ANEXO A

TRIBUNAL DE DISTRITO DE LOS ESTADOS UNIDOS
DISTRITO SUR DE NUEVA YORK

CHEVRON CORPORATION, Demandante, - contra - STEVEN R. DONZIGER y otros, Demandados.	:	Caso No. 11 Civ. 0691 (LAK)

TESTIMONIO DIRECTO DE PATRICK JUOLA, Ph.D.

Yo, PATRICK JUOLA, declaro por la presente, bajo pena de perjurio conforme al artículo 28 U.S.C. § 1746, que lo que sigue es correcto y veraz:

1. Soy profesor adjunto titular de Ciencias Informáticas de la Duquesne University, en Pittsburgh, Pensilvania. Soy experto en análisis forense e informático, en particular, con relación a análisis de textos y de autoría. He escrito 40 artículos revisados por homólogos, principalmente sobre conclusiones informáticas sobre autoría de documentos a través de análisis estadísticos de características lingüísticas. En mis investigaciones, me centro en el análisis forense e informático de características lingüísticas, y me especializo en el área de atribución de autoría.

2 Fui contratado por Gibson, Dunn & Crutcher, LLP (“Gibson Dunn”) en representación de Chevron Corporation (“Chevron”) en esta causa para establecer si ciertos documentos de los abogados y consultores de los demandantes en la causa *María Aguinda y otros* v. *Chevron Corporation* (también conocida como el caso de Lago Agrio) pueden encontrarse en el expediente judicial de primera instancia.

ANEXO DEL DEMANDANTE
3800 .
11 iv. 0691 (LAK)

CERT. MERRILL VER: JD

Resumen del dictamen

3. Sobre la base de mi análisis pericial de textos basado en cuestiones informáticas de documentos de los abogados y consultores de los demandantes identificados en la sentencia ecuatoriana y en el expediente de primera instancia de la causa Lago Agrio, llegué a la conclusión, con un grado de certeza razonable, de que el Memorando de Fusión, el Informe de Clapp, el Índice de Resúmenes, el correo electrónico sobre el Fideicomiso de Fajardo, el Borrador de Alegato y la Compilación de Datos de Selva Viva no se encuentran en el expediente de primera instancia. Asimismo, llegué a la conclusión, con un grado de certeza razonable, sobre la base de otras evaluaciones forenses, de que el análisis basado en cuestiones informáticas fue eficiente para identificar las fuentes de coincidencias lingüísticas entre los documentos producto del trabajo de los demandantes y la sentencia ecuatoriana, y de que dichas fuentes no existen en el expediente de primera instancia.

Antecedentes profesionales

4. Soy profesor adjunto titular de Ciencias Informáticas de la Duquesne University, en Pittsburgh, Pensilvania. Además, soy director de Evaluación de Variaciones en el Laboratorio de Idiomas, también de Duquesne.

5. Asimismo, soy fundador y director de Investigaciones de J Computing, Inc., (que opera con el nombre de Juola & Associates “J&A”), sociedad constituida en Pensilvania especializada en análisis de texto y autoría.

6. Completé una licenciatura en Ciencias en 1987 sobre ingeniería eléctrica en Johns Hopkins University, Baltimore, donde me especialicé en Matemáticas y en Ingeniería Eléctrica. También obtuve un M.S. en Ciencias Informáticas en la University of Colorado en 1991 y un M.S. en ciencias cognoscitivas, también en la University of Colorado, en 1993. Por último, completé mi doctorado en ciencias informáticas en la University of Colorado en 1995. Fui investigador posdoctoral adjunto del Departamento de Psicología Experimental de St. Hugh’s College y Lincoln College, Oxford University, entre 1995 y 1998.

7. En 1998, comencé en Duquesne, donde asumí el cargo de Profesor Ayudante en Matemáticas y Ciencias Informáticas. En el 2004, recibí la titularidad y pasé a ser Profesor Adjunto. En Duquesne doy clases, incluidas clases sobre procesamiento de lenguaje natural, programación lógica, ingeniería de programas informáticos y criptografía. En mis investigaciones, me centro en el análisis forense e informático de características lingüísticas. Dentro de ello, me especializo en el área de atribución de autoría.

8. Fui incluido en la Oficina de Salón de la Fama en Investigación de Duquesne en el 2009. También en el 2009 recibí el premio a la Excelencia Académica de McAnulty College Faculty en Duquesne.

9. Escribí más de 150 publicaciones en general, 40 de las cuales son artículos aprobados por sistema de revisión por homólogos. También soy autor de 2 libros y 9 capítulos de libros. La mayoría de mis publicaciones trata sobre conclusiones informáticas sobre autoría de documentos a través de análisis estadísticos de características lingüísticas.

10. Me desempeño con regularidad como revisor ad hoc en materias relacionadas con atribución de autoría, estilometría, humanidades digitales y análisis de textos para una serie de revistas, entre las que se incluyen LLC (antiguamente, Literary and Linguistic Computing), JASIST (Journal of the American Society for Information Systems Technology) y SPE (Software Practices and Experiments).

11. Mi empresa, Juola & Associates, se especializa en análisis forense e informático, en particular, con relación a análisis de textos y de autoría. Gestionamos proyectos que involucran atribución de autoría, descripción de características de autor, plagio y también búsqueda de documentos a gran escala.

12. Soy el principal arquitecto y diseñador de J GAAP (Java Graphical Authorship Attribution Program), sistema de análisis de autoría. Este sistema, fundado por la National Science

Foundation (NSF) por cerca de USD 2 millones, es un sistema de desarrollo y evaluación de nuevos métodos de atribución de autoría y determinación de buenas prácticas. Además de desarrollar buenas prácticas, la NSF también me encargó el desarrollo de un sistema de atribución de autoría de calidad forense (objetivo que hemos cumplido con la actual versión de JGAAP) así como la comercialización de tecnología desarrollada para ese fin.

Metodología

13. Primero revisé las conclusiones de otros peritos de la causa, incluidas las del Dr. Robert Leonard, quien identificó coincidencias lingüísticas entre los documentos producto del trabajo de los demandantes y la sentencia ecuatoriana. El Dr. Leonard identificó al menos nueve casos de coincidencias textuales sustanciales entre la sentencia ecuatoriana y los documentos producto del trabajo de los demandantes.[1]

14. El Dr. Leonard identificó coincidencias lingüísticas entre la sentencia ecuatoriana y los siguientes documentos producto del trabajo de los demandantes, que me fueron entregados por los abogados de Chevron:

a. Un documento titulado “Primer Borrador Memo Fusion JSP [Nov2007].doc” (en adelante, el “Memorando de Fusión”, PX 435);
b. Dos versiones de una hoja de cálculo no presentada titulada “pruebas pedidas en etapa de prueba.xls” y “GARR-HDD-003243” (en adelante, el “Resumen de índice de enero”, PX 433, y el “Resumen de índice de junio”, PX 434, y en conjunto los “Resúmenes de índices”);
c. Un borrador de escrito judicial, conocido como “alegato”, que contiene las afirmaciones y los argumentos de los Demandantes de Lago Agrio (en adelante, el “Borrador de Alegato”, PX 438);

[1] *Ver* Informe del Dr. Robert Leonard, 27 de junio del 2011, página 11.

d. Un correo electrónico de “Pablo Fajardo Mendoza” a tres personas, incluido “Steven Donziger” con el asunto “FIDEICOMISO” (en adelante, el “Correo electrónico de Fajardo sobre el fideicomiso”, PX 437);
e. Un correo electrónico(DONZ00025295.pdf) por el que se reenvía un informe que contiene el texto de DONZ00025296.doc (en adelante, el “Informe de Clapp”, PX 928);
f. Un documento que contenía datos de muestra conocidos como la “Compilación de Datos de Selva Viva” (PX 439-41).

15. Recibí el expediente judicial de primera instancia en dos etapas. En la primera etapa, aproximadamente el 13 de septiembre del 2011, el abogado de Chevron entregó a J&A aproximadamente 3.500 documentos electrónicos contenidos en aproximadamente 236.000 imágenes de páginas individuales, numeradas desde el CL0000-00000 hasta el CL2068-0216692, con la Sentencia del 14 de febrero del 2011 desde la página CL2065-0216338. Entendí del abogado de Chevron que esta versión del expediente que revisé es una fotocopia de la versión oficial del expediente conservado por la Corte Provincial de Justicia de Sucumbíos, y que dicha fotocopia fue realizada por el Secretario de la Corte Provincial de Justicia de Sucumbíos según los procedimientos judiciales normales, se le puso el sello en cada página para indicar la autenticidad de la copia, y fue entregado a Chevron en cuotas, según fue solicitado por el abogado ecuatoriano de la empresa durante el curso de la acción ante la primera instancia. Chevron escaneó las copias, creó archivos PDF, que luego fueron convertidos en documentos de una única página con formato TIFF y cargados en una plataforma electrónica, momento a partir del cual los archivos quedaron sujetos a un proceso de OCR automático.

16. El 30 de mayo del 2013, aproximadamente, J&A recibió de Chevron un disco duro que contenía los materiales electrónicos de 69 discos compactos (en adelante, la “Lista de contenidos del CD”) provistos por la Corte Nacional de Justicia de Ecuador. Entendí del abogado de Chevron que el Fiscal

de Ecuador había solicitado previamente copias de toda la información digital contenida en CD o DVD en el expediente judicial y que Chevron luego solicitó sus propias copias de esa información digital.

17. Dado el tamaño y la heterogeneidad de los datos recibidos, nuestra primera tarea fue convertir sistemáticamente todos los documentos a un formato común llamado UTF-16. Este formato es una variación del "texto plano", pero que permite las letras no inglesas o las letras con marcas diacríticas (acentos como "ó"). Esto ofrece una base de búsqueda fundada en palabras o caracteres utilizando una forma común de codificación que puede ser leído por máquinas. En el proceso de esta conversión, también eliminamos toda la puntuación y las diferencias de mayúsculas para maximizar las posibilidades de detectar coincidencias entre texto identificado en los documentos producto del trabajo de los abogados y el expediente del tribunal de primera instancia. Este es un procedimiento conservador que asegura que palabras que difieren solo en cuanto a las mayúsculas o a su puntuación sean correctamente comparadas.[2]

18. Dividimos cada uno de los documentos del expediente judicial en grupos de 5 palabras de longitudes (en adelante, "5-gramas" o en forma más general "n-gramas"). En palabras más simples, se trata sencillamente de grupos de cinco palabras consecutivas. Por ejemplo, la frase inglesa "El Presidente de la Corte Suprema de los Estados Unidos" constituye un 10-grama en sí misma y contiene dentro sí muchos 5-gramas, incluidos "El Presidente de la Corte", "de la Corte Suprema de" y "Suprema de los Estados Unidos".

19. A fin de precaver la posibilidad de errores de OCR, también creamos otro conjunto de n-gramas "parecidas" que trataron a todos los caracteres no latinos (incluidos los caracteres con marcas diacríticas o acentos) como iguales, por lo que el carácter "ó" se consideraría idéntico al carácter "ò" o

[2] El propio documento de la sentencia ecuatoriana es, desde ya, parte del expediente judicial, pero las comparaciones de la sentencia ecuatoriana consigo misma no habrían sido útiles para encontrar las fuentes de las que provino, y por tanto fue removida antes del análisis.

para el caso, el carácter "ö". Este tratamiento se aplica a fin de compensar los potenciales errores introducidos por el proceso de OCR. Los acentos están entre los aspectos más delicados de la escritura en relación con el OCR, ya que un poco de suciedad en el cristal o una impresora o un escáner de mala calidad pueden introducir, cambiar o eliminar marcas que podrían interpretarse como acentos. Al tratar igual a los caracteres que solo difieren en las marcas diacríticas, el efecto de dichos errores sobre el análisis se ve muy reducido o minimizado.

20. Las n-gramas son muy individuales; es poco común ver coincidencias de 7-gramas o más, excepto cuando las n-gramas sean parte de un vocabulario de frases que se superponen. "Presidente de la Corte Suprema de los Estados Unidos" es un ejemplo de ese tipo de frase, conocida por cualquier abogado. "Presidente de la Corte Superior de Justicia de Nueva Loja" es otro ejemplo, tal vez igualmente conocido para un abogado ecuatoriano. Las citas directas e indirectas, por supuesto, serían otra razón válida para que los documentos compartan n-gramas.

21. Después compilamos una lista de cada superposición lingüística específica entre los documentos del trabajo de los demandantes y la sentencia ecuatoriana identificada por el Dr. Leonard (en adelante, los "Ejemplos"). Dividimos los Ejemplos en 5-gramas también.

22. Una vez que los documentos fueron divididos en n-gramas de cinco palabras (5-gramas), utilizamos programas informáticos para identificar los 5-gramas que eran compartidos entre los Ejemplos y el expediente.

23. Sobre la base de las comparaciones, pudimos encontrar documentos en el expediente que contenían una coincidencia exacta (independientemente de las marcas diacríticas) de al menos cinco palabras con uno de los Ejemplos. Para cada uno de estos documentos, también pudimos identificar un área de similitud máxima, describiendo el grado aproximado de superposición y permitiéndonos mirar las

instancias específicas a fin de determinar si el resultado mostraba algún documento fuente para el texto superpuesto.

24. Si la computadora identificaba dichas coincidencias, primero verificábamos la coincidencia comparando visualmente la frase en cuestión con la parte pertinente del expediente judicial. Luego chequeábamos para ver si la coincidencia era una cita directa. Finalmente, analizamos la coincidencia para determinar si era una frase común o estereotipada, juzgando parcialmente la frecuencia y la distribución de la frase en los documentos y parcialmente sobre nuestro entendimiento del significado de la frase.

25. Como ejemplo ilustrativo, consideramos la similitud citada como Ejemplo 1 en el informe del 27 de junio del 2011 del Dr. Leonard.[3] El Ejemplo 1 es un bloque de texto con más de 90 palabras idénticas que aparecen en el Memo de Fusión y en la sentencia ecuatoriana. Al comparar esta superposición de 90 palabras con el expediente judicial, encontramos exactamente once coincidencias de cinco palabras o más en todos los más de tres mil documentos del expediente judicial. Ninguna de estas era sustancial; de hecho, las once eran de exactamente once palabras de largo, y ocho de las coincidencias eran la misma frase de cinco palabras “en el Ecuador como una”, frase común que apareció en varios documentos y contextos. Sobre la base de esta revisión, concluyo con un grado razonable de certeza científica que no hay documento en el expediente judicial de primera instancia que pueda ser una fuente del pasaje de 90 palabras identificado por el Dr. Leonard tal como aparece en el Memo de Fusión y la sentencia ecuatoriana.

26. Como otro ejemplo, consideramos la similitud citada como Ejemplo 2 en el informe del Dr. Leonard. El Ejemplo 2 es un bloque de texto con aproximadamente 150 palabras de superposición entre el Memo de Fusión y en la sentencia ecuatoriana. Encontramos un ejemplo de una superposición de diez palabras entre el texto del Ejemplo 2 y un documento del expediente identificado con el número CL0063-

[3] Informe del Dr. Robert Leonard del 27 de junio del 2011, p. 11.

0006644.txt. La superposición fue la siguiente: "bombas sumergibles en cinco pozos en el Campo Lago Agrio". El contexto en el que aparece la superposición es totalmente diferente del Ejemplo 2. Por lo tanto, sobre la base de nuestra revisión, ni CL0063-0006644.txt ni ningún otro documento del expediente de primera instancia es una posible fuente de la superposición de 150 palabras entre el Memo de Fusión y la sentencia ecuatoriana identificada en el Ejemplo 2 del Dr. Leonard.

27. Hicimos el mismo análisis para otras superposiciones del Memo de Fusión, así como para superposiciones entre la sentencia ecuatoriana y los documentos del producto de trabajo de los demandantes conocidos como el Informe de Clapp, los Resúmenes de Índice, el correo electrónico del Fondo de Fajardo, el Borrador del Alegato y la Compilación de Datos de Selva Viva. Sobre la base de nuestra revisión, ni el Memo de Fusión, ni el Informe de Clapp, los Resúmenes de Índice, el correo electrónico del Fondo de Fajardo, el Borrador del Alegato y la Compilación de Datos de Selva Viva aparecen en el expediente de primera instancia.

28. Después volvimos a examinar las instancias citadas individualmente de superposición o similitud que, según el Dr. Leonard, indicarían plagio. Aparte de las superposiciones atribuibles a las citas directas o a títulos de documentos específicos, nuestro análisis confirmó que ninguna de las nueve instancias individuales de superposición textual con la sentencia ecuatoriana identificada por el Dr. Leonard fueron encontradas en el expediente judicial. En mi opinión, muchas de estas superposiciones (por ejemplo, el Ejemplo de Leonard 2) serían suficientes *en sí* para demostrar plagio.

29. Los archivos que recibimos son copias electrónicas que habían estado sujetas a reconocimiento óptico de caracteres (“OCR”), proceso por el que se escanean y procesan copias en papel para crear archivos electrónicos que se puedan observar en la computadora. En abstracto, el OCR de mala calidad puede reducir el desempeño del análisis textual basado en la computadora de manera general, aunque la magnitud de la reducción varía según el tipo de análisis realizado, con la calidad de la imagen, y con la calidad del motor OCR usado.

30. A pesar de que nuestro análisis original intentó controlar esto, realizamos análisis adicionales para determinar los efectos de la calidad del OCR. A fin de llevar a cabo esto, comparamos todos los documentos del expediente judicial con "corpus estándar" en buen español. Para obtener este corpus, juntamos artículos de Wikipedia en español, el Corpus en Español MultiUN y, además, la versión en español del Corpus N-Grama de Google (1999-2010) para realizar un histograma de línea de base de todas las palabras en español y sus frecuencias, así como las frecuencias de los caracteres y los caracteres n-gramas que los comprenden.

31. Cada documento del expediente judicial de primera instancia fue comparado para determinar cuán cercana a su distribución estaba respecto de la línea de base, y por extensión cuáles eran las posibilidades de que fuera un español razonable. Un documento que encaja en todos los aspectos con el corpus en español se presume perfecto español en perfectas condiciones. La extensión de la diferencia es una medida tanto de la calidad de la escritura como del OCR. (Por ejemplo, un OCR malo producirá muchos errores, pero un documento mal escrito o un tipo inusual de documento, como una tabla, un mapa o un informe de mineralogía también sería inusual y divergente, por lo que este es un análisis conservador).

32. Sobre la base de mi revisión, la calidad total del escaneo fue bastante alta. Mi análisis comparativo de los documentos con el corpus de buen español indica que no más del 1-1,5 % de los documentos del expediente judicial no permitían búsquedas. La gran mayoría del expediente judicial era extremadamente similar a nuestro modelo de español obtenido de fuentes legibles por máquina y una cantidad relativamente baja de valores atípicos tenían un grado extraordinario de diferencias.

33. Recogimos todos estos documentos atípicos y los revisamos manualmente. Todos estos documentos fueron revisados a mano por al menos dos miembros de J&A, uno de los cuales fui yo mismo. Muchos de los documentos atípicos, después de la inspección manual, resultaron ser documentos sin texto, como imágenes fotográficas, tablas o mapas. Esta revisión manual de los documentos

atípicos no identificó ninguno que fuera el Memo de Fusión, el Informe de Clapp, los Resúmenes de Índice, el correo del Fondo de Fajardo, el Borrador del Alegato o la Base de Datos de Selva Viva y no ofreció una fuente de la superposición entre los documentos producto del trabajo de los demandantes y la sentencia ecuatoriana.

Conclusiones

34. Con la ayuda de computadoras, hemos buscado en todo el expediente judicial de primera instancia los documentos del producto del trabajo de los demandantes y potenciales fuentes de superposiciones textuales en la sentencia ecuatoriana identificada por el Dr. Leonard. Esta búsqueda informática cubrió todas las posibles fuentes de estos textos en el expediente judicial. Ninguna de las superposiciones textuales identificadas por el Dr. Leonard son explicables potencialmente por una fuente en el expediente judicial.

35. Respecto de una cantidad de superposiciones más largas, hemos realizado una búsqueda más general para encontrar "conatos de fallo" que podrían haber tergiversado el proceso de análisis (tal como el proceso de OCR). No encontramos ninguno.

36. Un análisis de las propiedades estadísticas de los documentos del expediente judicial indica que relativamente pocos documentos no pudieron ser analizados por computadora. Yo (así como otros miembros de mi equipo) analicé personalmente a mano estos documentos y no encontré fuentes de texto superpuesto.

37. Habiendo examinado personalmente fuentes posibles de estos textos por computadora y habiendo examinado personalmente los principales candidatos de errores de OCR, concluí que en un grado razonable de certeza el Memo de Fusión, el Informe de Clapp, los Sumarios de Índice, el correo electrónico del Fondo de Fajardo, el Borrador del Alegato y la Compilación de Datos de Selva Viva no se encuentran en el expediente judicial de primera instancia. Asimismo, concluí en un grado razonable de certeza científica

que las superposiciones textuales entre la sentencia ecuatoriana y los documentos producto del trabajo de los demandantes identificados por el Dr. Leonard no pueden derivar de fuentes del expediente judicial de primera instancia.

Declaro bajo pena de perjurio conforme a la legislación de los Estados Unidos de América que la información anterior es correcta y veraz. Firmado el [escrito a mano] 9 de octubre del 2013.

[firma]
Patrick Juola, Ph. D

JUOLA WITNESS DECLARATION.DOCX

MERRILL
CORPORATION
Merrill Communications LLC
225 Varick Street
New York, NY 10014 • (212) 620-5600

State of New York)
Estado de Nueva York
) ss:
) a saber:
County of New York)
Condado de Nueva York

Certificate of Accuracy
Certificado de Exactitud

This is to certify that the attached translation is, to the best of our knowledge and belief, a true and accurate translation from English into Spanish of the attached document.

Por el presente certifico que la traducción adjunta es, según mi leal saber y entender, traducción fiel y completa del idioma inglés al idioma español del documento adjunto.

Dated: November 6, 2013
Fecha: 6 de noviembre de 2013

Kate Alexander
Project Manager – Legal Translations
Merrill Brink International/Merrill Corporation

[firmado]
Kate Alexander
Gerente de Proyecto – Traducciones Legales
Merrill Brink International/Merrill Corporation

Sworn to and signed before
Jurado y firmado ante
Me, this 6th day of
mí, a los 6 días del
November 2013
~~mes de~~ noviembre de 2013

Notary Public
Notario Público

[firmado]
[sello]

ROBERT J. MAZZA
Notary Public, State of New York
No. 01MA5057911
Qualified in Kings County
Commission Expires April 1, 2014

EXHIBIT A

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

CHEVRON CORPORATION, Plaintiff, -against- STEVEN R. DONZIGER, et al., Defendants.	Case No. 11 Civ. 0691 (LAK)

DIRECT TESTIMONY OF PATRICK JUOLA, Ph.D.

I, PATRICK JUOLA, hereby declare under penalty of perjury pursuant to 28 U.S.C. § 1746, that the following is true and correct:

1. I am a tenured Associate Professor of Computer Science at Duquesne University, in Pittsburgh, PA. I am an expert in computational and forensic analysis, specifically related to text and authorship analysis. I have authored 40 peer-reviewed articles, primarily on the computational inference of document authorship via the statistical analysis of linguistic features. In my research, I focus on the computational and forensic analysis of linguistic features, and I specialize in the area of authorship attribution.

2. I have been retained by Gibson, Dunn & Crutcher, LLP ("Gibson Dunn") on behalf of Chevron Corporation ("Chevron") in this case to determine whether certain documents of the lawyers and consultants for the plaintiffs in the case of *Maria Aguinda y otros v. Chevron Corporation* (also known as the Lago Agrio case) can be found in the trial court record.

PLAINTIFF'S EXHIBIT
3800
11 Civ. 0691 (LAK)

Summary of Expert Opinion

3. Based on my expert computer-based textual analysis of the Lago Agrio plaintiffs' lawyers and consultants' work product identified in the Ecuadorian judgment and the trial court record in the Lago Agrio case, I have concluded, to a reasonable degree of certainty, that the Fusion Memo, the Clapp Report, the Index Summaries, the Fajardo Trust email, the Draft Alegato, and the Selva Viva Data Compilation are not in the trial court record. Moreover, I have concluded, to a reasonable degree of certainty, based on additional forensic evaluation, that the computer-based analysis was effective to identify sources for the linguistic overlaps between the plaintiffs' work product documents and the Ecuadorian judgment, and that those sources do not exist in the trial court record.

Background and Qualifications

4. I am a tenured Associate Professor of Computer Science at Duquesne University, Pittsburgh, PA. I am also the Director of the Evaluating Variations in Language Laboratory, also at Duquesne.

5. I am also the founder and Director of Research for J Computing, Inc., (dba Juola & Associates "J&A"), a Pennsylvania corporation specializing in text and authorship analysis.

6. I obtained a Bachelor of Science degree in 1987 in electrical engineering at the Johns Hopkins University, Baltimore, where I double majored in Mathematics and Electrical Engineering. I also obtained an M.S. degree in Computer Science from the University of Colorado in 1991 and an M.S. level certificate in cognitive science, also from the University of Colorado, in 1993. Finally, I received a Ph.D. in computer science from the University of Colorado in 1995. I was a postdoctoral research associate for the Department of Experimental Psychology at St. Hugh's and Lincoln Colleges, Oxford University, from 1995 to 1998.

7. In 1998, I started at Duquesne, where I took the position of Assistant Professor of Mathematics and Computer Science. In 2004, I was granted tenure and became an Associate Professor. At Duquesne, I teach classes, including classes in natural language processing, logic programming, software engineering, and cryptography. In my research, I focus on the computational and forensic analysis of linguistic features. Within that, I specialize in the area of authorship attribution.

8. I was inducted into the Office of Research Hall of Fame at Duquesne in 2009. Also in 2009, I received the McAnulty College Faculty Excellence in Scholarship Award from Duquesne.

9. I have authored over 150 publications in general, 40 of which have been peer-reviewed articles. I've also written 2 books and 9 book chapters. Most of my publications discuss the computational inference of document authorship via the statistical analysis of linguistic features.

10. I am a frequent ad-hoc reviewer on subjects pertaining to authorship attribution, stylometry, digital humanities, and text analysis for a number of journals, including LLC (formerly Literary and Linguistic Computing), JASIST (Journal of the American Society for Information Systems Technology), and SPE (Software Practices and Experiments).

11. My company, Juola & Associates, specializes in computational and forensic analysis, specifically with text and authorship analysis. We handle projects involving authorship attribution, profiling of author characteristics, plagiarism, and also large scale document searching.

12. I am the primary architect and designer of the JGAAP (Java Graphical Authorship Attribution Program) authorship analysis system. This system, funded by the National Science

Foundation (NSF) for nearly $2 million, is a system for developing and testing new methods of authorship attribution and determining best practices. In addition to developing best practices, the NSF has also charged me with the development of a forensic-quality authorship attribution system (a goal we have met with the current version of JGAAP) as well as commercializing the technology developed for this purpose.

Methodology

13. I first reviewed the findings of other experts in this case, including those of Dr. Robert Leonard, who had identified linguistic overlaps between the plaintiffs' work product documents and the Ecuadorian judgment. Dr. Leonard identified at least nine instances of substantial textual overlap between the Ecuadorian judgment and plaintiffs' work product documents.[1]

14. Dr. Leonard identified linguistic overlap between the Ecuadorian judgment and the following plaintiffs' work product documents, which I was provided by counsel for Chevron:

a. A document entitled "Primer Borrador Memo Fusión JSP [Nov2007].doc" (henceforth the "Fusion Memo," PX 435);

b. Two versions of an unfiled spreadsheet entitled "pruebas pedidas en etapa de prueba.xls" and "GARR-HDD-003243" (henceforth the "January Index Summary," PX 433, and the "June Index Summary," PX 434, and collectively the "Index Summaries");

c. A draft trial brief, known as an *alegato*, containing the allegations and arguments of the Lago Agrio Plaintiffs (henceforth the "Draft Alegato," PX 438);

[1] *See* June 27, 2011 Report of Dr. Robert Leonard, pg. 11.

d. An email from "Pablo Fajardo Mendoza" to three people including "Steven Donziger" on the subject of "FIDEOCOMISO" (henceforth the "Fajardo Trust email," PX 437);

e. An email (DONZ00025295.pdf) forwarding a report containing the text of DONZ00025296.doc (henceforth the "Clapp Report," PX 928);

f. A document containing sampling data known as the "Selva Viva Data Compilation" (PX 439-41).

15. I received the trial court record in two stages. In the first stage, on or about September 13, 2011, Chevron's counsel provided J&A with approximately 3500 electronic documents comprised of approximately 236,000 images of individual pages, numbered CL0000-00000 through CL2068-0216692, with the February 14, 2011 Judgment beginning at page CL2065-0216338. I understood from Chevron's counsel that this version of the record I reviewed is from a photocopy of the official version of the record maintained by the Provincial Court of Justice of Sucumbios, and that this photocopy was prepared by the Clerk of Court of the Provincial Court of Justice of Sucumbios per normal court procedures, stamped with a court seal on each page to indicate authenticity of the copy, and delivered to Chevron in installments as requested by the company's Ecuadorian trial counsel over the course of the lower court trial. Chevron scanned these copies, creating PDFs, which were then converted to single-page TIFF format and uploaded to an electronic platform, at which point the files were subjected to an automatic OCR process.

16. On or about May 30, 2013, J&A received from Chevron a hard drive containing the electronic contents of 69 compact discs (henceforth the "CD Content List") provided by the National Court of Justice of Ecuador. I understood from Chevron's counsel that the Attorney

General of Ecuador had previously requested copies of all digital information contained on CDs or DVDs in the court record and that Chevron had subsequently requested its own copies of this digital information.

17. Due to the size and heterogeneity of the data received, our first task was to systematically convert all documents to a common format called UTF-16. This format is a variation of "plain text" but that allows for non-English letters or letters with diacritical marks (accents such as "ó"). This provides a basis to search based on words or characters using a common encoding in machine-readable form. In the process of this conversion, we also stripped out all punctuation and capitalization distinctions to maximize the chance of detecting matches between text identified in the plaintiffs' work product documents and the trial court record. This is a conservative procedure in that it ensures that words that differ only in capitalization or punctuation will be correctly matched.[2]

18. We broke each of the documents in the court record into word groups of length 5 (henceforth "5-grams" or more generally "n-grams"). In layman's terms, these are simply groups of five consecutive words. For example, the English phrase "Chief Justice of the Supreme Court of the United States" constitutes a 10-gram in its own right and contains within it many overlapping 5-grams, including "Chief Justice of the Supreme," "of the Supreme Court of," and "Court of the United States."

19. To account for the possibility of OCR errors, we also created another set of "fuzzy" n-grams that treated all non-Latin characters (including characters with diacritical/accent marks) as alike, so the character "ó" would be considered to be identical to the character "ò" or

---

[2] The Ecuadorian judgment document itself is of course part of the court record, but comparisons of the Ecuadorian judgment with itself would not have been useful for finding the sources from which it derived, and hence it was removed prior to analysis.

for that matter, the character "ö." This treatment is done to help compensate for potential errors introduced by the OCR process. Accents are among the most fragile aspects of writing when subjected to OCR, as a bit of stray dirt on the lens or a bad printer/copier can easily introduce, change, or eliminate stray marks that will be interpreted as accent marks. By treating characters that differ only in diacritical marks as being the same, the effect of such errors on the analysis is greatly reduced or minimized.

20. N-grams are highly individual; it is uncommon to see matches of 7-grams or longer except in the cases where the n-grams are part of a common overlapping phrasal vocabulary. "Chief Justice of the Supreme Court of the United States" is an example of such a phrase, familiar to any lawyer. "President de la Corte Superior de Justicia de Nueva Loja" is another example, perhaps equally familiar to an Ecuadorian lawyer. Direct and attributed quotation, of course, would be another valid reason for two documents to share n-grams.

21. We then compiled a list of every specific linguistic overlap between the plaintiffs' work product documents and the Ecuadorian judgment identified by Dr. Leonard (henceforth "Examples"). We broke the Examples down into 5-grams as well.

22. Once documents were broken down into n-grams of five words (5-grams), we used computer software to identify any 5-grams that were shared between the Examples and the court record.

23. Based on these comparisons, we were able to find any documents in the court record that contained an exact match (without regard to diacritical marks) of at least five words with one of the Examples. For each of these documents, we also were able to identify an area of maximal similarity, describing the approximate degree of overlap and allowing us to look at the

specific instances to determine whether the result indicated a source document for the overlapping text.

24. If the computer identified any such matches, we first verified the match by visually comparing the matching phrase and the corresponding part of the court record. We then checked whether the match was a direct quotation. Finally, we analyzed the match to determine whether it was a common or stereotyped phrase, judging partially on the phrase's frequency and distribution across documents and partially on our understanding of the phrase's meaning.

25. As an illustrative example, we consider the similarity cited as Example 1 in the June 27, 2011 report of Dr. Leonard.[3] Example 1 is a block of text with more than 90 identical words appearing in the Fusion Memo and in the Ecuadorian judgment. Upon comparison of this 90-word overlap with the court record, we found exactly eleven matches of five words or more across the entire three thousand plus documents in the court record. None of these were substantial; in fact, all eleven were exactly five words long, and eight of the matches were of the same five-word phrase "en el Ecuador como una," a common phrase that appeared in numerous documents and contexts. Based on this review, I conclude with a reasonable degree of scientific certainty that there is no document in the trial court record that is a possible source for the 90-word passage identified by Dr. Leonard as appearing in both the Fusion Memo and the Ecuadorian judgment.

26. As another example, we consider the similarity cited as Example 2 in Dr. Leonard's report. Example 2 is a block of text with an approximately 150-word overlap between the Fusion Memo and the Ecuadorian judgment. We found one example of a ten word overlap between the text in Example 2 and a document in the record identified as number CL0063-

[3] June 27, 2011 Report of Dr. Robert Leonard, pg. 11.

0006644.txt. The overlap was as follows: "bombas sumergibles en cinco pozos en el Campo Lago Agrio." The context in which the overlap appears is entirely different than in Example 2. Therefore, based on our review, neither CL0063-0006644.txt nor any other document in the trial court record is a possible source for the 150-word overlap between the Fusion Memo and the Ecuadorian judgment identified in Dr. Leonard's Example 2.

27. We ran the same analysis for other overlaps in the Fusion Memo, as well as for overlaps between the Ecuadorian judgment and plaintiffs' work product documents known as the Clapp Report, Index Summaries, Fajardo Trust email, Draft Alegato and the Selva Viva Data Compilation. Based on our review, neither the Fusion Memo, the Clapp Report, the Index Summaries, the Fajardo Trust email, the Draft Alegato, nor the Selva Viva Data Compilation appear in the trial court record.

28. We then reexamined the individual cited instances of overlap or similarity that Dr. Leonard determined would indicate plagiarism. Aside from overlaps attributable to direct quotations or titles of specific documents, our analysis confirmed that none of the nine individual instances of substantial textual overlap with the Ecuadorian judgment identified by Dr. Leonard were found in the trial court record. In my opinion, many of these overlaps (e.g. Leonard Example 2) would be sufficient *by themselves* to indicate plagiarism.

29. The files we received were electronic copies which had been subjected to optical character recognition ("OCR"), which is a process by which hard copies are scanned and processed to create electronic files that can be viewed on the computer. In the abstract, poor quality OCR can reduce performance of computer-based text analysis generally, although the amount of performance reduction varies with the type of analysis performed, with the quality of the image, and with the quality of the OCR engine used.

30. Although our original analysis attempted to control for this, we ran additional analyses to determine the effects of OCR quality. To do this, we compared all of the documents in the court record with a "standard corpus" of well-curated Spanish. To obtain this corpus, we harvested Spanish Wikipedia articles, the MultiUN Spanish Corpus, and in addition the Spanish version of the Google N-Gram Corpus (from 1999-2010) to make a baseline histogram of all Spanish words and their frequencies as well as the frequencies of the characters and character n-grams that comprise them.

31. Each document in the trial court record was then compared to determine how close its distribution was to the baseline, and by extension how likely it was to be reasonable Spanish. A document that conforms in all respects to the Spanish corpus is presumed to be perfect Spanish in perfect condition. The extent of the difference is a measure both of the quality of writing and the quality of OCR. (E. g., a bad OCR will produce many errors, but a badly-written document or an unusual type of document such as a table, a map, or a mineralogy report would also be unusual and divergent, hence this is a conservative analysis.).

32. Based on my review, the overall scanning quality was quite high. My comparative analysis of the documents against the well-curated Spanish corpus indicates that no more than 1-1.5% of the documents in the court record were unsearchable. The vast majority of the court record was extremely similar to our model of Spanish obtained from machine-readable sources and a relatively few number of outliers had an extraordinary degree of difference.

33. We collected all of these outlier documents and reviewed them by hand. All of these documents were manually examined by at least two staff members of J&A, one of whom was myself. Many of the outlier documents proved, upon manual inspection, to be non-text documents such as photographic images, tables, or maps. This manual review of the outlier

documents failed to identify any that were the Fusion Memo, the Clapp Report, the Index Summaries, the Fajardo Trust email, the Draft Alegato, or the Selva Viva Database, and failed to provide a source for the overlap between the plaintiffs' work product documents and the Ecuadorian judgment.

Conclusions

34. With the aid of computers, we have searched the entire trial court record for the plaintiffs' work product documents and for potential sources for the textual overlaps in the Ecuadorian judgment identified by Dr. Leonard. This computer search has covered all possible sources of these texts within the court record. None of the textual overlaps identified by Dr. Leonard are potentially explainable by a source in the court record.

35. For a number of the longer overlaps, we have conducted a more general search looking for any "near-misses" that might have been garbled in process of analysis (such as by the OCR process). We have found no such near misses.

36. An analysis of the statistical properties of the documents in the court record indicates that a relative few of the documents could not be computer analyzed. I (as well as members of my staff) have personally hand-analyzed these documents and found no source for the overlapping text.

37. Having personally examined all possible sources for these texts by computer, and having personally examined by hand the primary candidates for OCR errors, I have concluded to a reasonable degree of scientific certainty that the Fusion Memo, the Clapp Report, the Index Summaries, the Fajardo Trust email, the Draft Alegato, and the Selva Viva Data Compilation are not in the trial court record. Moreover, I conclude to a reasonable degree of scientific certainty

that the textual overlaps between the Ecuadorian judgment and the plaintiffs' work product documents identified by Dr. Leonard cannot be derived from sources with the trial court record.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct. Executed on October 9, 2013.

Patrick Juola, Ph. D

JUOLA WITNESS DECLARATION.DOCX