad con la que lo hacen.

El utilizar ambos estadísticos lleva a unir elementos, no solo que están a distancias similares del resto en medias sino también en desviación. El propósito es que se clasifiquen con más fuerza los seis textos del mismo autor.

- **Análisis de la razón de verosimilitud**

En esta tesis, la hipótesis de partida (H_0 o hipótesis nula), es que la muestra dubitada y la muestra indubitada han estado producidas por el mismo autor, y la hipótesis alternativa (H_1) es que ambas muestras han estado producidas por diferentes autores. Esto se expresa en la Fórmula 5:

$$LR = \frac{p(\text{E}|\text{Hipótesis mismo escritor})}{p(\text{E}|\text{Hipótesis diferente escritor})}$$

Fórmula 5. Cálculo razón de verosimilitud.

LR es la razón de verosimilitud, es un cociente de probabilidades. En el numerador la probabilidad de que *E*, la evidencia, es decir, las propiedades lingüísticas de la muestra indubitada, en nuestro caso los verdaderos positivos o muestras propias del autor se le atribuyan a su autor. En el denominador la probabilidad de que *E* sea atribuida a otro autor. Esto permite contrastar si las propiedades lingüísticas del texto de la muestra dubitada, se le atribuyen al mismo autor que *E*, aceptándose la hipótesis nula, y rechazar la hipótesis de atribuírsela a un autor diferente, es decir, hipótesis alternativa.

En esta tesis se genera un ratio a partir de las probabilidades de clasificación, nombradas LR, dividiendo las probabilidades por el máximo de las probabilidades de las muestras correctamente clasificadas.

El siguiente paso es calcular las LR para cada individuo y cada una de las variables. En particular, en esta tesis existen dos maneras de medir la razón de verosimilitud, de manera positiva (LR+) expresada en la Fórmula 6; y, de manera negativa (LR-) expresada en la Fórmula 7.

$$LR+= \frac{Sensibilidad}{1-Especificidad} = \frac{\frac{VP}{(VP+FN)}}{1-\left(\frac{VN}{(VN+FP)}\right)}$$

Fórmula 6. Razón de verosimilitud positiva.

$$LR-= \frac{1-Sensibilidad}{Especificidad} = \frac{1-\left(\frac{VP}{(VP+FN)}\right)}{\left(\frac{VN}{(VN+FP)}\right)}$$

Fórmula 7. Razón de verosimilitud negativa.

LR+ varía entre cero e infinito (este valor se denota como >1000), cuanto mayor sea su valor, mayor es la probabilidad de acertar a la hora de clasificar correctamente la muestra dubitada. LR- varía entre 0 y 1, cuanto menor sea su valor, mayor será la probabilidad de clasificar correctamente la muestra dubitada.

Para que la muestra dubitada sea asignada al autor que le corresponde se deben dar las dos condiciones, tanto que el LR+ sea el mayor posible, como que el LR- sea el menor posible.

Capítulo 3:

Resultados sobre la distribución poblacional

En el capítulo 3 se muestran los resultados sobre la distribución poblacional. En iniciar el capítulo, se muestran los resultados de cada uno de los tres estudios. En primer lugar, los resultados generales sobre la distribución poblacional de las variables lingüísticas objeto de estudio (Apartado 3.1). En segundo lugar, se exponen los resultados sobre las variables que muestran una distribución poblacional diferenciada por el sexo del individuo con el fin de contribuir a la elaboración de perfiles lingüísticos (Apartado 3.2).

3. Resultados sobre la distribución poblacional

Los resultados de la distribución poblacional se dividen en dos subapartados de acuerdo con el segundo objetivo de esta tesis: 2.1.2. Creación de una distribución poblacional. En el Apartado 3.1. se crea la distribución poblacional de las variables lingüísticas divididas en los cuatro bloques de variables que se han propuesto en esta tesis: variables de complejidad, léxicas, pragmáticas y sintácticas.

En el Apartado 3.2, puesto que la muestra del estudio está dividida por sexo, se presentan los resultados de las pruebas estadísticas que muestran diferencias estadísticamente significativas entre ambos sexos. Estos resultados son de especial relevancia para los casos en que se debe realizar un perfil lingüístico para determinar el posible sexo del autor del texto dubitado.

3.1. Distribución poblacional de las variables

La distribución poblacional de las variables lingüística distribuidas por tipo se muestra en el Anexo II mediante tablas resumen de los estadísticos descriptivos principales, como son, la media (M), la desviación estándar (SD, del inglés *Standard Deviation*), los puntos mínimos y máximos, la mediana y la moda.

3.1.1. Variables de complejidad

El resultado de la distribución poblacional de las variables de complejidad del corpus de esta tesis se muestra de forma visual en el Gráfico 1. En este gráfico de cajas (*boxplot*, en inglés) se representa la distribución de cada variable teniendo en cuenta la mediana, los percentiles, los valores extremos y la existencia de observaciones atípicas y/o extremas.

La parte inferior de la caja indica los resultados que se encuentran en el percentil 25%, la línea señala el percentil 50% correspondiente a la mediana de la variable y la parte superior de la caja engloba los valores situados en el percentil 75%. La caja se ubica sobre un segmento (bigote) que tiene como extremos los valores mínimo y máximo de la variable. El extremo superior del bigote indicaría la frecuencia máxima de la variable y el extremo inferior la frecuencia mínima.

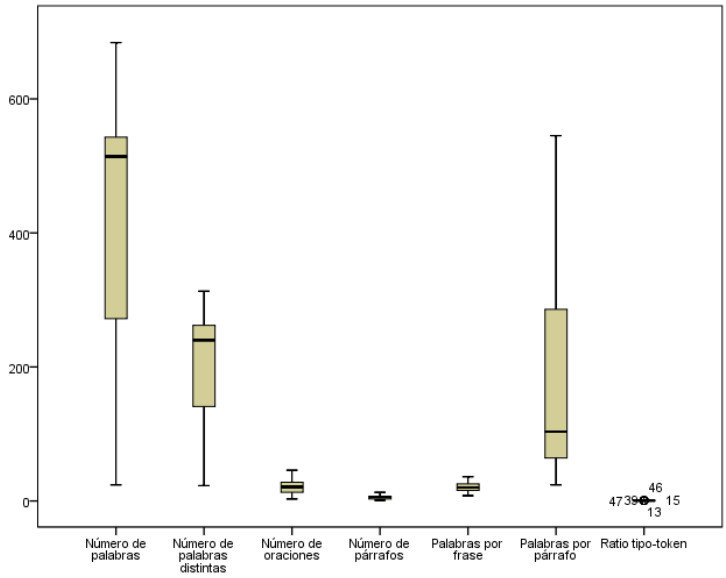

Gráfico 1. Distribución poblacional de las variables de complejidad.

A continuación, se destacan las variables que resultan más interesantes para tener en cuenta en la distribución poblacional puesto que no vienen delimitadas por las peticiones del investigador.

En primer lugar, se observa que la ratio type-token representativa de la riqueza léxica del escritor es de 0,55 de media, es decir, la mitad de las palabras no se repiten. Un valor bajo en esta variable indicaría un número alto de repeticiones dentro del texto y, podría indicar que el texto es menos rico desde el punto de vista del vocabulario. Las causas deberían de estudiarse en cada caso ya que, por ejemplo, podría deberse al nivel de estudios del escritor o al grado de especialización del texto.

Resulta muy interesante también, dado su uso en casos reales de peritajes lingüísticos, observar la longitud de la frase teniendo en cuenta el número de palabras por frase, la media poblacional es de 20,21; así como, la longitud de los párrafos definida por el número de palabras por párrafo tiene una media de 180,09 palabras.

Estos datos resultan de especial utilidad en futuros casos de atribución de autoría para destacar la relevancia de los resultados encontrados en ambas muestras. Unos resultados superiores o inferiores a estas medias podrían apuntar a esta variable como un rasgo idiosincrático del escritor.

3.1.2. Variables léxicas

En primer lugar, por lo que respecta al número de palabras malsonantes se observa que dicha variable muestra una media de 1 palabra malsonante por cada 400 palabras con una desviación estándar de 2,79 y con un mínimo de 0 y un máximo de 20 palabras. De este modo, se puede concluir que los resultados indican que los escritores con el perfil de este corpus –estudios universitarios, bilingües– tienden a utilizar un lenguaje no malsonante a pesar de estar escribiendo textos amenazantes.

En relación con el número de errores que produce el hablante (Gráfico 2) se observa que los errores más frecuentes son los errores en diacríticos (40,34%) seguidos por los errores de ortografía (16,34%), los errores gramaticales (12,84%), los errores producidos por el contacto de lenguas (12,32%) y las erratas (9,06%). Son menos frecuentes los errores de puntuación (5,19%), los errores en el uso de mayúsculas (3,11%) y, finalmente, los pleonasmos (0,78%).

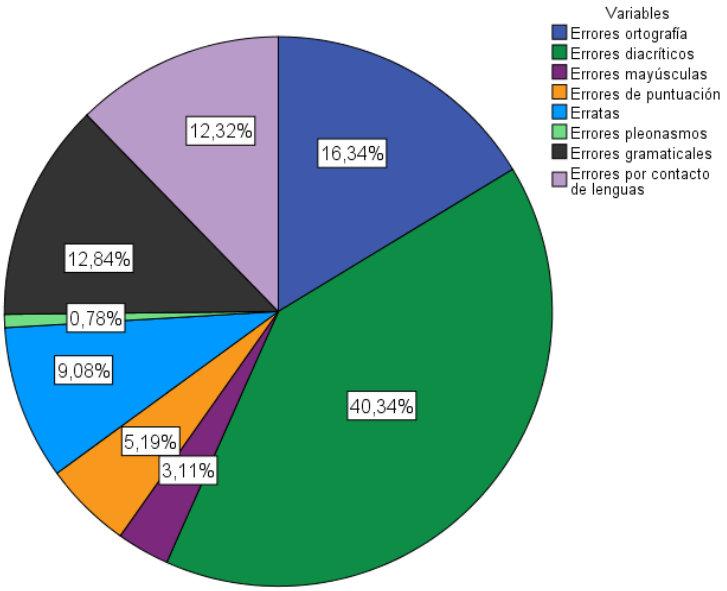

Gráfico 2. Distribución poblacional de la variable léxica número de errores.

En relación con la distribución poblacional de esta variable en la muestra de esta tesis (Gráfico 3), resulta destacable observar el comportamiento individual de ciertos escritores. Por ejemplo, se observa que el escritor 44 suele cometer un número significativamente superior de errores que la media poblacional, por ese motivo, en el gráfico se encuentra en los valores extremos[25] en el caso de errores de ortografía, diacríticos, pleonasmos y gramaticales. Otros casos que resultan de especial interés son

[25] Los valores extremos se indican con un asterisco y los valores atípicos con un círculo.

aquellos en que el escritor suele cometer un mayor número de errores pero de un solo tipo, por ejemplo, el escritor 35 muestra un número destacable de errores por contacto de lenguas en ambas muestras o en el caso del escritor 41 se presentan dificultades de puntuación.

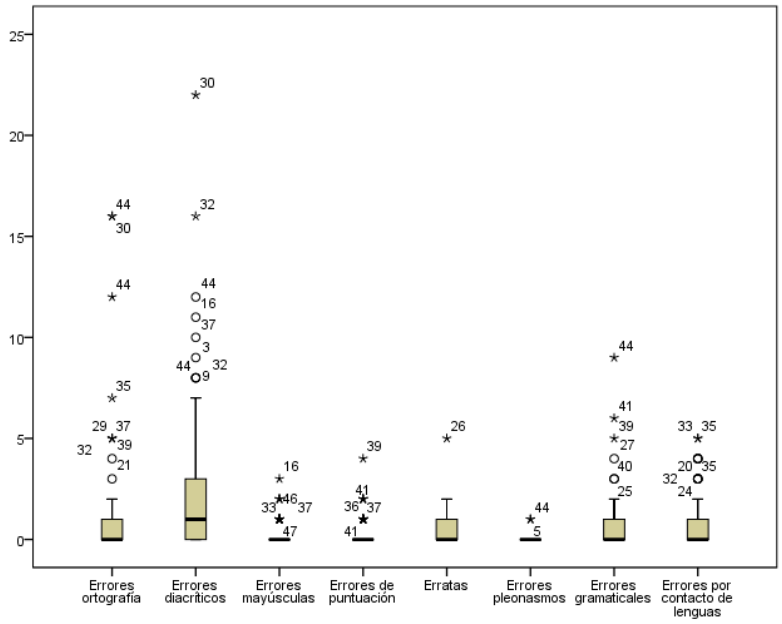

Gráfico 3. Distribución individual de la variable léxica errores.

Los hablantes, ante la posibilidad de expresar el futuro mediante el tiempo verbal de futuro o mediante la perífrasis verbal *ir a* + infinitivo, muestran preferencia por el tiempo verbal de futuro (60%) siendo esta la forma menos marcada.

En español la obligación (Gráfico 4) también se puede expresar de distintos modos: *tener que* + infinitivo, *deber* + infinitivo y *haber de* + infinitivo. Los resultados muestran que la forma más común en este corpus es *tener que* + infinitivo (57,28%), seguida de la forma *deber* + infinitivo (39,44%), y, en menor medida, se utilizaría la forma *haber de* + infinitivo (3,29%).

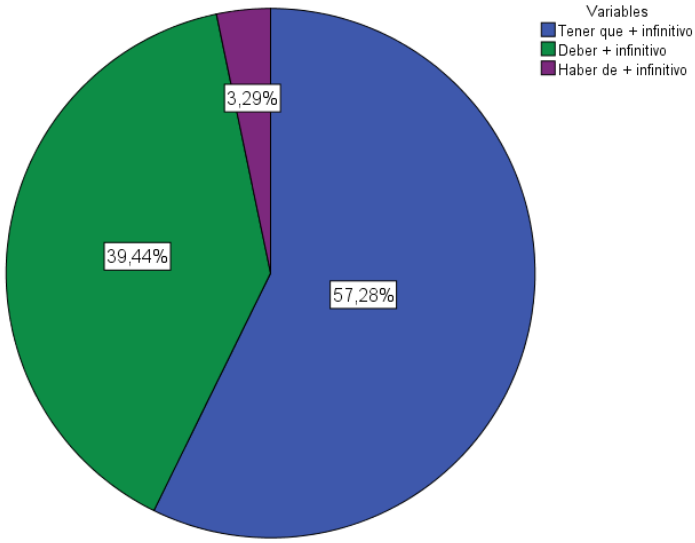

Gráfico 4. Distribución poblacional de la variable léxica expresión de la obligación.

En el Gráfico 5 se proyecta la distribución poblacional de la variable expresión de la obligación en español de esta muestra. Resulta de especial relevancia observar los casos de los escritores 32 y 28 que tienden a utilizar con mayor frecuencia *deber* y *haber de* + infinitivo, respectivamente.

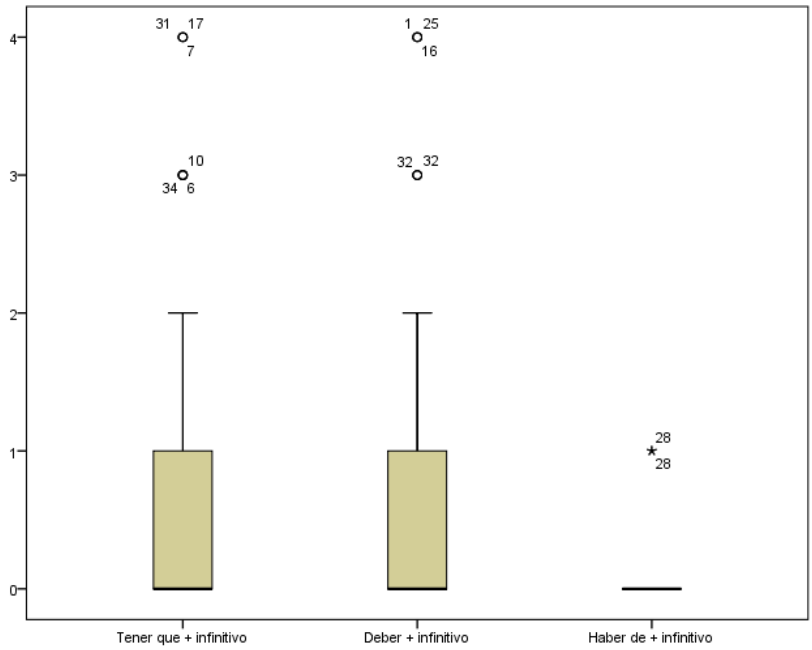

Gráfico 5. Distribución individual de la variable léxica expresión de la obligación.

Por lo que respecta a la expresión de la condición, los escritores muestran preferencia por introducir la condición mediante la conjunción *si* en el 90,67% de los casos frente la conjunción *como* seguida de subjuntivo que solo es utilizada en el 9,33% de las ocasiones.

El pretérito imperfecto de subjuntivo en español se puede utilizar con la terminación *–era* o con la terminación *–ese*, en este caso los hablantes utilizan más frecuentemente la terminación *–era* (87,57%) frente la terminación *–ese* (12,46%).

En los escritos no se ha observado un número abundante de abreviaturas. No obstante, se ha considerado relevante estudiar la abreviatura del sustantivo *euros* (Gráfico 6) por su frecuencia en este corpus y, también, por su frecuencia en cartas de extorsión en casos reales. En este caso los resultados muestran que la forma más común de escribir *euros* es en su forma no abreviada (64,56%), seguido de € (33,33%) y, finalmente, la abreviatura EUR (2,08%).

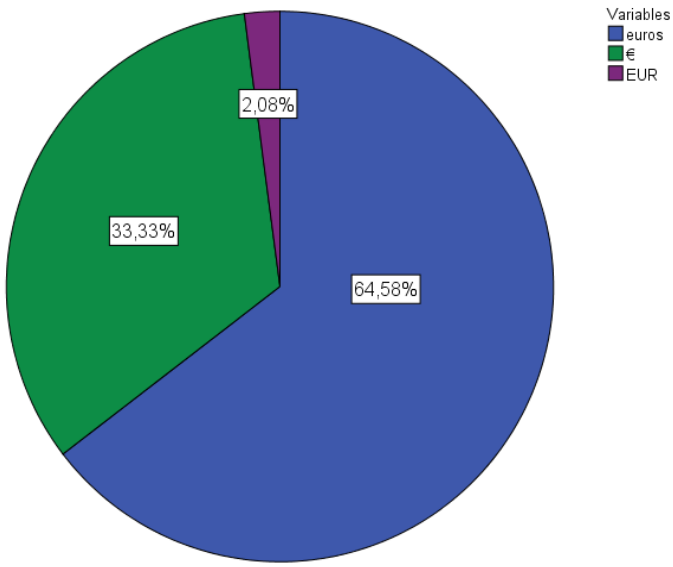

Gráfico 6. Distribución poblacional de la variable léxica uso de abreviaturas.

Finalmente, los escritores tienden a escribir por completo las palabras en sus escritos y, por este motivo, la media de palabras acortadas por documento se reduce a 0,1 por documento de 400 palabras aproximadamente.

3.1.3. Variables pragmáticas

Una de las variables que ha resultado muy relevante en los casos de autoría realizados en España, ha estado la variable intensificación del sujeto de primera persona del singular. Los resultados muestran que en español es menos frecuente atenuar el sujeto de primera persona mediante su ausencia en el discurso. Concretamente, la ausencia del *yo* representa el 91,75% de los casos.

Del mismo modo que la variable intensificación del sujeto de primera persona, también resulta muy relevante la intensificación del sujeto de primera persona del plural. El 93,68% de los autores tienden a no escribir el sujeto.

La forma en que un hablante expresa el énfasis también puede ser diversa. En este estudio se ha analizado la expresión del énfasis mediante el uso de mayúsculas, el uso de signos de puntuación y el uso de la repetición. Los resultados (Gráfico 7) muestran que la forma más común de expresar el énfasis en texto escrito es mediante el uso de mayúsculas (70,48%), en segundo lugar mediante el uso de signos de puntuación (19,05%) y, finalmente,

mediante el uso de la repetición de expresiones o palabras (10,48%).

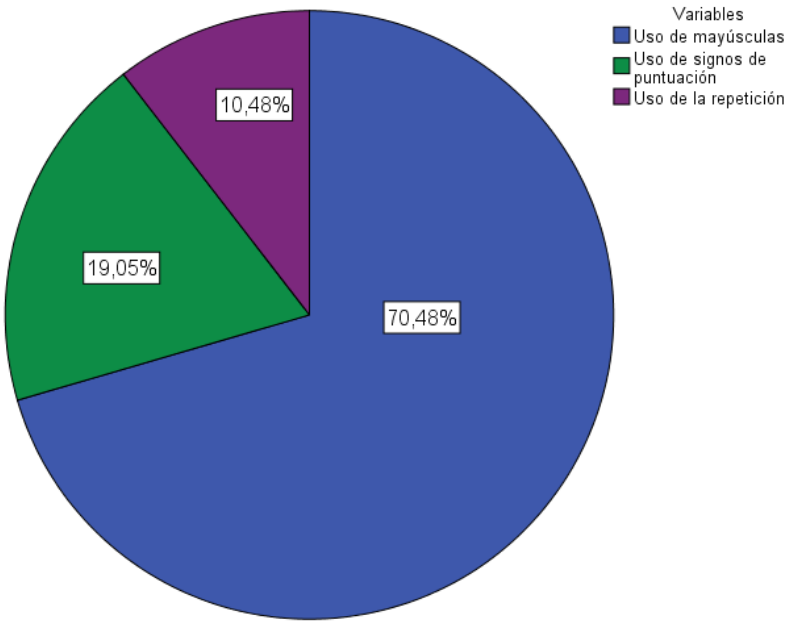

Gráfico 7. Distribución poblacional de la variable pragmática expresión del énfasis.

El número de preguntas que realiza un escritor en un documento suele ser superior al número de exclamaciones, una media de 1,38 y 0,59 por documento, respectivamente. En el caso del número de preguntas, resulta interesante observar el Gráfico 8, sobre la distribución individual, el caso de los autores 16 y 11 puesto que ambos autores tienen un número de realizaciones superior a la media poblacional –por tanto, una variación inter-escritor muy alta– y una variación intra-escritor muy baja.

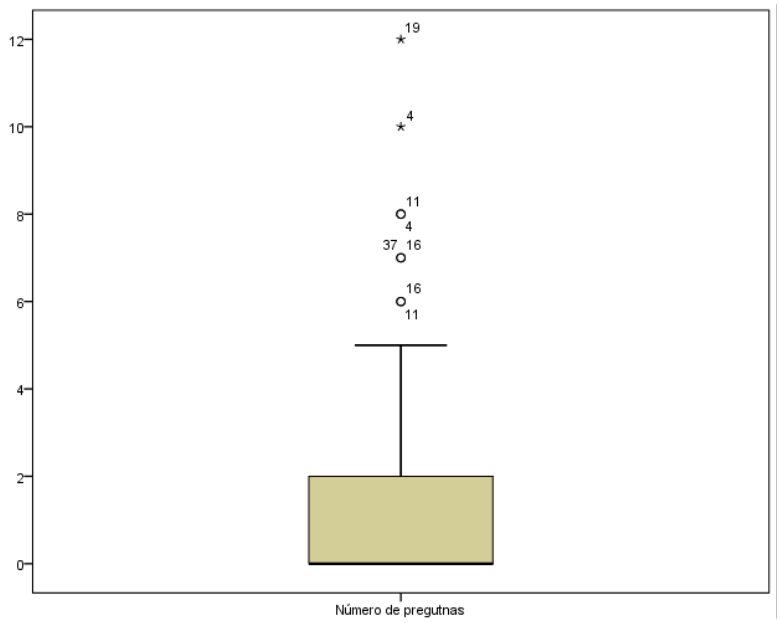

Gráfico 8. Distribución individual de la variable pragmática número de preguntas.

La lengua también permite escoger la forma en la que uno se dirige al destinatario. Se puede utilizar un trato formal como *usted* o un trato informal como el tuteo mediante el empleo de las formas verbales y pronominales de la segunda persona del singular. La forma más común en las cartas de amenazas analizadas resulta ser el trato formal con un 68,55% de los casos.

En el Gráfico 9 se ilustran los resultados por lo que concierne a la fórmula de salutación en las cartas. La fórmula más frecuente es *Estimado* (30,08%), seguida de *Hola* (20,33%), *Querido* (16,26%), *Buenos días/tardes* (12,20%) y *Señor* (10,57%). Las formas menos

frecuentes son *Apreciado* (2,44%), *A* (2,44%), *Sr* (1,63%), *BuenosXseñorX* (0,81%) y el nombre seguido de dos puntos (3,25%).

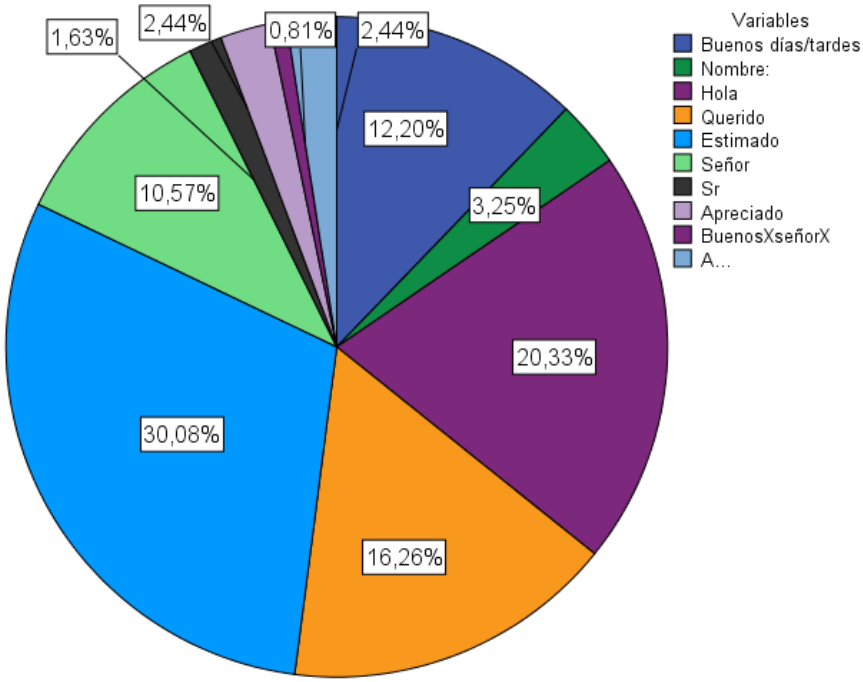

Gráfico 9. Distribución poblacional de la variable pragmática fórmula de salutación.

Las fórmulas de despedida del correo han mostrado ser muy variadas. En el Gráfico 10 se observa como la fórmula más frecuente por excelencia es *Atentamente* con un 45,83% de los casos. En un segundo plano se encuentran con un 5,56% el caso de *un saludo* y con un porcentaje del 4,17% los casos de *espero, gracias, cordialmente, reciba un cordial saludo o hasta nunca.*

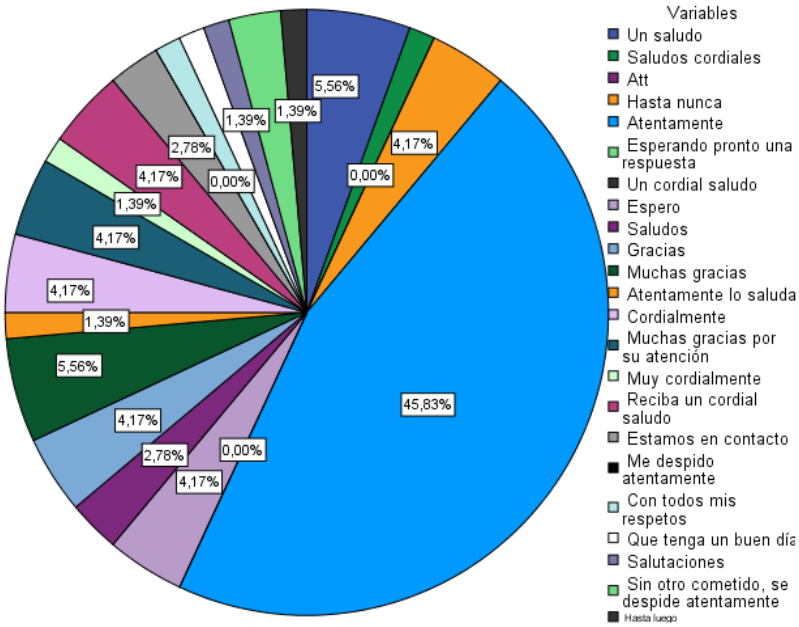

Gráfico 10. Distribución poblacional de la variable pragmática fórmula de despedida.

El uso de paréntesis como variable pragmática tiene una frecuencia media de 0,37 palabras por documento con una desviación estándar de 0,86. Por otro lado, el uso de guiones muestra una frecuencia de 0,02 con una desviación de 0,15. De este modo se observa que el uso de paréntesis y de guiones es relativamente bajo en un documento y, por tanto, un uso frecuente de paréntesis o guiones en un documento puede resultar un rasgo relevante para caracterizar el estilo idiolectal de un autor.

La variedad de marcadores discursivos tanto en la lengua escrita como oral es muy amplia. Mediante esta distribución poblacional se intenta observar cuales son los marcadores más y menos frecuentes de la lengua escrita. Los resultados (Gráfico 11) muestran que los marcadores más habituales son *¿verdad?* (35,48%) y *¿no?* (20,43%). También resulta bastante frecuente aunque con un porcentaje menor (12,90%) el marcador discursivo *¿no te parece?*.

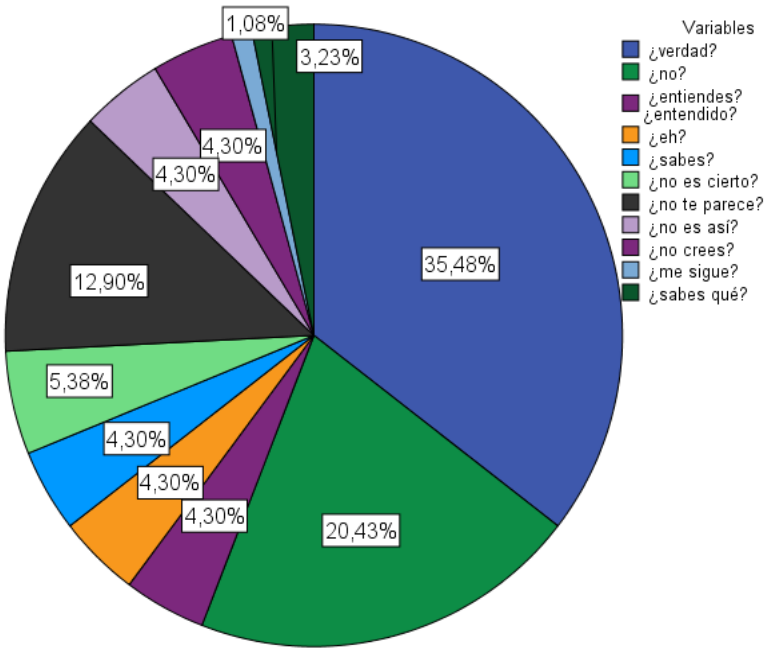

Gráfico 11. Distribución poblacional de la variable pragmática marcador discursivo.

Por último, se analiza como variable pragmática la aparición de los signos de puntuación. En español es obligatorio utilizar los signos de puntuación al inicio y al final, pero con la influencia de otros idiomas como el inglés y el catalán algunos escritores podrían perder el uso del signo de puntuación al inicio del enunciado en contextos informales. Los resultados muestran que en la mayoría de los casos (81,13%) los escritores mantienen los signos de puntuación al inicio y al final.

3.1.4. Variables sintácticas

Una de las variables sintácticas analizadas en esta tesis doctoral es el tipo de oración –simple o compleja– más frecuente en este corpus. Los resultados muestran que los hablantes estructuran su discurso en el 93,96% de los casos mediante oraciones complejas.

Por lo que concierne al tipo de oración compleja (Gráfico 12) se observa que los hablantes se expresan más frecuentemente a través de oraciones subordinadas (72,90%), en segundo lugar de oraciones coordinadas (18,92%) y, finalmente, de oraciones yuxtapuestas (8,16%).

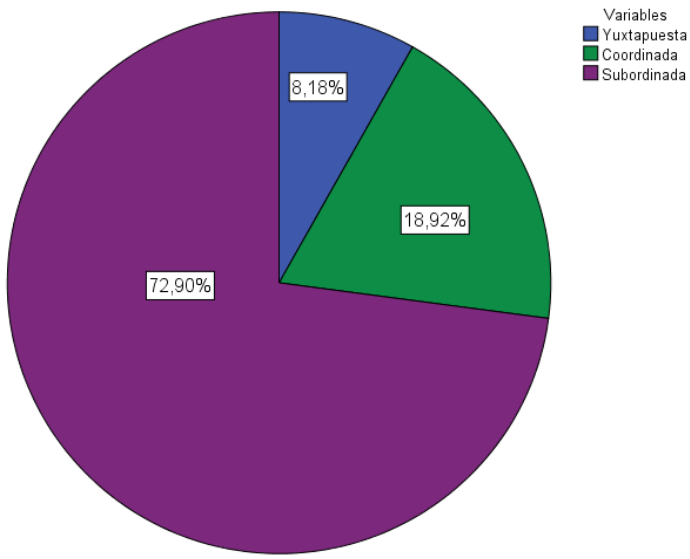

Gráfico 12. Distribución poblacional de la variable sintáctica tipo de oración compleja.

Dentro de cada tipo de oraciones complejas se analiza en primer lugar el tipo de oración yuxtapuesta. Los resultados proyectados en el Gráfico 13 muestran que en el 91,59% de los casos los escritores utilizan oraciones yuxtapuestas con coma, en el 6,67% de los casos con dos puntos y en el 1,75% restante con punto y coma.

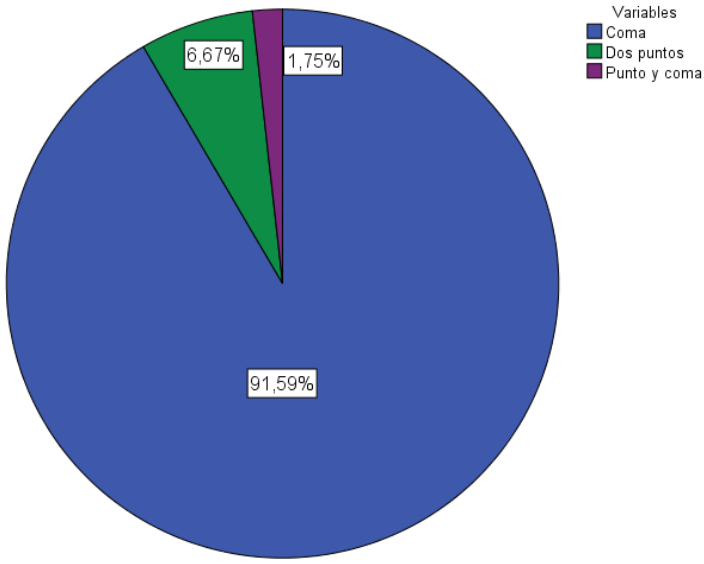

Gráfico 13. Distribución poblacional de la variable sintáctica tipo de oración yuxtapuesta.

En el caso de las oraciones coordinadas (Gráfico 14) se observa que la forma más común de coordinar oraciones es mediante las conjunciones copulativas *y/e* (71,86%). La coordinación mediante la conjunción adversativa *pero* ocurre en el 18,40% de los casos, mediante la conjunción disyuntiva *o/u* en el 6,99% y, por último, mediante la conjunción copulativa *ni* en el 2,74% de los casos.

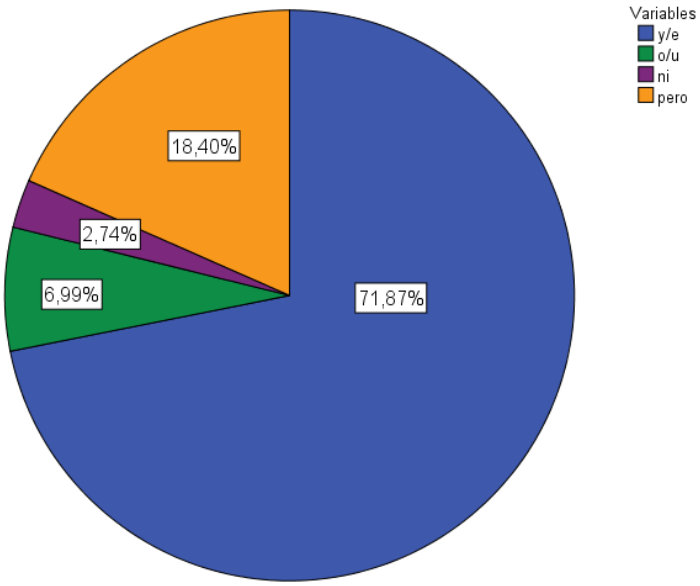

Gráfico 14. Distribución poblacional de la variable sintáctica tipo de oración coordinada.

Finalmente, se analiza el tipo de oración subordinada teniendo en cuenta el pronombre relativo –simple o compuesto– que la introduce. Esta variable resulta de especial interés en un corpus de hablantes bilingües catalán-español puesto que su frecuencia de uso puede convertirse en una marca de autoría[26]. Los resultados indican que en el 98,73% de los casos se utiliza el pronombre relativo simple *que* frente el pronombre relativo compuesto (1,27%).

[26] Véase Turell (2011: 80) sobre los resultados de esta variable en un caso real.

3.2. Variables con una distribución poblacional diferenciada por el sexo del individuo

En este apartado se introducen aquellas variables que muestran diferencias estadísticamente significativas entre mujeres y hombres. Los resultados, al igual que en los apartados anteriores, se presentan divididos por los diferentes tipos de variables lingüísticas. Estos resultados pueden ser de gran interés en la construcción de perfiles lingüísticos ya que pueden asesorar al lingüista forense en la determinación del sexo más probable del autor del texto dubitado.

3.2.1. Variables de complejidad

En la Tabla 9 se ilustran los resultados correspondientes a las tablas de contingencia de las variables de complejidad. En ésta se observa que únicamente dos de las once variables propuestas muestran diferencias estadísticamente significativas según el sexo del individuo, en concreto las variables de complejidad número de párrafos y palabras por frase.

Tabla 9. Medias de las variables de complejidad con una distribución poblacional diferenciada por el sexo del individuo.

Variables de complejidad	Hombre	Mujer	Diferencias entre sexos
Número de palabras	395,66	377,70	
Número de palabras distintas	196,75	191,21	
Número de oraciones	18,82	20,54	
Número de párrafos	6,20	4,22	*#
Palabras por frase	22,13	19,46	*#
Palabras por párrafo	156,02	189,63	
Caracteres por palabra	4,61	4,52	
Ratio type-token	,54	,55	

*Diferencia t-test 95%
#Diferencia Mann-Whitney 95%

Las diferencias por sexo en las variables de complejidad se dan en el mayor número de párrafos y de palabras por frases utilizadas por los autores hombres en sus textos. El resto de las variables son estadísticamente similares.

A continuación, se presenta un análisis más detallado de las variables de complejidad que han mostrado diferencias estadísticamente significativas entre ambos sexos y que, en un caso de perfiles lingüísticos, podrían ser de especial interés.

En primer lugar, la variable número de párrafos, ilustrada en el Gráfico 15, indica que los hombres tienden a dividir el texto mediante un número de párrafos superior (6,20 párrafos) que las mujeres (4,22 párrafos).

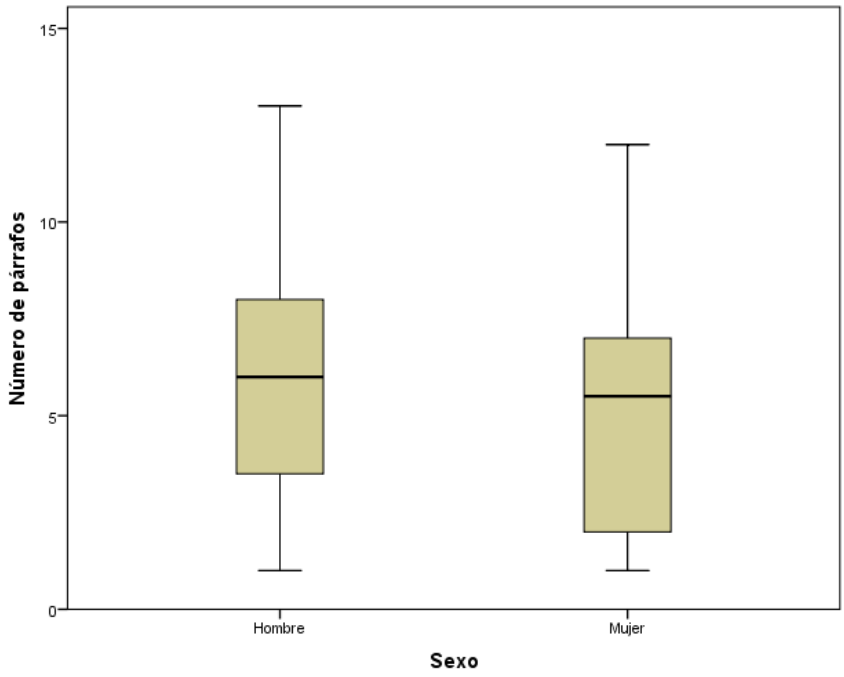

Gráfico 15.Resultado de la distribución poblacional por sexos de la variable de complejidad número de párrafos.

En segundo lugar, por lo que concierne a la variable número de palabras por frase (Gráfico 16) se observa que los hombres tienden a realizar oraciones más largas (22,13) que las mujeres (19,46). No obstante, se observa que en el caso de las mujeres hay algunas escritoras que destacan por su longitud de frase, superando las 40 palabras por frase.

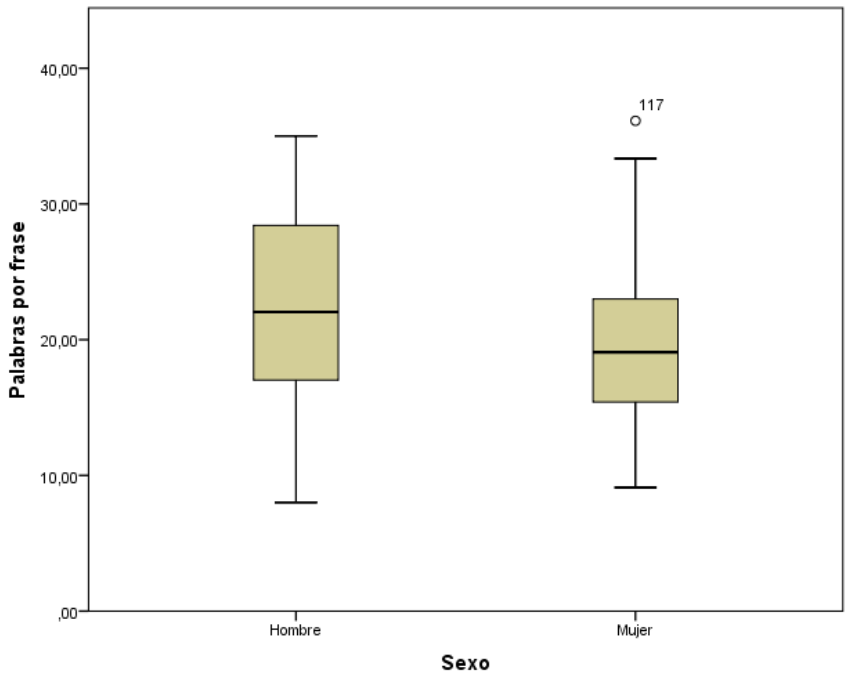

Gráfico 16. Resultado de la distribución poblacional por sexos de la variable de complejidad número de palabras por frase.

3.2.2. Variables léxicas

El análisis estadístico mediante la tabla de contingencia para determinar qué variables léxicas se diferencian entre sexos indica que las variables que muestran diferencias significativas entre ambos sexos son aquellas variables relacionadas con los errores por documento. El resultado del análisis estadístico se presenta de forma resumida en la Tabla 10.

Tabla 10. Medias de las variables léxicas con una distribución poblacional diferenciada por el sexo del individuo.

Variables léxicas		Hombre	Mujer	Diferencias entre sexos
Número de palabras malsonantes		1,61	,87	
Número de errores	Errores ortografía	2,02	,31	*#
	Errores diacríticos	3,41	1,34	*#
	Errores mayúsculas	,11	,16	
	Errores de puntuación	,43	,18	*#
	Erratas	,52	,39	
	Errores pleonasmos	,02	,04	
	Errores gramaticales	1,25	,37	*#
	Errores por contacto de lenguas	1,00	,42	*#

Expresión del futuro	Tiempo verbal de futuro	3,20	3,43	
	Perífrasis verbal *ir*	2,20	2,26	
Expresión de la obligación	*Tener que* + infinitivo	,50	,83	
	Deber + infinitivo	,55	,50	
	Haber de + infinitivo	,05	,05	
Expresión de la condición	*Si*	2,55	2,55	
	Como	,18	,29	
Expresión del pretérito imperfecto del subjuntivo	Terminación -*er*	,91	1,02	
	Terminación -ese	,25	,10	
Signos de puntuación	Aparece apertura y cierre de ¡! ¿?	,34	,59	*#
	Solo cierre de ¡! ¿?	,16	,11	
Abreviaturas	euros	,23	,18	
	€	,11	,09	

*Diferencia t-test 95%
#Diferencia Mann-Whitney 95%
Sombreado indica variables originalmente binarias

En este grupo de variables se puede destacar que el número de errores es mucho más elevado entre los hombres, siendo las diferencias estadísticamente significativas con las mujeres. La media de palabras malsonantes es casi del doble en el caso de los

hombres. No obstante, dicha diferencia no es estadísticamente significativa debida a una gran dispersión. Además, los autores femeninos destacan por un uso significativamente más alto de apertura y cierre de signos de puntuación.

En general, se puede observar que los hombres comenten más errores que las mujeres. Por lo que respecta a los errores de ortografía (Gráfico 17), se observa que los hombres cometen una media de 2 errores por documento y con un máximo de 16 errores por documento frente a una media de 0,31 en el caso de las mujeres y un máximo de 3.

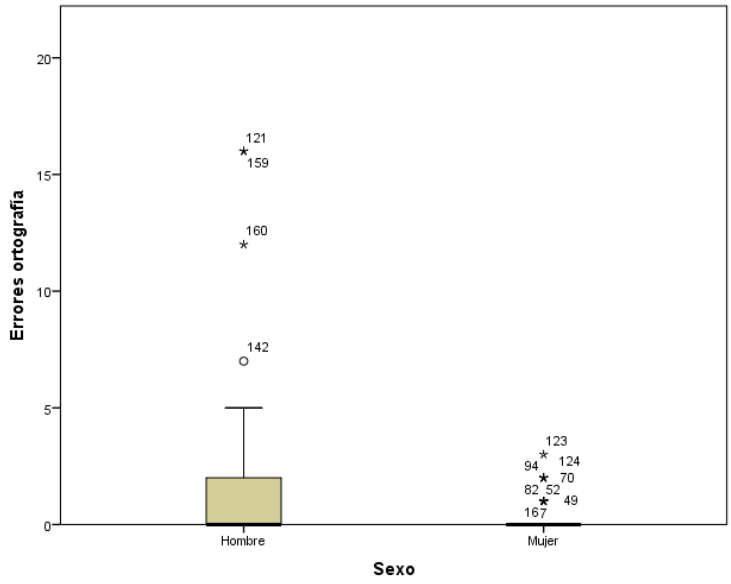

Gráfico 17. Resultado de la distribución poblacional por sexos de la variable léxica errores de ortografía.

Por lo que concierne a los errores diacríticos (Gráfico 18), los hombres realizan una media de 3,41 errores de ortografía por documento con un máximo de 22 errores, mientras que las mujeres tienen una media de 1, 34 errores con un máximo de 14 errores por documento.

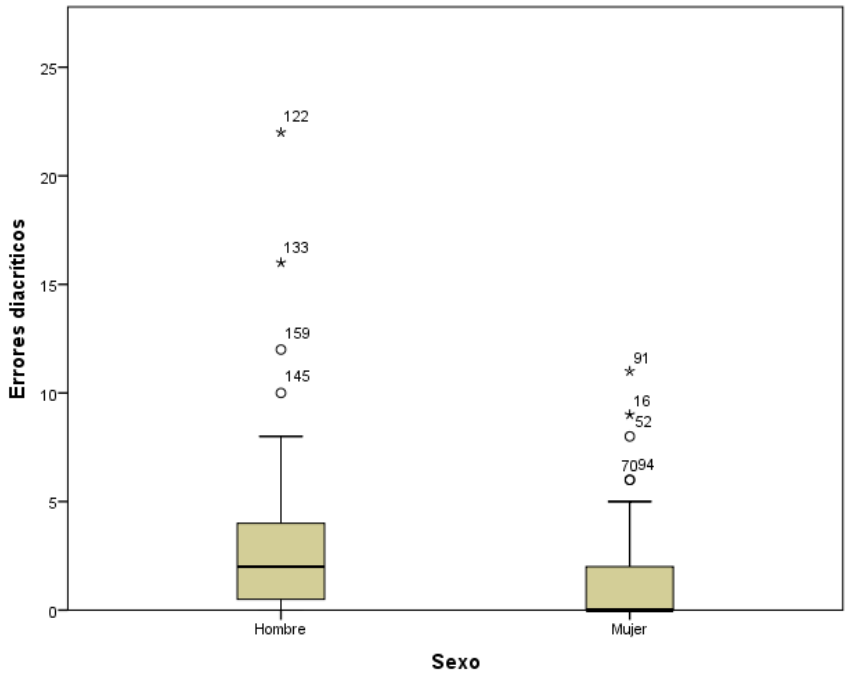

Gráfico 18. Resultado de la distribución poblacional por sexos de la variable léxica errores diacríticos.

Los errores gramaticales (Gráfico 19) son todavía menos frecuentes pero también muestran diferencias entre ambos sexos. Las mujeres cometen menos de un error gramatical por documento (0,37) y los hombres más de un error por documento, en concreto, 1,25 errores.

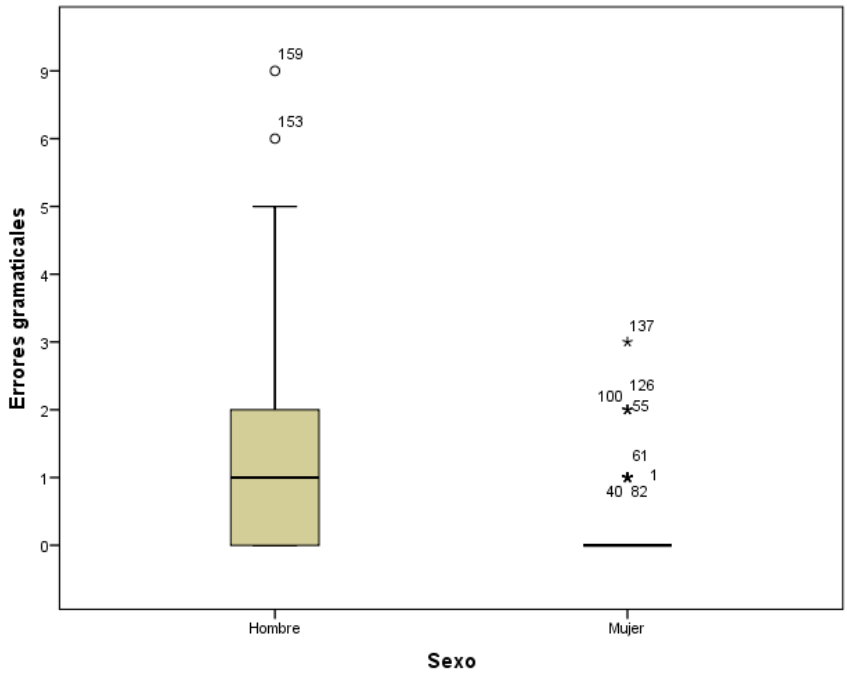

Gráfico 19. Resultado de la distribución poblacional por sexos de la variable léxica errores gramaticales.

Los errores de puntuación (Gráfico 20) son poco frecuentes en ambos sexos, no obstante, existen diferencias. En el caso de los hombres la media se sitúa en 0,43 errores por documento y en el caso de las mujeres en 0,18.

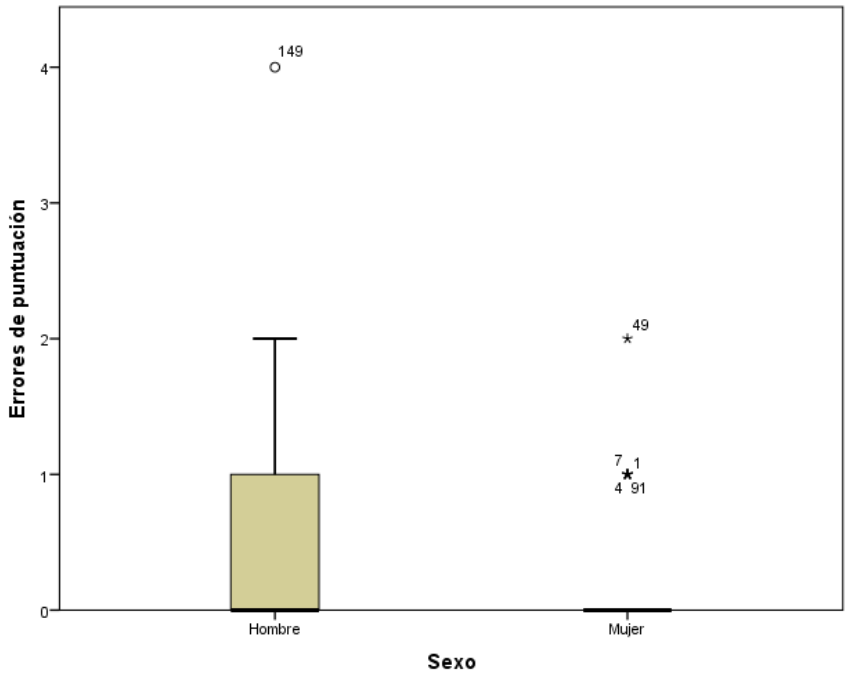

Gráfico 20. Resultado de la distribución poblacional por sexos de la variable léxica errores de puntuación.

El corpus de estudio se trata de una muestra de hablantes bilingües catalán-español, y, por este motivo, resulta de especial relevancia observar los errores producidos por el contacto de lenguas (Gráfico 21).

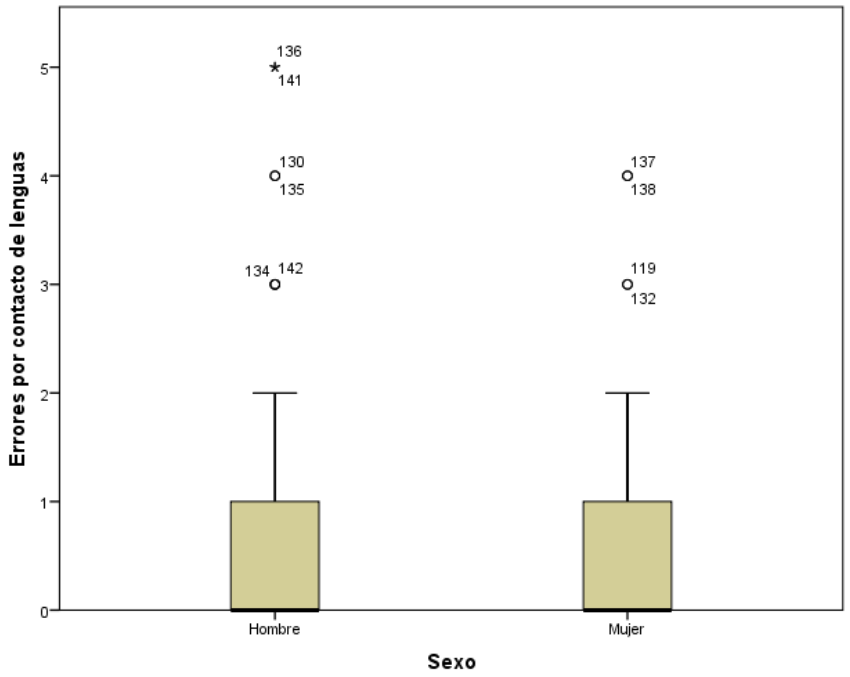

Gráfico 21. Resultado de la distribución poblacional por sexos de la variable léxica errores por contacto de lenguas.

Aunque a simple vista puede resultar muy similar el comportamiento de hombres y mujeres, el primer intercuartil presenta niveles más bajos en el caso de las mujeres, es decir, las mujeres acumulan un número de errores menor. Como se observa en el Gráfico 22 de barras apiladas, en el caso de los hombres el número de errores va en aumento, mientras que en el caso de las mujeres el número de errores va en descenso.

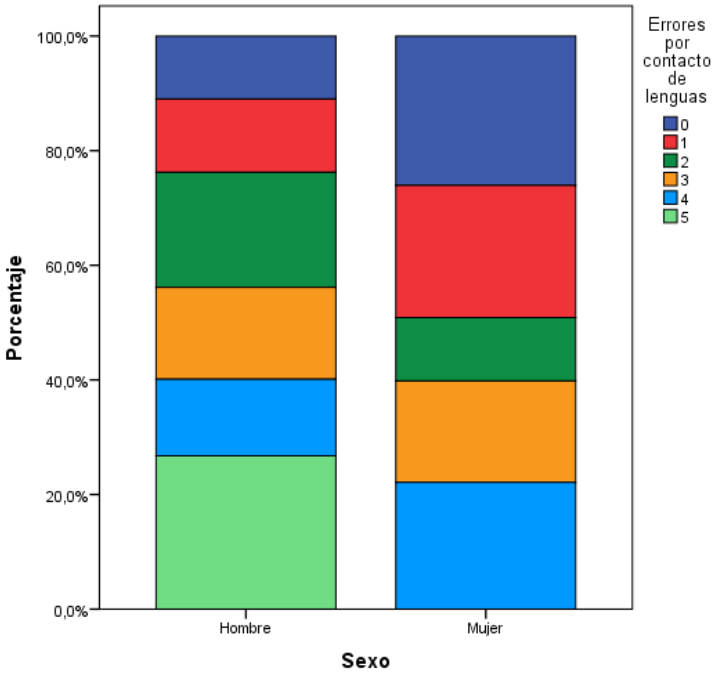

Gráfico 22. Resultado de la distribución poblacional por sexos de la variable léxica errores por contacto de lenguas mediante barras apiladas.

3.2.3. Variables pragmáticas

En la Tabla 11 se muestra de forma visual el resultado de las pruebas de contingencia para determinar si las variables pragmáticas muestran diferencias estadísticamente significativas entre ambos sexos. Se puede observar que el número de variables pragmáticas candidatas para ser útiles para indicar el sexo del escritor en casos de perfiles lingüísticos es muy bajo (12,5%).

Tabla 11. Medias de las variables pragmáticas con una distribución poblacional diferenciada por el sexo del individuo.

Variables pragmáticas		Hombre	Mujer	Diferencias entre sexos
Intensificación del sujeto 1º persona del singular	Ausencia del *yo*	14,55	14,22	
	Presencia del yo	1,07	1,37	
Expresión del énfasis	Uso de mayúsculas	,36	,48	
	Uso de signos de puntuación	,05	,15	
	Uso de la repetición	,05	,07	
Número de preguntas		,55$_a$	1,68	*#
Número de exclamaciones		,36$_a$	,67	
Trato	Formal	3,77	2,36	*#
	Informal	1,57	1,14	

Palabras entre paréntesis		,39	,37	
Intensificación del sujeto 1º persona del plural	Ausencia	,45	1,32	
	Presencia	,02	,09	
Marcadores discursivos	¿verdad?	,05	,26	*#
	¿no?	,02	,15	
	¿entiendes? ¿entendido?	,00	,03	
	¿eh?	,02	,03	
	¿sabes?	,00	,03	
	¿no es cierto?	,00	,04	
	¿no te parece?	,00	,10	
	¿no es así?	,00	,03	
	¿no crees?	,05	,02	
	¿me sigue?	,00	,01	
	¿sabes qué?	,00	,03	
Uso de guiones		,05$_a$	,02	

*Diferencia t-test 95%
#Diferencia Mann-Whitney 95%

Las diferencias por sexo en este caso están en que las mujeres hacen un mayor número de preguntas, utilizan con más frecuencia el marcador discursivo *¿verdad?* y son menos formales en su trato.

Por un lado, se puede observar que hay una diferencia estadísticamente significativa en el número de preguntas que realiza cada sexo. Tal y como se puede ver en el Gráfico 23, las mujeres realizan más preguntas que los hombres, en concreto las mujeres realizan una media de 1,68 preguntas por documento con un máximo de 12 preguntas y los hombres una media de 0,55 y un máximo de 7.

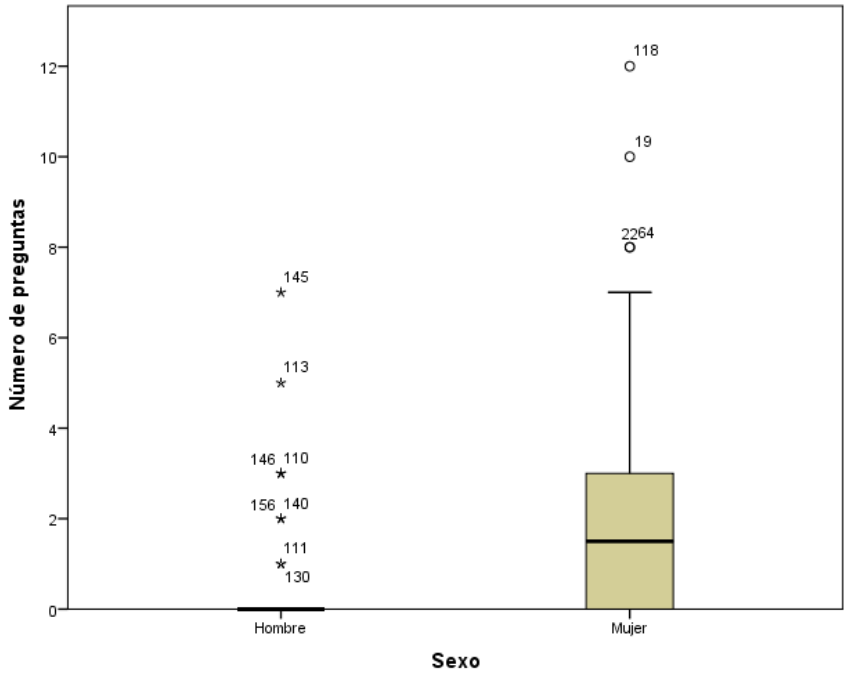

Gráfico 23. Resultado de la distribución poblacional por sexos de la variable pragmática número de preguntas.

Otro rasgo que muestra diferencias entre los sexos es cuántas veces se dirige el escritor al destinatario de una manera formal o informal, es decir, si utiliza más una forma formal o informal. En general se observa que los hombres prefieren dirigirse al lector más veces de forma formal que las mujeres (Gráfico 24).

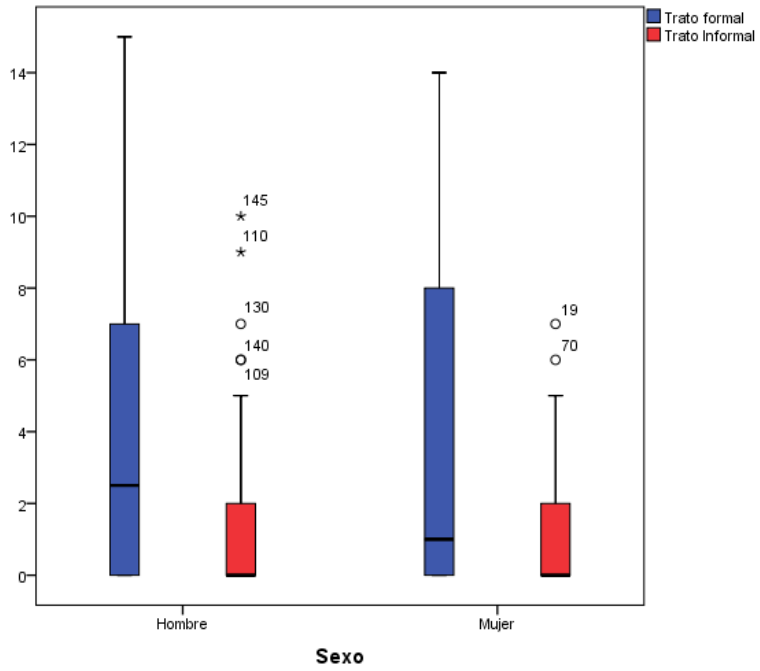

Gráfico 24. Resultado de la distribución poblacional por sexos de la variable pragmática trato.

En concreto y como se observa en el Gráfico 25 los hombres se dirigen al destinatario de forma formal un número de veces mayor (una media de 3,77 veces por documento) que las mujeres (una media de 2,36 veces).

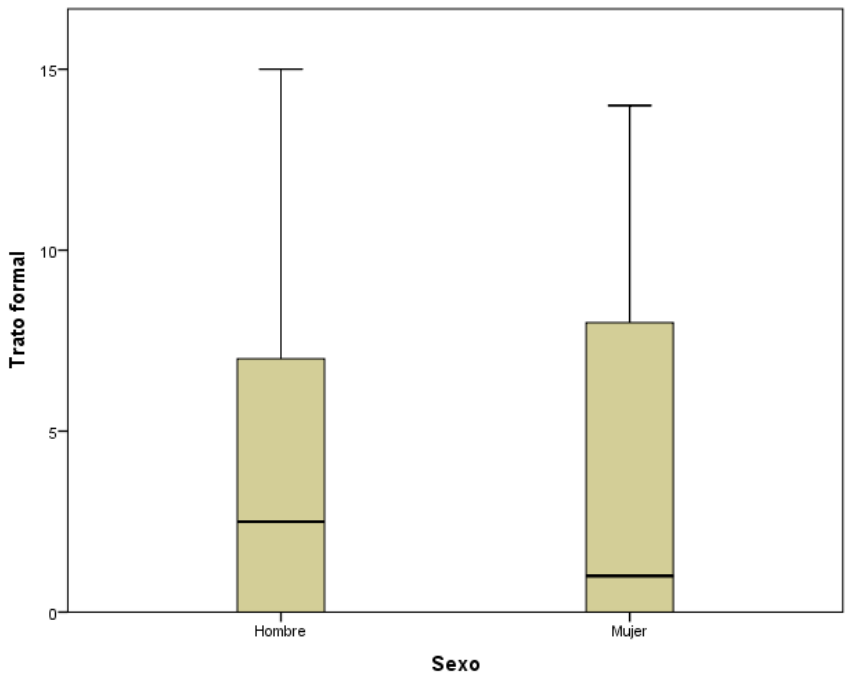

Gráfico 25. Resultado de la distribución poblacional por sexos de la variable pragmática trato formal.

Por lo que respecta al trato al destinatario de una forma informal (Gráfico 26), de nuevo son los hombres quienes se dirigen más veces al lector (1,57 los hombres, 1,14 las mujeres) aunque la diferencia es menor que en el caso de las formas más formales.

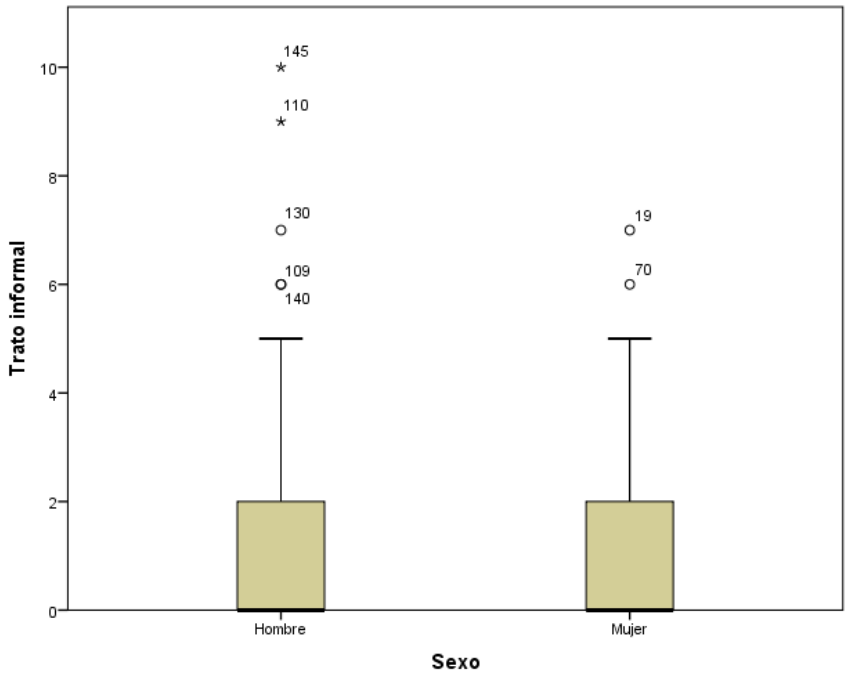

Gráfico 26. Resultado de la distribución poblacional por sexos de la variable pragmática trato informal.

En un último lugar, la opción que escoge cada grupo de individuos para escribir los signos de puntuación resulta también un rasgo diferenciador entre mujeres y hombres (Gráfico 27). En general, las mujeres tienden a seguir en mayor grado que los hombres la normativa sobre los signos de puntuación en español la cual indica que los signos de puntuación se deben escribir al principio y al final del enunciado.

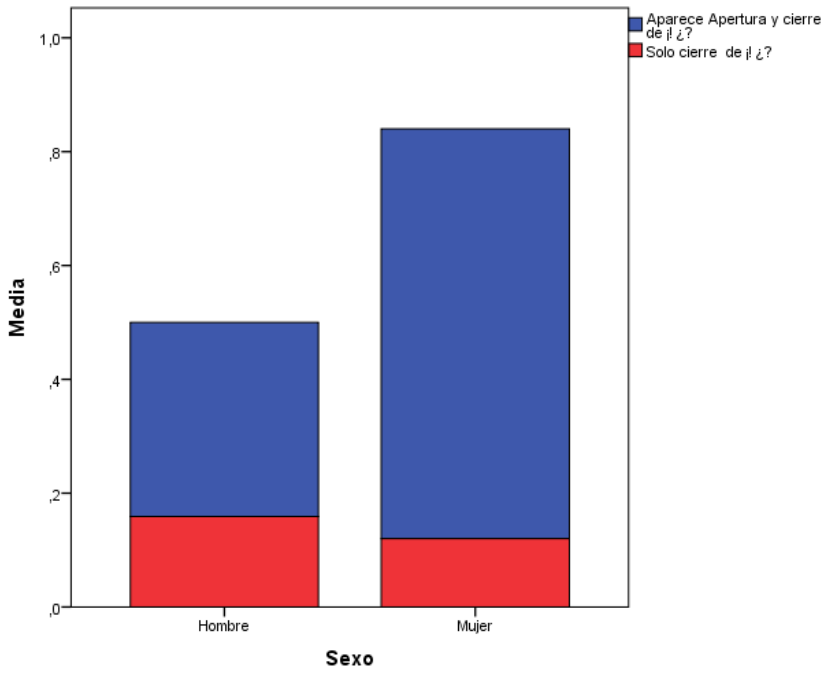

Gráfico 27. Resultado de la distribución poblacional por sexos de la variable pragmática signos de puntuación.

3.2.4. Variables sintácticas

El análisis estadístico para determinar posibles diferencias entre hombres y mujeres muestra que las variables sintácticas que indican diferencias entre mujeres y hombres son la frecuencia de uso de oraciones complejas, oraciones coordinadas introducidas por 'ni' y el tipo de oraciones subordinadas relativas introducidas por 'que'(véase Tabla 12).

Tabla 12. Medias de las variables sintácticas con una distribución poblacional diferenciada por el sexo del individuo.

Variables sintácticas		Hombre	Mujer	Diferencias entre sexos
Tipo de oración	Simple	2,14	2,93	
	Compleja	54,02	38,10	*#
Tipo de oración compleja	Yuxtapuesta	4,73	4,36	
	Coordinada	10,00	10,43	
	Subordinada	38,86	40,07	
Tipo de oración yuxtapuesta	Coma	4,07	3,32	
	Dos puntos	,27	,25	
	Punto y coma	,05	,08	
	Error de puntuación en yuxtapuestas	0,29	0,09	*#

Tipo de oración coordinada	*y/e*	6,59	7,18	
	o/u	,75	,66	
	ni	,43	,21	*#
	pero	1,61	1,87	
Tipo de oración subordinada	Relativas introducidas por *que*	5,07	3,95	*
	Relativas introducidas por *cual*	,11	,03	#

*Diferencia t-test 95%
#Diferencia Mann-Whitney 95%

En concreto, los hombres utilizan con mayor frecuencia oraciones complejas, aunque la proporción según el tipo de oración compuesta se mantiene igual entre hombres y mujeres. Dentro de las oraciones yuxtapuestas se encuentran más errores en los autores hombres y dentro de la oración coordinada hay un mayor uso de *ni* entre los hombres y entre las oraciones subordinadas las relativas introducidas por *que* o *cual*.

Por lo que concierne a la producción de oraciones complejas (Gráfico 28), las mujeres escriben menos oraciones complejas por documento que los hombres. En concreto, los hombres escriben una media de 54,02 oraciones complejas por documento mientras que las mujeres 38,01 oraciones.

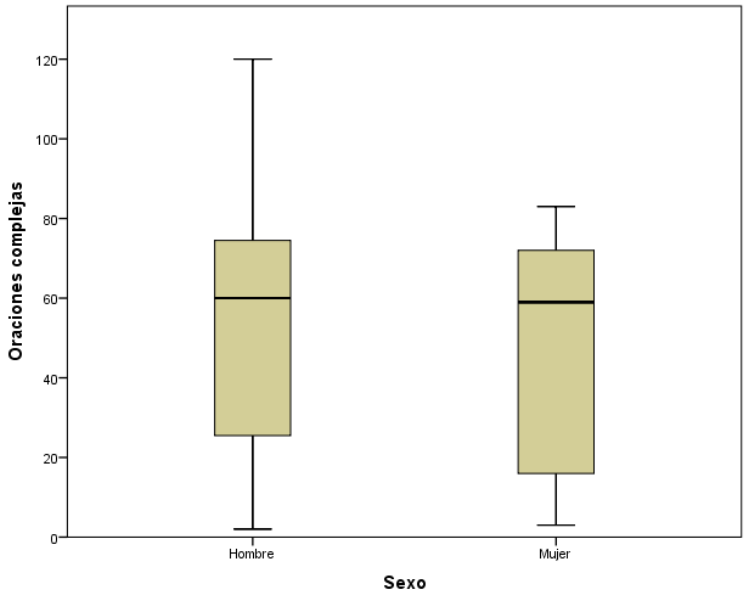

Gráfico 28. Resultado de la distribución poblacional por sexos de la variable sintáctica oraciones complejas.

En el caso de las oraciones coordinadas (Gráfico 29) introducidas por ‘ni’ se observa que los hombres utilizan con mucha mayor frecuencia la coordinación mediante ‘ni’.

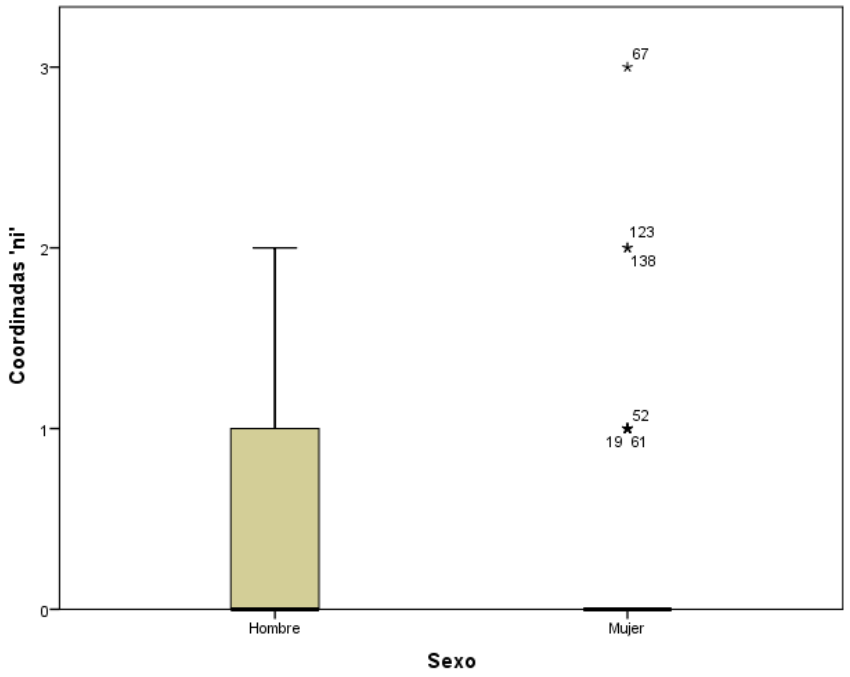

Gráfico 29. Resultado de la distribución poblacional por sexos de la variable sintáctica coordinadas ‘ni’.

Y, por último, las oraciones subordinadas relativas introducidas por el pronombre relativo simple o compuesto se muestran en el Gráfico 30.

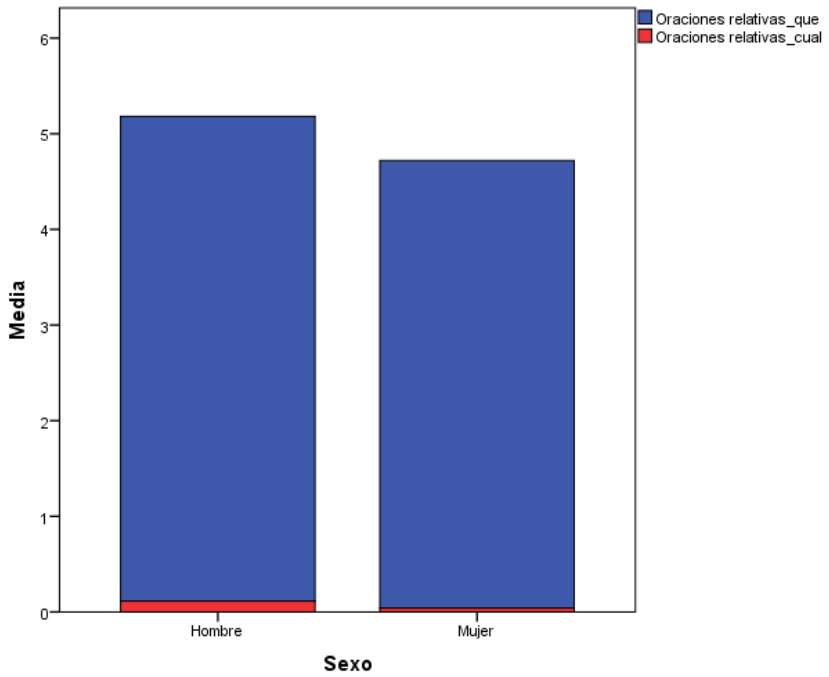

Gráfico 30. Resultado de la distribución poblacional por sexos de la variable tipo de oración subordinada.

Como es de esperar las oraciones subordinadas se introducen principalmente con el pronombre simple en ambos sexos. No obstante, se observa que las mujeres hacen un menor uso de las subordinadas relativas tanto de las introducidas por el pronombre relativo simple como por el compuesto.

En el caso de las relativas con pronombre relativos simple (Gráfico 31) se observa que su uso es más extendido en el caso de los hombres, en concreto, lo utilizan una media de 5,07 relativos simples por documento, que las mujeres con una media de 3,95 relativos simples.

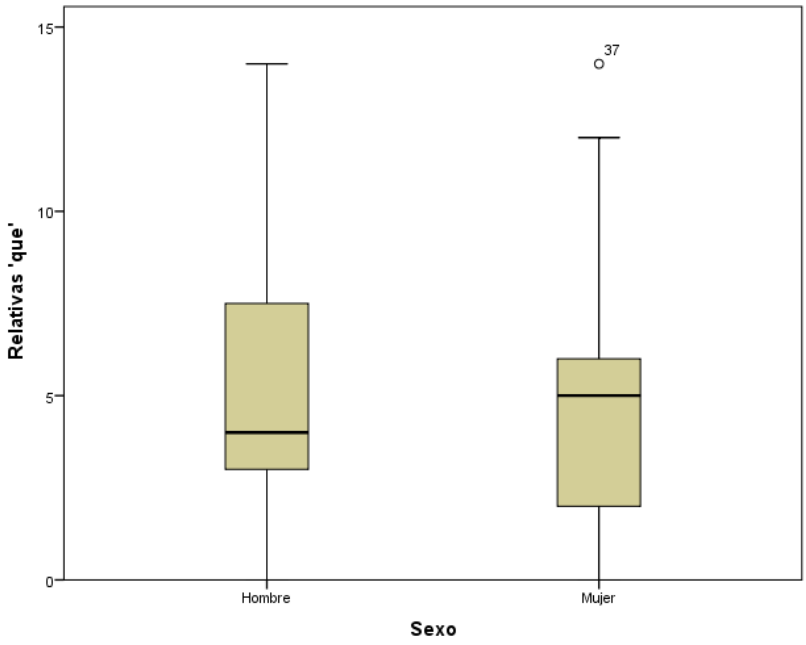

Gráfico 31. Resultado de la distribución poblacional por sexos de la variable sintáctica relativas 'que'.

En el caso del relativo compuesto, se observa que las mujeres tienden a utilizar de forma menos frecuente que los hombres las relativas compuestas. La representación gráfica (Gráfico 32) se ha llevado a cabo mediante un gráfico de barras acumuladas puesto que la mayoría de los datos se acomulaban en el 0 y el resto en 1 y un gráfico de cajas no permitiría una visualización clara de los resultados.

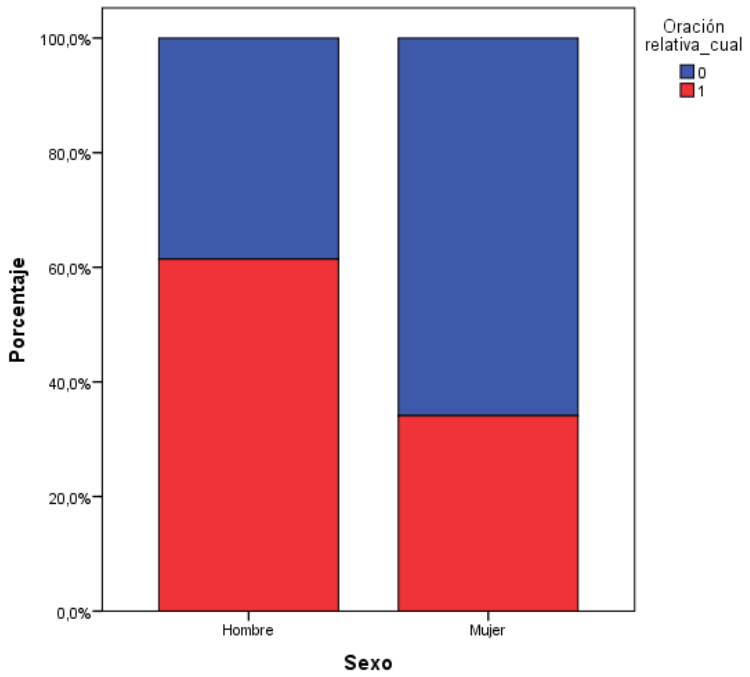

Gráfico 32 . Resultado de la distribución poblacional por sexos de la variable sintáctica relativas 'cual'.

3.2.5. Resumen

En la Tabla 13 se presenta un resumen de las variables lingüísticas que muestran diferencias entre sexos dividas por bloques de variables, es decir, complejidad, variables léxicas, pragmáticas y sintácticas. Estas variables son candidatas a la distinción entre hombres y mujeres en un caso real en que sea necesario realizar un perfil lingüístico de una carta dubitada.

Tabla 13. Resumen de variable con potencial discriminatorio por sexos.

Complejidad	Léxico	Pragmática	Sintácticas
Número de párrafos Palabras por frase	Errores ortográficos Errores diacríticos Errores de puntuación Errores de gramática Errores por contacto de lenguas Signos de puntuación ¡! ¿?	Número de preguntas Pronombres de trato Marcadores	Número de oraciones compuestas Número de oraciones coordinadas con ‘ni’ Error de puntuación en las oraciones yuxtapuestas Oraciones relativas

3.3. Resumen y discusión

En este capítulo 3 se ha realizado una distribución poblacional de una muestra dada puesto que no existía una medición de referencia sobre una población con estas características sociolingüísticas. Por este motivo, se han presentado los resultados desde una perspectiva frecuentista.

En el apartado 3.2.1 se muestran las tendencias generales de cada variable lingüística. Estos resultados permitirán en casos futuros de atribución de autoría determinar la expectativa de una determinada variable lingüística y, por tanto, su idiosincrasia.

Los resultados del apartado 3.2.2, son especialmente relevantes en casos de perfiles lingüísticos ya que, diversas variables han mostrado comportamientos distintos según el sexo del autor.

En gran parte de la literatura en lingüística forense se comete el error de utilizar la distancia de una media poblacional para decir algo acerca de un individuo. En esta tesis, se crea la base de un modelo y no se pretende establecer que si alguien utiliza una gran cantidad de palabras en una oración, deba ser un hombre.

La distribución poblacional desarrollada en este capítulo ha permitido ajustar el marco experimental de esta tesis doctoral y facilitar la implementación en el capítulo 4 de un modelo de cálculo de probabilidades con el fin de aproximar una medida de verosimilitud similar a las utilizadas en poblaciones conocidas.

Capítulo 4:

Resultados sobre los modelos de clasificación

En el capítulo 4 se muestran los resultados sobre los modelos de clasificación. En primer lugar, sobre el potencial discriminante de los distintos tipos de variables analizados en esta tesis doctoral (Apartado 4.1); seguidamente, los resultados sobre la identificación de patrones para variables binarias y continuas (Apartado 4.2); y, los resultados sobre las probabilidades de clasificación a posteriori (Apartado 4.3) tanto mediante el análisis discriminante directo como a partir de distancias. En último lugar, se muestra un resumen de los resultados y una breve discusión de los mismos.

4. Resultados sobre los modelos de clasificación

En este apartado se lleva a cabo un análisis discriminante para determinar qué variables de la distribución poblacional pueden ser capaces de diferenciar a unos individuos de otros, puesto que muestran una variabilidad mayor entre hablantes que dentro del mismo hablante. Es decir, se intenta encontrar aquellas variables que puedan ser capaces de reflejar y, por tanto, diferenciar el estilo idiolectal de los individuos.

En muchas investigaciones sociolingüísticas se tiende a dar una gran importancia a los valores medios y, de este modo, se tiende a encubrir casos de variación individual. No obstante, también puede observarse que desde los primeros estudios sociolingüísticos se han seleccionado casos individuales que se comentan con mayor extensión[27]. En el caso de la lingüística forense, es imprescindible el estudio del comportamiento del individuo y, por tanto, de las características lingüísticas que no obtienen valores medios para poder caracterizar su estilo idiolectal y diferenciarlo del resto.

[27] Véanse Labov, 1972: 99-107, 1994: 385-405.

4.1. Resultados sobre el potencial discriminante de las variables

A continuación se introducen los resultados de los análisis discriminantes para cada bloque de variables.

4.1.1. Variables de complejidad

Las variables de complejidad seleccionadas (Tabla 14) por su mayor potencial discriminante para la comparación forense de textos escritos son el número de párrafos, tokens, frases y la longitud de párrafo según el número de palabras.

Tabla 14. Análisis discriminante sobre variables de complejidad.

Variables introducidas/excluidas[a,b,c,d]

		Lambda de Wilks											
						F exacta				F aproximada			
Paso	Introducidas	Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.	Estadístico	gl1	gl2	Sig.
1	Número de párrafos	,130	1	46	107,000	15,568	46	107,000	,000				
2	Número de tokens distintos	,022	2	46	107,000	13,119	92	212,000	,000				
3	Número de frases	,006	3	46	107,000					10,083	138	315,597	,000
4	Palabras por párrafo	,002	4	46	107,000					8,749	184	417,649	,000

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

a. El número máximo de pasos es 22.

b. La F parcial mínima para entrar es 3.84.

c. La F parcial máxima para salir es 2.71

d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

Seguidamente, se presentan los resultados de la variación intra- e inter-escritor de las variables seleccionadas de forma gráfica. En primer lugar, se muestra la variable número de párrafos en el Gráfico 33. Dicha variable indica una mayor variación inter-autor que inter-autor. Por una lado, se observa de forma visual la remarcable variación inter-autor mediante la dispersión de los valores entre autores; por otro lado, se refleja una variación intra-escritor muy leve en el 77,77% de los casos ya que la distancia entre las seis muestras de un autor es relativamente pequeña (el rango intercuartílico Q3-Q1 es ≤ a 3 párrafos de diferencia entre las muestras).

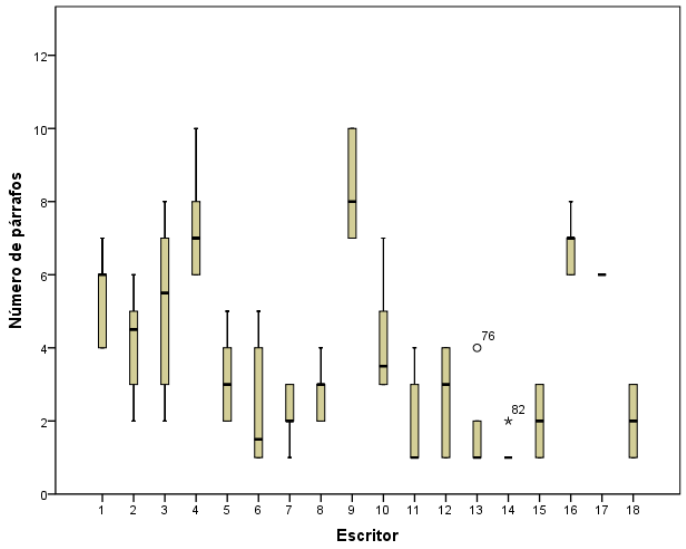

Gráfico 33. Variación intra- e inter-escritor variable número de párrafos.

En relación con la variable número de tokens, el Gráfico 34 refleja una alta variación inter-autor. Cada autor parece tener un rango de número de tokens por documento y suele ser relativamente constante en todas sus muestras. Más concretamente, solo 4 autoras presentan una variación media superior a los 150 tokens. Además, cabe destacar la relevancia de este resultado puesto que en este estudio esta variable estaba limitada a un espacio recomendado por el investigador.

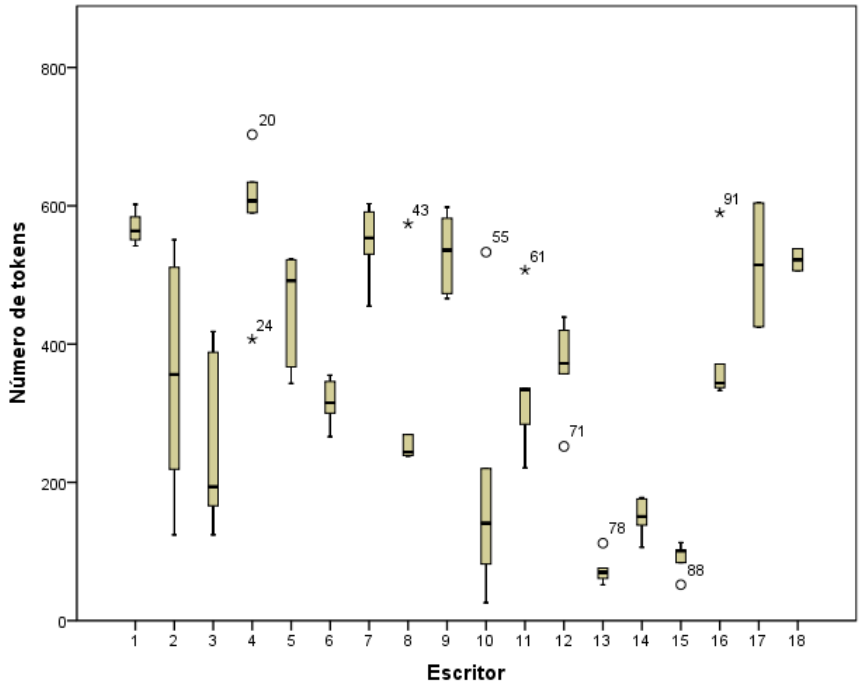

Gráfico 34. Variación intra- e inter-escritor variable número de tokens.

El Gráfico 35 muestra los resultados sobre la variable de complejidad número de frases. Se observa una variación intra-escritor muy leve en el 61,11% de los casos ya que la distancia entre los textos es relativamente pequeña (el rango intercuartílico Q3-Q1 es $\leq$ a 5 párrafos de diferencia entre las muestras). Además, también se observa una abundante dispersión en el gráfico entre los escritores y, por tanto, una elevada variación inter-autor.

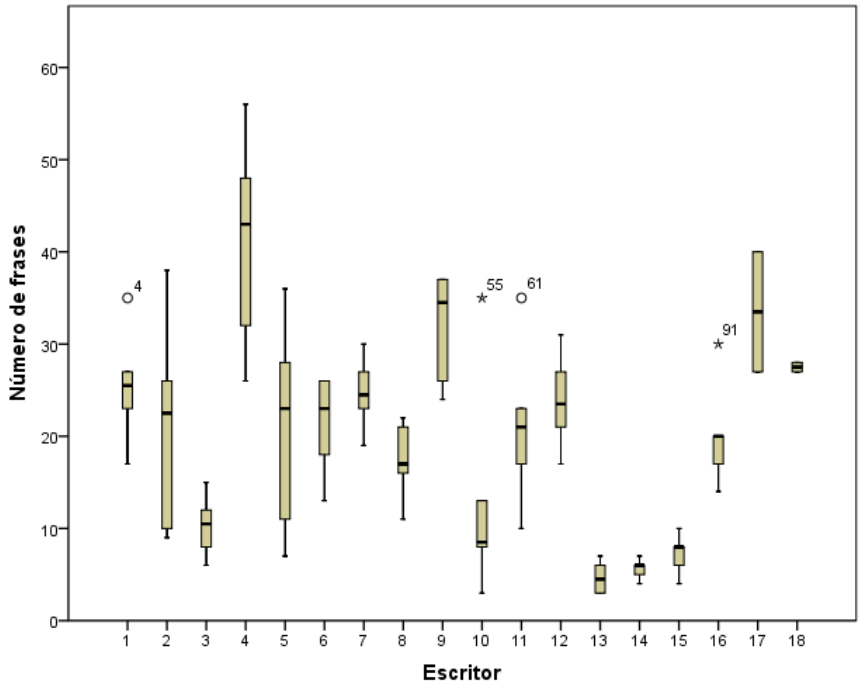

Gráfico 35. Variación intra- e inter-escritor variable número de frases.

El Gráfico 36 correspondiente a la longitud de párrafo por palabra se puede observar que en el 61,11% de los casos la variación intra-autor es muy pequeña –una moda de desviación de 15 palabras y un máximo de desviación de 45 palabras. No obstante, también existe una variación intra-autor en el 38,88% de los casos restantes muy elevada. Esta variación se refleja en el gráfico con una corta distancia entre las muestras de un mismo autor y la dispersión de los valores entre los autores.

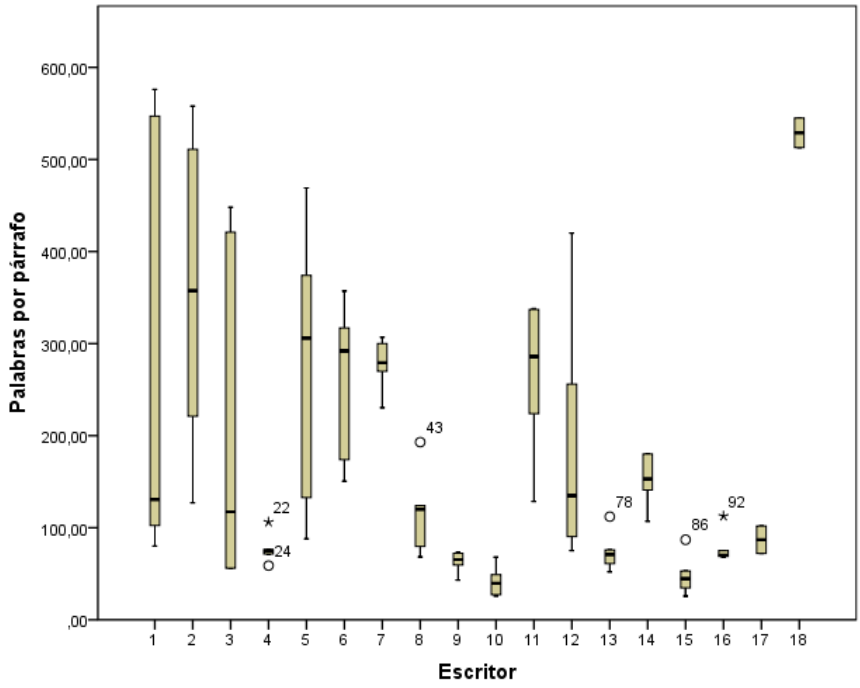

Gráfico 36. Variación intra- e inter-escritor variable palabras por párrafo.

4.1.2. Variables léxicas

Seguidamente se muestra la Tabla 15 con los resultados del análisis discriminante sobre las variables léxicas. Las variables seleccionadas como posibles marcadores de autoría son el número de errores de ortografía, de palabras malsonantes y, finalmente, de errores por contacto de lenguas.

Tabla 15. Análisis discriminante sobre variables léxicas.

Variables introducidas/excluidas[a,b,c,d]

		Lambda de Wilks											
						F exacta				F aproximada			
Paso	Introducidas	Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.	Estadístico	gl1	gl2	Sig.
1	Errores de ortografía	,277	1	46	116,000	6,571	46	116,000	,000				
2	Palabras mal sonantes	,089	2	46	116,000	5,902	92	230,000	,000				
3	Errores por contacto	,033	3	46	116,000					5,273	138	342,569	,000

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

a. El número máximo de pasos es 54.

b. La F parcial mínima para entrar es 3.84.

c. La F parcial máxima para salir es 2.71

d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

En el caso de los errores de ortografía (Gráfico 37) se observa que no son muy frecuentes en este tipo de autores –mujeres con estudios universitarios. En concreto, el 77,77% de los autores comete uno o ningún error ortográfico en el total de sus muestras. Además, las autoras que realizan errores ortográficos tienden a realizar el mismo número de errores en todas sus muestras. De este modo, se puede postular una clara tendencia a encontrar un número de errores ortográficos constante en las muestras de las autoras.

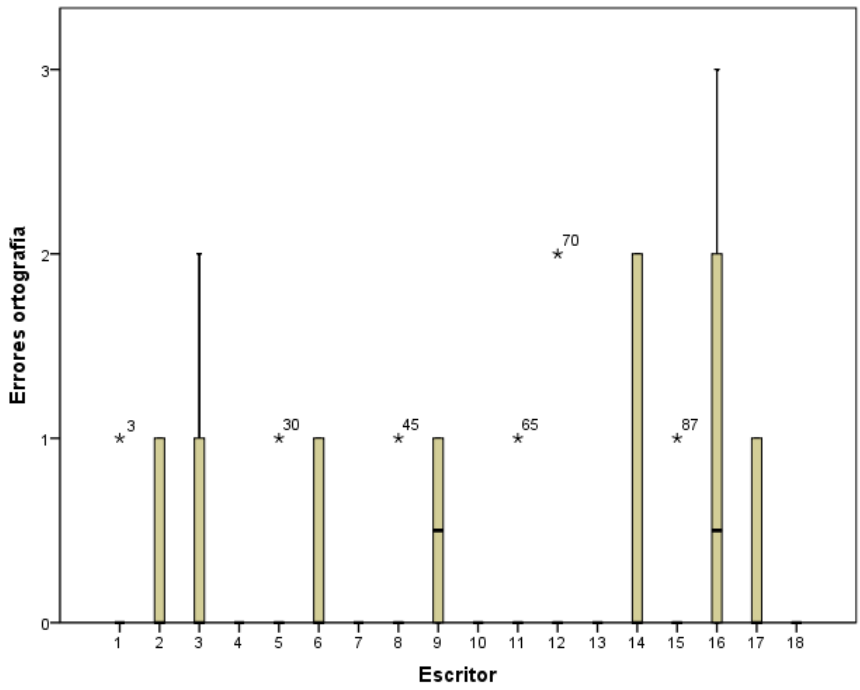

Gráfico 37. Variación intra- e inter-escritor variable errores de ortografía.

En segundo lugar, se analiza la variable palabras malsonantes (Gráfico 38). En el 33,33% de los casos las autoras no utilizan palabras malsonantes en ninguna de las muestras o una en una de sus muestras. No obstante, en el caso de utilizar palabras malsonantes se observa que cada escritora tiene una distribución particular de los valores de la variable observable en la mediana de la distribución o segundo cuartil de su caja.

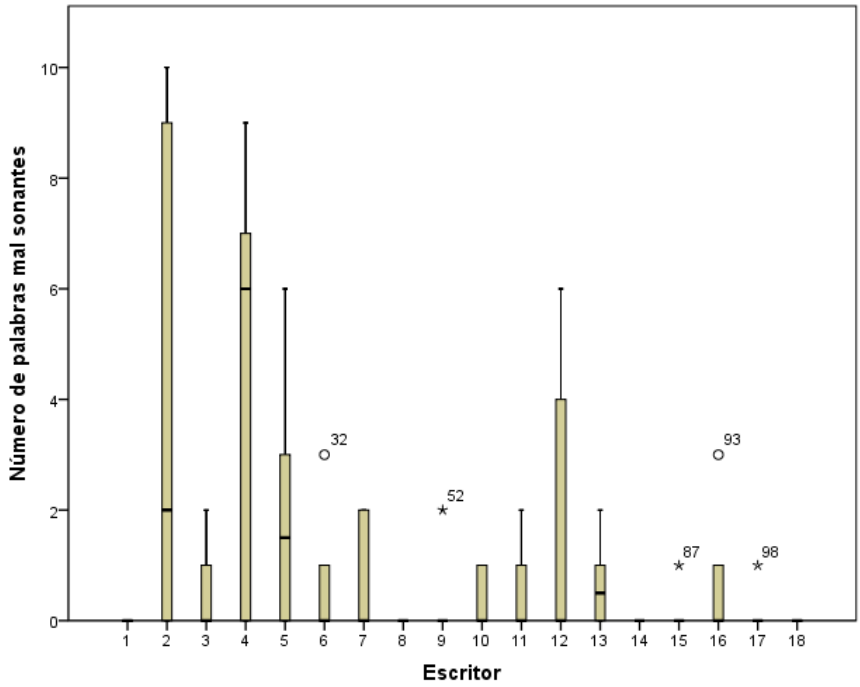

Gráfico 38. Variación intra- e inter-escritor variable palabras malsonantes.

Finalmente, por lo que respecta al grupo de variables léxicas seleccionadas por su potencial discriminante se muestra en el

Gráfico 39 la variable errores por contacto de lenguas. A pesar del bajo número de errores que comenten los individuos, se puede observar una menor variación intra-autor que variación inter-autor. Concretamente, este hecho se ve reflejado en que el 77,77% de los autores presentan o el mismo número de errores en todas las muestras o una desviación de un error entre las muestras.

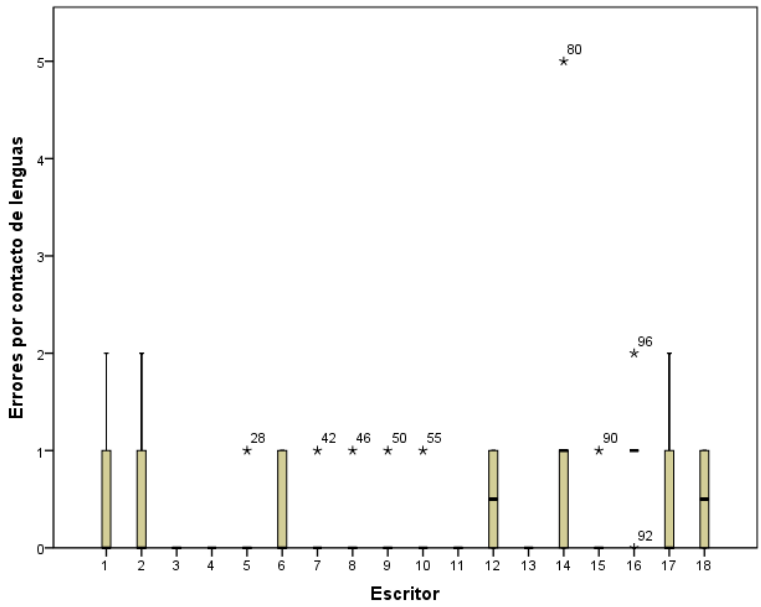

Gráfico 39. Variación intra- e inter-escritor variable errores por contacto de lenguas.

4.1.3. Variables sintácticas

El análisis discriminante arrojado sobre las variables sintácticas (Tabla 16) selecciona tres variables como posibles marcadores de autoría: las oraciones subordinadas, las oraciones simples y las oraciones complejas.

Tabla 16. Análisis discriminante sobre variables sintácticas.

Variables introducidas/excluidas[a,b,c,d]

		Lambda de Wilks											
						F exacta				F aproximada			
Paso	Introducidas	Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.	Estadístico	gl1	gl2	Sig.
1	Oraciones subordinadas	,183	1	46	72,000	7,006	46	72,000	,000				
2	Oraciones simples	,054	2	46	72,000	5,098	92	142,000	,000				
3	Oraciones complejas	,015	3	46	72,000					4,617	138	210,707	,000

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

a. El número máximo de pasos es 34.

b. La F parcial mínima para entrar es 3.0.

c. La F parcial máxima para salir es 2.50

d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

En el primer lugar, se observa que la variable oraciones subordinadas tiene una variación intra-escritor muy baja en algunos casos (una desviación inferior a 10 oraciones en el 38% de los autores) que permite distinguir fácilmente los patrones de comportamiento de los autores. No obstante, se observa una variación inter-escritor moderada ya que en algunos casos el comportamiento de dos autores resulta demasiado parecido para poder ser diferenciados mediante esta variable (véase por ejemplo la comparativa del autor 2 y 3 en el Gráfico 40).

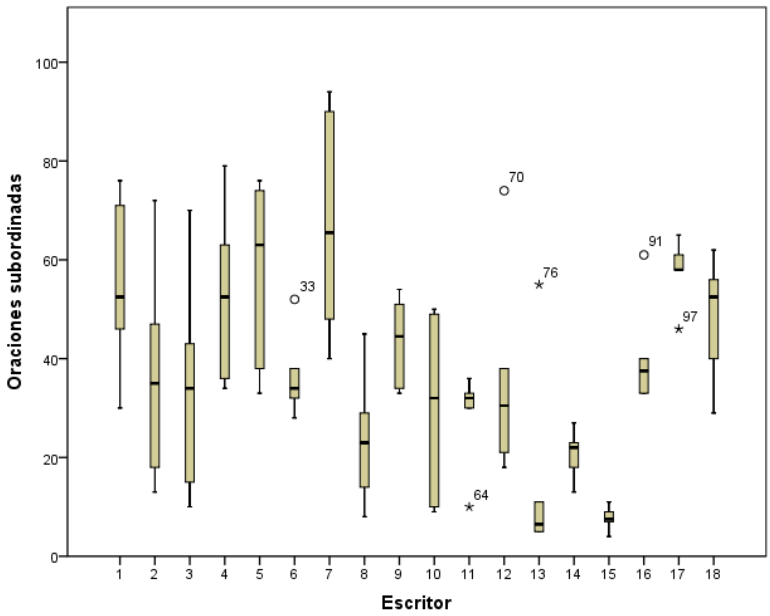

Gráfico 40. Variación intra- e inter-escritor variable oraciones subordinadas.

La segunda variable seleccionada en el análisis discriminante es el número de oraciones simples que produce el individuo. En el Gráfico 41 se ilustra una baja variación intra-escritor —en el 77,7% de los casos el rango intercuartílico Q_3- Q_1 del número de oraciones simples tiene una desviación ≤ 5 oraciones simples— frente una mayor variación inter-escritor.

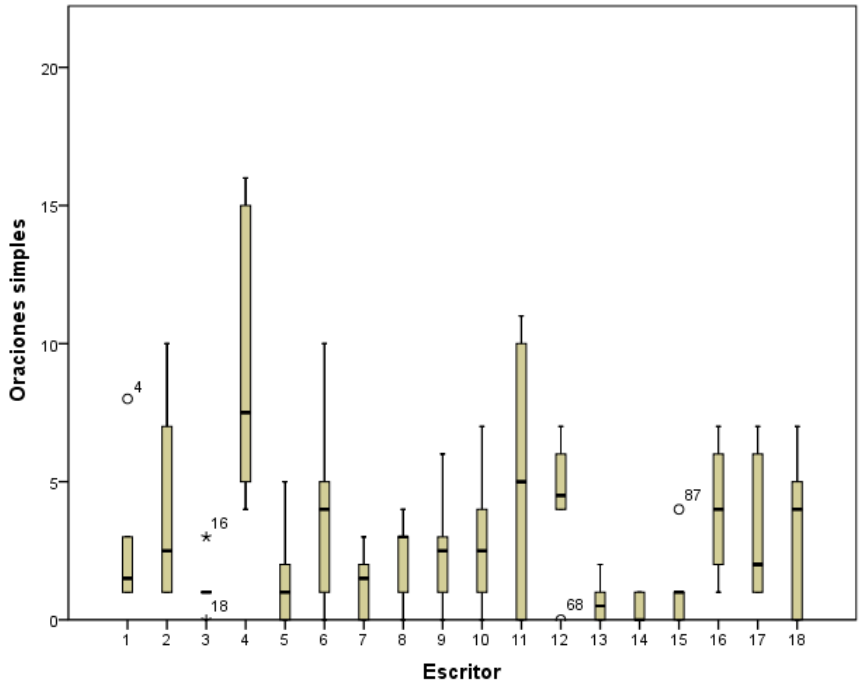

Gráfico 41. Variación intra- e inter-escritor variable oraciones simples.

En el Gráfico 42 se muestra la variación intra- e inter-escritor de los 18 autores. En este gráfico se pueden observar autores muy distanciados entre ellos, hecho que responde a la existencia de una variación inter-autor mayor que la variación intra-autor. No obstante, en algunos casos la variación intra-autor se excede de lo deseable (véase por ejemplo autor 1, 7 y 9) y en esos casos el proceso de atribuir la autoría sería dificultoso o imposible mediante esta variable.

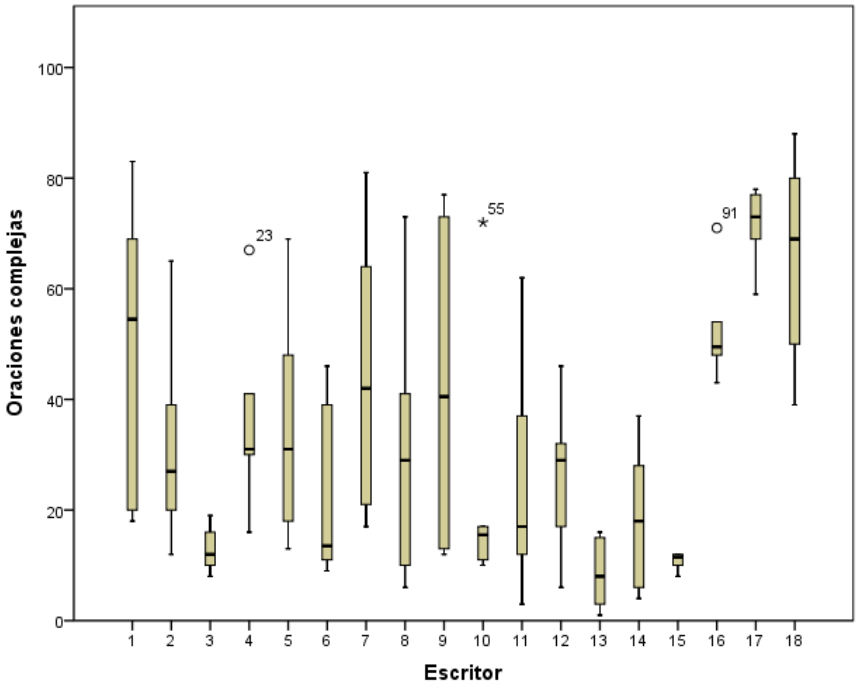

Gráfico 42. Variación intra- e inter-escritor variable oraciones complejas.

4.1.4. Variables pragmáticas

Las variables *cordialmente* y *la ausencia del sujeto de primera persona del singular* son las seleccionadas por el análisis discriminante tal y como se observa en la Tabla 17.

Tabla 17. Análisis discriminante sobre variables pragmáticas.

Variables introducidas/excluidas[a,b,c,d]

Paso	Introducidas	Lambda de Wilks							
						F exacta			
		Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	Cordialmente	,283	1	46	115,000	6,333	46	115,000	,000
2	Ausencia del sujeto *yo*	,111	2	46	115,000	4,977	92	228,000	,000

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

a. El número máximo de pasos es 118.

b. La F parcial mínima para entrar es 3.84.

c. La F parcial máxima para salir es 2.71

d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

El estadístico selecciona, en primer lugar, el adverbio *cordialmente* ya que permite diferenciar un individuo completamente, puesto que solo lo usa ese escritor.

La variable ausencia del sujeto *yo* (Gráfico 43) muestra una variación inter-escritor superior a la variación intra-escritor. Este hecho se puede ver reflejado en unas medianas muy distintas entre los autores y una concentración de los datos en los cuartiles 2 y 3 con distancias inferiores o iguales a las 10 realizaciones en el 83,33% de los casos.

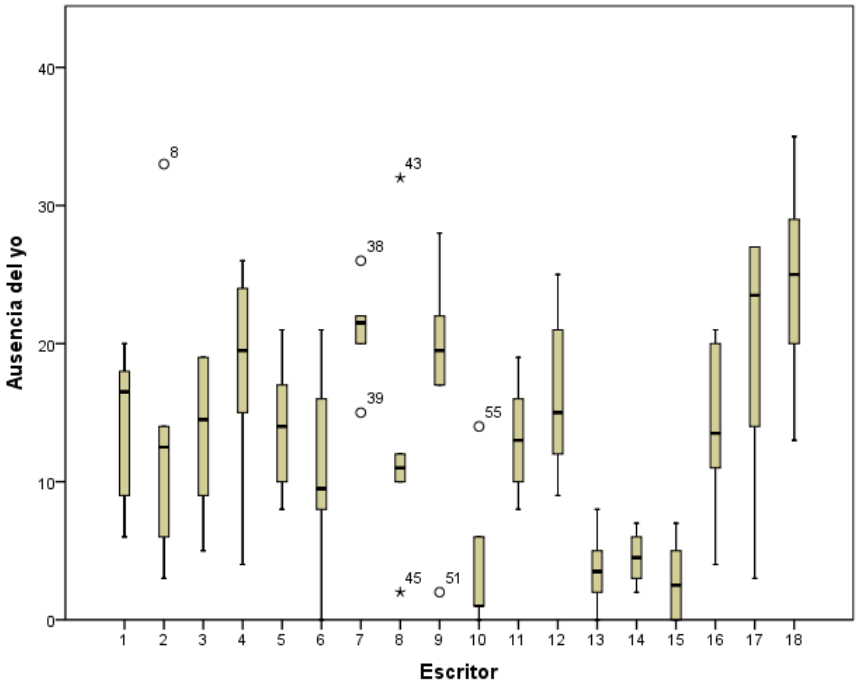

Gráfico 43. Variación intra- e inter-escritor variable ausencia del sujeto yo.

4.1.5. Resumen

En la Tabla 18 se muestran las variables con mayor potencial discriminante. Estas variables determinadas a partir del análisis discriminante son variables candidatas a marcadores de autoría ya que poseen una alta variación inter-autor y una baja variación intra-autor.

Tabla 18. Resumen variables discriminantes seleccionadas por bloques.

Complejidad	Léxicas	Sintácticas	Pragmáticas
Número de párrafos Número de tokens distintos Número de frases Palabras por párrafo	Errores de ortografía Palabras malsonantes Errores por contacto	Oraciones subordinadas Oraciones simples Oraciones complejas	Cordialmente Ausencia del sujeto *yo*

Este conjunto de variables es el utilizado para calcular las probabilidades de acierto y fallo en las clasificaciones posteriores mediante análisis discriminante. A partir de dichos resultados es posible establecer la sensibilidad y especificidad del método propuesto en esta tesis así como la razón de verosimilitud, es decir, cómo de probables son dichas clasificaciones.

4.2. Identificación de patrones

En este apartado se lleva a cabo una identificación de patrones (*pattern recognition*, en inglés). Es decir, se comprueba la distancia en que se encuentra cada una de las muestras del mismo individuo como de las demás.

El objetivo final de este apartado es identificar el modelo que ofrece mejores resultados para poder implementar las probabilidades a posteriori.

En este apartado se desarrollan paralelamente los análisis con variables binarias como con variables continuas en cuatro modelos separados: 1) binario directo, 2) continuo directo, 3) binario sobre matriz de distancias y 4) continuo sobre matriz de distancias. La explicación detallada de dichos análisis se encuentra en el apartado 2.5.2. Diseño experimental del análisis estadístico apartado c.

Para cada modelo se extrae una matriz de distancias concreta: para las variables binarias la matriz de concordancia simple y su distancia complementaria y para las variables continuas la distancia euclidea. Dada la extensión 108x108 de las matrices en los resultados de esta tesis, a continuación se muestra, a modo de ejemplo, un extracto 6x6 de la matriz de distancias simples del análisis 1. En la Figura 4 se puede observar, por ejemplo, que la distancia entre la muestra 1 del autor 1 (C1011) con la muestra 2 (C1012) del mismo autor es de 0,417 mientras que la distancia de la

muestra con muestra siempre es 0. Cuanto más se acerca a 0 más similares son las muestras, contra más se acerca a 1 más distintas.

Figura 4. Ejemplo Matriz de distancias simples.

Medida de coincidencia simple						
Caso	1:C1011	2:C1012	3:C1013	4:C1014	5:C1015	6:C1016
1:C1011	0,000	0,417	0,417	0,250	0,250	0,333
2:C1012	0,417	0,000	0,167	0,333	0,167	0,417
3:C1013	0,417	0,167	0,000	0,500	0,333	0,417
4:C1014	0,250	0,333	0,500	0,000	0,167	0,250
5:C1015	0,250	0,167	0,333	0,167	0,000	0,250
6:C1016	0,333	0,417	0,417	0,250	0,250	0,000

Una vez obtenidas las matrices de distancias se debe calcular el promedio de distancias y la desviación típica asociada de cada una de las muestras a su autor y al resto. De este modo se generan 36 variables, 18 variables Med y 18 variables SD. De este modo obtenemos una matriz de datos 108x36. A continuación en la Figura 5, se muestra un ejemplo de matriz de distancias euclídeas medias.

Figura 5. Ejemplo Matriz de distancias euclídeas medias.

Caso	Individuo	Muesta	Med1	SD1
C1011	1	1	14,9715	9,2814
C1012	1	2	12,959	8,9564
C1013	1	3	13,8138	8,0585
C1014	1	4	12,3697	7,3625
C1015	1	5	12,5447	7,5898
C1016	1	6	10,551	5,4888

En la Figura 5 se puede observar, por ejemplo, en Med1 el promedio de distancias para el individuo 1 de sus 6 muestras y en SD1 su correspondiente promedio de las desviaciones estándares.

A continuación con cada par de variables (por ejemplo, Med1 con SD1) se realiza un análisis discriminante, por tanto, un total 18 análisis. De cada análisis discriminante se extrae la "Pertenencia a grupo pronosticada", es decir, el grupo en el que clasifica el análisis a cada una de las muestras y la "Probabilidad de pertenencia al grupo", es decir, la probabilidad que le asigna el análisis a cada muestra de pertenecer al grupo en el que le ha clasificado (probabilidad a posteriori). Una muestra se va a clasificar en aquel autor en el que la probabilidad sea mayor, es decir, si la muestra C1011 se clasifica en el autor 1, es porque la probabilidad de clasificarse en ese autor es mayor que la probabilidad de clasificarse en cualquier otro autor.

De este modo, se generan nuevas matrices de trabajo sobre la probabilidad de autor de las muestras. En los análisis 1 y 2 estas matrices son de dimensión 108x19 (108 muestras x 19 variables: 1Grupo Pronostricado+18 Probabilidad_Posteriori). En los análisis 3 y 4 las matrices son de dimensión 108x36 (108 muestras x 36 variables: 18 Grupo Pronostricado+18 Probabilidad_Posteriori). De nuevo, dada la extensión de las matrices se incluye únicamente a continuación un extracto de la matriz correspondiente al análisis 4.

Figura 6. Ejemplo matriz de grupo pronosticado y probabilidad a posteriori asociada.

Caso	Individuo	Muestra	Grupo_Pornosticado1	Probabilidad_Posteriori1_1
C1011	1	1	1	0,955
C1012	1	2	1	0,919
C1013	1	3	1	0,764
C1014	1	4	1	0,791
C1015	1	5	1	0,813
C1016	1	6	1	0,305

En la Figura 6 se muestra en la columna individuo el autor real de la muestra y en el grupo pronosticado el autor más probable según el estadístico de haber producido la muestra así como su probabilidad en la columna probabilidad_posteriori. En este caso, los resultados son muy satisfactorios puesto que todas las muestras se han clasificado correctamente en el autor 1 y con una probabilidad muy próxima a 1.

Sobre esta matriz sobre la probabilidad de grupo de las muestras se obtienen los resultados finales que se muestran a continuación en los apartados 4.2.1 y 4.2.2. Mediante el análisis de las clasificaciones obtenidas se pretende comprobar:

1) en qué autor se clasifican las muestras teniendo en cuenta los siguientes estadísticos verdaderos positivos (VP), falsos negativos (FN), falsos positivos (FP) y verdaderos negativos (VN);

2) la validez de las clasificaciones mediante los conceptos de sensibilidad y especificidad;

3) cuánto de probables son estas clasificaciones, es decir, la razón de verosimilitud o LR[28].

Se debe tener en cuenta que se considera que las muestras de un autor están bien clasificadas, es decir, que ese autor está bien definido por sus muestras, cuando no solo el mayor número de muestras está bien clasificado (VP), sino que además se alcanza el valor máximo de LR+ (>1000) y el valor mínimo de LR-(0). Cuantos más requisitos cumplen las muestras de un autor, mejor clasificado (o discriminado) queda.

[28] Se puede encontrar una ampliación de los conceptos anteriores en el apartado 2.5.2 subapartado *d).*

4.2.1. Clasificación variables binarias

Se han realizado dos análisis discriminantes mediante variables binarias: el directo (Tabla 19) y, el análisis binario (Tabla 20). Se destacan en negrita los resultados más satisfactorios de LR.

Los resultados del análisis discriminante directo con variables binarias arrojan 3 autores con LR+ máximo y 2 autores de LR- mínimo pero ninguno de ellos reúne ambas condiciones.

El análisis de proximidades a partir de variables binarias muestra 1 autor con LR+ máximo y 7 autores con LR- mínimo. Ninguno recoge ambas condiciones aunque cabe destacar que el autor C116 se encuentra bastante próximo.

Tabla 19. Variables binarias discriminante directo.

ID	C101	C102	C103	C104	C105	C106	C107	C108	C109	C110	C111	C112	C113	C114	C115	C116	C117	C118	TOTAL
Verdadero Positivo	4	2	4	5	3	3	4	3	3	2	4	3	2	5	6	4	5	6	68
Falso positivo	2	2	0	2	0	2	2	4	2	1	1	1	1	5	6	0	6	3	40
Falso negativo	2	4	2	1	3	3	2	3	3	4	2	3	4	1	0	2	1	0	40
Verdadero negativo	100	100	102	100	102	100	100	98	100	101	101	101	101	97	96	102	96	99	62
SENSIBILIDAD	0.67	0.33	0.67	0.83	0.50	0.50	0.67	0.50	0.50	0.33	0.67	0.50	0.33	0.83	1.00	0.67	0.83	1.00	0.63
ESPECIFICIDAD	0.98	0.98	1.00	0.98	1.00	0.98	0.98	0.96	0.98	0.99	0.99	0.99	0.99	0.95	0.94	1.00	0.94	0.97	0.61
LR+	34.00	17.00	**>1000**	42.50	**>1000**	25.50	34.00	12.75	25.50	34.00	68.00	51.00	34.00	17.00	17.00	**>1000**	14.17	34.00	1.61
LR -	0.34	0.68	0.33	0.17	0.50	0.51	0.34	0.52	0.51	0.67	0.34	0.50	0.67	0.18	**0.00**	0.33	0.18	**0.00**	0.61

Tabla 20. Variables binarias discriminante a partir de proximidades.

ID	C101	C102	C103	C104	C105	C106	C107	C108	C109	C110	C111	C112	C113	C114	C115	C116	C117	C118	TOTAL
Verdadero Positivo	5	2	5	6	4	6	4	3	6	0	4	6	2	5	6	6	6	5	81
Falso positivo	6	6	2	2	2	3	3	1	6	0	4	4	2	5	7	1	4	3	61
Falso negativo	1	4	1	0	2	0	2	3	0	6	2	0	4	1	0	0	0	1	27
Verdadero negativo	96	96	100	100	100	99	99	101	96	102	98	98	100	97	95	101	98	99	41
SENSIBILIDAD	0.83	0.33	0.83	1.00	0.67	1.00	0.67	0.50	1.00	0.00	0.67	1.00	0.33	0.83	1.00	1.00	1.00	0.83	0.75
ESPECIFICIDAD	0.94	0.94	0.98	0.98	0.98	0.97	0.97	0.99	0.94	1.00	0.96	0.96	0.98	0.95	0.93	0.99	0.96	0.97	0.40
LR+	14.17	5.67	**42.50**	51.00	34.00	34.00	22.67	51.00	17.00	**>1000**	17.00	25.50	17.00	17.00	14.57	102.00	25.50	28.33	1.25
LR -	0.18	0.71	0.17	**0.00**	0.34	**0.00**	0.34	0.50	**0.00**	1.00	0.35	**0.00**	0.68	0.18	**0.00**	**0.00**	**0.00**	0.17	0.62

4.2.2. Clasificación variables continuas

Análogamente a los análisis realizados con las variables binarias, se realizan dos análisis discriminates con las variables continuas, un análisis directo (Tabla 21) y, el análisis continuo reflejado en la Tabla 22.

El análisis discriminante directo con variables continuas muestra 8 autores con LR+ máximo y 5 autores con LR- mínimos pero, tan solo 1 autor (C116) reúne ambas condiciones y otros dos están bastante próximos (C117 y C118).

El análisis de proximidades a partir de variables continuas indica que 5 autores poseen un LR+ máximo y 10 autores un LR- mínimo. De todos estos, 4 reúnen ambas condiciones y son C104, 107, 109 y 108.

Tabla 21. Variables continuas discriminante directo.

ID	C101	C102	C103	C104	C105	C106	C107	C108	C109	C110	C111	C112	C113	C114	C115	C116	C117	C118	TOTAL
Verdadero Positivo	6	3	4	4	2	3	6	5	5	2	5	3	5	5	3	6	6	6	79
Falso positivo	2	0	0	0	0	0	3	7	0	1	4	1	3	0	6	0	1	1	29
Falso negativo	0	3	2	2	4	3	0	1	1	4	1	3	1	1	3	0	0	0	29
Verdadero negativo	100	102	102	102	102	102	99	95	102	101	98	101	99	102	96	102	101	101	73
SENSIBILIDAD	1.00	0.50	0.67	0.67	0.33	0.50	1.00	0.83	0.83	0.33	0.83	0.50	0.83	0.83	0.50	1.00	1.00	1.00	0.73
ESPECIFICIDAD	0.98	1.00	1.00	1.00	1.00	1.00	0.97	0.93	1.00	0.99	0.96	0.99	0.97	1.00	0.94	1.00	0.99	0.99	0.72
LR+	51.00	**>1000**	**>1000**	**>1000**	**>1000**	**>1000**	34.00	12.14	**>1000**	34.00	21.25	51.00	28.33	**>1000**	8.50	**>1000**	102.00	102.00	2.57
LR -	**0.00**	0.50	0.33	0.33	0.67	0.50	**0.00**	0.18	0.17	0.67	0.17	0.50	0.17	0.17	0.53	**0.00**	**0.00**	**0.00**	0.38

Tabla 22. Variables continuas discriminante a partir de proximidades.

ID	C101	C102	C103	C104	C105	C106	C107	C108	C109	C110	C111	C112	C113	C114	C115	C116	C117	C118	TOTAL
Verdadero Positivo	6	1	4	6	1	6	6	4	6	3	5	3	6	2	6	6	6	6	83
Falso positivo	3	2	2	0	0	2	0	7	0	3	9	8	3	2	6	6	4	0	57
Falso negativo	0	5	2	0	5	0	0	2	0	3	1	3	0	4	0	0	0	0	25
Verdadero negativo	99	100	100	102	102	100	102	95	102	99	93	94	99	100	96	96	98	102	45
SENSIBILIDAD	1.00	0.17	0.67	1.00	0.17	1.00	1.00	0.67	1.00	0.50	0.83	0.50	1.00	0.33	1.00	1.00	1.00	1.00	0.77
ESPECIFICIDAD	0.97	0.98	0.98	1.00	1.00	0.98	1.00	0.93	1.00	0.97	0.91	0.92	0.97	0.98	0.94	0.94	0.96	1.00	0.44
LR+	34.00	8.50	34.00	**>1000**	**>1000**	51.00	**>1000**	9.71	**>1000**	17.00	9.44	6.38	34.00	17.00	17.00	17.00	25.50	**>1000**	1.38
LR -	**0.00**	0.85	0.34	**0.00**	0.83	**0.00**	**0.00**	0.36	**0.00**	0.52	0.18	0.54	**0.00**	0.68	**0.00**	**0.00**	**0.00**	**0.00**	0.52

Utilizado las variables de clasificación y probabilidad de cuatro análisis y testando las clasificaciones obtenidas mediante los calculos de sensibilidad, especificidad, LR+ y LR-, se observa que los modelos de clasificación obtenidos con variables continuas arrojan mejores resultados que los análisis realizados con variables binarias. No solo porque el número de autores con 6 muestras bien clasificadas (es decir VP) sea mayor 2 y 7 en los análisis de variables binarias, frente a 5 y 10 en sus paralelos continuos, sino porque si se tienen en consideración los parámetros de testeo calculados, sobre todo los LR+ y LR- arrojan también un mejor resultado para los de variables continuas.

Cabe destacar que estos resultados eran esperables, ya que la aproximación con variables continuas conserva más información. De acuerdo con estos criterios, se puede resumir cómo se evalúan las clasificaciones de los análisis expuestos en los puntos 4.2.1 y 4.2.2 en la Tabla 23 .

Tabla 23. Evaluación de las clasificaciones.

Análisis	Autores con 6 muestras VP	Autores con máximo LR+	Autores con mínimo LR-	Autores con máximo LR+ y mínimo LR-
Binario directo	2	3	2	0
Continuo directo	5	8	5	1
Binario distancias	7	1	7	0
Continuo distancias	10	5	10	4

Adicionalmente, en un intento de resumir toda la información de las tablas de clasificación (Tabla 19 a Tabla 22) de forma gráfica se pueden utilizar gráficos de dispersión. La expresión de resultados mediante gráficos resulta de especial interés en los juicios puesto que los hallazgos resultan más fáciles de entender tanto por los agentes judiciales como por las personas involucradas que no necesariamente deben tener conocimientos de estadística.

A continuación se muestra el Gráfico 44 y el Gráfico 45correspondientes al análisis 4 de distancias a partir de variables continuas ya que es el que ha obtenido los resultados mejores.

Cada posición en X representa un individuo, en el eje Y se refleja la probabilidad de cada una de sus muestras de ser de ese individuo. En el Gráfico 44 se observan en color verde las muestras que se clasifican correctamente a su autor (VP) y en color rojo las muestras que no se clasifican a su autor FN). De este modo se observa de forma gráfica la sensibilidad de este método, es decir, la probabilidad de detectar que la muestra se clasifique en el autor que le corresponde.

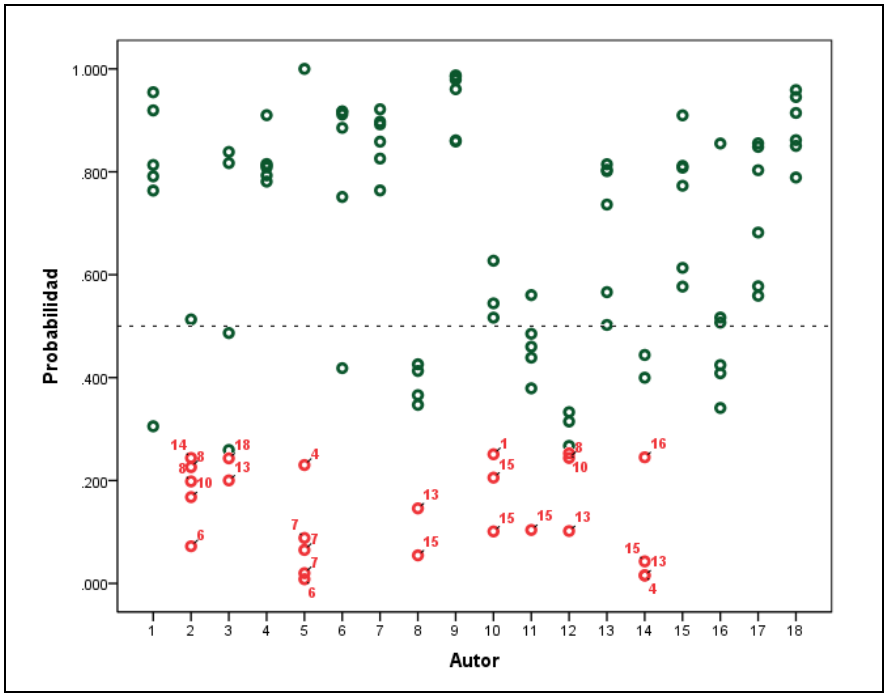

Gráfico 44. Clasificación con VP y FN.

Con respecto a la sensibilidad del método, se observa que existe un potencial de clasificación correcta del 76,85% y que, además, en el 60% de los casos la probabilidad de clasificación es superior al 50%. Cabe destacar también, que todos los casos de falso negativo se encuentran por debajo de una probabilidad del 25% y, que todos los verdaderos positivos se encuentran por encima de esa probabilidad.

En el Gráfico 45 se ilustra en color verde las muestras que se clasifican correctamente a su autor (VP), en color naranja las muestras que se clasifican en un grupo dado sin pertenecer a él (FP)

y el grupo real al que debería pertenecer la muestra y, la probabilidad de clasificación de todas las muestras.

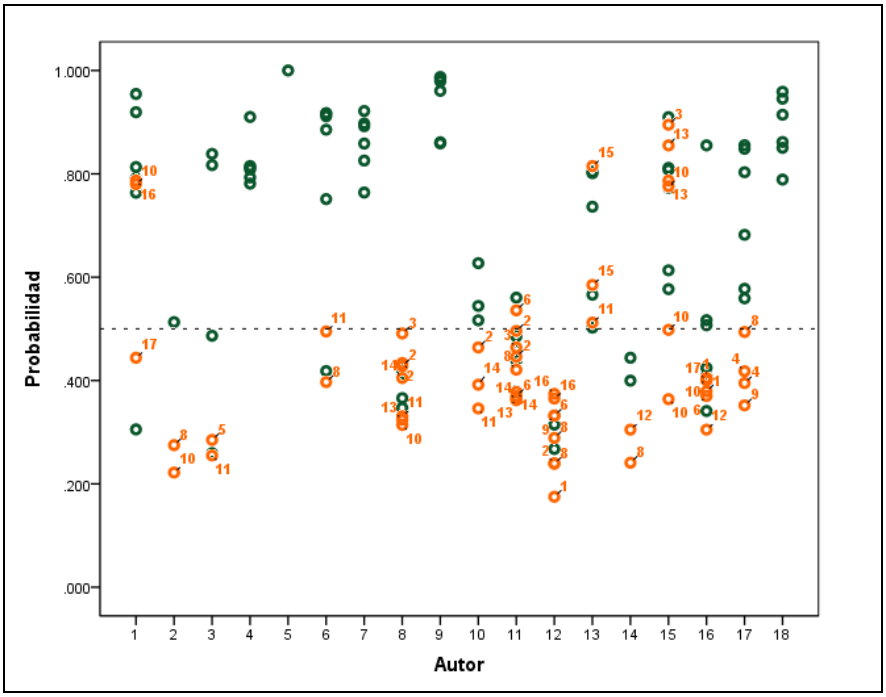

Gráfico 45. Clasificación con VP y FP.

Por lo que concierne a la especificidad del método, en este caso también se encuentran los verdaderos positivos en las probabilidades más altas. Además, la mayor parte de los verdaderos positivos (77,94%) se encuentra por encima del 50% de probabilidad y la mayor parte de los falsos positivos (82,45%) se encuentra por debajo de ese porcentaje. No obstante, los resultados no son tan satisfactorios como en el caso de los de la sensibilidad del método ya que se encuentran falsos positivos por encima del 25% e incluso del 50%.

Por tanto, en lo que a las clasificaciones se refiere parece que es el análisis con variables continuas a partir de proximidades el que ofrece mejores resultados. Teniendo en cuenta que las clasificaciones obtenidas en los análisis con variables binarias son menos satisfactorias, dichas variables quedan descartadas para estudiar las probabilidades a posteriori.

4.3. Probabilidades de clasificación a posteriori

La probabilidad de que una muestra propia sea adjudicada a su autor, debe ser mayor que la probabilidad de clasificación en el resto de autores. En otras palabras, se espera que los 6 textos de cada autor tengan una probabilidad muy alta en un rango minv-maxv (todos ellos) y que todos los demás 17x6 estén en un rango de probabilidades de pertenecer a este grupo minf-maxf que no esté contenido en el anterior, esto es maxf<minv. Cuanto mayor sea esta desigualdad, menor es el solapamiento.

Para comprobar con qué probabilidad se puede clasificar bien o mal a un individuo, se calculan los intervalos de las probabilidades de clasificación con los valores máximos y mínimos de las probabilidades del mismo autor frente a los valores mínimos y máximos de las probabilidades de ser clasificado en cualquier otro autor.

Resulta de especial interés que los intervalos no se solapen, o si lo hacen, que sea lo menos posible. Para medir el grado de solapamiento, es decir, en qué medida un intervalo contiene al otro, se calcula la diferencia entre el máximo de los dos mínimos y el mínimo de los dos máximos. No obstante, el verdadero interés reside en observar aquellos resultados en clave de no solapamiento, es decir, el complementario (1-solapamiento).

Se realiza una primera clasificación, CLAS A, teniendo en cuenta las muestras bien clasificadas (VP) y se calcula el peso como la proporción de muestras bien clasificadas en cada grupo (VP) sobre el total de muestras del grupo, 6.

Se realiza una segunda clasificación, CLAS B, ponderando el no solapamiento con el peso. Es decir, teniendo en cuenta los conceptos, tanto las muestras bien clasificadas como el solapamiento de los intervalos de probabilidades de clasificación.

Se establece como *criterio de máxima verosimilitud* aquel autor que tiene sus 6 muestras bien clasificadas, es decir, peso=1 y sus intervalos no están solapados, es decir, no solapamiento=2.

Por tanto el valor de CLAS B de máxima verosimilitud es 200, de acuerdo con este criterio, en la medida en que el valor de CLAS B de un autor ser aproxime a 200, será un autor mejor clasificado, es

decir, sus muestras se parecerán mucho y serán muy distintas a las demás.

4.3.1. Resultados discriminante directo

En la Tabla 24 se proyectan los resultados de los intervalos obtenidos a partir de las probabilidades a posteriori de los análisis discriminantes directos de las variables continuas.

Por un lado, en el caso de CLAS A, se indica en color verde que todas las muestras están bien clasificadas, en azul que se han clasificado correctamente 5 de 6 muestras, en color naranja se indica una clasificación débil con 4 muestras de 6 bien clasificadas y en color rojo una mala clasificación correspondiente a 3 o menos muestras bien clasificadas.

Por otro lado, en el caso de CLAS B, el color verde indica que todas las muestras están bien clasificadas y no hay solapamientos (>100), en color azul aquellos casos en que todas las muestras están bien clasificadas pero hay algunos solapados (75-100), en color naranja se destaca una clasificación débil ya que existe un solapamiento de los intervalos (50-75) y, en color rojo se muestra una mala clasificación (<50).

Tabla 24. Intervalos obtenidos a partir de las probabilidades a posteriori mediante análisis discriminante directo.

	Probabilidad del propio autor		Probabilidad de cualquier otro autor				**CLAS A**				**CLAS B**
Individuo	Mín.	Máx.	Mín.	Máx.	Def Mín: Máx de mín.	Def Max: Mín de máx	Número de muestras bien clasificadas (VP)	Peso	Solapa-miento	No Solapamien-to: distancia	No Solapa-miento Ponderado: puntación final
1	0.526	0.999	0.000	0.473	0.526	0.473	6	1.000	-0.053	1.053	**105.29**
2	0.001	1.000	0.000	0.779	0.001	0.779	3	0.500	0.778	0.222	**11.08**
3	0.007	1.000	0.000	0.480	0.007	0.480	4	0.667	0.473	0.527	**35.12**
4	0.085	1.000	0.000	0.671	0.085	0.671	4	0.667	0.586	0.414	**27.58**
5	0.120	1.000	0.000	0.537	0.120	0.537	2	0.333	0.416	0.584	**19.45**
6	0.071	0.926	0.000	0.639	0.071	0.639	3	0.500	0.567	0.433	**21.63**
7	0.641	0.989	0.000	0.122	0.641	0.122	6	1.000	-0.518	1.518	**151.84**
8	0.007	0.754	0.000	0.963	0.007	0.754	5	0.833	0.747	0.253	**21.10**
9	0.295	0.999	0.000	0.673	0.295	0.673	5	0.833	0.377	0.623	**51.88**
10	0.000	0.892	0.000	0.993	0.000	0.892	2	0.333	0.891	0.109	**3.62**
11	0.065	0.929	0.000	0.403	0.065	0.403	5	0.833	0.338	0.662	**55.20**

12	0.095	0.952	0.000	0.720	0.095	0.720	3	0.500	0.625	0.375	**18.77**
13	0.472	0.674	0.000	0.501	0.472	0.501	5	0.833	0.029	0.971	**80.95**
14	0.053	1.000	0.000	0.455	0.053	0.455	5	0.833	0.402	0.598	**49.82**
15	0.294	0.712	0.000	0.698	0.294	0.698	3	0.500	0.404	0.596	**29.80**
16	0.606	1.000	0.000	0.269	0.606	0.269	6	1.000	-0.337	1.337	**133.69**
17	0.982	1.000	0.000	0.009	0.982	0.009	6	1.000	-0.973	1.973	**197.29**
18	0.984	1.000	0.000	0.010	0.984	0.010	6	1.000	-0.975	1.975	**197.46**
Máxima Verosimilitud	**1.000**	**1.000**	**0.000**	**0.000**	**1.000**	**0.000**	**6**	**1.000**	**-1.000**	**2.000**	**200.00**

En el caso de CLAS A, se destacan en color verde cinco casos en que todos los textos (6/6) del autor están bien clasificados y se obtienen un total de cinco autores; cinco autores se muestran en color azul con 5 muestras de 6 bien clasificadas; se observan dos autores en color naranja con una clasificación débil, es decir, se atribuyen 4 de sus 6 muestran a un autor; y, en color rojo se marcan los cuatro casos con una mala clasificación cuando el número de muestras clasificadas es igual o inferior a 3. En la Tabla 25 se observan de forma resumida los resultados de CLAS A.

Tabla 25. Resumen intervalos discriminante directo CLAS A.

CLAS A: Nº de grupos	
5	Verde: Todos bien clasificados 6/6
5	Azul: Bien Clasificados 5/6
2	Naranja: Clasificación débil 4/6
4	Rojo: Mala clasificación (≤3)/6

En el caso de los resultados CLAS B, en color verde se observan 5 grupos con una buena clasificación y sin solapamientos (>100); en color azul se observa un caso en que todos los textos están bien clasificados con poco solapamiento (75-100); en color naranja aparecen dos casos reflejando de este modo una clasificación débil debido al solapamiento de los intervalos (50-75); y, finalmente, en color rojo aparecen 10 casos con una mala clasificación (<50).

Estos resultados se muestran de forma resumida en la Tabla 26.

Tabla 26. Resumen intervalos discriminante directo CLAS B.

CLAS B: Nº de grupos	
5	Verde: Todos bien clasificados y no solapados (>100)
1	Azul: Todos Bien Clasificados y poco solapados (75-100)
2	Naranja: Clasificación débil, debido al solapamiento de los intervalos (50-75)
10	Rojo: Mala clasificación (<50)

4.3.2. Resultados discriminante a partir de distancias

En Tabla 27 se proyectan los intervalos obtenidos a partir de las probabilidades a posteriori mediante análisis discriminante a partir de distancias con variables continuas.

Tabla 27. Intervalos obtenidos a partir de las probabilidades a posteriori mediante análisis discriminante a partir de distancias.

	Probabilidad del propio autor		Probabilidad de cualquier otro autor				**CLAS A**				**CLAS B**
Individuo	Mín.	Máx.	Mín.	Máx.	Def Mín: Máx de mín.	Def Max: Mín de máx.	Número de muestras bien clasificadas	Peso	Solapa-miento	No Solapamien-to: distancia	No Solapa-miento Ponderado: puntación final
1	0.305	0.955	0.000	0.405	0.305	0.405	6	1.000	0.099	0.901	**90.08**
2	0.072	0.513	0.000	0.496	0.072	0.496	1	0.167	0.423	0.577	**9.61**
3	0.200	0.839	0.000	0.895	0.200	0.839	4	0.667	0.639	0.361	**24.10**
4	0.781	0.910	0.000	0.418	0.781	0.418	6	1.000	-0.363	1.363	**136.28**
5	0.008	1.000	0.000	0.285	0.008	0.285	1	0.167	0.276	0.724	**12.06**

6	0.418	0.917	0.000	0.536	0.418	0.536	6	1.000	0.118	0.882	**88.24**
7	0.764	0.922	0.000	0.077	0.764	0.077	6	1.000	-0.687	1.687	**168.68**
8	0.054	0.426	0.000	0.494	0.054	0.426	4	0.667	0.371	0.629	**41.91**
9	0.859	0.987	0.000	0.352	0.859	0.352	6	1.000	-0.507	1.507	**150.74**
10	0.101	0.627	0.000	0.780	0.101	0.627	3	0.500	0.526	0.474	**23.69**
11	0.104	0.560	0.000	0.512	0.104	0.512	5	0.833	0.408	0.592	**49.33**
12	0.102	0.333	0.000	0.305	0.102	0.305	3	0.500	0.203	0.797	**39.83**
13	0.502	0.815	0.000	0.855	0.502	0.815	6	1.000	0.312	0.688	**68.77**
14	0.015	0.444	0.000	0.405	0.015	0.405	2	0.333	0.390	0.610	**20.34**
15	0.577	0.910	0.000	0.815	0.577	0.815	6	1.000	0.239	0.761	**76.14**
16	0.341	0.855	0.000	0.787	0.341	0.787	6	1.000	0.446	0.554	**55.43**
17	0.559	0.856	0.000	0.444	0.559	0.444	6	1.000	-0.115	1.115	**111.51**
18	0.789	0.959	0.000	0.065	0.789	0.065	6	1.000	-0.724	1.724	**172.39**
Máxima Verosimilitud	**1.000**	**1.000**	**0.000**	**0.000**	**1.000**	**0.000**	**6**	**1.000**	**-1.000**	**2.000**	**200.00**

En la columna CLAS A se observan 10 autores en color verde con una clasificación correcta total (6/6); 1 autor en color azul con 5 muestras bien clasificadas; en color naranja 2 autores con una clasificación débil; y, en color rojo 4 autores ya que tienen 3 o menos muestras bien clasificadas.

En la columna CLAS B se muestran 5 autores en color verde reflejando que todos los textos están bien clasificados y no solapados; en color azul 3 autores ya que todos sus textos están bien clasificados y poco solapados; en color naranja, 3 autores con una clasificación débil; y, finalmente, en color ojo 7 autores con una mala clasificación.

En la Tabla 28 se muestra un resumen de los resultados de ambos análisis.

Tabla 28. Resumen intervalos discriminante distancias CLAS A y CLAS B.

CLAS A: Nº de grupos	
10	Verde: Todos bien clasificados 6/6
1	Azul: Bien Clasificados 5/6
2	Naranja: Clasificación débil 4/6
4	Rojo: Mala clasificación (<3)/6
CLAS B: Nº de grupos	
5	Verde: Todos bien clasificados y no solapados (>100)
3	Azul: Todos Bien Clasificados y poco solapados (75-100)
3	Naranja: Clasificación débil, debido al solapamiento de los intervalos (50-75)
7	Rojo: Mala clasificación (<50)

4.4. Resumen y discusión

Los resultados de los análisis realizados en este capítulo se resumen en la Tabla 29. En esta tabla se incluye el tipo de análisis llevado a cabo, sus resultados de clasificación y las probabilidades a posteriori obtenidas.

Tabla 29. Resumen resultados sobre los modelos de clasificación.

Análisis	Clasificación	Probabilidades
Análisis discriminante directo con variables binarias	3 grupos con LR+ máximo	
	2 grupos de LR- mínimo	
	ninguno que reúna ambas condiciones	
Análisis discriminante directo con variables continuas	8 grupos con LR+ máximo	**CLAS A**
	5 grupos de LR- mínimo	5 grupos Verde: Todos bien clasificados 6/6
	1 grupo de ellos reúne ambas condiciones y otros dos bastante próximos	5 grupos Azul: Bien Clasificados 5/6
		2 grupos Naranja: Clasificación débil 4/6
		4 grupos Rojo: Mala clasificación (<3)/6
		CLAS B
		5 grupos Verde: Todos bien clasificados y no solapados (>100)
		1 grupos Azul: Todos Bien Clasificados y poco solapados (75-100)
		2 grupos Naranja: Clasificación débil, debido al solapamiento de los intervalos (50-75)

		10 grupos Rojo: Mala clasificación (<50)
Análisis de distancias a partir de variables binarias	1 grupo con LR+ máximo	
	7 grupos de LR- mínimo	
	Ninguno reúne ambas condiciones	
Análisis de distancias a partir de variables continuas	5 grupos con LR+ máximo	**CLAS A**
	10 grupos de LR- mínimo	10 grupos Verde: Todos bien clasificados 6/6
	4 grupos de ellos reúnen ambas condiciones	1 grupos Azul: Bien Clasificados 5/6
		2 grupos Naranja: Clasificación débil 4/6
		4 grupos Rojo: Mala clasificación (<3)/6
		CLAS B
		5 grupos Verde: Todos bien clasificados y no solapados (>100)
		3 grupos Azul: Todos Bien Clasificados y poco solapados (75-100)
		3 grupos Naranja: Clasificación débil, debido al solapamiento de los intervalos (50-75)
		7 grupos Rojo: Mala clasificación (<50)

Teniendo en cuenta los resultados de las clasificaciones obtenidas en los análisis y las probabilidades a posteriori, se puede concluir que los mejores resultados se obtienen con el análisis de distancias a partir de variables continuas.

De este modo el análisis estadístico que se propone para futuras implementaciones de la razón de verosimilitud en la comparación forense de textos escritos es:

1) Obtención de la matriz de distancias euclídeas mediante un análisis cluster de las variables discriminantes.

2) Obtención de la matriz de distancias euclídeas medias mediante el cálculo del promedio de distancias (MED) y su desviación típica asociada (SD).

3) Generación una nueva matriz a partir de las variables "pertenencia a grupo pronosticada" y "Probabilidad de pertenencia al grupo" del análisis discriminante de cada par MED y SD.

4) Obtención de los resultados de clasificación (VP, FN, FP y VN) a partir de la matriz anterior.

5) Comprobación del método mediante los conceptos de sensibilidad (cociente entre VP y la suma VP+FN) y especificidad (cociente entre VN y la suma VN+FP),

6) Aplicación de la razón de verosimilitud mediante LR+ y LR-.

7) Cálculo de los intervalos de las probabilidades de clasificación a posteriori.

Capítulo 5:

Conclusiones

En el capítulo 5 se lleva a cabo, por un lado, un sumario de las principales conclusiones sobre los resultados obtenidos en esta tesis doctoral; y, por otro lado, se ponen de manifiesto las principales contribuciones de la misma a la comunidad científica y las futuras líneas de investigación.

5. Conclusiones

La prueba pericial presentada ante los tribunales hasta el momento ha sido valorada, por una lado, por la comunidad científica quien debía marcar los estándares para la validación de los métodos científicos aplicados a la pericial; y, por otro lado, por los agentes judiciales quienes se centraban en valorar las cuestiones más de procedimiento y las cualificaciones de los firmantes de la pericial. En ningún momento, se ha creído que los jueces debían valorar la prueba según los estándares exigidos por el método científico sino que debían atribuirle, exclusivamente, un valor jurídico. Este modo de valoración de la prueba pericial es consecuencia de la concepción tradicional de que la ciencia y el derecho son entes no relacionados.

Con el planteamiento de la necesidad de nuevos estándares científicos en la sentencia del caso Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993 (509 U.S. 579) y la propuesta del paradigma de la razón de verosimilitud, está relación distante entre la ciencia y el derecho debe necesariamente estrecharse. La valoración de las periciales en un juicio no puede ni debe realizarse a partir de la puesta en escena de las partes y de los mismos peritos sino que la decisión judicial debe basarse en unos estándares de metodología científica robustos y que sean comprensibles por los jueces. Además, el juez debe poder verificar la validez científica de la información ya que dicha información constituirá parte de la base

de su decisión sobre los hechos. Como se indica en las Reglas Federales, el juez tiene la tarea de garantizar que el testimonio del perito se basa en los estándares de validez científica acordados y aceptados por la comunidad científica.

Esta tesis doctoral tenía como principal objetivo realizar un primer acercamiento de la metodología de la comparación forense de textos escritos en español al resto de ciencias forenses que aplican actualmente la relación de verosimilitud. Una de las principales dificultades para la implementación de técnicas estadísticas en lingüística forense es el número de muestras y, por tanto, la construcción de distribuciones poblacionales de variables lingüísticas. Por este motivo, este estudio se planteó superar este obstáculo real y diseñar un diseño estadístico que pueda trabajar con un número relativamente bajo de muestras y de autores para aplicar la razón de verosimilitud.

Los objetivos propuestos al inicio de esta tesis se han alcanzado puesto que se ha conseguido 1) el diseño de una propuesta metodológica para la selección y el análisis de variables lingüísticas con potencial discriminatorio; 2) se ha creado una distribución poblacional de variables lingüísticas en texto escrito reducida para poder ser comparable a una situación real; y 3) se ha conseguido diseñar un método estadístico con el que poder implementar la razón de verosimilitud.

Finalmente, se ha establecido que la comparación forense de textos escritos se puede llevar a cabo mediante un análisis estadístico comparable al resto de ciencias forenses. Dicho análisis se basa, por un lado, en las clasificaciones de las muestras de un autor teniendo en cuenta los estadísticos de verdaderos positivos (VP), falsos negativos (FN), falsos positivos (FP) y verdaderos negativos (VN). Por otro lado, dichas clasificaciones se deben validar mediante el cálculo de la probabilidad de detectar muestras propias (sensibilidad) y la probabilidad de muestras que no son propias (especificidad). Finalmente, con todos estos datos se aplica el cálculo de las razones de verosimilitud para obtener la probabilidad de que una muestra propia se asigne a su autor en comparación a que una muestra que no es propia también se le asigne a dicho autor (LR+); y, la probabilidad de que una muestra propia no se le asigne a su autor en comparación a que las muestras ajenas se asignen al resto de los autores (LR-).

De este modo, las muestras de un autor se clasificarán correctamente cuando cumplan los requisitos de: un mayor grupo de muestras que se clasifican en el autor al que pertenecen (VP); un valor máximo de LR+ (>1000); y, un valor mínimo de LR-(0). Cuantos más requisitos se cumplan mayor es la probabilidad de clasificar correctamente las muestras dubitadas de un autor.

Con esta técnica se ha conseguido una clasificación correcta superior al 75% de los casos y el 60% de ellos posee una

probabilidad de clasificación superior al 50%. Cabe destacar además que hay una barrera de sensibilidad en el 25% por la que todos los textos que se clasifican en su verdadero autor se encuentran por encima de ese porcentaje y todos los textos que se clasifican erróneamente a otro autor se encuentran por debajo.

Esta tesis contribuye a la lingüística forense y, en particular, al campo de la comparación forense de textos escritos, ya que la metodología implementada puede ser de gran utilidad en el momento de resolver casos de atribución de autoría o de perfiles lingüísticos. También es una contribución a otras ciencias forenses que pueden implementar el marco de la razón de verosimilitud.

En un plano más detallado, las contribuciones más importantes de esta tesis tienen que ver con su carácter innovador, original y transmisible y con la obtención de unos resultados más fiables. Esta investigación será de gran utilidad para el campo de la comparación forense de textos escritos ya que aporta: a) una base de datos unificada de textos en español peninsular que permite obtener una distribución poblacional de diferentes variables lingüísticas; b) un método estadístico común basado en técnicas de estadística multivariante y el marco de la razón de verosimilitud; c) una primera aproximación a un código de buenas prácticas en la comparación forense de textos escritos puesto que se tienen en cuenta factores de control en el momento de la compilación de los datos, se llevan a cabo procedimientos de muestreo, se

implementan técnicas cuantitativas y se testa la validez de la mismas. Es un nuevo código de buenas prácticas que puede proporcionar resultados más fiables y concluyentes en la atribución de autoría. Y, finalmente, d) esta propuesta metodológica puede ser transferida a otras lenguas y otros sistemas judiciales.

De este modo se puede concluir que esta tesis está en los primeros pasos de poder cumplir con los estándares de la decisión de Daubert en la comparación forense de textos escritos ya que 1) ofrece una metodología probada y que podría ser replicable, 2) no informa un margen de error puesto que el objeto de estudio (la lengua) es muy variable pero sí que puede ofrecer niveles de confianza y, 3) existen unos estándares para controlar la fiabilidad de la técnica. Restaría pendiente que la comunidad de lingüistas forenses acepte la técnica, se realicen publicaciones y que éstas fueran revisadas.

Los resultados de esta tesis pretenden aportar datos al debate sobre la idoneidad del marco de la razón de verosimilitud a la lingüística forense —en concreto, a la comparación forense de textos escritos para la atribución de autoría— y las dificultades de su aplicación. Como ya se ha comentado anteriormente, la metodología de la razón de verosimilitud aplicada en la comparación forense de voz es un buen modelo de partida para la comparación forense de textos escritos. Dichos estudios, comparten la misma casuística por lo que concierne a la dificultad de disponer de datos e insisten en la creación de distribuciones poblaciones de referencia relevantes para las comparaciones forenses.

En esta tesis se aborda el enorme problema de reunir datos de población relevante y, este punto diferencia esta tesis de muchos trabajos anteriores. Una cosa es establecer que una variable estilística es inusual después de una búsqueda en Google y, otra muy distinta es hacerlo después de analizarla en un corpus de escritura de personas de perfiles sociolingüísticos muy similares.

Como líneas futuras la metodología propuesta se debería ampliar a corpus más grandes, a otras lenguas y a otros sistemas judiciales y, además, se debería estudiar la presentación de los resultados numéricos de la razones de verosimilitud ante el tribunal para la correcta comprensión de los mismos.

BIBLIOGRAFÍA

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, *20*(5), 67–75.

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, *26*(2), 7.

Abercrombie, D. (1969). Voice qualities. In N. N. Markel (Ed.), *Psycholinguistics: an Introduction to the Study of Speech and Personality*. London: The Dorsey Press.

Aitken, C., Berger, C. E. H., Buckleton, J. S., Champod, C., Curran, J. M., Dawid, A. P., … Zadora, G. (2011). Expressing evaluative opinions: A position statement. *Science & Justice*, *51*, 1–2.

Aitken, C. G. G., Berger, C. E. H., Buckleton, J. S., Champod, C., Curran, J. M., Dawid, A. P., … Jackson, G. (2011). Expressing evaluative opinions: a position statement. *Science & Justice*, *51*(1), 1–2.

Aitken, C. G. G., & Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists* (Vol. 2). Chichester, England: John Wiley & Sons.

Argamon, S. M., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in fromal written texts. *TEXT*, *23*, 321–346.

Argamon, S. M., Koppel, M., Pennebaker, J., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Commun*, *52*(2), 119–123.

Baayen, R. H., van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, *11*(3), 121–131–121–131.

Bel, N., Queralt, S., Spassova, M. S., & Turell, M. T. (2012). The use of sequences of linguistic categories in forensic written text comparison revisited. In *Proceedings of The International Association of Forensic Linguists' Tenth Biennial Conference* (Vol. Birmingham, pp. 192–209).

Berger, C. E. H., Buckleton, J. S., Champod, C., Evett, I. W., & Jackson, G. (2011). Evidence evaluation: A response to the court of appeal judgment in R v T. *Science & Justice*, *51*(2), 43–49.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics :investigating language structure and use*. Cambridge etc.: Cambridge University Press.

Britain, D. (1991). *Dialect and Space: A Geolinguistic Analisis of Speech Variables in the Fens*. Universidad de Essex.

Burrows, J. F. (1987). Word patterns and story shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, *2*, 61–70.

Burrows, J. F. (2003). Questions of Authorship: Attribution and Beyond A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York. *Computers and the Humanities*, *37*(1), 5–32.

Camacho Rosales, J. (2002). *Estadística con SPSS (versión 11) para Windows*. Madrid : RA-MA.

Campoy, J. M. H., & Almeida, M. (2005). *Metodologia de la investigacion sociolingüistica*.

CEH, B. (2010). Criminalistiek is terugredeneren [Criminalistics is reasoning backwards]. *Nederlands Juristen Blad*, 784–789.– 784–789.

Chambers, J. K. (1995). *Sociolinguistic theory: linguistic variation and its social significance*. Cambridge, Mass.: Blackwell.

Champod, C., Baldwin, D., Taroni, F., & Buckleton, J. S. (2003). Firearm and tool marks identification: the Bayesian approach. *AFTE JOURNAL*, *35*(3), 307–316.

Chaski, C. E. (1996). *Lingusitic methods of determining authorship*. Naschville, TN.

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, *8*, 1–65.

Cicres Bosch, J. (2007). Aplicació de l'Anàlisi de l'entonació i de l'alienació tonal a la identificació de parlants en fonètica forense.

Committee on Identifying the Needs of the Forensic Sciences Community, N. R. C. (2009). Strengthening Forensic Science in the United States: A Path Forward.

Cook, R., Evett, I. W., Jackson, G., Jones, P. J., & Lambert, J. A. (1998). A hierarchy of propositions: deciding which level to address in casework. *Science & Justice*, *38*(4), 231–239.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *Forensic Lingusitics*, *1*, 26–43.

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, *25*(4), 431–447.

Coulthard, M., Grant, T., & Kredens, K. (2011). Forensic Linguistics. *The SAGE Handbook of Sociolinguistics*, *36*, 529–545.

Coulthard, M., & Johnson, A. (2007). *An Introduction of Forensic Linguistics: language in evidence*. London and New York: Routledge.

Curran, J. M., Champod, T. N. H., & Buckleton, J. S. (2000). *Forensic interpretation of glass evidence*. CRC.

Davis, L. M. (1990). *Statistics in dialectology*. Tuscaloosa, Ala. : University of Alabama Press.

De Vel, O., Anderson, A. M., Corney, M., & Mohay, G. (2002). E-mail authorship attribution for computer forensics. *Applications of Data Mining in Computer Security*, *6*.

Delgado, C. (2001). Comentarios sobre el contexto actual de la identificación forense de locutores.

Dreher, J. J., & Young, E. (1969). Chinese author identification by segment distribution. In L. Dolezel & R. W. Bailey (Eds.), *Statistics and Style (Mathematical Linguistics and Automatic Language Processing, 6)*. New York: Elsevier.

Eckert, P. (1997). Age as a sociolinguistic variable. In F. Coulmas (Ed.), *The Handbook of Sociolinguistics* (pp. 151–167). Oxford: Basil Blackwell.

Ellegard, A. (1962). *A statistical method for determining authorship: The Junius Letters*. Gothenburg: University of Gothenburg.

Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice, 38*(3), 198–202.

Evett, I. W., & Weir, B. S. (1998). *Interpreting DNA evidence:statistical genetics for forensic scientists*. Sunderland, MA: Sinauer Associates.

Exteberria, J., Joaristi, L., & Lizasoaín, L. (1990). *Programación y análisis estadísticos básicos con SPSS/PC*. Madrid : Paraninfo.

Farrell, M. G. (1993). Daubert v. Merrell Dow Pharmaceuticals, Inc.: Epistemilogy and Legal Process. *Cardozo L.Rev., 15*, 2183.

Fenton, N. (2011). Science and law: Improve statistics in court. *Nature*, *479*(7371), 36–37.

Fenton, N., & Neil, M. (2012). On limiting the use of Bayes in presenting forensic evidence, 1–27.

Flick, U. (1998). *An Introduction to qualitative research*. London: SAGE.

Gains, J. (1999). Electronic mail - a new style of communication or just a new medium? An investigation into the text features of e-mail. *English for Specific Purposes*, *18*(1), 81–101.

Gamon, M. (2004). Linguistic correlates of style: Authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics*, 611–617.

Gavaldà Ferré, N. (2011). Sociolingüística de la variació i lingüística forense. *Llengua, Societat I Comunicació*, (9), 49–59.

Gibbons, J., & Turell, M. T. (2005). *Dimensions of forensic linguistics*. Amsterdam: John Benjamins.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic

speaker recognition. *Odyssey 2004: The Speaker and Language Recognition Workshop Odyssey-04 Odyssey 2004: The Speaker and Language Recognition Workshop*, *20*(2–3), 331–355. http://doi.org/10.1016/j.csl.2005.08.005

Grant, T. (2007). Quantifying evidence for forensic authorship analysis. *International Journal of Speech Language and the Law*, *14*(1), 1–25.

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons & M. T. Turell (Eds.), *Dimensions of Forensic Linguistics* (pp. 215–229). Amsterdam; Philadelphia: John Benjamins.

Grant, T., & Baker, K. L. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *International Journal of Speech, Language and the Law*, *8*(1), 66–79.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, *22*(3), 251–270.

Gummesson, E. (1991). *Qualitative methods in management research*. Newbury Park (Calif.): Sage.

Gumperz, J. J., & Hymes, D. H. (1986). *Directions in sociolinguistics: the ethnography of communication*. Oxford: Basil Blackwell.

Guy, G. (1980a). Variation in the group and in the individual. In W. Labov (Ed.), *Locating language in time and space* (pp. 1–36). Nueva York: Academic Press.

Guy, G. (1980b). Variation in the group and the individual. In W. Labov (Ed.), *Locating language in time and space* (pp. 1–36). New York: Academic Press.

Guy, G. (1993). The quantitative analysis of linguistic variation. In D. R. Preston (Ed.), *American Dialect Research* (pp. 223–249). Amsterdam y Filadelfia: John Benjamins.

Halliday, M. A. K., & Hasan, R. (1989). Language, context, and text: Aspects of language in a social-semiotic perspective.

Hänlein, H. (1999). *Studies in authorship recognition : a corpus-based aprroach*. Frankfurt am Main: Peter Lang.

Hepler, A. B., Saunders, C. P., Davis, L. J., & Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International*.

Hirst, G., & Feiguina, O. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, *22*(4), 405–417.

Holmes, D. I. (1994). Authorship Attribution. *Computers and the Humanities*, *28*(2), 87–106.

Holmes, D. I. (2003). Stylometry and the Civil war: the case of the Pickett letters. *Chance*, *16*(2), 18–25.

Hoover, D. L. (2003a). Another perspective on vocabulary richness. *Computers and the Humanities*, *37*(2), 151–178.

Hoover, D. L. (2003b). Frequent Collocations and Authorial Style. *Literary and Linguistic Computing*, *18*(3), 261–286.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer.

Johnstone, B. (1996). *The linguistic individual: Self-expression in language and linguistics*. Oxford University Press, USA.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, *1*(3), 233–334.

Juola, P. (2007). Future trends in authorship attribution. In P. Craiger & S. Shenoi (Eds.), *Advances in Digital Forensics III* (pp. 119–132).

Juola, P. (2008). *Authorship Attribution*. Hanover: Now Publishers.

Juola, P., & Baayen, R. H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, *20*(Suppl), 59–67.

Koehler, J. J. (2013). Linguistic confusion in Court: Evidence from the forensic sciences. *Brooklyn Law School's Journal of Law & Policy*, *21*(2), 515–540.

Koppel, M., Akiva, N., & Dagan, I. (2007). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information and Science Technology*, *57*(11), 1519–1525.

Koppel, M., Schler, J., & Argamon, S. M. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*(1), 9–26.

Kredens, K. (2001). Language Corpora in Forensic Linguistics. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC'99-Practical Applications in Language Corpora*. Main: Peter Lang.

Labov, W. (1963). The social motivation of a sound change. *Word, 19*, 273–309.

Labov, W. (1966). *The Social stratification of english in New York City*. Cambridge: Cambridge University Press.

Labov, W. (1972). *Sociolinguistic patterns*. Oxford: Basil Blackwell.

Labov, W. (1982). Perspectives on historical linguistics. In W. P. Lehmann & Y. Malkiel (Eds.), (pp. 79–92). Amsterdam: Benjamin.

Labov, W. (1994). *Principles of linguistic change*. Oxford: Blackwell.

Labov, W. (2003). Some sociolinguistic Principles. In C. B. Paulston & G. R. Tucker (Eds.), *Sociolingusitics: The Essential Readings* (pp. 234–250). Oxford: Basil Blackwell.

Lavid, J. (2005). *Lenguaje y nuevas tecnologías : nuevas perspectivas, métodos y herramientas para el lingüística del siglo XXI*. Madrid: Cátedra.

LePage, R. B., & Tabouret-Keller, A. (1985). *Acts of identity: Creole-based approaches to language and ethnicity.* Cambridge: Cambridge UP.

López Morales, H. (1994). *Métodos de investigación lingüística*. Salamanca: Ediciones Colegio de España.

Love, H. (2002). *Attributing authorship :an introduction*. Cambridge: Cambridge University Press.

Mannion, D., & Dixon, P. (1997). Authorship attribution: the case of Oliver Goldsmith. *The Statistician*, *46*, 1–18.

McDermott, S. D., Willis, S. M., & McCullough, J. P. (1999). The evidential value of paint. Part II: a Bayesian approach. *Journal of Forensic Sciences, 44*, 263–269.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (Vol. 2). Edinburgh: Edinburg University Press.

McMenamin, G. R. (1993). *Forensic Stylistics*. Amsterdam: Elsevier.

McMenamin, G. R. (2001). Style markers in authorship studies. *International Journal of Speech Language and the Law*, *8*(2), 93–97.

Meuwly, D., & Drygajlo, A. (2001). Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM). En *2001: A Speaker Odyssey-The Speaker Recognition Workshop*.

Milroy, J. (1983). On the Sociolinguistic History of /h/-dropping in English. In M. A. Davenport, E. Hansen, & H. F. Nielsen (Eds.), *Current Topics in English Historical Linguistics* (pp. 37–57). Odense: Odense University Press.

Milroy, L. (1987). *Observing and analysing natural language : a critical account of sociolinguistic method*. Oxford: Basil Blackwell.

Milroy, L., & Gordon, M. (2003). *Sociolinguistics: Method and Interpretation*. Oxford: Basil Blackwell.

Miret, P., Salvador, A., Serracant, P., & Soler, R. (2008). Enquesta a la Joventut de Catalunya 2007. *Generalitat de Catalunya, Departament d'Acció Social I Ciutadania, Secretaria de Joventut.*

Moreno Fernández, F. (1990). *Metodología sociolingüística*. Madrid: Gredos.

Morrison, G. S. (2009a). Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework. *Australian Journal of Forensic Sciences*, *41*(2), 155–161.

Morrison, G. S. (2009b). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*(4), 298–308.

Morrison, G. S. (2012). The likelihood-ratio framework and forensic evidence in court: A response to R v T. *The international journal of evidence & proof*, *16*(1), 1–29.

Mortera, J. U. L. I. A., Tre, U. R., & Dawid, A. P. (2006). Probability and evidence.

Neumann, C., Evett, I. W., & Skerrett, J. (2012). Quantifying the weight of evidence from a forensic fingerprint comparison: a

new paradigm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(2), 371–415.

Nolan, F. (1983). *The Phonetic bases of speaker recognition*. Cambridge etc.: Cambridge University Press.

Nordgaard, A., & Rasmusson, B. (2012). The likelihood ratio as value of evidence—more than a question of numbers. *Law, Probability and Risk*.

Orebaugh, A., & Allnutt, J. (2009). Classification of Instant Messaging: Communicatinos for Forensic Analysis. *The International Journal of Forensic Computer Science*, *4*(1), 22–28.

Ortí, A. (1999). La confrontación de modelos y niveles epistemológicos en la génesis e historia de la investigacion social. In J. M. Delgado & J. Gutiérrez (Eds.), *Métodos y técnicas cualitativas de investigacion en Ciencias sociales* (pp. 85–95). Madrid: Síntesis.

Paul, H. (1889). *Principles of the history of language*. New York: MacMillan.

Picornell-García, I. (2014). La aplicación de la atribucion de autoria en la investigacion e inteligencia: la aplicación practica (y su problematica). In E. Garayzábal-Heinze, M. Jimenez-Bernal,

& M. Reigosa-Riveiros (Eds.), *Lingüística Forense: la Lingüistica en el ámbito Legal y Policial* (pp. 79–94). Euphonia Ediciones.

Providers, A. of F. S. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, *49*(3), 161–164.

Queralt, S., Spassova, M. S., & Turell, M. T. (2011). L'ús de les combinacions de seqüències de categories gramaticals com a nova tècnica de comparació forense de textos escrits. *LSC-Llengua, Societat I Comunicació*, *9*, 59–67.

Queralt, S., & Turell, M. T. (2012). Testing the discriminatory potential of sequences of linguistic categories (n-grams) in Spanish, Catalan and English Corpora. In *The regional conference of the International Association of Forensic Linguists* (Vol. Kuala Lump).

Raghvan, S., Kovashka, A., & Mooney, R. (2010). Authorship Attribution Using Probabilistic Context-Free Grammars. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 38–42).

Rangel, F., & Rosso, P. (2013a). On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style. In CEUR-WS.org (Ed.), *Proceedings of*

the ESSEM Workshop on Emotion and Sentiment in Social and Expressive Media (pp. 34–46).

Rangel, F., & Rosso, P. (2013b). Use of Language and Author Profiling: Identification of Gender and Age. In *Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science* (p. 177).

Redmayne, M., Roberts, P., Aitken, C. G. G., & Jackson, G. (2011). Forensic science evidence in question. *Criminal Law Review.5*.

Rico-Sulayes, A. (2011). Statistical Authorship Attribution of Mexican Drug Trafficking Online Forum Posts. *International Journal of Speech, Language and the Law*, *18*(1), 53–74.

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2011). Extending the confusion about Bayes. *The Modern Law Review*, *74*(3), 444–455.

Rose, P. (2002). *Forensic speaker identification*. London; New York: Taylor & Francis.

Rudman, J. (1998). The state of Authorship Attribution Studies: Some problems and solutions. *Computers and the Humanities*, *31*, 351–365.

Ruiz Olabuénaga, J. I. (1999). *Metodología de la investigación cualitativa*. Bilbao: Universidad de Deusto.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892–895.

Sankoff, G. (1980). A quantitative paradigm for the study of communicative competence. In G. Sankoff (Ed.), *The Social Life of Language* (pp. 295–310). Filadelfia: University of Pennsylvania Press.

Schmied, J. (1993). Qualitative and quantitative research approaches to English relative constructions. *Souter and Atwell*, *1993*, 85–96.

Sevigny, M. (1981). Triangulated Inquiry-A Metjodology for the analysis of classroom interaction. In J. Green & C. Wallat (Eds.), *Ethnography and Language in Educational Settings* (pp. 65–85). Norwood, N.J.: Ablex.

Skerrett, J., Neumann, C., & Mateos-Garcia, I. (2011). A Bayesian approach for interpreting shoemark evidence in forensic casework: Accounting for wear features. *Forensic Science International*, *210*(1), 26–30.

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*, 145–152.

Smith, W. (1994). Computers, statistics and disputed authorship. In J. Gibbons (Ed.), *Language and the Law*. New York: McGraw-Hill.

Solan, L., & Tiersma, P. M. (2005). *Speaking of crime: the language of criminal justice*. Chicago: University of Chicago Press.

Spassova, M. S. (2009). El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español.

Spassova, M. S., & Grant, T. (2008). Categorizing Spanish Written Texts by Author Gender and Origin by Means of Morpho-Syntactic Trigrams: some observations on method's feasibility of application for linguistic profiling. In *Curriculum, language and the law Inter-University Centre* (Vol. University).

Spassova, M. S., & Turell, M. T. (2007). The use of morpho-syntactically annotated tag sequences as forensic markers of authorship attribution. In M. T. Turell, M. S. Spassova, & J. Cicres (Eds.), *Proceedings of the Second European IAFL Conference on Forensic Linguistics, Language and the Law* (pp. 229–237). Barcelona: Publicacions de l'IULA.

Stamatatos, E. (2009a). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556.

Stamatatos, E. (2009b). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001a). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, *35*(2), 193–214.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001b). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, *35*, 193–214.

Svartvik, J. (1968). *The Evans statements*. University of Goteburg.

Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge University Press.

Thompson, W. C. (2012). Bad cases make bad law: Reactions to R v T. *Law, Probability and Risk*.

Trudgill, P. J. (1975). Sociolinguistics and Scots Dialects. In J. D. Mc Clure (Ed.), *The Scots Language in Education: Association for Scottish Literary Studies. Occasional Ppaers No3.* (pp. 28–34). Association for Scottish Literary Studies.

Trudgill, P. J. (1983a). *On dialect: social and geographical perspectives*. New York: New York University Press.

Trudgill, P. J. (1983b). *Sociolinguistics : an introduction to language and society*. Harmondsworth: Penguin.

Trudgill, P. J. (1986). *Dialects in contact*. (Basil Blackwell, Ed.). New York.

Turell, M. T. (2004a). Textual kidnapping revisited: the case of plagarism in literary translation. *International Journal of Speech Language and the Law*, *11*(1), 1–26.

Turell, M. T. (2004b). The disputed authorship of electronic mail: linguistic, stylistic and pragmatic markers in short texts. In *First European IAFL Conference on Forensic Linguistics, Language and Law* (Vol. Cardiff Un).

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech Language and the Law*, *17*(2), 211–250.

Turell, M. T. (2011). La tasca del lingüísta detectiu en casos de deteccio de plagi i determinacio d'autoria de textos escrits. *Llengua, Societat I Comunicació*, *9*, 67–83.

Turell, M. T., & Gavaldà Ferré, N. (2013). *Towards an Index of Idiolectal Similitude (or Distance) in Forensic Authorship Analysis*. *Brooklyn Law School's Journal of Law & Policy*.

Woolls, D., & Coulthard, M. (2007). Tools for the trade. *International Journal of Speech Language and the Law*, *5*(1), 33–57.

Wright, D. (2013). Stylistic variation within genre conventions in the enron email corpus. *The International Journal of Speech, Language and the Law*, *20*(1), 45–75.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378–393.

Zheng, R., Qin, Y., Huang, Z., & Chen, H. (2003). Authorship analysis in cybercrime investigation. In *ISI'03 Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics* (pp. 59–73).

Zipf, G. K. (1932). Selected studies of the principle of relative frequency in language.

CASOS CITADOS

Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993 (509 U.S. 579)

Carmichael v. Samyang Tire, Inc., 131 F.3d 1433 (1997)

Kumho Tire Co. v. Carmichael 526 U.S. 137 (1999)

Lapper v R 2005 (NZCA 259)

Regina v GK 2001 (NSWCCA 413)

R v Berry y Wenitong 2007 (VSCA 202)

R v Doheny y Adams 1997 (1 Cr App R 369, CA)

Anexos

I. ANEXO 1: Lista de palabras malsonantes

A) Se incluyen como palabras malsonantes:

Imbécil, putaditas, cabrón, puta/puto, mierdas, jodas, cojones, cabronazo, ratero de mierda, gilipollas, chorizo de mierda, malditos, grandísimo desgraciado, jodida fianza, idiota, zorrón de mucho cuidado, zorra, cabronaza, cabrear, jodido, maldita fulana, follamigos, follar, mierdas, tocar los cojones, descojonar, mamón, hijo de puta, me suda la polla, coño, te corto los huevos, meter por el culo, pedazo de cabrón, polla, manta de hostias, maricón, joder, cabreado, somanta de hostias.

B) No se incluyen como palabras malsonantes:

Insultos del tipo "ladrón", "tonta"; palabras vulgares como "chucho"; miserable mundo, miserable y asqueroso céntimo, estafador, maloliente, malcriadas, puñetera, sinvergüenza, suelta de cuidado, inútil, cornuda, caraculo, pimpollo, desgraciado, pervertido, criminal, bastardo, escoria repulsiva, cerdo, depravado, escoria, energúmeno, chupasangres.

II. ANEXO 2: Estadísticos descriptivos de la distribución poblacional

Tabla 30. Variables de complejidad.

Variables de complejidad	Media	SD	Mín.	Máx.	Mediana	Moda
Número de palabras	382,76	188,16	24	716	421,50	540
Número de palabras distintas	192,78	80,10	23	351	211,00	241
Número de oraciones	20,05	11,16	3	56	21,00	23
Número de párrafos	4,78	3,10	1	13	4,00	1
Palabras por frase	20,21	6,90	8,00	53,43	19,29	13,00
Palabras por párrafo	180,09	159,40	24,00	576,00	107,20	49,00
Ratio type-token	,55	,10	,41	,96	,50	,49

Tabla 31. Variables léxicas.

Variables léxicas		Media	SD	Mín.	Máx.	Mediana	Moda
Número de palabras malsonantes		1,07	2,79	0	20	,00	0
Número de errores	Errores ortografía	,77	2,20	0	16	,00	0
	Errores diacríticos	1,90	3,30	0	22	,50	0
	Errores mayúsculas	,15	,46	0	3	,00	0
	Errores de puntuación	,24	,63	0	4	,00	0
	Erratas	,43	,74	0	5	,00	0
	Errores pleonasmos	,04	,19	0	1	,00	0
	Errores gramaticales	,60	1,25	0	9	,00	0

	Errores por contacto de lenguas	,58	1,09	0	5	,00	0
Expresión del futuro	Tiempo verbal de futuro	3,37	3,25	0	18	2,00	2
	Perífrasis verbal *ir a*	2,24	3,33	0	20	1,00	0
Expresión de la obligación	*Tener que* + infinitivo	,74	1,02	0	5	,00	0
	Deber + infinitivo	,52	,86	0	4	,00	0
	Haber de + infinitivo	,05	,22	0	1	,00	0
Expresión de la condición	*Si*	2,55	2,27	0	10	2,00	0
	Como	,26	,65	0	4	,00	0
Expresión del pretérito imperfecto del subjuntivo	Terminación - *era*	,99	1,44	0	8	,00	0
	Terminación - ese	,14	,49	0	4	,00	0
Abreviaturas	euros	,19	,39	0	1	,00	0
	€	,10	,30	0	1	,00	0
	EUR	,01	,08	0	1	,00	0
Acortar palabras		,01	,16	0	2	,00	0

Tabla 32. Variables pragmáticas.

Tipo	**Variable**	**Media**	**SD**	**Mín.**	**Máx.**	**Moda**
Intensificación del sujeto 1º persona del singular	Ausencia del *yo*	14,30	9,11	0	35	12
	Presencia del *yo*	1,29	1,69	0	10	0
Expresión del énfasis	Uso de mayúsculas	,45	1,42	0	11	0
	Uso de signos de puntuación	,12	,48	0	3	0
	Uso de la repetición	,07	,27	0	2	0

Número de preguntas		1,38	2,25	0	12	0
Número de exclamaciones		,59	1,49	0	11	0
Trato (Address form)	Formal	2,74	3,78	0	15	0
	Informal	1,26	2,00	0	10	0
Salutación	Greetings	,74	,44	0	1	1
	Buenos días/tardes	,09	,29	0	1	0
	Name:	,02	,15	0	1	0
	Hola	,15	,36	0	1	0
	Querido	,12	,33	0	1	0
	Estimado	,23	,42	0	1	0
	Señor	,09	,28	0	1	0
	Sr	,01	,11	0	1	0
	Apreciado	,02	,13	0	1	0
	BuenosXseñorX	,01	,08	0	1	0
	A…	,02	,13	0	1	0
Despedida	Farewells	,45	,50	0	1	0
	Un saludo	,02	,15	0	1	0
	Saludos cordiales	,01	,08	0	1	0
	Att	,00	,00	0	0	0
	Hasta nunca	,02	,13	0	1	0
	Atentamente	,21	,41	0	1	0
	Esperando pronto una respuesta	,00	,00	0	0	0
	Un cordial saludo	,00	,00	0	0	0
	Espero	,02	,13	0	1	0
	Saludos	,01	,11	0	1	0
	Gracias	,02	,13	0	1	0
	Muchas gracias	,02	,15	0	1	0
	Atentamente lo saluda	,01	,08	0	1	0

	Cordialmente	,02	,13	0	1	0
	Muchas gracias por su atención	,02	,13	0	1	0
	Muy cordialmente	,01	,08	0	1	0
	Reciba un cordial saludo	,02	,13	0	1	0
	Estamos en contacto	,01	,11	0	1	0
	Me despido atentamente	,01	,08	0	1	0
	Con todos mis respetos	,01	,08	0	1	0
	Que tenga un buen día	,01	,08	0	1	0
	Salutaciones	,01	,08	0	1	0
	Sin otro cometido, se despide atentamente	,01	,11	0	1	0
	Hasta luego	,01	,08	0	1	0
Palabras entre paréntesis		,37	,86	0	5	0
Intensificación del sujeto 1º persona del plural	Ausencia	1,09	2,86	0	20	0
	Presencia	,07	,36	0	3	0
Marcadores discursivos	¿verdad?	,20	,53	0	3	0
	¿no?	,12	,49	0	4	0
	¿entiendes? ¿entendido?	,02	,15	0	1	0
	¿eh?	,02	,19	0	2	0
	¿sabes?	,02	,19	0	2	0
	¿no es cierto?	,03	,17	0	1	0
	¿no te parece?	,07	,86	0	11	0
	¿no es así?	,02	,19	0	2	0
	¿no crees?	,02	,19	0	2	0

	¿me sigue?	,01	,08	0	1	0
	¿sabes qué?	,02	,13	0	1	0
Uso de guiones		,02	,15	0	1	0
Signos de puntuación	Aparece apertura y cierre de los signos de exclamación	,52	,50	0	1	1
	Solo aparece cierre de los signos de exclamación	,12	,33	0	1	0

Tabla 33. Variables sintácticas.

Variables sintácticas		**Media**	**SD**	**Mín.**	**Máx.**	**Mediana**	**Moda**
Tipo de oración	Simple	2,71	3,18	0	16	1,00	1
	Compleja	42,37	27,32	1	120	39,00	12
Tipo de oración compleja	Yuxtapuesta	4,46	6,15	0	61	3,00	1
	Coordinada	10,32	6,02	0	28	10,00	9
	Subordinada	39,74	20,09	2	94	41,00	47
Tipo de oración yuxtapuesta	Coma	3,52	3,96	0	20	2,00	1
	Dos puntos	,26	,55	0	3	,00	0
	Punto y coma	,07	,30	0	2	,00	0
Tipo de oración coordinada	*y/e*	7,02	4,18	0	17	6,50	6
	o/u	,68	1,03	0	6	,00	0
	ni	,27	,64	0	3	,00	0
	pero	1,80	1,61	0	7	1,00	1
Tipo de oración subordinada	Relativas introducidas por *que*	4,25	3,21	0	14	4,00	1
	Relativas introducidas por *cual*	,05	,23	0	1	,00	0