Universidad Nacional Autónoma de México

Facultad de Filosofía y Letras
Colegio de Letras Hispánicas

Análisis estilométrico para la detección de plagio

T E S I S

que, para obtener el título de
Licenciado en Lengua y Literaturas Hispánicas,
presenta

Alejandro Rosas González

Asesor: Dr. Gerardo Eugenio Sierra Martínez

Ciudad Universitaria, 2011

A mis padres; mis primeros y mejores maestros.

GRANIZA TÁNTO, como para que yo recuerde
y acreciente las perlas
que he recogido del hocico mismo
de cada tempestad.

César Vallejo

El escritor, claro, usa una lengua, pero al usarla
ya no es la lengua de todos, sino el habla suya.

Enrique Amberson Imbert

Agradecimientos

A mis padres. Soy lo que soy gracias a ustedes. Los amo.

Mamá: Esta tesis busca la justicia porque de ti aprendí eso. Gracias.

Papá: Esta tesis está bien escrita y trata del estilo porque de ti aprendí eso. Gracias.

Papás: Amo la lengua porque de ustedes dos aprendí eso. Muchas gracias.

A mi hermano. Eres mi cómplice, mi apoyo, mi mejor amigo. Eres el mejor hermano que se puede tener. Gracias por todo.

A toda mi familia. No recuerdo a algún miembro de mi familia no me haya ayudado en algo alguna vez y tampoco me imagino alguno que no se alegrará de verme terminar mi carrera. Gracias.

A mi asesor, el Dr. Gerardo Sierra Martínez. Nadie creyó en mí como tú; a veces ni yo mismo. Has sido no solo mi tutor y mi maestro, sino mi amigo en mis dos vidas: la académica y la personal. Me enseñaste que en un grupo de trabajo no hay que preocuparse cuando te piden más y mejores cosas, sino cuando ya no te piden que las hagas. Nunca lo voy a olvidar. Gracias.

A la Dra. Fernanda López Escobedo. Eres, extraoficialmente, mi co-asesora y ha sido un honor trabajar contigo desde el principio. Gracias por tanta ayuda. Nos resta mucho por hacer; ¡cuenta conmigo!

A la Mtra. Margarita Palacios Sierra. Maestra, es usted una inspiración no solo para mí, sino para todos quienes la conocen. Espero, sinceramente, llegar a mi madurez y tener algo de su entereza y sabiduría. Gracias.

Al Dr. Alfonso Medina Urrea. Alfonso, gracias por la confianza que implicó decirme: "Me parece que no eres tan wey; que tienes cierto nivel". Espero que siempre tengas esa impresión de mí.

Al Mtro. Javier Cuétara Priede. Javier, nunca voy a olvidar que gracias a ti conocí la fonética, la ingeniería lingüística y la lingüística forense. Gracias por todo.

A Claudia. Merci par m'avoir rendu le Pausilippe et la mer d'Italie, la fleur qui plaît tant à mon coeur désolé, et la treille où la pampre à la rose s'allie. Merci car je continue à rêver dans la grotte où nage la **sirène**. Je t'aime.

A mi tío Jorge. Gracias por poner en mis manos, por primera vez, la herramienta que al parecer nunca dejaré: la computadora.

A mis compañeros y amigos. Brenda Castro, Irasema Cruz, Jorge Lázaro, Teresita Reyes, Octavio Sánchez, César Antonio Aguilar, Rodrigo Alarcón, Antonio Reyes, Víctor Mijangos, Pavel Soriano, Iria da Cunha, Josh Careaga, Azury Aparicio, Luis Cabrera, Jessica Méndez, José Luis Vieyra, Paulina Aguado y a todos quienes son y han sido parte del Grupo de Ingeniería Lingüística. Esta tesis tiene algo de todos ustedes. ¡GIL, baby!

A los proyectos DGAPA-IN403108 y CONACyT-82050 por haber financiado la finalización de mis estudios y la realización de esta, mi tesis de licenciatura.

A la máxima casa de estudios y mi segunda casa, la UNAM, por haberme dado, de manera gratuita, educación que vale oro; azul y oro. ¡México, Pumas, Universidad!

Índice

Índice de figuras

Índice de tablas

Introducción

La lengua está en todas partes. La inevitable e inquisidora pregunta: ¿de qué vas a vivir si estudias lingüística? actualmente no es más que un cuestionamiento anacrónico. Hoy en día existen caminos nuevos y emocionantes, aunque sinuosos, que el lingüista está en posibilidad de recorrer. La lingüística computacional, la neurolingüística, la adquisición de lenguaje y, más recientemente, la lingüística forense son solo algunos ejemplos de opciones interesantes y viables para ejercer la profesión de la lengua.

En el ámbito de las ciencias forenses cada vez es más necesaria la participación de la lingüística. Ésta, a su vez, se abre cada vez más a técnicas, métodos y herramientas estadísticos y computacionales que no solo optimizan el procesamiento de datos y la obtención de resultados, sino que confieren cierto grado de confianza a los mismos. Así pues, la interacción del conocimiento lingüístico y el conocimiento estadístico y computacional puede ayudar a resolver cuestiones legales de manera mucho más completa de lo que sería si actúan de manera independiente.

Una de las formas más comunes en que se aprovecha tal interacción dentro del ámbito legal es la emisión de dictámenes u opiniones expertas por parte de profesionales de la lengua. Entre las distintas áreas con las que pueden estar relacionados tales dictámenes se encuentra la detección de plagio; es decir, la localización de copia en lo sustancial de obras ajenas, dándolas como propias. Para tal fin, existen diversos sistemas computacionales que representan un apoyo importante en los trabajos forenses. Estas herramientas realizan búsquedas de similitud textual con base en la localización de fragmentos textuales determinados;

sin embargo, estas búsquedas no resultan suficientes si la copia se ha realizado por medio de la paráfrasis, por ejemplo. En este punto el tema cobra importancia y se hace necesario realizar investigación al respecto.

El Grupo de Ingeniería Lingüística (GIL) realiza, desde octubre de 2008, investigación relacionada con detección de plagio y clasificación automática de documentos. Uno de los métodos que han sido puestos a prueba para detectar similitud entre dos textos es la estilometría.

El objetivo de esta tesis es aplicar el método de la estilometría para analizar y clasificar el estilo lingüístico y literario; es decir las elecciones lingüísticas que hacen los autores y que son independientes del contenido. Por medio de este método se obtendrán datos estadísticos que representan determinados aspectos del estilo de los autores, mismos que pueden ser comparados para establecer el grado de similitud entre dos textos.

La tesis está organizada de forma que se vaya de lo general a lo particular. Así pues, después de esta breve introducción, en el segundo apartado se presentará una explicación general de lo que es la lingüística forense, seguida de la exposición de cuáles son las áreas que la conforman y de una explicación de cómo interactúan con ella diversas áreas, tales como la lingüística, el derecho y la ingeniería lingüística. En el tercer apartado se expondrá, brevemente, el problema que representa el plagio en la sociedad y se desarrollará, más ampliamente, qué es el plagio, los niveles y tipos de plagio existentes y las técnicas actualmente utilizadas para su detección. Uno de esos métodos es la estilometría, que tendrá dedicado el apartado cuarto para explicar qué es y cómo funciona, además de argumentar a favor de su utilización en casos de búsqueda de similitud. El quinto apartado, por su parte, contendrá la puesta a prueba del método de la estilometría

en un caso concreto de detección de plagio, así como un reporte detallado de la forma en que se llevó a cabo este experimento, de cómo se conformó el corpus utilizado y de la prueba denominada de "consistencia de estilo" de los autores. Finalmente, el apartado sexto incluirá las conclusiones obtenidas después de la realización de esta investigación y los posibles trabajos futuros que se planteen a raíz de ella.

Capítulo 1

Lingüística forense

En México el término *lingüística forense* es poco conocido. La primera noción que viene a la mente de quien escucha tal concepto suele estar relacionada con cadáveres e investigaciones policiacas. Esto sucede porque a través de los medios de comunicación, especialmente la televisión, a menudo se nos presentan casos en que los investigadores visitan a quien llaman "el forense" para aclarar algún asunto que puede ser la causa de muerte de alguna víctima, los efectos que provocó una droga en algún detenido o la razón por la que se encontró determinada sustancia en un cuerpo, entre otras cosas. De hecho, el cargo de la persona a quien visitan los investigadores de la televisión es *médico forense* y se encarga de dar opiniones basadas en su experticia médica no solo con respecto a la muerte, sino a diversas cuestiones legales.

Es necesario, entonces, aclarar que la palabra forense no se refiere a muertos, sino a lo perteneciente o relativo a un foro; es decir al "sitio en que los tribunales oyen y determinan las causas" (RAE, 2001), o a la *curia* que es el "conjunto de abogados, escribanos, procuradores y empleados en la Administración de Justicia" (RAE, 2001). Ahora bien, en la misma forma que la medicina, hay una importante cantidad de ciencias que encuentran una forma de aplicación en el ámbito forense: la informática, la psicología, la química y la lingüística, entre otras.

Hecha la aclaración, se puede definir el término *lingüística forense*. Según la IAFL (International Association of Forensic Linguists), se trata de la interfaz entre

lengua y derecho. Esta definición es la más aceptada y difundida, sin embargo hace falta clarificarla para tener un mejor entendimiento del término. Una interfaz es una relación que existe entre diferentes elementos. En este caso, la relación existe "entre el lenguaje, el crimen y la ley, donde la ley incluye la aplicación de la ley, los asuntos judiciales, la legislación, las discusiones o actas en la ley, y aún donde sólo potencialmente se implica alguna infracción a la ley o alguna necesidad para buscar un remedio legal" (Olsson, s.f.). En otras palabras, la *lingüística forense* es la aplicación de conocimiento lingüístico (métodos, herramientas, técnicas, etc.) en cuestiones legales.

1.1 Áreas de la lingüística forense

Turell (2005: 13) habla de una clasificación amplia de la lingüística forense que "cubre todas las áreas en las que la Lengua y el Derecho se interrelacionan", y de una clasificación restrictiva que "se refiere especialmente a la utilización de pruebas lingüísticas en los juicios y, por tanto, a la actuación de los lingüistas en contextos jurídicos y judiciales".

En cuanto a la clasificación amplia, la *International Association of Forensic Linguists* (2006) registra tres grandes áreas de la disciplina:

- el lenguaje jurídico y legal (*Language of the Law*), que se ocupa de la redacción de leyes, del análisis del lenguaje administrativo y jurídico, y de la interpretación de textos públicos oficiales.

- el lenguaje del proceso legal (*Language of the Legal Process*), que se encarga del análisis del discurso que se produce en diversos contextos, tales como los interrogatorios policiales, la interacción lingüística que se produce entre los diferentes actores que intervienen en un juicio.
- el lenguaje de evidencia o probatorio (*Language as Evidence*), que atiende al uso de evidencia lingüística de diversas índoles en la comparación forense de textos orales y escritos para poder alcanzar diferentes objetivos de la práctica forense.

En lo concerniente a la segunda forma de clasificación, la restrictiva, se considera ocho principales campos de acción de la lingüística forense: la autenticidad de declaraciones, el significado e interpretación de leyes y textos legales, el significado de contratos y estatutos, la dialectología forense, los problemas de habilidad lingüística, el estado de marcas registradas como palabras o frases en el lenguaje, la fonética forense y la estilística forense.

Para este trabajo, se tomará esta segunda forma de clasificación, ya que tiene por principal característica el enumerar y describir cada uno de los campos de acción de la lingüística forense, además de incluir descripciones de ejemplos y casos reales en que se requieren acciones y análisis lingüísticos específicos. A continuación se enumera y se describe, brevemente, dichas áreas:

1.1.1 Autenticidad de declaraciones

A menudo se presentan casos en que las declaraciones de quienes están envueltos en cuestiones legales no son confiables. Para explicar esto, hay varias causas

posibles: desde la malicia con la que miente un criminal, hasta una confesión que se obtuvo por medio de la tortura o las condiciones de estrés en las que normalmente declaran los implicados. En este caso, la labor del lingüista forense consiste en hacer observaciones acerca de la autenticidad o falsedad de las declaraciones; ello, por supuesto, con base en sus conocimientos de análisis conversacional, análisis del discurso, teoría de actos del habla, psicolingüística, etcétera.

1.1.2 Significado e interpretación de leyes y textos legales

Uno de los problemas más comunes que se presentan en cuestiones legales es la ambigüedad en documentos oficiales. Este problema se debe, principalmente, al hecho de que el lenguaje legal posee particularidades de diversa índole que deriva en dificultad para entender o interpretar las leyes. Las confusiones pueden deberse, también, a la común pluralidad de interpretaciones que se presentan en algunas formas de la lengua; en este sentido, Olsson (s.f.: 8) apunta:

> the work of Roger Shuy, and other US linguists, has encompassed many areas of civil and criminal practice, but right from the beginning, the law itself was, as it were, subject to questioning: what does this law mean? How do different people perform when asked if they 'understand' their rights?

Sin embargo es necesario exponer, también, que gran parte de las situaciones de ambigüedad es provocada por la falta de conocimiento de técnicas adecuadas de redacción y escritura, por parte de quienes elaboran este tipo de escritos. Levi

(1993) reporta un caso en que testificó acerca de la complejidad sintáctica que presentaba una notificación, cuyo propósito era informar a un ciudadano la forma en que debería reclamar ciertos beneficios. La misiva estaba tan mal escrita que su objetivo, el de informar acerca de los derechos de una persona, no fue alcanzado, de tal forma que el afectado, con ayuda de un testigo experto en lingüística, hizo un reclamo a la corte remitente de la notificación. Levi encontró rasgos sintácticos que probablemente interfirieron con el entendimiento del texto; por ejemplo, múltiples negativos, integraciones complejas, nominalizaciones confusas, verbos pasivos sin sujeto y combinaciones difíciles de operadores lógicos como *y*, *o*, *si* y *a menos que* (Coulthard & Johnson, 2007).

1.1.3 Significado de contratos y estatutos

Tiersma y Solan (2002) hablan del significado de contratos, estatutos y pólizas de seguro como espacios de aplicación para la lingüística forense. En el caso de estos textos es común que se presente dificultad de entendimiento y ambigüedad en el contenido. Cuando uno de estos problemas se presenta, es necesario acudir a un experto en el uso y el significado del lenguaje para contar con una opinión confiable, aunque cabe señalar que el mismo lingüista deberá poner en claro que ser "experto en el significado del lenguaje" no siempre quiere decir que se tenga la última palabra en lo que a una interpretación se refiera; por el contrario, la mayoría de veces, esto significa, simplemente, avalar que es posible interpretar algún texto de varias formas. Así pues, los lingüistas actúan sólo como guía en pasajes difíciles, usando análisis lingüístico para explicar cómo es que varias

interpretaciones de un texto legal son posibles; de tal forma, es más probable que sea aceptado su testimonio por parte de las cortes que si tratan de decirle a los jueces lo que significa un texto legal (Tiersma & Solan, 2002).

1.1.4 Dialectología Forense

En la película *Los dioses deben estar locos* aparece un suceso en el que un nativo africano es acusado de robo y daño en propiedad ajena por haber perseguido y matado a una gacela. El problema fue que la gacela pertenecía al ganado de un ciudadano, y el lugar donde aconteció el hecho era el rancho de dicho residente. El juez, por medio de un intérprete improvisado, pregunta al nativo si él mató a la gacela, a lo que el aborigen responde que sí, pero que no entiende por qué eso es una razón para que lo hayan llevado a un lugar oscuro y con barrotes, tomando en cuenta que toda su vida él había hecho eso sin más consecuencias que acabar con el hambre que aquejaba a su familia. El juez hace caso omiso de esta declaración y, sin rodeos, le pregunta si se declara culpable de los cargos. Ante esa pregunta, el intérprete expresó al juez su impotencia para hacer esa pregunta al acusado, a razón de que no encontraba forma de traducir la palabra *culpable* a la lengua del nativo. Al final, la condena fue de encarcelamiento por un determinado periodo de tiempo que el infractor pasó dentro de la carcel sin saber por qué razón lo habían encerrado. Como cita Olsson (s.f.) de Gibbons (1994), "*the…system…around interrogation in the courtroom is alien to Aboriginal culture*". Este es un ejemplo típico de un caso en el que se requiere de un lingüista para que realice la labor de intérprete especializado; es decir, que además de traducir, posea

determinados conocimientos en lo que a cuestiones legales se refiere; de esa forma, será capaz de ayudar en asuntos que, por diferencias dialectales, sean difíciles de entender, resolver o comunicar.

1.1.5 Problemas de habilidad lingüística

Esta área de la lingüística forense atiende a la pregunta: ¿es capaz tal o cual persona de entender una advertencia policiaca? A diferencia de las cuestiones dialectales, los problemas de habilidad lingüística pueden presentarse entre personas que hablan el mismo idioma y el mismo dialecto. En este caso, las causas de situaciones tales como la falta de entendimiento de una advertencia policiaca o la aceptación de algún cargo sin entender las implicaciones pueden ser neurológicas o incidentales. Una persona con algún problema lingüístico, tal como una afasia, una lesión, o retraso mental, puede estar impedida para captar el significado de lo que escucha o de lo que lee; es decir, posee una habilidad lingüística limitada, lo cual obstaculiza su participación en cualquier asunto legal. El declarante, además, puede ser analfabeto, hablante de otra lengua, muy joven o tener una desventaja étnico-lingüística. De la misma forma, un individuo que es sometido a un arresto, a un secuestro, a una violación o a cualquier situación que genere niveles elevados de estrés, puede ver afectada su habilidad para comprender o producir lenguaje en cualquiera de sus manifestaciones. Así lo explica Olsson (s.f.) cuando dice que la relación entre figuras de autoridad (la policía) y el defendido o acusado, puede resultar en una declaración escrita, una grabación o un video que puede variar considerablemente de lo que se hubiera

obtenido del mismo sujeto si se le hubiera dado la oportunidad de hacer su declaración en un ambiente no coercitivo o menos amenazante.

1.1.6 Marcas registradas como palabras o frases en el lenguaje

Una de las áreas de la lingüística forense en la que más se solicita opiniones expertas es el estado de marcas registradas como palabras o frases en el lenguaje. Cuántas veces hemos escuchado acerca de demandas entre compañías por sacar a la venta un producto o un servicio con un nombre que parece emular a otro. La realidad es que situaciones sociales tales como la piratería y la falta de cultura del respeto a la propiedad intelectual generan una gran cantidad de problemas legales, mismos que, en su mayoría, implican a la lengua y, por tanto, a los lingüistas. Olsson (s.f.) nos da un ejemplo claro en este caso: la empresa multinacional de comida rápida *Mc Donald's* se vio envuelta en un caso en el que *Quality Inns International*, una firma de hotelería, anunció la apertura de una cadena de hoteles que se llamaría *McSleep*. La compañía de comida rápida arguyó que el prefijo '*Mc*' ya era utilizado por ellos en diversos productos, tales como *McFries* o *McNuggets*, y que, por tanto, cualquier uso de ese prefijo remitiría a un posible consumidor a la empresa *Mc Donald's,* la cual, según sus argumentos, había originado el proceso de unir sustantivos al prefijo ya mencionado. Este fue un problema que requirió de la opinión experta de lingüistas para determinar si el prefijo en disputa podría ser atesorado por una marca comercial como propiedad exclusiva y, en consecuencia, se debería prohibir el uso del mismo por parte de otra empresa. Al final, la trasnacional involucrada ganó el caso por razones que poco tienen que ver con la

lingüística y mucho con la economía. La resolución de casos como este depende de varios factores; lo interesante es que uno de ellos es la injerencia del quehacer lingüístico.

1.1.7 Fonética Forense

La fonética forense es el uso de técnicas fonéticas aplicadas a investigaciones criminales o judiciales (Olsson 2008: 156). Desde hace algunos años, en México, existe un incremento de casos en los que el audio o el video de un hacho concreto son una prueba crucial, de tal forma que la fonética forense juega un rol importante en ellos. En lo que respecta a las técnicas que se utilizan en fonética forense, estas son muy variadas. Las más utilizadas son el análisis de espectrogramas y el estudio de los rangos de entonación humana y de las distintas características de los formantes en el caso de grabaciones de teléfonos celulares (Olsson, 2008: 155). A menudo, el fonetista forense se encuentra en la situación de tener que analizar no sólo sonido de voces humanas, sino diversos sonidos ambientales de una grabación. Esto tiene que ver con el análisis de sonidos, acústicos o electrónicos, que pueden llegar a producir ciertos dispositivos. Hollien (1990: 11) explica que, en su labor de fonetista forense, ha visto casos en los que el objeto de interés en una grabación es el sonido de armas de fuego, turbinas de avión, o algún objeto o aparato colocado en la habitación desde donde se realiza una llamada telefónica. En este caso, la experticia lingüística consiste en aislar las voces humanas; no en analizarlas.

Es necesario mencionar que la fonética forense se divide en varias sub-áreas, de las cuales las principales son tres: la comparación forense de voz, el contenido en disputa de grabaciones y la construcción de perfiles lingüísticos. A continuación se describen brevemente cada una de ellas.

1.1.7.1 Comparación forense de voz

La comparación forense de voz es la tarea más común de los fonetistas forenses. Consiste en comparar las grabaciones de voz y establecer un grado de similitud entre las voces. Las técnicas utilizadas para este fin han ido evolucionando. En un inicio se utilizaba la comparación gráfica de espectrogramas, tal como se hace con huellas digitales o secuencias de ADN, lo cual no resultó útil porque en el caso del ADN

> …cada muestra es idéntica y exhaustiva; es decir, contienen toda la información necesaria para la identificación de un individuo. Por el contrario, las características que se obtienen de una muestra de voz dependen de una diversidad de factores y no resulta viable añadir el mismo valor identificativo que el de los métodos mencionados. (López, 2010: 24)

Además, los rasgos de la voz, tales como entonación, sonoridad, etc. pueden ser alterados conscientemente por el hablante. A pesar de que las voces puedan ser similares y los rasgos puedan estar alterados, siempre existe cierto grado de variación entre voces por muy parecidas que estas sean. Por eso no es posible comparar muestras de voz solamente con imágenes espectrográficas y se necesitan técnicas más eficaces.

Una técnica más eficaz para hacer comparación forense de voz es el análisis acústico. En este caso, se estudian características y representaciones gráficas de la voz humana que pueden ser visualizadas y medidas por una computadora, tales como la frecuencia fundamental o entonación, el espectro, los formantes, etc. Estos elementos son medidos y se utilizan técnicas estadísticas para comparar diferentes voces y, así, aportar una opinión en cuanto al grado de similitud entre una voz y otra.

1.1.7.2 Resolución de contenido en disputa de grabaciones

Este punto tiene que ver con la posibilidad de que alguien falsifique evidencia o manipule información. Para que el contenido de una grabación sea auténtico debe cumplir ciertas características: a) debe estar completa b) no debe haber sido interrumpida de ninguna manera c) ninguna de sus partes debe haber sido removida d) nada debe haber sido agregado a la grabación. Para autentificar una grabación se deben seguir procesos determinados, tales como escucharla y hacer un examen del estado de la misma, en caso de que se trate de una cinta o grabación magnetofónica. Actualmente, la mayoría de las grabaciones se hacen en formato digital, lo cual facilita los análisis acústicos que pueden realizarse en ellas, sin embargo también facilita una posible manipulación malintencionada. Hollien (1990: 185) dice que es común encontrar casos en los que alguno de los actores de un juicio legal argumenta que alguna prueba ha sido alterada y otro defiende la autenticidad de la misma; por tal motivo, ante desacuerdos tan delicados, Hollien destaca, que el fonetista forense debe abordar el problema de la autentificación de

contenidos de una forma sistemática, abarcadora y ética. Cabe mencionar que el campo de la autentificación de contenidos de grabaciones actualmente se inclina por utilizar ingenieros acústicos para realizar la autentificación, de una manera más exacta, aunque también inciden en ella lingüistas forenses.

1.1.7.3 Construcción de perfiles lingüísticos

Según Coulthard y Johnson (2007: 148), el fonetista forense debe ser capaz de construir el perfil lingüístico de un criminal, a partir de una grabación. Esto es extraer información sociolingüística, semántica, léxica etcétera y, así, determinar un perfil lingüístico que puede ser utilizado como elemento de prueba en un caso legal. Con esta labor se pueden subsanar las necesidades policiales en cuanto a buscar sospechosos que encajen con un determinado perfil.

1.1.8 Estilística Forense

La estilística forense es la aplicación de la ciencia de la lingüística estilística en contextos forenses (McMenamin, 2002). La lingüística estilística se ocupa del análisis científico de marcadores de estilo individuales observados y descritos en el idiolecto de un solo autor, así como de marcadores de estilo generales identificados en el lenguaje o dialecto de grupos de autores. Los asuntos forenses comúnmente relacionados con la estilística son la atribución de autoría y la detección de plagio.

La atribución de autoría es el proceso de identificar quién escribió un texto (Dickinson, 2007). Esto puede servir para determinar si una muerte, en la que se encontró una carta en la que el fallecido exime de cualquier responsabilidad a todos, en realidad fue un suicidio o se trató de un asesinato. El lingüista forense puede determinar, con base en diversos estudios y técnicas, en qué medida es posible que la persona muerta en realidad haya escrito el texto o si es posible que lo haya hecho alguien más.

Por otro lado, la noción de atribución de autoría está intrínsecamente relacionada con la de detección de plagio, que es el proceso de determinar si una obra, o parte de ella, constituye un plagio; es decir, que el autor haya hecho pasar el trabajo de otra persona por propio. En este escenario, es posible que el lingüista forense ofrezca su opinión con base en pruebas de diversas índoles que, por ser el tema de este trabajo, se describen en el apartado 3.

1.2 Relación de la lingüística forense con otras áreas

Es conveniente aclarar la relación que guarda la lingüística forense con distintas áreas. Este trabajo, se enfocará en describir el nexo de la lingüística forense con la lingüística, con el derecho y con la ingeniería lingüística.

1.2.1 Relación con la lingüística

La relación entre la lingüística general y la lingüística forense va más allá de lo estrictamente relacionado con aplicación. Actualmente se realizan investigaciones

lingüísticas en distintas partes del mundo, las cuales no deben su génesis al proceder normal de la lingüística teórica, sino a la necesidad de esclarecer asuntos legales. De esta forma, los resultados obtenidos de esas investigaciones subsanan, por un lado, la necesidad de encontrar solución a dificultades legales y, por otro, generan conocimientos que amplían los horizontes teóricos de la lingüística. Tal caso puede observarse en las consideraciones finales de Niamh McCombe (2002), quien propone que, tal como hay posibilidad de utilizar "palabras clave" en los métodos de identificación de autoría, es posible emplear "letras clave" en la medición y en la comparación de estilos autorales. Esto, eventualmente, podría dirigir investigaciones acerca de la importancia de la letra como unidad lingüística propia del estilo de un autor.

1.2.2 Relación con el derecho

La incidencia de la lengua en el ámbito legal es clara. Las leyes, las declaraciones, determinado tipo de pruebas; todo pasa por el interés de la lingüística. Por tal motivo, como nos señalan Tiersma y Solan (2002), cada vez es más común que el lingüista se encuentre en la posición de dar su opinión como "testigo experto", lo que en México es más comúnmente conocido como "peritaje".

En materia de derecho, el documento oficial que norma la labor forense a nivel federal es el *Código federal de procedimientos penales* (2009). En este documento podemos encontrar un capítulo dedicado a la forma en que puede actuar un perito en un caso legal. Un perito es la "persona que, poseyendo determinados conocimientos científicos, artísticos, técnicos o prácticos, informa, bajo

juramento, al juzgador sobre puntos litigiosos en cuanto se relacionan con su especial saber o experiencia" (Real Academia Española, 2001).

Ahora bien, es ineludible señalar que existe renuencia por parte de los juzgados para aceptar tanto pruebas como opiniones lingüísticas. Esta situación es entendible si se toma en cuenta la naturaleza subjetiva de la lengua, sin embargo se debe considerar que lingüística no es lo mismo que lengua; lingüística es la ciencia encargada de estudiar el lenguaje y se trata de una ciencia tan válida como las demás:

> La ciencia moderna se ha mostrado muy eficaz en el estudio de las realidades materiales, pero siempre ha esquivado las realidades mentales. La ciencia cognitiva, ciencia multidisciplinar, quiere rellenar este vacío. La ciencia cognitiva se convierte en la ciencia de la mente, y en ella converge el estudio de la inteligencia artificial, la neurología, antropología, etología, lógica, filosofía de la mente, psicología, y por último, la lingüística.
> …La lingüística así pasa a ser una ciencia del orden superior, puesto que está integrada en la macrociencia de la cognición. (Martínez del Castillo, 2004: 107)

1.2.3 Relación con la ingeniería lingüística

Finalmente, conviene mencionar que los métodos de análisis y descripción lingüística han evolucionado. El uso de nuevas tecnologías computacionales facilita la emisión de opiniones expertas en todas las áreas en que se realiza labor forense. De la misma forma en que ciencias como la medicina, la balística, la dactiloscopia, etc. se auxilian de avances tecnológicos e informáticos, la lingüística se apoya en diversas técnicas y herramientas que no solo facilitan el trabajo del

lingüista forense, sino que, además, le confieren validez y provocan un mayor grado de confianza de parte de los juzgados hacia dicha labor.

La ingeniería lingüística es la aplicación de conocimientos sobre el lenguaje para el desarrollo de técnicas y sistemas computacionales que sean capaces de reconocer, comprender, interpretar, describir y generar lenguaje humano en cualquiera de sus formas. Es un área interdisciplinar que conjuga los conocimientos de lingüistas, computólogos, ingenieros en cómputo e informáticos principalmente. La ingeniería lingüística actúa como soporte de acciones forenses, en cuanto a que se encarga del diseño de herramientas informáticas que permiten el manejo y el análisis de pruebas lingüísticas con gran precisión. Tal es el caso de los programas de análisis acústico del habla, que se utilizan en fonética forense, y de los sistemas de atribución automática de autoría y análisis estilístico, que se utilizan en la detección de plagio.

Capítulo 2

Detección de plagio

El plagio es un problema que se presenta de varias formas y en diversos sectores de la sociedad. La mayoría de las ocasiones en que se comete plagio se hace en forma consciente y alevosa, tratando de hacer pasar por propio el trabajo de alguien más; situación que, sin duda, es censurable e implica un acto de injusticia para el autor original. Ahora bien, otra posibilidad es caer en tal error por descuido y desinformación. Como cuando se olvida referenciar alguna cita o, simplemente, se ignora que cualquier obra o trabajo ajenos a que se haga alusión deben estar correctamente referenciados, situación común en la esfera estudiantil, donde la falta de hábitos correctos de investigación y de estudio conlleva a infringir constantemente el principio básico de respeto al trabajo y dedicación del otro. El plagio, además, es un problema que puede ocurrir en diversos ámbitos, tales como los sistemas de cómputo, la música, el medio editorial (tanto en el comercial como en el académico) y en el terreno de la investigación.

En la investigación académica es importante promover la innovación a partir de la producción original de conocimiento y en el terreno docente es necesario inculcar en los alumnos las nociones básicas de integridad y honestidad estudiantil. Estos dos campos, sin embargo, se han visto considerablemente afectados por el problema del plagio, sobre todo tomando en cuenta situaciones tales como la facilidad que existe actualmente para acceder a recursos de información en Internet, la insuficiente tipificación del plagio como falta grave y la delgada línea

que existe entre utilizar obras anteriores para sustentar trabajos originales y utilizar información ajena en forma indebida.

2.1 Qué es el plagio

Es pertinente definir, lo más exactamente posible, el concepto de plagio para continuar utilizándolo. Esta labor no es sencilla, pues implica, en primer lugar, hacer una exploración de las diversas formas en que se ha definido el concepto y, en segunda instancia, reflexionar en torno a estas posibilidades para encontrar una definición apropiada. La RAE define plagio como la acción o efecto de plagiar, que, a su vez, significa "copiar en lo sustancial obras ajenas, dándolas como propias" (RAE, 2001); esta definición hace surgir la duda de qué es "lo sustancial". ¿Se refiere a las ideas o al texto? Es decir; ¿solo se incurre en plagio al llevar a cabo una copia exacta de fragmentos textuales ajenos? En tal caso, el hecho de realizar una paráfrasis, es decir tomar las ideas, reescribirlas, no representaría un plagio; sin embargo la *Ley federal de derechos de autor* (2006) sí prohíbe, en su artículo 27, fracción VI, "la divulgación de obras derivadas, en cualquiera de sus modalidades, tales como la traducción, adaptación, paráfrasis, arreglos y transformaciones". Esta ley nos responde la primera duda al poner en claro que una reproducción o copia por paráfrasis sí está contemplada como falta para el sistema legal, lo cual significa que "lo sustancial" a lo que se refería el diccionario es, en efecto, el uso indebido de las ideas.

Ahora bien, es necesario, aclarar que el plagio es un acto fraudulento en que se implican dos cosas: robar el trabajo de alguien más y mentir acerca de ello. Sin

embargo se debe especificar que este robo puede llevarse a cabo de varias formas y en varios niveles como nos lo dicen Maurer et. al. (2006: 1050) cuando afirman que el plagio puede constituirse:

- tomando el trabajo de alguien más como propio
- copiando palabras o ideas de alguien más sin dar el debido crédito
- cometiendo un error al poner una cita entre comillas
- dando información incorrecta acerca de la fuente de una cita
- cambiando las palabras, pero manteniendo la estructura de oración de una fuente sin dar crédito
- copiando demasiadas palabras o ideas de una fuente, de tal forma que esto ocupe la mayoría del trabajo, aunque se dé el crédito correspondiente. (Plagiarism.org, 2006)

De la lista anterior llaman la atención tanto el tercer punto como el último, pues se menciona, por un lado, que el plagio se puede cometer por error y por otro, que un plagio puede existir incluso dando el crédito correspondiente. Para dejar esto en claro, es necesario exponer lo tipos de plagio que existen.

2.2 Niveles y tipos de plagio existentes

El plagio no siempre se comete de forma intencional y, contrario a lo que podría pensarse, tampoco implica en todas las ocasiones robar el trabajo de alguien más. En primer lugar, se puede incidir en plagio de forma accidental debido a carencia

de conocimiento con respecto al tema y por desconocimiento o falta de entendimiento de normas de referencia o citado (Maurer, et.al., 2006). Esto sucede comúnmente en ámbitos estudiantiles, donde las reglas de referencia suelen variar de un método a otro o incluso entre distintas instituciones. A lo anterior se le suma el hecho de que no existe conciencia del mal que se puede causar en prejuicio de otro o de sí mismo por parte del estudiante promedio, quien demerita la posibilidad de incurrir en una falta al no exigirse rigor en la realización de sus trabajos escolares. En segundo lugar, puede darse el caso de que la gran cantidad de información provoque confusión y olvido con respecto a la procedencia de una idea y esto, a su vez, induzca a pensar que una idea es propia cuando en realidad se ha leído o escuchado de alguna fuente. En tercer lugar está el plagio intencional; copiar el total o una parte de un trabajo ajeno y asumirse como el autor original. Finalmente está el autoplagio que consiste en usar trabajo propio publicado anteriormente sin referir el original (Maurer et.al., 2006).

Aunado a estas categorías de plagio, existen métodos que comúnmente se utilizan para realizar la copia. El más común, por la facilidad que existe actualmente para acceder a medios de información electrónica, es el denominado *copy-paste*, cuyo nombre viene de la acción, tan común en el uso de computadoras, de copiar textualmente fragmentos exactos de un escrito y pegarlos en el trabajo propio sin referenciar la fuente. El plagio de ideas, por su lado, es usar conceptos similares u opiniones que no son conocimiento común, lo cual no debe ser confundido con la paráfrasis que consiste en cambiar la sintaxis, utilizar palabras con significado igual o parecido, reordenar oraciones o expresar el contenido de un trabajo original con diferentes palabras. La paráfrasis es válida, siempre y cuando se haga la debida referencia; en caso contrario, constituye un plagio. Otras

categorías son: el plagio de código, que se presenta en el ámbito del desarrollo de software computacional; el uso inapropiado de signos de puntuación y marcas de referencia tipográficas, tales como comillas, dos puntos, tamaños de letra, márgenes, etc., lo cual impide identificar exactamente las partes ajenas que se han usado en un trabajo propio y, por último, el plagio de traducciones que se refiere a usar información que se ha traducido de alguna lengua a la del trabajo propio y no dar referencia al trabajo original (Maurer, et. al., 2006).

En vista de las múltiples posibilidades que existen para incurrir en plagio, y de los abundantes casos que se presentan a diario en los ámbitos ya mencionados, se ha implementado varias formas para identificar textos que hayan incidido en la tan nombrada falta; incluso para determinar en qué partes del texto se ha llevado a cabo el plagio. Se trata de una gran gama de opciones que vale la pena mencionar y describir.

2.3 Herramientas y métodos existentes para la detección de plagio

Ante una sospecha, es necesario corroborar y demostrar si existe plagio. Después de leer un texto se puede tener la sensación de haberlo leído anteriormente o, como sucede muy a menudo en el ámbito académico, un profesor puede dudar de la autoría de algún trabajo, basado en el conocimiento que tiene del rendimiento de sus alumnos. Lo ideal en estos casos es recordar la fuente plagiada y presentarla como prueba para comprobar que existe plagio. Sin embargo, en caso de no tener presente dicha fuente, es necesario cotejar el texto en cuestión con un grupo de

documentos posiblemente copiados; es decir con determinados documentos fuente, lo cual implica no solo un considerable esfuerzo humano y la inversión de cantidades considerables de tiempo, sino una cobertura limitada de posibles fuentes plagiadas. Es en este punto donde resulta clave contar con ayuda informática que permita comprobar la sospecha o, en su caso, desecharla.

Existen diversos métodos que se usan para detectar y medir similitud textual. Se puede mencionar tres principales, por ser los más comunes (Maurer et. al., 2006): a) la comparación de un documento sospechoso con un grupo de documentos en los que, posiblemente, se encuentra el documento fuente b) la búsqueda de un fragmento dentro de un documento mediante un motor de búsqueda en internet y c) el análisis y comparación manual de estilos de autores. La mayoría de los sistemas disponibles se basan en el primer y segundo métodos; sin embargo, aunque en menor medida, también existen opciones que facilitan la comparación de estilo a nivel manual.

Entre las herramientas para detección de plagio que se han desarrollado hasta ahora hay algunos ejemplos que vale la pena mencionar. Para tal fin, en este apartado se clasifican los diversos sistemas con base en los métodos que utilizan para detectar similitud textual y el tipo de plagio que buscan detectar.

En primer lugar está la comparación de cadenas textuales. Se trata de buscar un grupo de segmentos de texto en determinados documentos o bien de buscar lo que Maurer et. al. (2006) llaman "fingerprints"; es decir, cadenas textuales de longitud determinada, obtenidas de un documento sospechoso. En este caso, la búsqueda se realiza en la totalidad de documentos de referencia o en índices de "fingerprints" extraídos de los mismos.

Otro método que se utiliza comúnmente en la detección de similitud es la búsqueda de similitud mediante n-gramas. Un *n*-grama es una secuencia de *n* elementos, donde *n* es un número determinado. Estos elementos "pueden ser unidades de diversa índole, dependiendo del nivel extralingüístico o lingüístico en el que se observan. Pueden representar combinaciones de palabras, caracteres alfanuméricos, signos de otra tipología, etc." (Spassova, 2009: 59-60). En el caso de la detección de similitud se trabaja con n-gramas de palabras. Así un monograma es una sola palabra, un bigrama es la secuencia de dos palabras; un trigrama, la secuencia de tres palabras y así sucesivamente. La técnica consiste en buscar cada n-grama del texto en cuestión en uno o varios textos de referencia y presentar un reporte final de las coincidencias, lo cual ayuda a identificar fragmentos plagiados.

WCopyfind (WCopyfind, 2010) es un ejemplo de herramienta que utiliza estos métodos de búsqueda de similitud. El sistema detecta palabras o frases en repositorios de documentos tanto por medio de "fingerprints" como utilizando *n*-gramas. WCopyfind realiza la búsqueda en un conjunto local de documentos; es decir, en un grupo de archivos que se encuentren almacenados en la misma computadora en la que se encuentre instalado el programa y, además, es un sistema gratuito.

Otro ejemplo es EVE2 (Eve2, 2000). Se trata de un sistema de paga que realiza la búsqueda no solo en repositorios locales, sino en documentos de internet y no se limita a una sola cadena textual, sino que presenta un conjunto de coincidencias entre el documento en cuestión y diversos documentos de la web, lo cual resulta provechoso, tomando en cuenta la gran cantidad de información disponible en línea y la abundante cantidad de plagios que se realizan con ella.

Este último sistema utiliza dos de los métodos más comunes: la búsqueda de cadenas textuales y la búsqueda en internet por medio de un motor de búsqueda.

La búsqueda de similitud textual por cadenas textuales y por *n*-gramas resulta efectiva cuando el plagio es del tipo *copy-paste* (ver 3.2) y presenta la ventaja de que puede ser utilizada en diversos lenguajes. Las herramientas desarrolladas con base en los dos métodos anteriormente descritos fueron hechas para el idioma inglés, sin embargo resultan eficientes también para el español, ya que la comparación y búsqueda de cadenas textuales es independiente a la lengua, pues lo que se busca es una consecución de caracteres iguales en uno y otro texto, sin importar qué lengua se utilice.

Ambos métodos ven mermada su efectividad cuando las oraciones en cuestión son cortadas irregularmente, de tal forma que existe copia, pero no se trata de una copia exacta. Lo anterior sucede cuando existe disparidad de caracteres entre los textos, como sucede en el caso de faltas ortográficas o uso de caracteres especiales. En dichos casos la comparación de cadenas textuales y la comparación de *n*-gramas resultan poco útiles y es necesario acudir a otros métodos.

En copias hechas con paráfrasis es necesario aplicar métodos de detección de similitud más avanzados. En este caso, la detección de cadenas textuales no resulta de utilidad, pues la paráfrasis es, en esencia, cambiar palabras de un texto por otras sin cambiar el significado. Así, la comparación automática de palabras no traerá un resultado satisfactorio en este caso, sin embargo el plagio existirá de todos modos, pues tomar ideas ajenas y darlas como propias es inadecuado e ilegal, aunque se cambie las palabras e, incluso, la estructura. La estilometría es una opción viable en casos en los que la copia se realiza por medio de la paráfrasis. Se

trata de un método que mide el estilo de los autores, basado en marcas personales que se plasma en los escritos. En el siguiente apartado de este trabajo se profundiza en la estilometría. Por ahora es suficiente mencionar que no existe una herramienta como tal que otorgue resultados terminantes; los sistemas de detección automática de plagio solamente representan una ayuda para el dictamen; es decir, no responden categóricamente a la pregunta ¿existe plagio?; simplemente representan una ayuda.

Los sistemas antes mencionados ayudan a conferir validez o invalidez a una sospecha, pero quien toma la decisión final es siempre el usuario, aunque cabe aclarar que en un contexto legal la situación cambia. El lingüista forense es quien se vale de estas herramientas para fundamentar objetivamente una opinión experta y, aunque podría hacerlo, no resuelve categóricamente si existe plagio o no, ya que su labor en un escenario legal se limita a testificar si existe similitud entre dos textos y, en caso de haberla, en qué medida la hay y en qué medida esa similitud es producto del azar.

Capítulo 3

Estilometría

La estilometría es un método por el que se puede optar para detectar similitud textual. No se limita a encontrar similitud exacta entre fragmentos textuales, sino que representa una opción en la búsqueda de similitud y copia cuando se ha utilizado algún procedimiento avanzado para el plagio, tal como la paráfrasis, o cuando la situación presenta un caso de plagio de ideas o, incluso, de plagio de código (ver 3.2). En este apartado se exponen las diferentes características de la estilometría, así como algunas de las posturas que han surgido a favor y en contra de este método. Asimismo, se presentan las razones por las que se eligió a la estilometría para realizar el análisis del caso particular que se presentará al final de este trabajo.

3.1 Estilística

La estilometría tiene sus orígenes en la estilística, que nace como una corriente de la crítica literaria. En primera instancia, la estilística tenía como objeto de estudio la utilización del lenguaje en sus facetas artísticas; es decir, se ocupaba de los usos literarios que los autores hacían del lenguaje para provocar determinada situación estética. Ahora bien, esta disciplina nace en el seno de la teoría y crítica literaria,

pero, como dice Stanley Fish (1994) desde su génesis, toma la postura de ser diferente a lo acostumbrado en estudios literarios:

> La estilística nació de una reacción a la subjetividad e imprecisión de los estudios literarios. Los estilistas pretendieron sustituir los éxtasis apreciativos de la crítica impresionista, por descripciones lingüísticas precisas y rigurosas, y moverse de estas descripciones a interpretaciones de las que pudieran obtener medidas de objetividad. La estilística, en suma, es un intento de otorgar base científica a la crítica literaria. (Fish, 1994: 103)

Las descripciones y análisis lingüísticos que propuso la estilística se encaminaron a estudiar el estilo. Desde el punto de vista literario, puede definirse el estilo como, "las <<selecciones>> disponibles que se le presentan al escritor, su modo de elegir entre las posibilidades lingüísticas que se ofrecen a su exigencia expresiva" (Amberson, 1979: 123). Por otro lado, en este trabajo interesa más la definición lingüística del estilo, en cuyo sentido David Viñas Piquer (2002: 392) propone que el estilo se compone de "los detalles formales que ponen de manifiesto innovaciones estilísticas"; él mismo propone una analogía para explicar lo anterior: "igual que un hablante parte de las oraciones nucleares simples y les aplica las reglas derivacionales para conseguir distintas construcciones sintácticas, un escritor tiene también predilección por determinadas estructuras y construcciones gramaticales y estas predilecciones marcan su estilo" (Viñas, 2002: 395).

Tomando en cuenta el objetivo de esta tesis, la definición de estilo que tomaremos es la que se propone en la literatura relacionada con lingüística forense. Esta definición no dista mucho de la que propone la lingüística, pues coincide en decir que el estilo se conforma por las elecciones lingüísticas de los

autores; sin embargo, aporta la idea de que el estilo también se conforma por las diferencias sutiles pero regulares que existen incluso entre textos que comparten lenguaje, género y tema, pero difieren en autoría, género del autor o en parámetros similares (Golcher, 2007: 1).

Retomando a la estilística, cabe aclarar que este trabajo se orienta a la parte lingüística de la cuestión. Así, se dejan de lado ideas tales como la búsqueda de descripciones de naturaleza expresiva y de interpretación por medio de la estilística. Únicamente se hace mención de dichas situaciones como antecedentes que resultan ineludibles al realizar un estudio de estilometría. En síntesis, la particularidad que se toma de la estilística para este trabajo es que "la estilística no se propone explicar, sino describir. No nos da el *porqué* de una obra, sino el *qué es* y *cómo* está construida" (Amberson Imbert, 1979: 129). De ahí que, como dice Fish (1994), los estudios estilísticos sean esencialmente comparativos, como será el caso del estudio específico que se expone en el siguiente apartado.

En este punto es pertinente relacionar lo que se ha definido como estilo con una referencia obligada por su abundante aparición en los estudios de lingüística forense: el idiolecto.

3.2 Idiolecto

La noción de idiolecto se refiere a la idea de que exista un dialecto personal. Gerald Mc Menamin dice que no hay dos individuos que usen y perciban el lenguaje de la misma manera; existen, por lo menos, mínimas diferencias en la forma de hablar y de escribir de las personas. El idiolecto es la combinación

inconsciente y única del conocimiento lingüístico, las asociaciones cognitivas y las influencias extralingüísticas. (McMenamin, 2002)

Ahora bien, dos importantes trabajos en lingüística forense se oponen al concepto de idiolecto. Tanto el libro de Malcolm Coulthard y Alisson Johnson como el de John Olsson demeritan la posibilidad de utilizar la noción de idiolecto para realizar estudios concernientes a la detección de similitud y a la atribución de autoría. El primero habla del problema que representa buscar marcadores distintivos que determinen el idiolecto (Coulthard & Johnson, 2007: 170) y el segundo complementa la oposición arguyendo que la manera en que se adquiere el lenguaje y las modificaciones que este sufre con paso del tiempo y con los diversos factores que lo transforman son circunstancias que se oponen completamente a la existencia del idiolecto (Olsson, 2008: 62).

No obstante la clara oposición a la noción de idiolecto por parte de los trabajos mencionados, coinciden los autores en proponer la idea de variación e, incluso, la de estilo.

3.3 Variación

Olsson (2008: 31) distingue entre dos tipos de variación: la variación intra-autor, que se refiere a las diferencias que existen entre textos escritos por el mismo autor, y la variación inter-autor, es decir las diferencias que existen entre un autor y otro.

Existen diferentes causas que propician la variación intra-autor. La primera es el género textual, la cual sugiere que no importa si dos textos pertenecen al

mismo autor; puede existir variación entre ellos si dichos escritos pertenecen a distintos géneros o tipologías textuales. En segundo lugar, la condición pública o privada de los textos puede desembocar en diferencias intra-autor, pues es considerablemente diferente escribir un artículo que tenga como finalidad ser difundido a escribir textos de carácter personal, tales como cartas o diarios. Otra causa de este tipo de variación puede ser, y comúnmente lo es, el tiempo en que fue escrito el texto; entre más tiempo haya pasado entre la escritura de un texto y otro por parte del mismo autor, la variación será mayor. Finalmente, situaciones como el cambio de circunstancias en que se escribe, los parámetros sociales de los lectores a quienes va dirigido el texto y la intencionalidad son otras causas que pueden propiciar la variación intra-autor.

Para terminar con el tema de la variación, es necesario exponer que existen principios generales en lo que a variación intra-autor se refiere, mismos que se derivan de las causas de variación. Se puede asumir que todos los autores exhiben variación entre sus textos, pero esta variación disminuye si sus textos son del mismo género. En cambio, si comparamos textos de diferentes autores que comparten género y temática, se hará presente el hecho de que la variación inter-autor es de más probable aparición. Aunado a esto, las causas de esta variación pueden ser conocidas si tenemos suficientes referencias de los textos que se analiza y el grado de variación es en cierto grado predecible, ya que es muy probable que la variación inter-autor sea mayor que la variación intra-autor si se trata, como en el caso que se analizará, de textos del mismo género, pero de diferentes autores.

Se ha visto hasta aquí, que la existencia de un idiolecto es dudosa, sin embargo está claro que existe variación entre textos de diferentes autores más que

entre textos del mismo autor y que esta variación es, no solo identificable, sino cuantificable. Este trabajo no se opone a que es muy difícil, tal vez imposible, determinar marcadores de estilo inherentes a un ser humano en específico. Por otro lado, apoya, por un lado, la idea de que es posible clasificar trabajos escritos con base en la medición estadística de la variación de dichos textos y, por otro, que explorar métodos de medición y descripción de índices de variabilidad entre textos es necesario para las tareas de detección de similitud que requiere la lingüística forense. En este sentido, esta tesis se apoya en la aseveración hecha por Coulthard y Johnson (2007: 161), misma que cito de forma textual y en su idioma original:

> …the situation is not as bad as might at first seen, because such texts usually contain information or clues, which massively restrict the number of possible authors. Thus the task of the linguistic detective is never one of identifying an author from millions of candidates on the basis of the linguistic evidence alone, but rather of selecting (and, of course, *des*electing) from a very small number of candidate authors, usually fewer than a dozen and in many cases only two.

Así pues, se adopta, a partir de este punto, la idea de que la labor del lingüista forense no es identificar un autor de entre un grupo de millones de autores, sino de seleccionar y clasificar textos de un pequeño número de candidatos. Esta clasificación es posible y resulta factible realizando un análisis cuantitativo de frecuencias, de tal manera que se obtenga el grado de diferencia, o de variabilidad, entre dos textos y se pueda medir en qué grado esta diferencia es significativa.

3.4 Marcadores estilísticos

Para realizar la medición mencionada es necesario establecer parámetros en los que estará basado el análisis estadístico. En la estilometría, estos parámetros son los marcadores estilísticos, mismos que deben atender a determinados criterios. En primer lugar, los marcadores a analizar deben incluir tanto parámetros que puedan ser elegidos conscientemente por el autor como parámetros que el autor plasme de manera inconsciente, es decir que no sean observados a simple vista por el autor, pero que el lingüista pueda observar y analizar. Aunado a esto, los parámetros que se utilicen deben ser comunes a los autores que se analizan y, en el particular caso que se trata en este trabajo, serán ajenos tanto al género textual como a la temática de los textos. Así pues, el hecho aparentemente débil de que el análisis estilístico se mantiene al margen de realizar afirmaciones sobre el significado y el valor de los textos resulta en un beneficio agregado al propósito de esta tesis. Lo anterior quedará suficientemente aclarado en el apartado 5.1 donde se hace la descripción del caso específico de detección de similitud y se enumeren las condiciones del corpus analizado.

Los parámetros de análisis en estilometría se llaman marcadores estilísticos. Estos marcadores pueden y deben ser de diversa índole, pues como dice McMenamin (2002) los marcadores de estilo únicos, que actúan solos, son raros; la atribución de autoría requiere de la identificación de varios marcadores; de hecho, él habla de trescientos diferentes marcadores de estilo. Ahora bien, esta gran cantidad de marcadores se agrupa en diversos conjuntos dependiendo del tipo al que pertenecen y el nivel del estilo que atienden. Así, hay marcadores del

formato del texto, de abreviaturas, de puntuación, de sintaxis, de discurso e incluso de errores y de corrección.

Es necesario delimitar el trabajo que se realiza con fines forenses y, por tanto, el análisis de todos los marcadores estilísticos existentes resultaría tardado y no sería posible de realizar con el sistema computacional que se propone, además, existen marcadores estilísticos que no resultan útiles para determinados estudios. Un ejemplo de marcador estilístico que se dejó de lado en esta tesis son las palabras de contenido; es decir, las palabras que poseen un significado de mayor peso en un escrito que las palabras funcionales. Las palabras de contenido pertenecen a clases abiertas, tales como los adjetivos, los verbos y los sustantivos; mientras que las palabras funcionales corresponden a clases cerradas, tales como las preposiciones, los artículos, etc. Las palabras de contenido no fueron analizadas aquí porque su uso depende del contenido de los textos y el contenido, a su vez, depende del tema, cuestión que, como aclaramos desde el principio, queda fuera de lo que concierne a la estilometría. De la misma forma, se discriminó una lista amplia de marcadores de estilo por tratarse de parámetros cuyo análisis arrojaría resultados dependientes del contenido.

La longitud de oración es el primer marcador estilístico que se utilizará en este trabajo. La razón es que presenta menos variación entre escritos del mismo autor que entre escritos de diferentes autores, sobre todo cuando los textos pertenecen al mismo género, y en ese sentido, coincide con Peng & Hengartner (2001: 5), quienes proponen que, al elegir las unidades de análisis, se debe usar aquellas que presenten una importante variación iter-autor y poca variación intra-autor. De esa forma, aumenta la posibilidad de que la clasificación de textos por el criterio de longitud de oración tienda a separar elementos de diferentes autores,

pues, como afirman Coulthard y Johnson (2007: 165), la longitud de oración puede estar atenida a la función que las oraciones desempeñan como punto de interacción entre el lector y el escritor. Esto se refiere a la posibilidad que tiene el autor de decidir usar oraciones cortas o largas. Se pueden usar oraciones cortas para facilitar el entendimiento por parte del lector de las ideas que el autor busca transmitir y se pueden usar oraciones largas si así se prefiere. La longitud de oración, entonces, puede originarse en una decisión consciente, aunque también puede hacerlo en una inconsciente si es que el autor no repara en ella. En ambos casos, funciona correctamente como un marcador estilístico que sirve para determinar variación entre textos.

La longitud de palabra es un marcador estilístico que suele estar fuera del control consiente del autor y, por tanto, será utilizado es este trabajo. A diferencia de lo que sucede con la longitud de oración, los escritores rara vez reparan en el tamaño de las palabras que utilizan; sin embargo, la frecuencia con la que aparecen palabras de determinada longitud en los textos de un autor puede ser una marca característica de escritura. Grieve (2005: 8-11) realiza una recopilación de los más conocidos experimentos en atribución cuantitativa de autoría que se han realizado usando la longitud de palabra como marcador estilístico. Se ha cuantificado desde la frecuencia de aparición de palabras de dos, tres y cuatro caracteres, hasta la longitud de palabras sin usar el carácter como unidad de medida, sino las sílabas. En este trabajo, se usará la longitud de palabra sin restringirla a un número específico de caracteres; esto es que no se utilizará determinada longitud de palabra para cuantificarla, sino que se tomarán en cuenta longitudes variadas, desde palabras de un carácter hasta palabras de 14 o más caracteres.

El siguiente marcador estilístico que se usará es la longitud de párrafo. Esta medida está intrínsecamente relacionada con la longitud de oración; el tamaño de los párrafos varía dependiendo de la longitud de oración que determinado autor emplea. Aunado a lo anterior, Clough (2000: 6) considera que existen algunos rasgos simples que pueden ser usados para técnicas estadísticas; entre ellos menciona la longitud de oración y la longitud de párrafo.

Finalmente se toma la frecuencia de uso de signos de puntuación como marcador estilístico. En este caso se tomó la decisión atendiendo a que Bilge, Büsra Celikkaya, & Hatun, (s.f.) proponen que los marcadores de estilo más recomendables para ser evaluados, si se requiere mantener el estudio al margen del contenido, son, entre otros, la longitud de palabra, la longitud de oración y el uso de signos de puntuación. En adición a esto, Grieve (2005: 19) dice que la frecuencia de signos de puntuación puede ser un poderoso parámetro estilístico en textos modernos, en los que existe una gran gama de posibilidades en los que el autor puede utilizar estos caracteres gramáticos.

3.5 Signature

La herramienta que se utilizará para realizar el análisis es el sistema computacional llamado *Signature* (Millican, 2003). Este programa fue elaborado por Peter Millican, afiliado a la Universidad de Oxford, y permite hacer análisis estilométrico utilizando cinco marcadores estilísticos: la longitud de palabra, la longitud de oración, la longitud de párrafo, uso de letras del alfabeto y uso de signos de puntuación. En el caso específico de este trabajo no se describe el

análisis de uso de letras del alfabeto como marcador estilístico por no ser un elemento estilístico que haya sido mencionado en los estudios que se revisaron en el trabajo de investigación de esta tesis.

Signature ofrece al usuario gráficas y tablas que representan la frecuencia de los marcadores de estilo mencionados. Permite seleccionar diferentes archivos de texto y da a escoger entre la opción de comparar dichos documentos entre ellos y la posibilidad de manejar un grupo de varios documentos como un solo corpus. La posibilidad de visualizar gráficas comparativas es útil para observar niveles de variación entre diferentes textos, mientras que la opción de agrupar archivos puede servir para hacer una representación estilográfica de varios textos de un mismo autor. Las gráficas dan una idea visual de las similitudes y diferencias que existen entre los textos; además es posible visualizar tablas de distribución de frecuencias absolutas y relativas; es decir, por porcentaje. Sin embargo, estas vistas no resultan suficientes para realizar un análisis estilométrico; apenas nos dan una aproximación estilográfica.

Una de las características que tiene *Signature* es que ofrece la posibilidad de sustituir la mera comparación visual de gráficas, misma que puede resultar engañosa, por una medida estadística objetiva que exprese, en términos numéricos, el grado de similitud que existe entre dos textos (Millican, 2010). Para este fin, el sistema da la opción de mostrar una prueba de comparación estadística llamada prueba χ^2, misma que será explicada en el siguiente apartado que corresponde a la descripción de un caso específico.

Capítulo 4

Estilometría para detectar similitud textual en un caso concreto

Cada parte de este trabajo se ha realizado con la finalidad de exponer los procesos que se ejecutaron y las decisiones que se tomaron en un caso específico de detección de plagio.

Así, el primer apartado de la tesis tiene dos funciones: por un lado, ubicar este trabajo en un campo de estudio y aplicación específicos de la lingüística: la lingüística forense; y, por otro, poner en claro que, por estar íntimamente relacionada con el ámbito legal, esta disciplina implica determinados procedimientos de confidencialidad. Dichos procedimientos impiden revelar en esta tesis algunos datos tales como los autores involucrados y las referencias de los trabajos en cuestión; incluso, se tiene una prohibición imperiosa de publicar cualquier fragmento, por mínimo que sea, de los textos que se analizaron. En este entendido, el presente trabajo no incluye un apéndice con el corpus utilizado, por tratarse este de un corpus privado y de accesibilidad y difusión restringida. De tal forma, se pone de manifiesto que en la descripción del caso específico que atañe a este trabajo no habrá referencias explícitas o implícitas a la autoría de los textos ni se plasmará fragmento alguno involucrado.

Existen, sin embargo, datos que sí pueden ser expuestos y que, en realidad, son los que realmente interesan al análisis. Tales datos son la información de frecuencias de las tablas, las gráficas de datos obtenidas, el tamaño y el género de

los textos en cuestión, así como la descripción del formato que tienen los documentos.

El breve estado de la cuestión en lo que a detección de plagio se refiere sirve para ubicar en qué formas y niveles se puede cometer plagio y, así, justificar la utilización del método de la estilometría. En casos con dificultad mayor, como el que aquí se trata, es difícil encontrar solución utilizando técnicas como la búsqueda de cadenas textuales. Sin duda, de haberse encontrado copia indebida hecha de manera textual, no hubiera existido el requerimiento por parte de una autoridad para escrutar el caso usando métodos estadísticos objetivos.

Finalmente, el apartado concerniente a la estilometría constituye la argumentación central de este trabajo. El argumento busca validar el método de la estilometría para la clasificación de textos. Es prudente recordar la aclaración de que este trabajo no se basa en la idea de la existencia del idiolecto ni de la huella digital textual, sino en el hecho de que la variación puede ser medida tomando como criterios de medición ciertos parámetros de estilo.

4.1 Descripción del caso

El análisis que se expone a continuación responde a una solicitud recibida por el Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería de la UNAM por parte de un solicitante de servicios. La solicitud fue procesar y comparar de manera automática, matemática y objetiva dos textos, y así realizar un estudio enfocado a definir si existía o no similitud textual entre ellos y en qué medida esta similitud era producto del azar.

Los textos en cuestión pertenecen a autores que realizaron, de manera independiente, dos documentos extensos. Ambos documentos, por ser publicaciones ya difundidas, comparten la característica de tener apartados que se encuentran alejados de lo analizable en términos estilísticos, tales como la bibliografía, diversas citas textuales, referencias personales e institucionales, etc. Aunado a esto, los dos escritos pertenecen al mismo género textual; el léxico que se utiliza en cada uno de los documentos es altamente parecido y, por si esto fuera poco, los dos abordan exactamente la misma temática, pero con diversos enfoques. Así pues, el equipo de trabajo del GIL se encontró con la problemática de tener que clasificar y medir las diferencias entre textos que comparten léxico, temática y género.

Los escritos se encuentran envueltos en una polémica. Uno de los dos autores, evidentemente quien publicó primero su texto, acusa al otro de haber plagiado su trabajo. El acusado niega tal imputación arguyendo que las coincidencias entre los textos se deben precisamente a la cercanía genérica y temática.

El trabajo entonces, del GIL, consistió en analizar la situación y proponer una posible solución que es la que se expone en el siguiente subapartado.

4.2 Descripción y tratamiento del corpus

El corpus de análisis que recibió el GIL consistía en dos libros. El primer paso, entonces, para poder procesar el corpus fue convertir ese corpus textual físico en un corpus textual informatizado, ya que todos los procesos computacionales que

se realizan en el GIL parten de tener un corpus que, de acuerdo con Sierra (2008: 445): sea un "conjunto de textos elegidos y anotados con ciertas normas y criterios para el análisis lingüístico, de forma que se sirve de la tecnología y de las herramientas computacionales para generar resultados más exactos".

La construcción de este corpus implica todo un proceso de compilación que se sigue por pasos. En primer lugar, todo corpus debe estar construido de acuerdo con el objetivo que se persigue. En este caso, el objetivo es realizar comparación textual para clasificar documentos. Los siguientes pasos son la selección y la obtención de los textos, mismos que parecían haber sido cubiertos ya en su totalidad, pues los textos a comparar le fueron otorgados al GIL desde un principio.

Los pasos finales son la definición del equipo de trabajo y la digitalización de los documentos. En cuanto a la definición del equipo de trabajo fue necesario asignar diversas tareas a diversas personas, pues la solicitud del análisis del caso incluía una fecha de entrega muy próxima a la fecha de solicitud. Se designó, principalmente, a digitalizadores, revisores de digitalización e informáticos de apoyo que se encargaron de buscar herramientas y métodos que resultaran útiles para el objetivo planteado.

En primer lugar se realizó el escaneo de todas y cada una de las páginas de los libros, de tal forma que se obtuvieron imágenes de cada página. Los libros presentan diversos problemas para ser digitalizados, tales como la dificultad de obtener una imagen clara de las páginas, muchas veces por la imposibilidad de abrir completamente el libro sobre el escáner. Esto puede parecer pueril, sin embargo en procesos de recopilación de corpus textuales informatizados este tipo de problemas es común y representa la inversión de grandes cantidades de tiempo

y esfuerzo. Posteriormente, se utilizó un reconocedor óptico de caracteres (OCR, por sus siglas en inglés) para extraer información textual de las imágenes que se tenían. Al final, se realizó un proceso de revisión de los archivos obtenidos con el reconocedor de caracteres, pues este tipo de programas suelen ser falibles cuando se trata de reconocer letras en imágenes provenientes de escaneos.

Finalmente se obtuvo un corpus compuesto por dos textos; uno de 10 873 palabras, el presuntamente plagiario, y otro de 27 176 palabras el supuesto plagiado. Ambos documentos fueron guardados en archivos de texto plano, con extensión *.txt*, pues este es el formato que requieren todos los sistemas enfocados a análisis textual informatizado.

4.3 Metodología propuesta para el caso

En este apartado se describe la metodología propuesta para la detección de similitud y se detalla la propuesta de una prueba más; la de consistencia de estilo.

4.3.1 Prueba de similitud entre los textos en controversia

En un principio, la solicitud hecha al GIL fue clara y concisa. Había que analizar y comparar dos textos para encontrar y medir sus diferencias y sus similitudes. Para este fin, se realizó un estudio completo de las posibilidades que había para buscar similitud de manera automática o semiautomática y, así, se llegó a la conclusión de que había que utilizar diferentes métodos: la comparación de cadenas textuales, el

análisis de palabras funcionales y el análisis estilométrico y estilográfico. Este trabajo se centra en el último.

El análisis estilométrico y estilográfico tenía que realizarse solamente entre los dos textos en controversia y, por tanto, la situación parecía resuelta en ese momento; sin embargo, el Grupo de Ingeniería Lingüística, en su afán por explorar las diferentes posibilidades que ofrece el procesamiento de textos, no se limitó a realizar esta comparación entre dos textos, sino que fue más allá y planteó una metodología específica para el caso. Además de la comparación textual entre dos documentos, atendiendo a las nociones de estilo y variación intra-autor e inter-autor ya expuestas en este trabajo, se realizó una prueba de consistencia de estilo de los autores en cuestión.

4.3.2 Prueba de consistencia de estilo

Así pues, hubo que ampliar el corpus de dos a cuatro documentos. Si en un principio se tenían dos documentos, uno de cada autor, que de ahora en adelante serán llamados textos en controversia, se decidió recopilar dos textos más; uno de cada autor, diferentes a los que ya se tenían y que de ahora en adelante serán llamados textos de referencia.

Atendiendo a metodología forense, los textos de referencia debían cumplir con determinadas características. En primer lugar, la autoría de los textos debería ser indudable; es decir, que los textos debían haber sido escritos por los autores y se debía tener seguridad y constancia de ello. En adición, los textos de referencia debían haber sido escritos en una fecha anterior a la de la controversia; de tal

forma, servirían como referencia del estilo de los autores. En segundo lugar, la fecha de producción de dichos textos debía ser lo más cercana posible a la de la controversia para que la posible variación estilística por cuestiones temporales no incidiera en la prueba. Finalmente, debía existir contemporaneidad entre los textos; es decir, que entre la escritura de uno y otro, no debía haber pasado más de uno o dos años.

Estos textos de referencia servirían para dos fines:

1. Establecer un nivel de variación inter-autor esperado con base en la comparación de los dos textos de referencia. Estos textos son de autoría definida; es decir, que no son dubitados. Por tanto, el nivel de variación que presenten entre ellos se puede tomar como el nivel mínimo de variación esperado entre textos de diferentes autores. Así pues, si el nivel de variación entre los dos textos en controversia es menor al establecido como el esperado inter-autor, entonces se incrementará la sospecha de plagio. De otra forma, si el nivel de variación se mantiene igual o por encima de lo esperado, la sospecha de plagio se verá disminuida, pues eso significaría que no hay similitud entre los textos.
2. Determinar si los autores son consistentes en su propio estilo. Los dos textos de la supuesta víctima son de autoría definida y el nivel de variación entre ellos se puede tomar como el nivel máximo de variación esperado entre textos del mismo autor. De tal forma que, si el nivel de variación entre los textos del supuesto plagiario es mayor que el establecido como esperado intra-autor, la sospecha de plagio aumentará, pues eso significaría que el supuesto plagiario no es consistente en su estilo.

El esquema 1 muestra lo anterior de manera gráfica.

En la parte inferior del esquema se ilustra cómo el resultado de la comparación entre los textos de referencia es un nivel esperado de variación inter-autor.

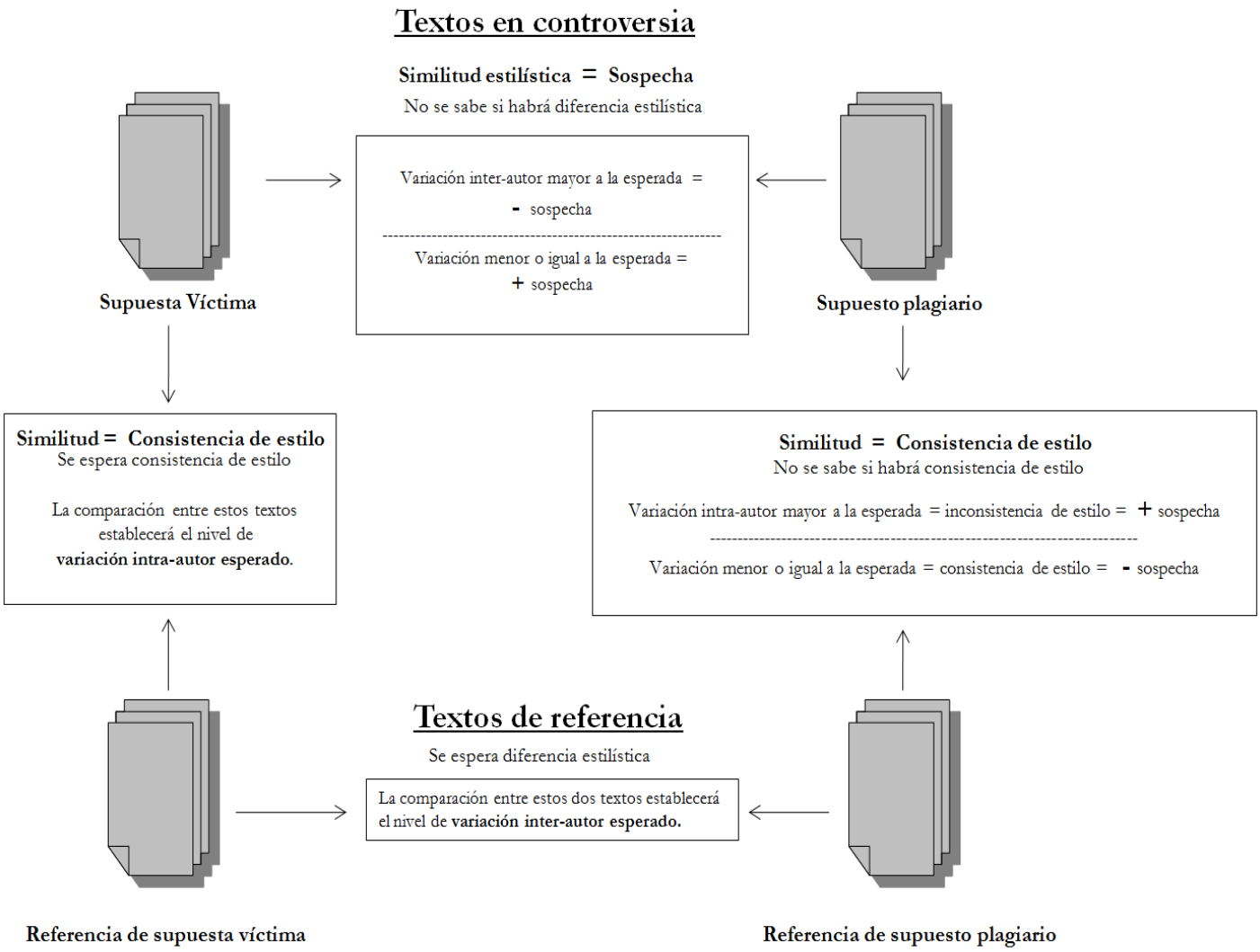
En la parte superior de la figura se puede observar que la comparación entre los textos en controversia presenta dos opciones:

1. Variación inter-autor mayor a la esperada, que es igual a mayor sospecha de plagio, y
2. Variación inter-autor menor o igual a la esperada, que es igual a menor sospecha de plagio.

En la parte izquierda del esquema se muestra que se espera consistencia de estilo y que el resultado de la comparación entre los dos textos de la supuesta víctima supondrá un nivel esperado de variación intra-autor.

Finalmente, en la parte derecha de la figura, se dice que no se sabe si habrá consistencia de estilo y se dan dos posibilidades:

1. Variación intra-autor mayor a la esperada, que es igual a inconsistencia de estilo y, por tanto, a mayor sospecha y
2. Variación intra-autor menor a la esperada, que es igual a consistencia de estilo y, por consiguiente, a menor sospecha de plagio.



Esquema1: Pruebas de similitud estilística y de referencia

Una vez explicado el procedimiento de comparación y análisis que se seguirá, se puede exponer los datos obtenidos del análisis realizado.

4.4 Análisis estilográfico y estilométrico para la detección de plagio

El análisis estadístico realizado se compone de dos partes: el análisis estilográfico o descriptivo y el análisis estilométrico o estadístico. Por tanto, el apartado de cada marcador estilístico incluye una descripción de puntos observados en las gráficas y tablas de frecuencias. El análisis estadístico de todos los marcadores se concentra en el apartado 4.4.2 y es seguido de la descripción de resultados obtenidos.

4.4.1 Análisis descriptivo

Este análisis consta de la descripción de los datos obtenidos por medio del programa *Signature*. Dicha descripción se realiza por medio de gráficas y tablas que muestren el comportamiento y las tendencias de los datos.

4.4.1.1 Longitud de palabra

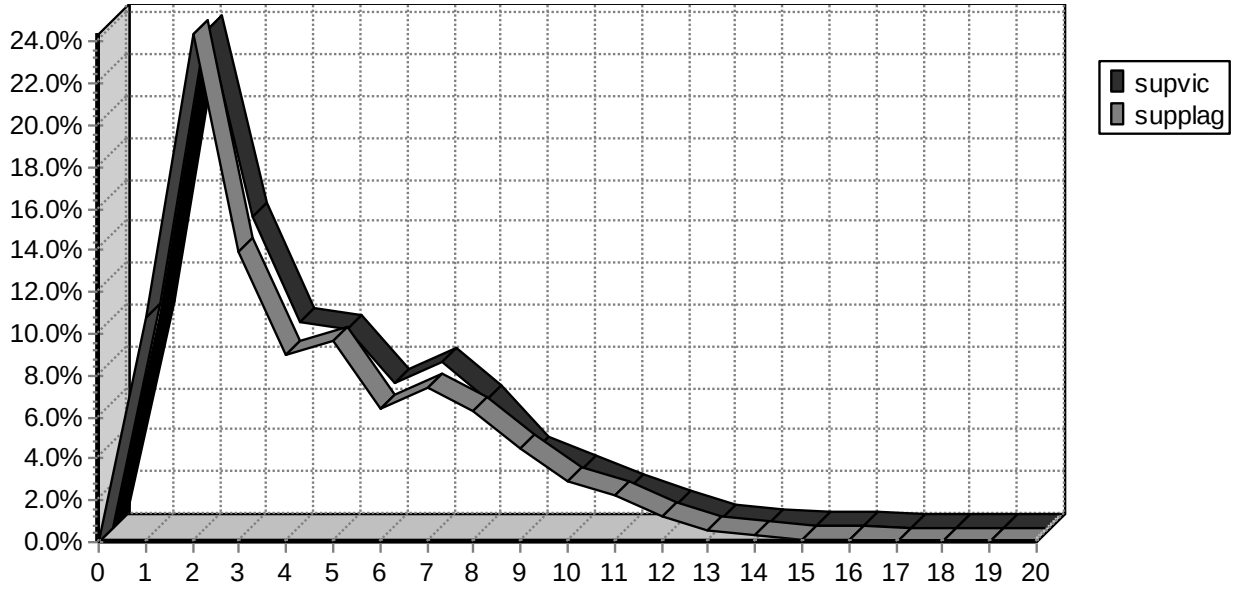
En cuanto a las obras en controversia, en la tabla 1 se puede observar que ambos autores manejan porcentajes similares de longitud de palabra. La constante en los porcentajes de longitud de palabra es que en ambos autores los porcentajes se mantienen con una variación constante de no más de un punto porcentual, sin embargo existen variaciones mayores que pueden destacarse, tales como las de las

longitudes 3 y 9, en las que la diferencia porcentual es de .99 y .84, respectivamente.

Tabla 1: Distribución de frecuencias de uso, absoluta y porcentual, de longitud de palabra de textos en controversia y de referencia.

Longitud de palabra	Textos en controversia				Diferencia %	Textos de referencia			
	Frecuencia absoluta de uso		Porcentaje de uso			Frecuencia absoluta de uso		Porcentaje de uso	
	supplag	supvic	supplag	supvic		supplagref	supvicref	supplagref	supvicref
1	1168	2769	10.74%	10.19%	0.55%	1796	701	11.23%	11.20%
2	2650	6513	24.37%	24%	0.37%	3681	1423	23.02%	22.80%
3	1514	4052	13.92%	14.91%	**0.99%**	2492	966	15.59%	15.50%
4	983	2672	9.04%	9.83%	0.79%	1291	502	8.08%	8.04%
5	1048	2601	9.64%	9.57%	0.07%	1575	615	9.85%	9.84%
6	701	1881	6.45%	6.92%	0.47%	1242	431	7.77%	6.90%
7	810	2168	7.45%	7.98%	0.53%	1197	395	7.49%	6.32%
8	684	1669	6.29%	6.14%	0.15%	899	369	5.62%	5.91%
9	494	1005	4.54%	3.70%	**0.84%**	652	254	4.08%	4.07%
10	318	764	2.92%	2.81%	0.11%	528	257	3.30%	4.11%
11	241	534	2.22%	1.96%	0.26%	260	135	1.63%	2.16%
12	137	302	1.26%	1.11%	0.15%	178	115	1.11%	1.84%
13	67	117	0.62%	0.43%	0.19%	112	32	0.70%	0.51%
14	33	53	0.30%	0.20%	0.10%	53	29	0.33%	0.46%
15	14	36	0.13%	0.13%	0.00%	14	13	0.09%	0.21%
16	7	26	0.06%	0.10%	0.04%	14	7	0.09%	0.11%
17	3	6	0.03%	0.02%	0.01%	1	2	0.01%	0.03%
18	0	5	0.00%	0.02%	0.02%	2	1	0.00%	0.02%
19	1	2	0.01%	0.01%	0.00%	0	0	0.00%	0%
20	0	1	0.00%	0.00%	0.00%	0	0	0.00%	0%
Total	**10873**	**27176**	**100.00%**	**100.00%**		**15987**	**6247**	**100.00%**	**100.00%**

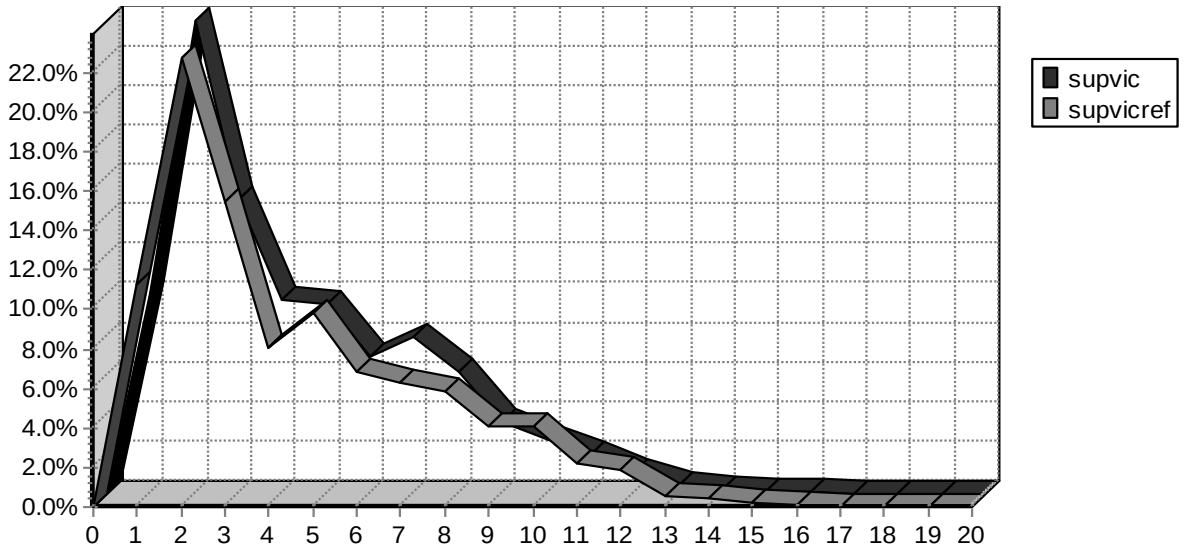
Esto se refuerza observando la gráfica 1 en la que la curva de frecuencia de ambos textos es muy parecida, pero las variaciones en las longitudes 3 y 9 también se hacen notorias. Otro punto que se debe señalar es que la frecuencia absoluta de longitud de palabra del texto de la supuesta víctima muestra que este autor utilizó, en total, 40 palabras de más palabras de 16 caracteres (26 de 16, 6 de 17, 5 de 18, 2 de 19 y 1 de 20), mientras que el supuesto plagiario utilizó solo 11 (7 de 16, 3 de 17 y 1 de 19) de estas palabras que pueden ser consideradas como largas.



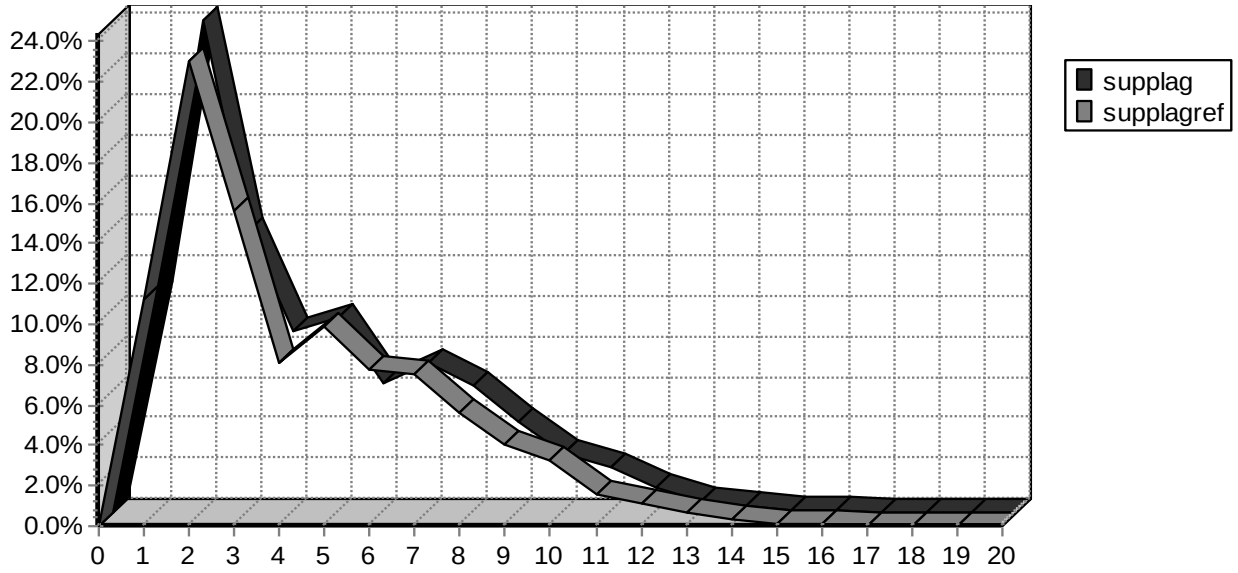
Gráfica 1: Distribución de frecuencias de longitud de palabra entre obras en controversia.

Las gráficas 2 y 3 nos muestran la parte descriptiva de la prueba de consistencia de estilo de los autores. La supuesta víctima presenta variaciones en las longitudes de palabra 4, 7 y 10, con respecto al supuesto plagiario, aunque la curva de frecuencia de las obras es parecida. El supuesto plagiario muestra una curva de frecuencias parecida a la de la supuesta víctima, con variaciones en las longitudes 3, 4, 6 y 10. En el caso de la longitud de palabra es poco lo que puede

observarse en el análisis descriptivo; sin embargo, se recomienda realizarlo para efectuar después el análisis estadístico que muestra más detalladamente las igualdades y diferencias entre los textos en controversia y de referencia.



Gráfica 2: Distribución de frecuencias de longitud de palabra entre obras de la supuesta víctima.



Gráfica 3: Distribución de frecuencias de longitud de palabra entre obras del supuesto plagiario.

4.4.1.2 Longitud de oración

En lo que respecta a este marcador estilístico, se presentan dos tablas de frecuencia. La tabla 2 muestra las frecuencias absoluta y relativa de la longitud de oración de las obras en controversia y de las de referencia. La tabla 3, por su parte, muestra solamente la frecuencia absoluta de las longitudes de oración del texto en controversia de la supuesta víctima y su referencia, pero de una manera categorizada; esto es, se agruparon las categorías de la variable longitud de oración en grupos de frecuencias para tener mejor manejo de los datos.

Tabla 2: Distribución de frecuencias de uso, absoluta y porcentual, de longitud de oración entre obras en controversia y de referencia

Longitud de oración	Textos en controversia				Textos de referencia			
	Frecuencia absoluta de uso		Porcentaje de uso		Frecuencia absoluta de uso		Porcentaje de uso	
	supplag	supvic	supplag	supvic	supplagref	supvicref	supplagref	supvicref
1	4	10	1.46%	1.48%	0	0	0%	0%
2	0	2	0%	0.30%	1	1	0.22%	0.76%
3	3	6	1.09%	0.89%	0	0	0%	0%
4	4	6	1.46%	0.89%	0	1	0%	0.76%
5	0	5	0%	0.74%	0	0	0%	0%
6	0	2	0%	0.30%	4	1	0.87%	0.76%
7	1	12	0.36%	1.77%	1	1	0.22%	0.76%
8	0	4	0%	0.59%	0	0	0%	0%
9	2	5	0.73%	0.74%	4	0	0.87%	0%
10	1	7	0.36%	1.03%	3	0	0.65%	0%
11	1	3	0.36%	0.44%	0	3	0%	2.29%
12	2	16	0.73%	2.36%	2	1	0.43%	0.76%
13	2	9	0.73%	1.33%	9	2	1.95%	1.53%
14	1	9	0.36%	1.33%	4	3	0.87%	2.29%
15	5	12	1.82%	1.77%	9	0	1.95%	0%
16	6	16	2.18%	2.36%	3	3	0.65%	2.29%
17	4	7	1.46%	1.03%	7	1	1.52%	0.76%

18	3	15	1.09%	2.22%	9	2	1.95%	1.53%
19	3	10	1.09%	1.48%	8	0	1.73%	0%
20	7	17	2.55%	2.51%	7	3	1.52%	2.29%
21	5	7	1.82%	1.03%	11	2	2.38%	1.53%
22	4	14	1.46%	2.07%	12	2	2.60%	1.53%
23	3	11	1.09%	1.63%	15	8	3.25%	6.11%
24	2	8	0.73%	1.18%	11	2	2.38%	1.53%
25	9	8	3.27%	1.18%	11	1	2.38%	0.76%
26	5	14	1.82%	2.07%	16	1	3.46%	0.76%
27	6	10	2.18%	1.48%	14	4	3.03%	3.05%
28	3	12	1.09%	1.77%	21	0	4.55%	0%
29	4	17	1.46%	2.51%	12	4	2.60%	3.05%
30	9	11	3.27%	1.63%	14	3	3.03%	2.29%
31	4	10	1.46%	1.48%	13	2	2.81%	1.53%
32	7	10	2.55%	1.48%	18	0	3.90%	0%
33	7	8	2.55%	1.18%	10	4	2.17%	3.05%
34	8	11	2.91%	1.63%	14	1	3.03%	0.76%
35	5	17	1.82%	2.51%	8	2	1.73%	1.53%
36	6	8	2.18%	1.18%	5	2	1.08%	1.53%
37	4	12	1.46%	1.77%	11	1	2.38%	0.76%
38	7	13	2.55%	1.92%	16	1	3.46%	0.76%
39	3	10	1.09%	1.48%	10	5	2.17%	3.82%
40	6	8	2.18%	1.18%	7	1	1.52%	0.76%
41	7	10	2.55%	1.48%	14	1	3.03%	0.76%
42	8	8	2.91%	1.18%	5	1	1.08%	0.76%
43	5	6	1.82%	0.89%	9	2	1.95%	1.53%
44	4	8	1.46%	1.18%	7	2	1.52%	1.53%
45	6	4	2.18%	0.59%	6	0	1.30%	0%
46	7	12	2.55%	1.77%	10	3	2.17%	2.29%
47	5	9	1.82%	1.33%	5	0	1.08%	0%
48	2	10	0.73%	1.48%	6	3	1.30%	2.29%
49	3	10	1.09%	1.48%	5	0	1.08%	0%
50	5	10	1.82%	1.48%	8	1	1.73%	0.76%
51	2	12	0.73%	1.77%	9	0	1.95%	0%
52	5	7	1.82%	1.03%	2	2	0.43%	1.53%
53	2	11	0.73%	1.63%	10	2	2.17%	1.53%
54	4	6	1.46%	0.89%	6	1	1.30%	0.76%
55	3	3	1.09%	0.44%	2	0	0.43%	0%
56	4	5	1.46%	0.74%	3	2	0.65%	1.53%
57	0	6	0%	0.89%	2	0	0.43%	0%
58	4	6	1.46%	0.89%	2	2	0.43%	1.53%
59	3	7	1.09%	1.03%	4	2	0.87%	1.53%
60	2	4	0.73%	0.59%	2	2	0.43%	1.53%
61	1	4	0.36%	0.59%	4	0	0.87%	0%
62	3	7	1.09%	1.03%	1	3	0.22%	2.29%
63	3	4	1.09%	0.59%	0	0	0%	0%
64	0	4	0%	0.59%	3	0	0.65%	0%
65	2	4	0.73%	0.59%	1	1	0.22%	0.76%

66	1	4	0.36%	0.59%	0	1	0%	0.76%
67	3	2	1.09%	0.30%	2	4	0.43%	3.05%
68	3	6	1.09%	0.89%	1	0	0.22%	0%
69	2	4	0.73%	0.59%	0	2	0%	1.53%
70	2	10	0.73%	1.48%	1	1	0.22%	0.76%
71	0	1	0%	0.15%	1	0	0.22%	0%
72	1	0	0.36%	0%	1	1	0.22%	0.76%
73	1	3	0.36%	0.44%	1	0	0.22%	0%
74	1	5	0.36%	0.74%	2	1	0.43%	0.76%
75	0	2	0%	0.30%	0	2	0%	1.53%
76	0	2	0%	0.30%	1	1	0.22%	0.76%
77	0	5	0%	0.74%	0	0	0%	0%
78	0	2	0%	0.30%	1	1	0.22%	0.76%
79	2	3	0.73%	0.44%	0	1	0%	0.76%
80	0	3	0%	0.44%	0	0	0%	0%
81	0	4	0%	0.59%	1	1	0.22%	0.76%
82	2	3	0.73%	0.44%	0	0	0%	0%
83	0	1	0%	0.15%	0	0	0%	0%
84	0	1	0%	0.15%	1	1	0.22%	0.76%
85	1	3	0.36%	0.44%	0	1	0%	0.76%
86	0	1	0%	0.15%	0	0	0%	0%
87	0	1	0%	0.15%	0	0	0%	0%
88	1	2	0.36%	0.30%	0	2	0%	1.53%
89	0	4	0%	0.59%	0	1	0%	0.76%
90	1	0	0.36%	0%	0	0	0%	0%
91	0	4	0%	0.59%	0	0	0%	0%
92	0	4	0%	0.59%	0	0	0%	0%
93	1	2	0.36%	0.30%	0	0	0%	0%
94	1	3	0.36%	0.44%	1	1	0.22%	0.76%
95	1	0	0.36%	0%	0	0	0%	0%
96	1	0	0.36%	0%	0	0	0%	0%
97	0	2	0%	0.30%	0	0	0%	0%
98	0	0	0%	0%	0	0	0%	0%
99	1	1	0.36%	0.15%	1	0	0.22%	0%
100	3	20	1.09%	2.95%	1	11	0.22%	8.40%
Total	**275**	**677**	**100.00%**	**100.00%**	**462**	**131**	**100.00%**	**100.00%**

La razón de haber realizado esta categorización únicamente para la prueba de consistencia de estilo de la supuesta víctima fue que la prueba estadística de la ji cuadrada, de la que se hablará más delante, no permite que los valores de longitud de oración calculados sean menores al 5 % y, como puede observarse en la tabla 2, casi todas las longitudes porcentuales de oración son inferiores a dicho porcentaje.

La categorización de la tabla 3 se hizo con base en la observación de las diferentes frecuencias de longitud de oración observadas en los textos de la supuesta víctima. Se ahondará en la categorización en la parte del análisis descriptivo de la prueba de consistencia de estilo de la supuesta víctima.

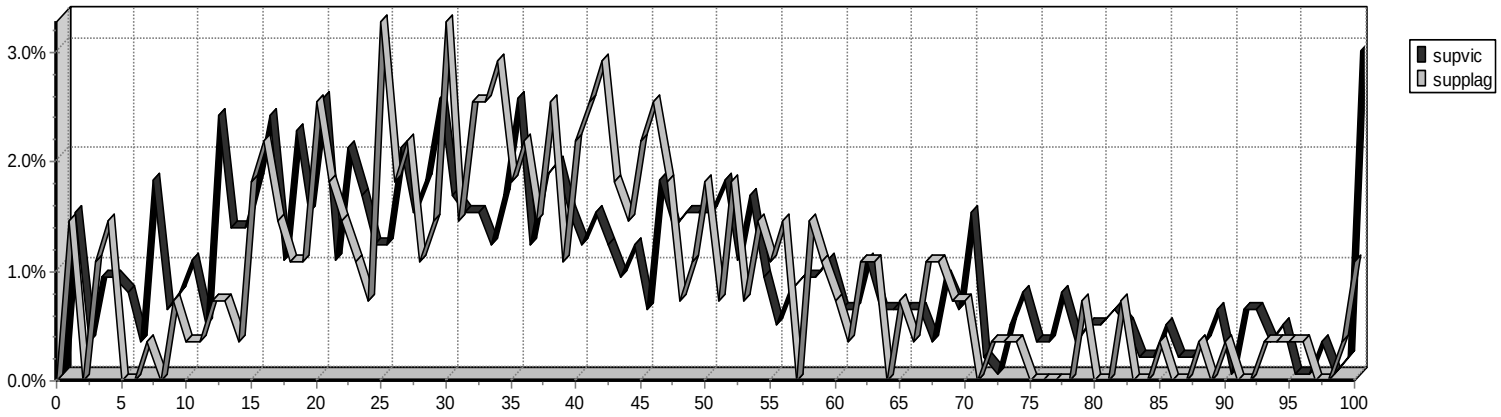
Tabla 3: Distribución de frecuencias absolutas de uso de longitud de oración entre obras de la supuesta víctima.

Longitud de oración		supvic	
		supvic	supvicref
Categoría	De 1 a 10	59	4
	De 11 a 30	226	45
	De 31 a 50	194	32
	De 51 a 70	116	25
	De 71 a 90	46	13
	De 91 o más	36	12
Total		**677**	**131**

Hecha la aclaración acerca de la tabla 3, se puede empezar a realizar el análisis descriptivo.

En primer lugar, en cuanto a los textos en controversia, llama la atención que la mayoría de los porcentajes de longitud de oración se mantienen por debajo del 2.5% y los que superan este nivel son las longitudes 25 y 30 en el caso del supuesto plagiario, y la 100 en el caso de la supuesta víctima. Aunado a eso, se puede observar que ambos autores presentan porcentajes altos de longitud de oración particularmente en el rango que va de la longitud 15 a la 50, pero la supuesta víctima los presenta también en la longitud 100. Dicha situación se cumple de la misma forma en los textos de referencia de los autores; incluso, el

porcentaje de uso de longitud de oración mayor a 100 en la obra de referencia de la supuesta víctima es mucho más claro, pues llega a 8%.

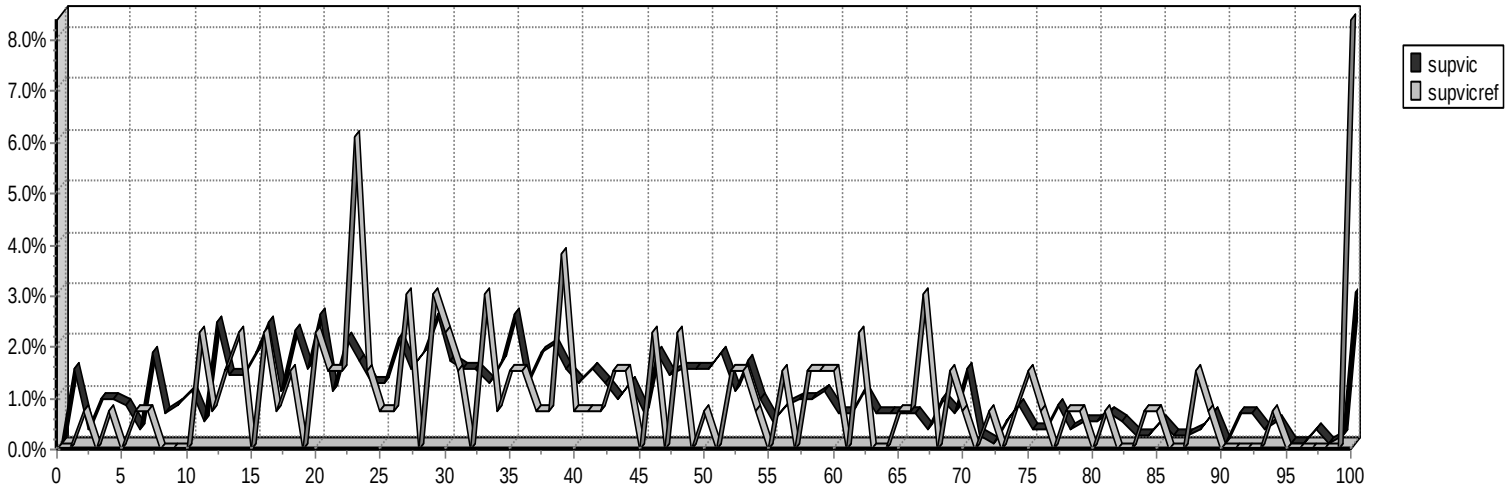


Gráfica 4: Distribución de frecuencias de longitud de oración entre obras en controversia.

Lo anterior lleva a concluir que el supuesto plagiario utiliza más frecuentemente oraciones que van de cortas (de 15 palabras) hasta medianas (de 50 palabras), mientras que la supuesta víctima tiende a utilizar oraciones que van de cortas (de 10 palabras) a largas (de 100 palabras). Esta observación se cumple tanto en los textos en controversia como en los textos de referencia.

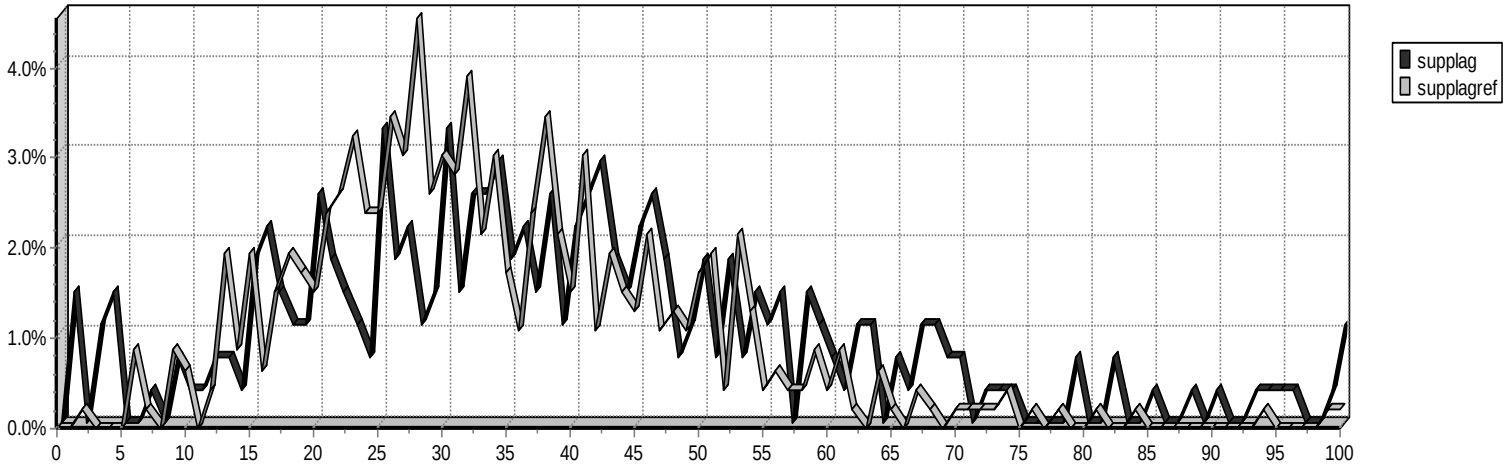
Las gráficas de longitud de oración refuerzan lo dicho. En la gráfica 4, puede observarse que las frecuencias de la supuesta víctima se mantienen al mismo nivel o por debajo de la curva del supuesto plagiario en la mayoría de los picos; sin embargo, la longitud 100 muestra un dramático incremento por parte de la supuesta víctima, mientras que el supuesto plagiario presenta un incremento también, pero de menor longitud. La gráfica 5, la de consistencia de estilo de la supuesta víctima, muestra que el incremento de la longitud 100 se da tanto en la

obra en controversia como en la obra de referencia de la supuesta víctima, lo cual sugiere consistencia estilística al menos en lo concerniente a este marcador.



Gráfica 5: Distribución de frecuencias de longitud de oración entre obras de la supuesta víctima.

La gráfica 6 muestra las longitudes de oración del supuesto plagiario y en ella se puede observar que el estilo de utilizar oraciones que van de cortas a medianas se mantiene.



Gráfica 6: Distribución de frecuencias de longitud de oración entre obras del supuesto plagiario.

4.4.1.3 Longitud de párrafo

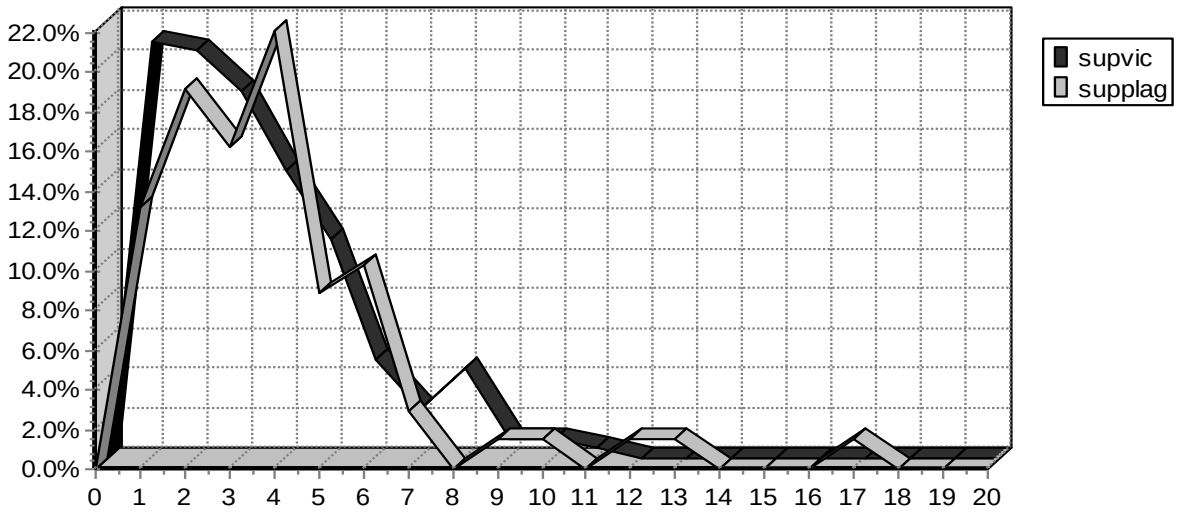
En la tabla 4 puede observarse datos que refuerzan lo observado en el apartado referente a longitud de oración. En los textos en controversia, la distribución de longitudes de párrafo de la supuesta víctima se da entre las longitudes 1 y 11 en su mayoría, mientras que la del supuesto plagiario se extiende hasta 17, lo cual sugiere que la supuesta víctima utiliza más frecuentemente párrafos con pocas oraciones, mientras que el supuesto plagiario maneja párrafos con mayor cantidad de oraciones. Por su parte, los textos de referencia cumplen con la lo observado anteriormente. En el caso del texto de referencia de la supuesta víctima, de hecho, no hay párrafos de más de 7 oraciones, mientras que el texto de referencia del supuesto plagiario presenta párrafos de longitud 8, 9 y 12.

Tabla 4: Distribución de frecuencias de uso, absoluta y porcentual, de longitud de párrafo en obras en controversia y de referencia

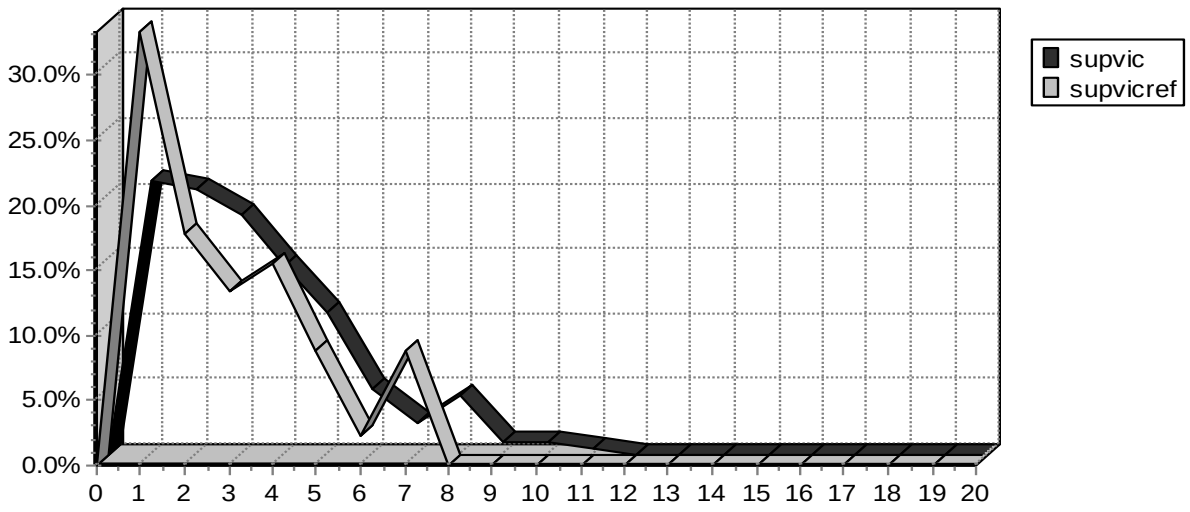
Longitud de párrafo	Textos en controversia				Textos de referencia			
	Frecuencia absoluta de uso		Porcentaje de uso		Frecuencia absoluta de uso		Porcentaje de uso	
	supplag	supvic	supplag	supvic	supplagref	supvicref	supplagref	supvicref
1	9	42	13.24%	21%	17	15	13.60%	33.33%
2	13	41	19.12%	20.50%	24	8	19.20%	17.78%
3	11	37	16.18%	18.50%	22	6	17.60%	13.33%
4	15	29	22.05%	14.50%	19	7	15.20%	15.56%
5	6	22	8.82%	11%	21	4	16.80%	8.89%
6	7	10	10.29%	5%	14	1	11.20%	2.22%
7	2	5	2.94%	2.50%	3	4	2.40%	8.89%
8	0	9	0%	4.50%	3	0	2.40%	0%
9	1	2	1.47%	1%	1	0	0.80%	0%

10	1	2	1.47%	1%	0	0	0%	0%
11	0	1	0%	0.50%	0	0	0%	0%
12	1	0	1.47%	0%	1	0	0.80%	0%
13	1	0	1.47%	0%	0	0	0%	0%
14	0	0	0%	0%	0	0	0%	0%
15	0	0	0%	0%	0	0	0%	0%
16	0	0	0%	0%	0	0	0%	0%
17	1	0	1.47%	0%	0	0	0%	0%
18	0	0	0%	0%	0	0	0%	0%
Total	**68**	**200**	**100.00 %**	**100.00 %**	**125**	**45**	**100.00%**	**100.00%**

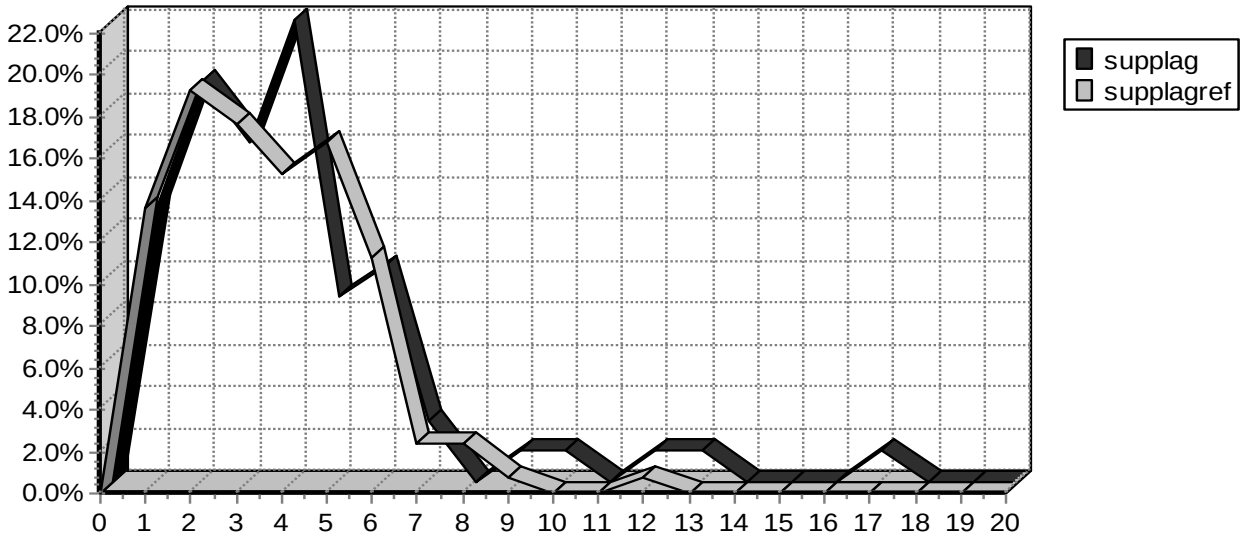
Las gráficas de longitud de párrafo muestran más claramente lo observado en la tabla 4. En primer lugar, la gráfica 7 presenta ocurrencia de párrafos de longitud mayor a 12 en el caso del supuesto plagiario, mientras que la supuesta víctima no los tiene. En segundo lugar, la gráfica 8 muestra una longitud máxima de 12 en la supuesta víctima y la gráfica 9 revela la existencia de párrafos con longitud mayor a 12 en el caso del supuesto plagiario. Cabe aclarar que el porcentaje de aparición de párrafos con longitudes mayores a 12 en el supuesto plagiario es mínimo, sin embargo se presenta tanto en el texto en controversia como en el texto de referencia. De la misma forma, la longitud de párrafo menor a 12 se presenta en ambos textos de la supuesta víctima.



Gráfica 7: Distribución de frecuencias de longitud de párrafo entre obras en controversia.



Gráfica 8: Distribución de frecuencias de longitud de párrafo entre obras de la supuesta víctima.



Gráfica 9: Distribución de frecuencias de longitud de párrafo entre obras del supuesto plagiario.

4.4.1.4 Uso de signos de puntuación

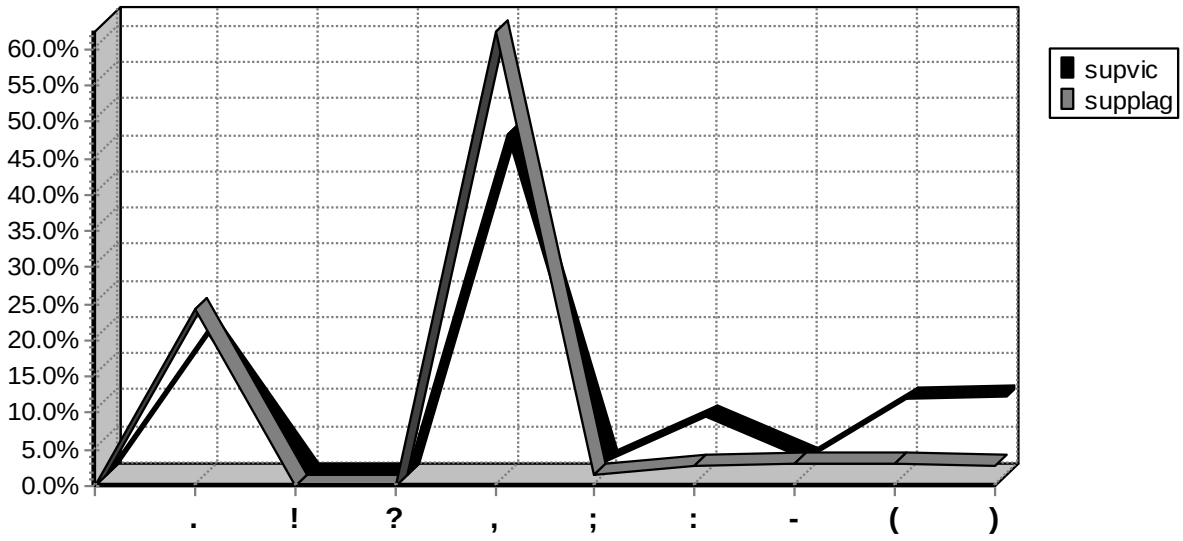
El uso de signos de puntuación se mide de una forma diferente a los marcadores anteriores. En este caso, en lugar de tratarse de una medición de frecuencia de longitud, se trata de un conteo de frecuencia de uso directo de los caracteres de puntuación: punto, coma, signos de admiración y de interrogación, punto y coma, dos puntos, guión y paréntesis.

En la tabla 5 se muestran las frecuencias por cada signo de puntuación y puede observarse que el porcentaje de uso de punto, coma y punto y coma es mayor en el supuesto plagiario que en la supuesta víctima, mientras que los dos puntos y los paréntesis son más frecuentemente utilizados por la supuesta víctima, esto tanto en los textos en controversia como en los textos de referencia.

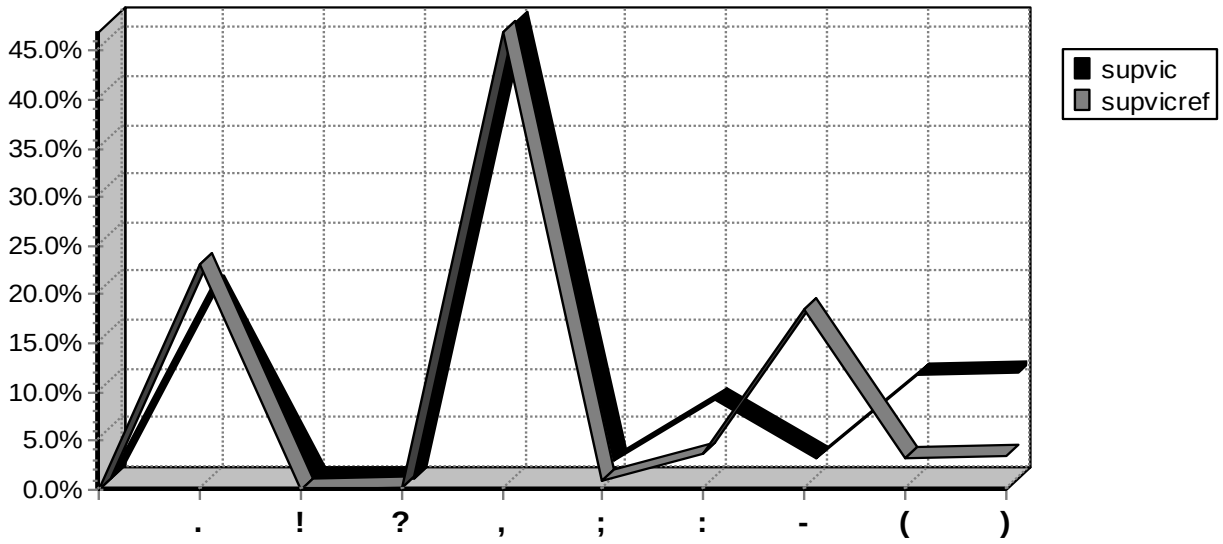
Tabla 5: Distribución de frecuencias de uso, absoluta y porcentual, en signos de puntuación entre obras en controversia y de referencia.

Uso de signos de puntuación	Textos en controversia				Textos de referencia			
	Frecuencia absoluta de uso		Porcentaje de uso		Frecuencia absoluta de uso		Porcentaje de uso	
	supplag	supvic	supplag	supvic	supplagref	supvicref	supplagref	supvicref
.	260	674	24.41%	19.70%	506	150	31.76%	23.22%
!	1	0	0.09%	0%	5	0	0.31%	0%
?	0	0	0%	0%	1	1	0.06%	0.15%
,	665	1595	62.44%	46.61%	923	303	57.94%	46.90%
;	16	67	1.50%	1.96%	16	6	1.00%	0.93%
:	28	277	2.63%	8.10%	47	24	2.95%	3.72%
-	32	75	3.01%	2.19%	46	119	2.89%	18.42%
(	32	362	3.01%	10.58%	25	21	1.57%	3.25%
)	31	372	2.91%	10.86%	24	22	1.51%	3.41%
Total	**1065**	**3422**	**100.00%**	**100.00%**	**1593**	**646**	**100.00%**	**100.00%**

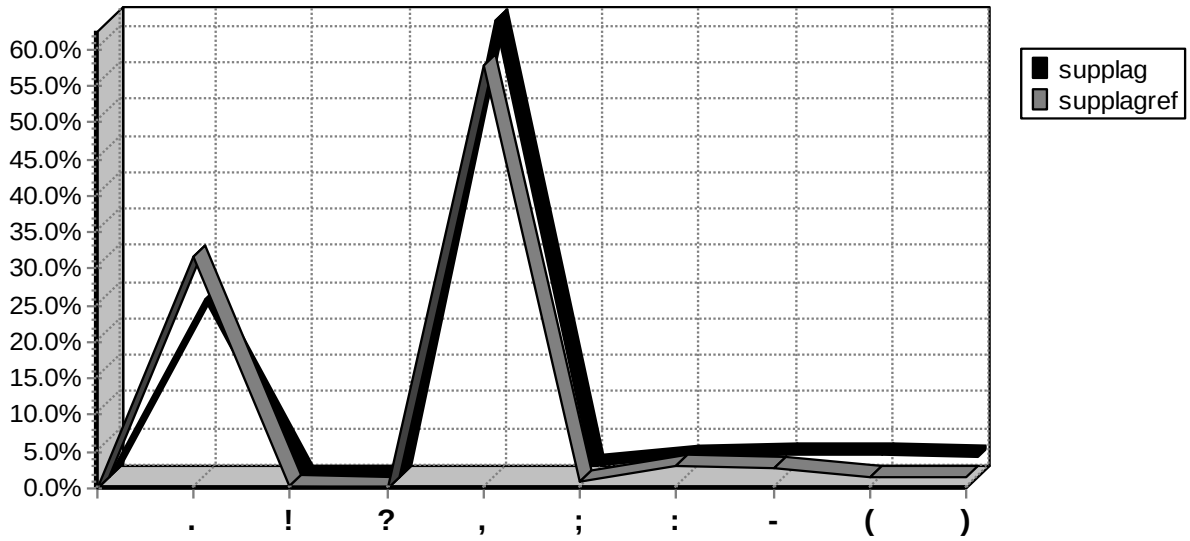
La gráfica 10, la de comparación de uso de signos de puntuación entre los textos en controversia, no demuestra similitud a simple vista, sin embargo la gráfica 12, la de consistencia de estilo del supuesto plagiario, en lo que a puntuación se refiere, sí sugiere consistencia de estilo. La gráfica 11, por su parte, muestra consistencia de la supuesta víctima solo en el uso de puntos, comas y puntos y comas; pero muestra diferencias, al menos gráficas, en dos puntos, guiones y paréntesis.



Gráfica 10: Distribución de frecuencias de signos de puntuación entre obras en controversia.



Gráfica 11: Distribución de frecuencias de signos de puntuación entre obras de la supuesta víctima.



Gráfica 12: Distribución de frecuencias de signos de puntuación entre obras del supuesto plagiario.

Finalmente, es necesario aclarar que lo llevado a cabo hasta el momento es una prueba descriptiva de los datos y en ningún momento debe tomarse como resultado. A continuación se detalla la segunda parte de la prueba realizada en este trabajo; la de ji cuadrada. Para que sea posible ofrecer conclusiones acerca de si existe similitud textual entre los textos en controversia y consistencia de estilo entre los autores es necesario llevar a cabo dicha prueba y contrastar los resultados que arroje con los observados en la prueba descriptiva, de tal forma que las consideraciones finales sean resultado de un equilibrio entre lo observado por la experiencia del lingüista y lo obtenido de datos estadísticos.

4.4.2 Análisis estadístico

Para complementar el análisis descriptivo, es necesario realizar un análisis estadístico exhaustivo que confiera un mayor grado de rigor al análisis. En el

siguiente apartado se describe el análisis estadístico exhaustivo realizado en este caso específico.

4.4.2.1 Descripción de resultados de las pruebas estadísticas

Además de mostrar gráficas y tablas de frecuencia para realizar análisis descriptivos, el sistema *Signature* ofrece la posibilidad de obtener resultados estadísticos. Estos resultados se obtienen con base en la prueba de la ji cuadrada, que será descrita de manera breve en el siguiente apartado, de tal forma que sea posible entender lo que la prueba hace y la interpretación de los resultados que la prueba otorga.

4.4.2.2 Descripción de la prueba estadística $\chi 2$

En estadística existen dos tipos de datos: los numéricos y los categóricos. Los datos numéricos son todos aquellos que se refieren a conteos o numeraciones que van de cero a un número finito o a un número infinito, mientras que los datos categóricos registran cualidades o categorías. Así, el conteo de autos que circulan sobre una vialidad es un dato numérico y la marca de esos mismos autos constituye un dato categórico.

Los datos que se manejan en este estudio son de tipo categórico. Las variables categóricas son los textos a analizar y tienen dos categorías dependiendo del análisis; la primera serían los textos que se comparan (supvic y supplag ó

supvic y supvicref, etc.) y la segunda variable categórica es el marcador estilístico que se analiza (la longitud de oración, la longitud de palabra, etc.) y sus categorías dependen del número de palabras en la oración o del número de letras en las palabras.

Ahora bien, existen pruebas estadísticas diferentes para cada tipo de datos. En el caso de los datos categóricos, una de las pruebas para medir igualdad de proporciones entre dos variables es la ji cuadrada (χ^2). Esta prueba mide la igualdad entre dos variables y para ello plantea, de entrada, la hipótesis de que no existen diferencias entre las dos variables categóricas. A esta aseveración se le llama hipótesis nula y se rechaza en caso de que el valor de ji cuadrada resulte significativo (Pardo Merino & Ruiz Díaz, 2002). La significancia, normalmente en estadística, se mide con referencia a un *P-value* que es la probabilidad de encontrar un valor más extremo al que se ha obtenido; si el valor de ji cuadrada es significativo, entonces se rechazará la hipótesis nula; es decir, se rechazará que no existe diferencia entre las dos variables categóricas y se aceptará la hipótesis alternativa que denota que sí existen diferencias significativas entre las variables. En síntesis, la hipótesis nula es igualdad entre los textos y la hipótesis alternativa es que no existe igualdad.

4.4.3 Descripción de resultados de las pruebas de significancia estadística

Cada uno de los marcadores estilísticos analizados en este estudio presenta una tabla de contingencia de valores observados (tablas 1 a 5). Con base en dichas

tablas, el sistema *Signature* realiza la prueba ji cuadrada de manera automática, de tal forma que nos ofrece valores con marcas específicas de significancia.

En la tabla 6 se simplifica los valores ji cuadrada y se hace una distinción básica entre igualdad (=) y diferencia (≠). Cabe aclarar que esta distinción se realizó con base en la interpretación de valores de ji cuadrada, mismos que no se especifican, ya que se trata de valores numéricos cuyo valor absoluto conllevaría cierto grado de confusión en el análisis a simple vista. Basta con tener presente que donde hay una marca de diferencia (≠) la hipótesis nula o idea de igualdad se rechaza; es decir, que no existe similitud de estilo, mientras que donde se observa una marca de igualdad (=) la hipótesis nula, o idea de similitud, se acepta y, por tanto, existe similitud estilística.

Tabla 6: Resultados de $\chi 2$ correspondientes a las comparaciones de rasgos estilométricos

Rasgo estilométrico	Obras en controversia	Consistencia de estilo entre autores	
	Resultado χ^2 calculado	Resultados χ^2 calculados	
		supplag	supvic
Longitud de palabra	≠	≠	≠
Longitud de oración	=	=	=
Longitud de párrafos	≠	=	≠
Puntuación	≠	≠	≠

≠ Evidencia estadística de diferencias entre las obras.
= Evidencia estadística de similitudes entre las obras.

Los resultados de la prueba estadística para determinar la similitud de estilo entre las obras en controversia muestran que existen diferencias significativas; es decir, que son particularmente diferentes en 3 rasgos estilométricos:

- Longitud de palabra
- Longitud de párrafos
- Uso de puntuación

Por otra parte, en lo concerniente a longitud de oración, la prueba de la ji cuadrada muestra que la diferencia no es significativa, por tanto puede aseverarse que existe similitud textual solamente en uno de los cuatro marcadores discursivos analizados, mientras que hay diferencias significativas en tres de ellos, lo cual apoya la idea, hasta el momento, de que no existe similitud estilística entre los textos en controversia.

La prueba estadística de consistencia de estilo de los autores muestra que el supuesto plagiario no mantiene consistencia de estilo en dos de los marcadores (longitud de palabra y uso de signos de puntuación), mientras que la supuesta víctima no es consistente en tres de los marcadores (longitud de palabra, longitud de párrafo y uso de signos de puntuación), situación que apoya lo observado en la prueba descriptiva. En las gráficas de consistencia de estilo en longitud de palabra (gráficas 2 y 3) y en uso de signos de puntuación (gráficas 11 y 12) no fue posible observar consistencia en ninguno de los dos autores. Por su parte, en las gráficas de consistencia de estilo en longitud de párrafo (gráficas 8 y 9), se anticipó que existía consistencia en los autores. En este caso, la prueba estadística no apoya

totalmente lo observado en la prueba descriptiva; muestra que la supuesta víctima no es consistente, pero reafirma que el supuesto plagiario sí lo es.

Finalmente, la prueba estadística referente a la longitud de oración, sí coincide con lo visto en la prueba descriptiva (4.4.1.2) (gráficas 5 y 6). En cuanto a este marcador, la prueba ji cuadrada muestra que existe consistencia de estilo tanto en la supuesta víctima como en el supuesto plagiario.

5 Conclusiones y trabajo futuro

Como todas las técnicas utilizadas con fines forenses, la estilometría no da una respuesta concreta a la pregunta ¿existe plagio?; sin embargo contribuye a reforzar o a disminuir la sospecha, pues ofrece evidencia de niveles de igualdad o de diferencia.

Con base en los análisis estilográfico y estilométrico presentados en este trabajo, se puede concluir que no existe similitud estilística en tres de los cuatro marcadores de estilo analizados en los textos en controversia; es decir que no se observó variación inter-autor significativa, por tanto la sospecha de copia ilegal o plagio disminuye.

La prueba de consistencia de estilo muestra que el supuesto plagiario mantiene consistencia estilística en dos de los cuatro marcadores de estilo, mientras que la supuesta víctima es inconsistente en tres de los cuatro marcadores de estilo analizados. La variación intra-autor en el caso de los textos del supuesto plagiario fue menor que la observada en los textos de la supuesta víctima.

En el presente caso la sospecha de plagio se ve disminuida, pues la prueba de similitud entre los textos en controversia no muestra que exista similitud de estilo. Aunado a lo anterior, la prueba de consistencia de estilo indica que el supuesto plagiario mantuvo consistencia de estilo entre su texto de referencia y su texto en controversia.

Con respecto al sistema *Signature* se observó que presenta ciertas deficiencias. En primer lugar, no aclara al usuario que necesita un mínimo de porcentaje en las variables para poder realizar la prueba ji cuadrada. En segundo lugar, en este

mismo caso, no permite categorizar los datos. La categorización y la obtención de la ji cuadrada de longitud de oración en la prueba de consistencia de estilo se realizaron con sistemas diferentes. Finalmente, el *Signature* no muestra los pasos de la realización de la prueba ji cuadrada.De lo anterior, se puede concluir que *Signature* no puede ser siempre utilizado; es necesario desarrollar una herramienta que lo supere en la presentación de los datos estadísticos y en la manipulación de los datos textuales.

En lo concerniente a los datos manejados en este estudio se encontró la posibilidad de categorizar las frecuencias de los marcadores estilísticos, longitud de palabra y longitud de párrafo, utilizando criterios lingüísticos, tal como se hizo en el apartado 5.4.1.2, relativo a longitud de oración.

Con respecto al método utilizado, se puede concluir que la estilometría ofrece ventajas en relación con los métodos y las técnicas computacionales que existen actualmente para detectar similitud textual y copia. En primer lugar, permite la clasificación textual y la medición de similitud en un nivel más profundo que el textual; es decir, que es capaz de detectar similitud en textos, que puede deberse a una copia hecha por medio de la paráfrasis. Aunado a esto, la estilometría, comparte ventajas con sistemas de detección de similitud en varios aspectos:

- Es rápida
- Se sustenta en bases estadísticas y numéricas
- Es objetiva
- Puede abarcar una gran cantidad de texto para el análisis

Por tanto, es posible que sea tomada en cuenta por el foro legal, ya que una de las razones por las cuales los dictámenes forenses lingüísticos carecen de confianza es el hecho de que se consideran subjetivos y poco confiables, como mencionan Tiersma y Solan (2002):

> Reasons for judicial reluctance to admit linguistic expertise include concerns that it is not sufficiently reliable, the belief that issues like the meaning of a text can just as well be decided by a jury, and sometimes even institutional and political considerations.

La estilometría, por su parte, no es un método subjetivo; los resultados que ofrece no están supeditados a la interpretación de quien la utilice.

Es necesario aclarar que este trabajo no busca sustituir los métodos existentes de detección de similitud con la estilometría, sino mostrar la efectividad de una nueva opción que puede, y debe, ser utilizada en conjunto con otras.

En lo concerniente a trabajo futuro, la realización de este estudio dentro del ámbito de la lingüística forense abrió dos posibilidades para continuar con investigación de este tipo.

Primero, en lo particular, el trabajo futuro más importante será medir una mayor cantidad de marcadores estilísticos. A mayor cantidad de datos, mayor exactitud en los resultados y, por tanto, debe explorarse la posibilidad de estudiar marcadores estilísticos tales como palabras de contenido, palabras funcionales, marcas tipográficas y estructuras sintácticas y discursivas; todo esto atendiendo, por supuesto, a las peculiaridades de cada caso.

En lo general sería muy interesante explorar la posibilidad de realizar estilometría con base en información morfosintáctica y sintáctica. Como puede observarse a lo largo de este trabajo, la medición de estilo se realizó bajo criterios

textuales superficiales. El uso de etiquetas de partes de la oración, también conocido como etiquetado morfosintáctico del texto, puede dar pauta a realizar mediciones parecidas a la de longitud de oración o longitud de párrafo, pero ya no tomando como unidad de conteo la palabra o la frase computacional, sino las frases y oraciones sintácticas. Estos constituyentes pueden obtenerse, en el caso de las frases, de un proceso de Shallow Parsing; es decir, de análisis de constituyentes sintácticos básicos. Por su parte, las oraciones pueden adquirirse por medio de un proceso de Parsing que "es el análisis completo de constituyentes de la oración y sus relaciones sintácticas, siguiendo las reglas de una gramática".

Otra posibilidad es realizar comparación de estilos con fines forenses usando ya no corpus textuales, sino orales transcritos. Para este fin, es necesario contar con corpus debidamente anotados en niveles fónicos muy específicos, lo cual implicaría, primeramente, todo un trabajo de recopilación de muestras lingüísticas, acompañado de una valoración de sistemas de etiquetado en textos orales transcritos, de acuerdo con diversos parámetros y características de la lengua oral. Lo importante sería la posibilidad de tratar documentos resultantes del etiquetado de corpus orales transcritos con métodos estadísticos que en su origen fueron desarrollados para análisis en textos. Lo anterior podría contribuir a tareas tales como la identificación de hablantes o la delimitación de zonas lingüísticas con base en rasgos estilísticos orales. Esta idea puede parecer atrevida, pero ya existe la intención en el Grupo de Ingeniería Lingüística de la UNAM de recopilar este tipo de muestras y etiquetarlas para, de tal forma, contar con corpus de estudio con los que se pueda poner a prueba iniciativas tales.

Finalmente, una de las conclusiones a las que se llega en este trabajo es que existe la necesidad de explorar e impulsar el área de la lingüística forense. A

diferencia de otros países en los que esta área se encuentra no solo en desarrollo, sino ya establecida, en México la labor de los lingüistas al interior del sistema legal es mínima y, cuando ocurre, la falta de conocimiento con respecto a procedimientos provoca que esta labor se realice de manera irregular y sin bases éticas. En este sentido, esta tesis hace tres aportaciones principales. Primero, ubica la labor del lingüista en el medio forense, por medio de la descripción del área de la lingüística forense. Segundo, mantiene una postura cuidadosa de las afirmaciones que se realizan, de tal forma que, si bien el título habla de la detección de plagio, a lo largo del trabajo se hace énfasis en que el lingüista forense no determina si hay plagio o no; simplemente aporta pruebas que deberán ser analizadas e interpretadas por autoridades competentes. Finalmente, se propone el uso de un método estadístico objetivo en una labor que tradicionalmente dependía de opiniones subjetivas. Se pretende que dicha propuesta sea un exhorto para que todos aquellos que incursionen en la lingüística forense apliquen técnicas, metodologías y herramientas que, lejos de ahuyentar la confianza del sistema legal en la labor lingüística, la refuercen, la validen y la incrementen.

6 Referencias

Amberson, E. (1979). *La crítica literaria y sus métodos.* México: Alianza Editorial.

Bilge, L., Büsra Celikkaya, E., & Hatun, K. (s.f.). *Stylometry in information retrieval sistems.* Obtenido de: http://www.cs.bilkent.edu.tr/~canf/CS533/CS533Spr07stuPresent/stylometry07.ppt

Bloomfield, L. (2010). *The plagiarism resource site.* Recuperado el 9 de febrero de 2011, de: http://plagiarism.phys.virginia.edu/Wsoftware.html

Butcher, A. (s.f.). *Forensic Phonetics: Issues in speaker identification evidence.* Recuperado en marzo de 2009, de Department of Speech Pathology and Audiology: http://www.flinders.edu.au/speechpath/Prato.pdf

Clough, P. (2000). *Plagiarism in natural and programming languages: an overview of current tools and technologies.* Sheffield: Department of Computer Science, University of Sheffield.

Código federal de procedimientos penales. (2009). *Nuevo Código publicado en el Diario Oficial de la Federación el 30 de agosto de 1934* . México: Cámara de diputados del H. Consejo de la Unión.

Coulthard, M., & Johnson, A. (2007). *An Introduction to forensic linguistics: language in evidence.* New York: Routledge.

Dickinson, M. (2007). Text Classification. *Language and Computers* . Georgetown, U. S. A. Recuperado en marzo de 2009 de: http://www9.georgetown.edu/faculty/mad87/06/261/slides/searching.pdf

Eve2. (2000). Recuperado el 9 de febrero de 2011, de: http://www.canexus.com/index.shtml

Fish, S. (1994). ¿Qué es la estilística y porqué se dicen cosas terribles de esta? En Alberto Vital (Ed.), *Conjuntos. Teorías y enfoques literarios recientes* (pp. 101-130). México: Instituto de Investigaciones Filológicas, UNAM.

Fletcher, P., Hughes, A., & Woods, A. (1986). *Statistics in language studies.* Cambridge: Cambridge University Press.

ForensicLab. (2003). *Laboratori de Lingüística Forense.* Recuperado el Marzo de 2011, de: http://www.iula.upf.edu/forensiclab/fpreses.htm

Golcher, F. (2007). A new statistical measure and its application to stylometry. *Proceedings of Corpus Lingüística* , 1-26.

Grieve, J. W. (2005). *Quantitative autorship attribution: a history and an avaluation. Tesis de Maestría.* Vancouver: Simon Fraser University.

Hollien, H. (1990). *The acoustics of crime. The new science of forensic phonetics.* New York: Plenum Press.

International Association of Forensic Linguists. (2006). *International Association of Forensic Linguists*. Recuperado el 26 de febrero de 2009, de: http://www.iafl.org/

Levi, J. (1993). Evaluating jury comprehension of the Illinois capital sentencing instructions. *American Speech* , 20-49.

López Escobedo, F. (2010). *El análisis de las características dinámicas de la señal de habla como posible marca para la comparación e identificaciónforense de voz: Un estudio para el español de México. Tesis de Doctorado.* Barcelona, España: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Martínez del Castillo, J. G. (2004). La lingüística, ciencia del hombre. *Language Design* , 103-138.

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism. A Survey. *Journal of Universal Computer Science* , 1050-1084.

McCombe, N. (2002). *Methods Of Author Identification. Tesis de Licenciatura* . The University of Dublin. Recuperado el 17 de febrero de 2010, de: http://www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/mccombe0102.pdf

McMenamin, G. R. (2002). *Forensic Linguistics. Advances in forensic stylistics.* New York: CRC PRESS.

Millican, J. (2010). *Introduction to Textual Analysis using Signature.* Recuperado el 10 de diciembre de 2010, de Philocomp.net: http://www.philocomp.net/humanities/signature

Millican, J. (2003). *Signature.* Recuperado el 2010, de Philocomp.net: http://www.philocomp.net/humanities/signature

Olsson, J. (2008). *Forensic Linguistics.* London / New York: Continuum.

Olsson, J. (s.f.). *What is Forensic Linguistics?* Recuperado el 26 de febrero de 2009, de The home of Forensic Linguistics: http://www.thetext.co.uk/docs/what_is.doc

Pardo Merino, A., & Ruiz Díaz, M. Á. (2002). *SPSS 11. Guía para el análisis de datos.* Madrid: McGraw Hill.

Peng, R., & Hengartner, N. (2001). *Department of Statistics Papers.* Recuperado el 31 de Mayo de 2010, de Department of Statistics Papers: http://escholarship.org/uc/item/9w56v25f#page-1.

Plagiarism.org. (1996). *Plagiarism.org*. Recuperado el 18 de agosto de 2010, de Plagiarism.org: http://www.plagiarism.org

Real Academia Española. (2001). *Diccionario de la Real Acedemia Española.* España: Espasa Calpe.

Sierra, G. (2008). Diseño de corpus textuales para fines lingüísticos. *Memorias del IX Encuentro Internacional de Lingüística en el Noroeste* (págs. 445-462). Hermosillo, Sonora: Universidad de Sonora.

Sierra, G., & Rosas, A. (2010). Una clasificación de corpus lingüísticos informatizados. En Rosa María Ortiz Ciscomani (Ed.), *Análisis lingüístico: enfoques sincrónico, diacrónico e interdisciplinario* (págs. 335-355). Hermosillo, Sonora: Universidad de Sonora.

Spassova, M. S. (2009). *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español. Tesis de doctorado.* Barcelona, España: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Tiersma, P., & Solan, L. (2002). The linguist on the witness stand: forensic linguistics in american courts. *Language* , 221-239.

Turell, M. T. (2005)(Ed.). *Lingüística forense, lengua y derecho. Conceptos, métodos y aplicaciones.* (págs. 13-16). Barcelona: Documenta Universitaria.

Ley federal de derechos de autor (1996). México, Distrito Federal, México: Diario oficial de la federación. H. Congreso de la Unión.

Viñas Piquer, D. (2002). *Historia de la crítica literaria.* Barcelona: Ariel.

WCopyfind. (2010). Recuperado el 9 de febrero de 2011, de http://plagiarism.phys.virginia.edu/Wsoftware.html