PRESEEA y su aporte a la creación de perfiles lingüísticos en Lingüística forense

PRESEEA and its application to Linguistic Profiling in Forensic Linguistics

Resumen

Uno de los campos de estudio más recientes en el ámbito hispánico es la Lingüística forense, caracterizada por el uso de técnicas lingüísticas para investigar delitos. Entre sus técnicas más relevantes para la determinación de la autoría de un texto se destaca el estudio de la variación lingüística, por lo que su conocimiento es de gran importancia. El proyecto PRESEEA tiene como objeto la recopilación de un corpus sociolingüístico representativo de habla del mundo hispánico y poder estudiar la variación en español. Por su magnitud y organismos implicados en su desarrollo, puede tener importantes repercusiones en la emergencia y desarrollo de la Lingüística forense para lengua española. Este trabajo presenta el estado actual, las técnicas más comunes y la aplicación de la información de PRESEEA a la tarea de creación automática de perfiles sociolingüísticos. Los resultados muestran gran precisión en la determinación del género y la procedencia, y parecen indicar límites léxicos difusos en los estratos de edad y educación.

Palabras Clave

Lingüística forense, perfiles lingüísticos, sociolingüística, identificación de autoría, creación de corpus, análisis sociolingüístico.

Abstract

One of the most recent areas of interest in Spanish studies is Forensic Linguistics, distinguished by the use of linguistic analysis to investigate crime. Among the most relevant analysis for Authorship Recognition is the study of linguistic variation of a certain text. PRESEEA project aiming at developing a representative corpus of spoken world-wide Spanish can have a significant impact on the development and establishment of Forensic Linguistics for Spanish. This work presents the State of Art, the most common techniques and the application of PRESEEA to the automatic creation of automatic linguistic profiling. Results get a high precision in determining sex and origin and they show how age and education variables seem not to have clear lexical limits among them.

Key words

Forensic linguistics, linguistic profiling, sociolinguistics, authorship attribution, corpus development, sociolinguistic analysis.

1. Introducción

Una de las disciplinas lingüísticas más recientes en el ámbito hispánico es la Lingüística forense (Jiménez Bernal, Reigosa Riveiros y Garayzábal Heinze 2012). La Asociación Internacional de Lingüistas Forenses (IAFL)[1] la define como la interfaz entre Lengua y Derecho. A este respecto, Santana Lario y Falces Sierra (2002) hablan de tres grandes vertientes en este campo de estudio, como son el lenguaje de la ley o textos jurídicos, el lenguaje en los procesos legales o la argumentación judicial y, finalmente, el lenguaje como prueba o evidencia lingüística. Es, en este último aspecto, donde más interés está adquiriendo recientemente. Watt (2010) establece que la mayoría de los trabajos realizados por lingüistas forenses, al menos en el Reino Unido, tiene como objetivo este valor probatorio, bien para identificar al emisor o bien para hacer un posible perfil lingüístico de este.

La Lingüística forense hace uso de cualquier técnica lingüística para investigar delitos (Crystal 1987). El conocimiento sobre Geografía dialectal, Lexicografía o Sociolingüística puede ayudar en la investigación de un crimen. Las cartas de amenaza, las notas de secuestro, las llamadas telefónicas, las conversaciones grabadas etc. pueden contener una rica fuente de información lingüística (Jordan 2002; Coulthard y Johnson 2007; Grant 2008; McMenamin 2010; Olsson 2004). A este respecto, Carlos Delgado (2004), Jefe de la Sección de Acústica Forense de la Comisaría General de Policía Científica, destaca los componentes dialectales del habla entre los principales rasgos que deben ser considerados en el análisis para la detección de emisor. No obstante, los investigadores policiales no suelen estar familiarizados con la variación lingüística (Jiménez et al. 2012).

Un perfil lingüístico se lleva a cabo cuando ningún sospechoso ha sido identificado y se necesita tener una imagen aproximada de las características sociales del emisor. Tal sería el caso de grabaciones de llamadas telefónicas anónimas de secuestradores o avisos de bombas (Watt 2010). Su objetivo es reducir la población de posibles sospechosos asociando sus rasgos lingüísticos con ciertos grupos geográficos y sociales. Actualmente se está prestando mucha atención científica a la elaboración automática de perfiles lingüísticos, destacándose por grado de interés el estudio de redes sociales (Argamon, Koppel, Pennebaker y Schler 2009; Rao, Yarowsky, Shreevats y Gupta 2010; Mikros, 2012; Nguyen, Gravel, Trieschnigg y Meder 2013; Bamman, Eisenstein y Schnoebelen 2014; Dunn, Argamon, Rasooli y Kumar 2015; Schwartz et al. 2013; Stamatatos, Potthast, Rangel, Rosso y Stein 2015). Estos trabajos tratan de determinar cualquier tipo de variable social a partir del análisis lingüístico de textos provenientes de medios como Facebook o Twitter. Para ello se analiza el tipo de discurso de los emisores en estos medios y se asocian con las características personales que ellos mismos declaran poseer tales como género, edad, educación, etc.

Entre los estudios actuales más relevantes para conocer la variación lingüística en español se destaca el proyecto PRESEEA[2]. Se trata de un proyecto fundamental para conocer la diversidad del español y puede tener importantes repercusiones en la emergencia y desarrollo de la Lingüística forense en el ámbito hispánico dado que el estudio de la variación de una lengua es un requisito esencial para poder realizar perfiles lingüísticos y

[1] http://www.iafl.org/
[2] http://preseea.linguas.net/

detección de emisor. Respecto a las directrices metodológicas de PRESEEA, Moreno Fernández (1996) destaca que el objetivo principal de este proyecto es el estudio del habla de comunidades urbanas bien establecidas de todo el mundo hispánico a partir de variables sociales como el *género*, la *edad* y el *grado de instrucción*. Además, se aconseja atender a factores como la profesión, los ingresos económicos y las condiciones de alojamiento en el proceso de postestratificación, ya que pueden tener impacto en el uso de la lengua.

Este trabajo presenta una propuesta inicial de aplicación de los datos de PRESEEA a la determinación automática de variables sociolingüísticas de emisor tal como se está realizando en el ámbito de la Lingüística forense. De esta manera, mediante el software estilométrico Stylo[3] (Eder y Rybicki 2011; Eder, Rybicki, y Kestemont 2014) se ha realizado un análisis cuantitativo del *léxico* proveniente de las transcripciones de todo el corpus PRESEEA accesible públicamente[4]. Stylo es una herramienta de experimentación estilométrica escrita en R (Pallmann 2015) que incluye algunos de los métodos más actuales de aprendizaje y clasificación automática como *Delta Distances*, *K-Nearest Neighbour* o *Support Vector Machines,* la rápida manipulación de frecuencias de *unidades léxicas* y *caracteres* de gran uso en estilometría, y la clasificación de un texto determinado en una serie de categorías previamente definidas. Este trabajo estudia si los datos léxicos que provee PRESEEA pueden ayudar en la tarea de determinación de *género*, *edad*, *procedencia* y *educación*. Tal como se puede observar en los resultados, algunas de estas variables pueden llegar a predecirse con bastante fiabilidad.

2. Metodología

En la mayor parte de los estudios de creación automática de perfiles lingüísticos se usan los mismos tipos de rasgos y formas de clasificación que en la identificación del autor del texto (Abbasi y Chen 2005; El Manar El Bouanani y Kassou 2014; Stamatatos 2009; Tamboli y Prasad 2013). No existe aún una lista clara de características lingüísticas que se puedan analizar para llevar a cabo esta tarea de una manera fehaciente, aunque se prefieren aquellas que tengan una gran frecuencia de aparición, que sean inmunes a su supresión voluntaria y que puedan encontrarse en la mayor parte de los individuos de la población de estudio (McMenamin 2010; Picornell García 2012). Entre los más usados se encuentran las frecuencias y distribuciones de *caracteres* y de *palabras* ya que ambas se pueden usar en todas las lenguas, no se ven afectadas por errores gramaticales y se pueden computar muy fácilmente.

En los trabajos de clasificación automática lingüística, un sistema aprende a partir de un conjunto de textos usados de entrenamiento. Para ello primero se computan frecuencias y patrones lingüísticos y se asocian a un determinado grupo social. De esta manera el programa puede aprender cuáles son las características propias de una clase o grupo social e intenta prever cómo serán nuevos textos producidos por ellos. Finalmente, la tarea del algoritmo de clasificación será asignar una clase (*autor, género, procedencia*, etc.) a nuevos textos no vistos

[3] https://sites.google.com/site/computationalstylistics/stylo

4 http://preseea.linguas.net/Corpus.aspx Fecha de consulta: 17 de mayo de 2016

con anterioridad (Aggarwal 2014). La experimentación se realiza en varios pasos: selección del corpus, elección de rasgos o patrones lingüísticos, método de aprendizaje, test y evaluación de resultados.

Formalmente, el proceso de aprendizaje automático se describe como un par consistente, por un lado, en un *vector de rasgos* con cada una de las características lingüísticas que se quiere estudiar ($x_1, x_2, x_3, \ldots x_n$) y por otro, el resultado esperado para tal vector (una determinada *clase*, un determinado *autor*, *género*, *edad*, etc.) A modo de ejemplo, la figura 1 muestra el uso de un vector tridimensional en la clasificación de un texto con solo dos clases posibles. A la izquierda se observan los rasgos del texto estudiados y su posterior representación en un eje de coordenadas tridimensional. A la derecha podemos ver cada texto representado en cada uno de los puntos. Como se aprecia, los textos se agrupan claramente en dos clases de acuerdo con sus características.

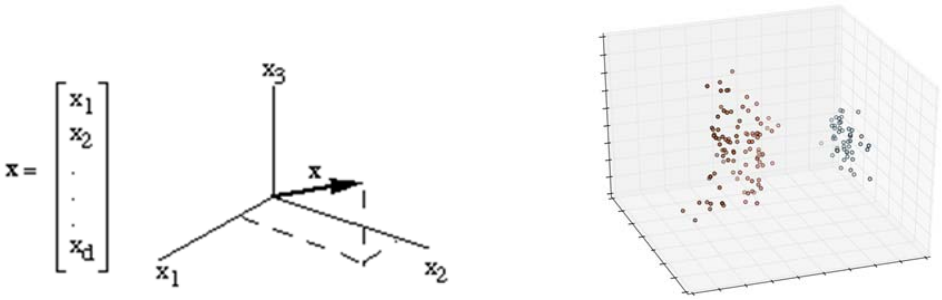


Figura 1. Representación tridimensional de un vector

Este trabajo hace uso de frecuencias y distribuciones de *palabras* proporcionados por el software Stylo sobre los textos de PRESEEA para crear un vector de rasgos cuya clase o resultado esperado es alguna de las variables mencionadas: *género*, *edad*, *procedencia* y *educación*. Como veremos a continuación, se han realizado tres tipos de experimentos diferentes: 1) con el *léxico* de PRESEEA tal como aparece en su versión web; 2) con los *lemas* del léxico de PRESEEA y 3) con las *clases de palabras* subyacentes al léxico de PRESEEA. Esta triple distinción será explicada en el siguiente apartado.

2.1. Corpus

El proyecto PRESEEA[5] es actualmente el proyecto sociolingüístico más importante y ambicioso del español. Consta actualmente de 168 entradas que incluyen transcripción, etiquetado y audio. Tales entradas están clasificadas por *ciudades*, *género*, *edad* y *nivel de estudios*. Actualmente están representadas 10 ciudades del ámbito hispánico: Alcalá de Henares, Caracas, La Habana, Lima, Madrid, Medellín, Monterrey, Montevideo, Santiago y Valencia. Tal como se recoge en la metodología de trabajo de este proyecto (Moreno Fernández 1996), se ha querido simplificar los estratos de *edad* y se han establecido tres franjas posibles: de 20 a 34 años, de 35 a 54 años y de 55 años en adelante. Finalmente, el nivel de estudios comprende tres niveles: un primer

[5] http://preseea.linguas.net. Fecha de consulta: 17 de mayo de 2016

nivel que incluye a los analfabetos, individuos sin estudios y con estudios de enseñanza Primaria (hasta los 10-11 años de edad aprox.), es decir, hasta unos 5 años aproximadamente de escolarización; un segundo nivel con individuos que hayan realizado la Enseñanza Secundaria (hasta los 16-18 años de edad aprox.), es decir, unos 10-12 años aproximadamente de escolarización; un último grupo que contiene la Enseñanza Superior (hasta los 21-22 años de edad aprox.), es decir, unos 15 años aproximadamente de escolarización. La figura 2 muestra cómo se presenta la interfaz al usuario:

Figura 2. Interfaz de PRESEEA[6]

Esta interfaz permite el acceso a las transcripciones y los audios del proyecto ordenados por sus características sociolingüísticas. De esta manera, todas las trascripciones disponibles fueron descargadas, excluyendo las de "Caracas" ya que este subcorpus constaba de menos archivos que los demás, haciendo un total de 162 entradas.

De todos los archivos descargados, 36 archivos fueron extraídos aleatoriamente de este corpus siguiendo un muestreo estratificado para componer el *corpus test* (test), es decir, se extrajo aproximadamente un 20% del total del corpus de PRESEEA para hacer pruebas y ver si el algoritmo de aprendizaje era capaz de determinar el *género*, *procedencia*, *nivel de estudios* y *edad* satisfactoriamente. Los 126 restantes se usaron para entrenar el sistema, es decir, para que el sistema pudiera aprender de ellos y se incluyeron en el *corpus de entrenamiento* (train). La media de tamaño de archivos es aproximadamente seis mil palabras por archivo de transcripción. La figura 3 muestra la distribución de variables sociolingüísticas en nuestro *corpus train* y *test*:

[6] http://preseea.linguas.net/Corpus.aspx

PROCEDENCIA (TRAIN/TEST):	GÉNERO (TRAIN/TEST):	NIVEL DE ESTUDIOS (TRAIN/TEST):
Montevideo: 18 (14/4)	Hombres: 81 (63/18)	1: 54 (42/12)
Madrid: 18 (14/4)	Mujeres: 81(63/18)	2: 54 (42/12)
Valencia: 18 (14/4)		3: 54 (42/12)
Alcalá: 18 (14/4)	**EDAD:**	
Monterrey: 18 (14/4)	1: 54(42/12)	
Medellín: 18 (14/4)	2: 54(42/12)	
La Habana: 18 (14/4)	3: 54(42/12)	
Santiago de Chile: 18 (14/4)		
Lima: 18 (14/4)		

Figura 3. Corpus seleccionado para la experimentación.

Una vez descargadas todas las transcripciones de PRESEEA, el conjunto de textos se transformó en tres corpus diferentes:

a) Un primer corpus de *tokens* o palabras tal como habían sido transcritas en el corpus original. Se eliminó cualquier tipo de anotación paralingüística, intervención de los entrevistadores o marcas adicionales de la transcripción o discurso del informante.

b) Un segundo corpus *lematizado*. Con ayuda de la herramienta *TreeTagger*[7] (Schmid 1995), los palabras fueron *lematizadas*, es decir, se elimina y unifica la flexión verbal, nominal y adjetival (*pensamos* > *'pensar'*, *coches* > *'coche'*, *'el' | la'| los'*> *'el'*). TreeTagger, disponible para 20 lenguas diferentes, permite anotar las palabras de un texto con su tipo de palabra y su lema o forma primigenia. Göhring (2009) establece que la precisión de esta herramienta llega al 93% en las anotaciones de *clases de palabras* en español. Treetagger comprende hasta 75 etiquetas diferentes tal como muestra la tabla 1:

[7] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

ADJ	*Adjectives*	*NC*	*Common nouns*	*UMMX*	*measure unit*
ADV	*Adverbs*	*NEG*	*Negation*	*VCLIger*	*clitic gerund verb*
ALFP	*Plural letter of the alphabet*	*NMEA*	*measure noun*	*VCLIinf*	*clitic infinitive verb*
ALFS	*Singular letter of the alphabet*	*NMON*	*month name*	*VCLIfin*	*clitic finite verb*
ART	*Articles*	*NP*	*Proper nouns*	*VEadj*	*Verb 'estar'. Past participle*
CARD	*Cardinals*	*PAL*	*Portmanteau word formed by 'a' and 'el'*	*VEger*	*Verb 'estar'. Gerund*

Tabla 1. Ejemplo de etiquetas de la herramienta TreeTagger

c) Un grupo solo con *clases de palabras*. Exclusivamente se tomaron las etiquetas sobre tipos de palabras otorgadas por *TreeTagger* y se eliminaron los *tokens* o palabras de la transcripción original. La figura 4 ejemplifica como resulta el corpus tras su *lematización* y anotación con *clases de palabras*. A la izquierda se pueden observar los lemas provistos *TreeTagger* y a la derecha la *clase de palabra* a la que pertenecen.

De esta manera, del corpus original del PRESEEA obtenemos tres corpus diferentes: 1) uno con las palabras originales pero sin marcas ni discurso del entrevistador; 2) un segundo corpus de *lemas* o palabras a las que se les ha eliminado la *flexión* para unificarlas, y 3) un último corpus solo de las *clases de palabras* del corpus. Cada uno de estos tres corpus será llevado al software Stylo para su análisis y así poder apreciar cuál proporciona mejores resultados en la determinación de las variables sociolingüísticas de cada uno de los textos.

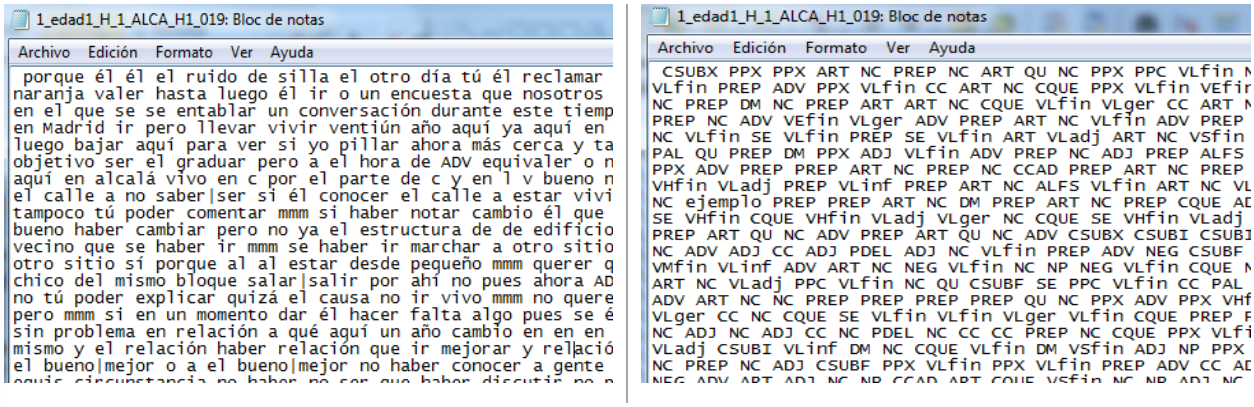


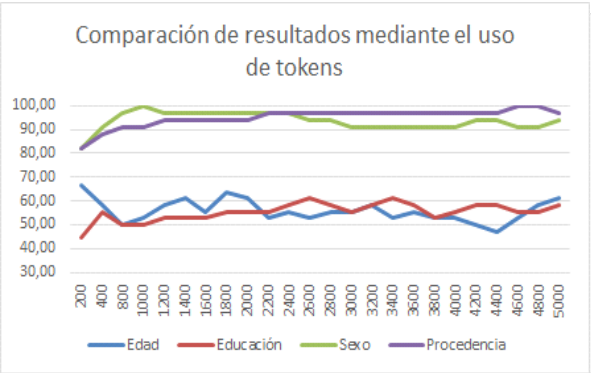
Figura 4. Corpus lematizado (izda.) y etiquetado (dcha.)

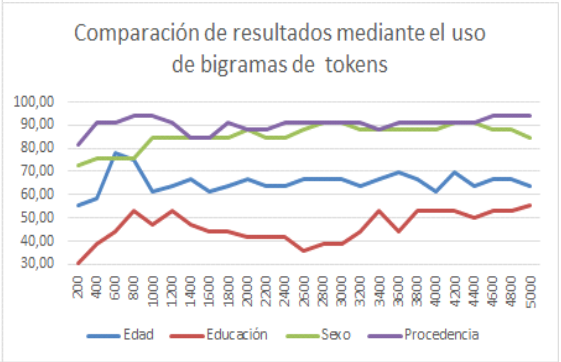
3. Análisis y resultados

Como se decía anteriormente, una vez creado el corpus de análisis, lo integramos en el software Stylo. De este programa, que viene incorporado con diferentes algoritmos de aprendizaje, se ha escogido por su precisión el conocido como *Support Vector Machines*[8]. Stylo incorpora fácilmente a este algoritmo tanto *frecuencias de unidades léxicas* como de *gramas* (combinaciones de 1, 2, 3, etc. unidades).

Se realizaron diferentes tipos de experimentos. En primer lugar se trabajó con el corpus 1 o corpus de *tokens* con el que se experimentó tanto con palabras aisladas o *unigramas*, como con *bigramas* o combinaciones de dos unidades. En segundo lugar, se usó el corpus 2 o corpus de *lemas*, con los que también se experimentó usando *unigramas* y *bigramas*. Finalmente, se utilizó el corpus de *clases de palabras* o corpus 3 con el que se analizaron distribuciones de bigramas y trigramas (tres elementos) al ser solo 75 los unigramas posibles.

Stylo crea una matriz de rasgos a partir de tales frecuencias cuyas dimensiones se pueden restringir fácilmente. Se partió de las 200 *unidades* más frecuentes de cada corpus hasta los 5000 incrementando el vector en 500 rasgos adicionales. En las figuras 4, 5 y 6 se pueden observar los resultados medios de precisión en la clasificación de variables sociolingüísticas usando *tokens* (figura 4), *lemas* (figura 5) y *clases de palabras* (figura 6). En el eje de abscisas se observa el tamaño del vector y en el de ordenadas la precisión alcanzada con ese tamaño.



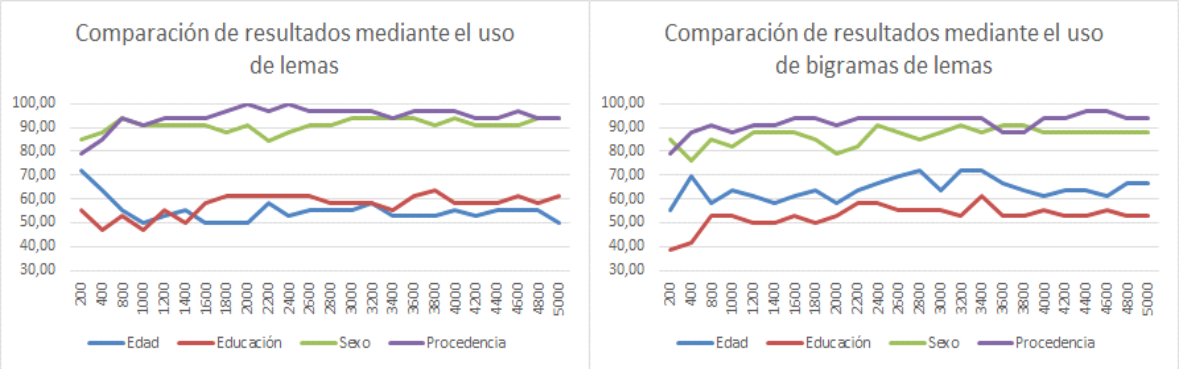


EDAD	*EDUCACIÓN*	*GÉNERO*	*PROCEDENCIA*
1 token	***1 token***	***1 token***	***1 token***
Media: 56,0	Media: 55,5	Media: 93,6	Media: 93,6
Desviación: 4,5	Desviación: 3,7	Desviación: 3,8	Desviación: 3,8
Bigrama de tokens	***Bigrama de tokens***	***Bigrama de tokens***	***Bigrama de tokens***
Media: 65,4	Media: 46,1	Media: 85,3	Media: 85,3
Desviación: 4,6	Desviación: 6,6	Desviación: 5,1	Desviación: 5,1

Figura 4. Porcentaje de precisión por tokens o palabras del corpus.

[8] Conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik.

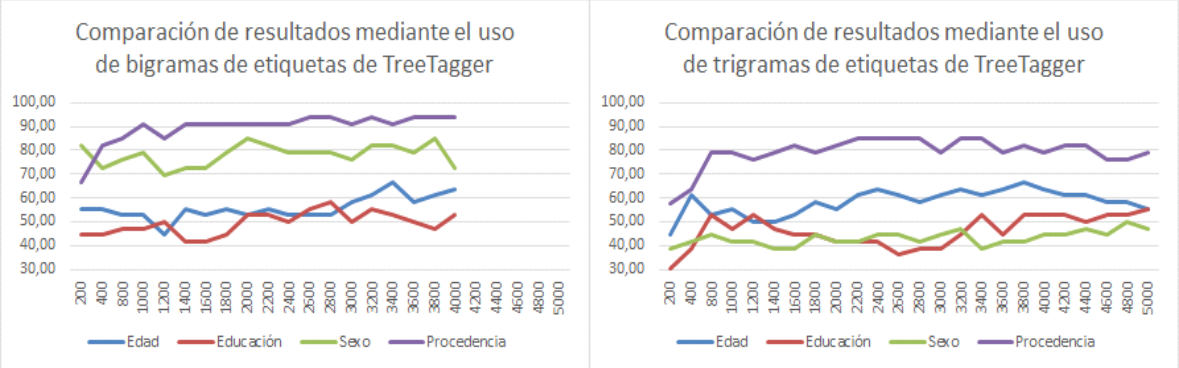
Como se aprecia en la figura 4, se llega a un 93% medio de acierto en la clasificación de *género* y de *procedencia,* y se observa una diferencia significativa en la media de resultados al compararlas con la *edad* y la *educación*. Ambas variables se mantienen en torno al 50-60% de precisión con el uso de *unigramas*. Es muy llamativo que la única clase cuya precisión mejora al usar *bigramas* sea la *edad*, llegándose casi a 80% de precisión en algunos momentos tal como se observa en la gráfica. Esto parece indicar que existen combinaciones de unidades especiales en función de la variable diageneracional.



EDAD	*EDUCACIÓN*	*GÉNERO*	*PROCEDENCIA*
1 lema	***1 lema***	***1 lema***	***1 lema***
Media:54,9	Media:57,1	Media:91	Media:94,4
Desviación: 4,8	Desviación: 4,6	Desviación: 2,7	Desviación: 4,5
Bigrama de lemas	***Bigrama de lemas***	***Bigrama de lemas***	***Bigrama de lemas***
Media:64	Media:52,7	Media:86,5	Media:91,9
Desviación: 4,8	Desviación: 4,6	Desviación: 3,7	Desviación: 3,9

Figura 5. Porcentaje de precisión por lemas extraídos del corpus con TreeTagger.

Como se puede apreciar en la figura 5, no existen grandes diferencias entre los resultados provenientes del uso de *tokens* y el uso de *lemas*. Observamos que, al igual que en el caso anterior, la precisión en la *edad* mejora con el uso de *bigramas de lemas*, lo que refuerza la idea de que existen determinadas combinaciones de elementos asociadas a determinado grupos generacionales. Respecto al *género*, es significativo que, una vez eliminadas las marcas morfológicas de género y número se sigan obteniendo muy buenos resultados, lo que indica una diferenciación diasexual en cuanto al vocabulario.



EDAD	*EDUCACIÓN*	*GÉNERO*	*PROCEDENCIA*
Bigrama de etiquetas	***Bigrama de etiquetas***	***Bigrama de etiquetas***	***Grama de etiquetas***
Media:55,8	Media:49,4	Media:77,9	Media:89,4
Desviación: 4,8	Desviación: 4,7	Desviación: 4,3	Desviación: 6,3
Trigrama de etiquetas	***Trigrama de etiquetas***	***Trigrama de etiquetas***	***Trigrama de etiquetas***
Media:57,8	Media:43	Media:78,3	Media:78,9
Desviación: 5,7	Desviación: 3	Desviación: 3,8	Desviación: 6,4

Figura 6. Porcentaje de precisión por clases de palabras extraídas del corpus con TreeTagger.

La figura 6 muestra los resultados en la determinación de *edad*, *educación*, *género* y *procedencia* simplemente analizando frecuencias de *clases de palabras*: *sustantivo, verbo, adjetivo, preposición*, etc. Como se explicó en el apartado 2.1, TreeTagger llega a identificar hasta 75 clases de palabras diferentes[9]. Aunque la precisión mediante clases de palabras empeora respecto a los casos anteriores, es significativo que *género* y *procedencia* aún sigan obteniendo buenos resultados, lo que parece sugerir que existen determinadas *clases de palabras* más propias de un determinado *género* y de un determinado *lugar*.

Tal como se ha indicado antes, el tamaño de los vectores de frecuencias varía de 200 a 5000. Como se observa en todas las gráficas, en general el aumento del tamaño del vector no es determinante a la hora de clasificar con mayor éxito, lo que parece indicar que las diferencias léxicas de *género*, *edad*, *procedencia* y *educación* se encuentran en el léxico más frecuentemente usado.

4. Discusión

La tabla 2 muestra una matriz de error de los resultados por *unigramas de tokens*. A su izquierda observamos el error cometido al clasificar la *edad* y a la derecha al clasificar *educación*. En la primera columna de cada uno

[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt

de los dos grupos se muestra la *edad* y la *educación* real, es decir, la que aparece en el corpus PRESEEA, y en las tres centrales la sugerida por el clasificador automático.

Respecto a la clasificación de la *edad*, los principales errores de clasificación se observan de los *grupos 1* y *3* hacia el *2*. Los grupos de edad de PRESEEA comprenden las edades 1) de 20 a 34 años, 2) de 35 a 54 años y 3) de 55 años en adelante y, observando los resultados, parece que las fronteras léxicas entre los grupos extremos (1 y 3) y el central (2) podrían no estar totalmente claras.

A la derecha podemos ver la matriz de error en la clasificación del *nivel educativo*. Se observa que el *grupo 1* es el que obtiene peores resultados; el clasificador incluye en el *grupo 2* a nueve de los doce textos que deberían haberse clasificado en el *grupo 1*. Parece que la diferencia léxica entre aquellos con estudios primarios, *grupo 1* con 5 años de escolarización, y aquellos con estudios secundarios, *grupo 2* con 10-12 años de escolarización, se confunde.

	Edad clasificada		
Edad real en el corpus por grupos (1, 2 y 3)	**1**	**2**	**3**
1		X	
1			X
1		X	
1		X	
1		X	
1		X	
2			X
2	X		
3		X	
3	X		
3		X	
3		X	
3		X	

	Educación clasificada		
Educación real en el corpus por grupos (1, 2 y 3)	**1**	**2**	**3**
1		X	
1		X	
1		X	
1			X
1		X	
1		X	
1			X
1		X	
1		X	
2	X		
2			X
3		X	
3		X	
3		X	

Tabla 2. Matriz de errores de clasificación.

5. Conclusiones

Este estudio presenta un primer uso del corpus de PRESEEA a la tarea de identificación automática de variables sociolingüísticas propia de la Lingüística forense. Para ello hemos utilizado una de las herramientas más conocidas en este ámbito como es Stylo. Esta herramienta nos ha permitido obtener frecuencias de *unigramas*, *bigramas* y *trigramas* de *tokens*, *lemas* y *clases de palabras* y su posterior aplicación al algoritmo de aprendizaje *Support Vector Machines.* De esta manera, los resultados muestran que los rasgos que ofrecen mejores resultados son los *unigramas de tokens* aplicados a la detección del *género* y *procedencia* con una precisión media de 93%. Los valores obtenidos en la determinación de *edad* y *educación* se situación en torno al 50% de precisión media, excepto para la estimación de la *edad* con el uso de *bigramas de tokens*, aspecto este que deberá ser ampliado en estudios futuros.

Queremos destacar que PRESEEA es un proyecto en desarrollo por lo que, a medida que se vaya ampliando el corpus, se podrá llegar a mejores resultados y más robustos. Otra limitación que habrán que solventarse será la lista de rasgos lingüísticos que han de usarse de manera inequívoca así como el tamaño del vector en la determinación de cada una de las variables sociolingüísticas ya que, como ha podido observarse, los rasgos léxicos no han funcionado igual para *procedencia, género, educación* y *edad*.

Finalmente queremos hace hincapié que todo análisis forense debería mostrar de alguna manera la probabilidad o verosimilitud de que un determinado texto pertenezca a un determinado grupo diasexual, diageneracional, diatópico o diastrático. Por el momento, Stylo solo muestra aquel grupo más cercano a las características del texto sin ningún tipo de indicación adicional. La finalidad de la Lingüística forense es determinar el autor y las características de este más allá de toda duda y, aunque el análisis automático puede dar buenas pistas de los usos dialectales, es necesario un análisis manual cualitativo que lo refuerce.

Mario Crespo Miguel
mario.crespo@uca.es
Profesor Ayudante Doctor
Instituto de Investigación en Lingüística Aplicada
Universidad de Cádiz

Referencias bibliográficas

Abbasi, Ahmed y Chen, Hsinchun (2005): "Applying authorship analysis to extremist-group web forum messages", *Intelligent Systems*, IEEE, 20.5, pp. 67-75.

Aggarwal, Charu C. (2014): *Data classification: algorithms and applications*, New York: CRC Press.

Argamon, Shlomo, Koppel, Moshe, Pennebaker, James W. y Schler, Jonathan (2009): "Automatically profiling the author of an anonymous text", *Communications of the ACM*, 52.2, pp. 119-123

Bamman, David, Eisenstein, Jacob y Schnoebelen, Tyler (2014): "Gender identity and lexical variation in social media", *Journal of Sociolinguistics*, 18.2, pp. 135-160.

Coulthard, Malcolm y Johnson, Alice (2007): *An Introduction to Forensic Linguistics: Language in Evidence*, London and New York: Routledge.

Crystal, David (1987): *The Cambridge Encyclopedia of Language*, Cambridge: Cambridge University Press

Delgado Romero, Carlos (2004): *La identificación de locutores en el ámbito forense*, Tesis Doctoral, Madrid: Universidad Complutense. http://eprints.ucm.es/tesis/inf/ucm-t25153.pdf.

Dunn, Jonathan, Argamon, Shlomo, Rasooli, Amin y Kumar, Geet (2016): "Profile-based authorship analysis", *Digital Scholarship in the Humanities*, 31.4, pp. 689-710

Eder, Maciej y Rybicki, Jan (2011): "Stylometry with R", *Digital Humanities 2011: Conference Abstracts,* Stanford: Stanford University, pp. 308-11.

Eder, Maciej, Rybicki, Jan y Kestemont, Mike (2014): "Stylometry with R: A package for computational text analysis", *R Journal*, 16.1, pp. 107-121

El Manar El Bouanani, Sara y Kassou, Ismail (2014) "Authorship analysis studies: A Survey". *International Journal of Computer Applications*, 86.12, pp. 22–29.

Göhring, Anne (2009): *Spanish Expansion of a Parallel Treebank*, Tesis doctoral, Zürich: University of Zürich.

Grant, Tim (2008): "Approaching questions in forensic authorship analysis", John Gibbons y María Teresa Turell (eds.), *Dimensions of forensic linguistics*, Amsterdam: John Benjamins, pp. 215–229.

Jiménez Bernal, Míriam, Reigosa Riveiros, Mercedes y Garayzábal Heinze, Elena (2012): "La lingüística forense: licencia para investigar la lengua", Elena Garayzábal Heinze, Míriam Jiménez Benal y Mercedes Reigosa Riveiros (eds.), *Lingüística forense: la lingüística en el ámbito legal y policial*, Madrid: Euphonia, pp. 28–50.

Jordan, Sherilynn Nidever (2002): *Forensic Linguistics: The Linguistic Analyst and Expert Witness of Language Evidence in Criminal Trials*, Tesis doctoral, Los Angeles: Biola University.

McMenamin, Gerald R. (2010.): "Forensic stylistics: theory and practice of forensic stylistics". IMalcolm Coulthard y Alice Johnson (eds.), *The Routledge handbook of forensic linguistics*, London: Routledge, pp. 473–486.

Mikros, George K. (2012): "Authorship Attribution and Gender Identification in Greek Blogs", *Methods and Applications of Quantitative Linguistics*, 21, Belgrade: University of Belgrade Academic Mind, pp. 21-32

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

Nguyen, Dong, Gravel, Rilana, Trieschnigg, Dolf y Meder, Theo (2013): "'How old do you think I am?' A study of language and age in Twitter", *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM 2013, Palo Alto, CA: AAAI Press, pp. 439-448

Olsson, John (2004): *Forensic Linguistics. An Introduction to Language, crime and the law*. London: Continuum.

Pallmann, Philip (2015): "Applied meta-analysis with R". *Journal of Applied Statistics*, 42.4, pp. 914–915.

Picornell García, I. (2012): "La aplicación de atribución de autoría en la investigación e inteligencia: La aplicación práctica", Elena Garayzábal Heinze, Míriam Jiménez Benal y Mercedes Reigosa Riveiros (eds.), *Lingüística forense: la lingüística en el ámbito legal y policial*, Madrid: Euphonia, 7, pp. 8-93.

PRESEEA (2014): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá.

Rao, Delip, Yarowsky, David, Shreevats, Abhishek y Gupta, Manaswi (2010): "Classifying latent user attributes in twitter", *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, New York: ACM, pp. 37-44.

Santana Lario, Juan y Falces Sierra, Marta (2002): "Any statement you make can be used against you in a court of law: Introducción a la Lingüística Forense", *A Life in Words. A Miscellany Celebrating Twenty-Five Years of Association between the English* Department of Granada University and Mervyn Smale (1977-2002), Granada: Editorial de la Universidad de Granada, pp. 267-280.

Schmidt, Helmut (1995): "Treetagger, a language independent part-of-speech tagger". *Institut für Maschinelle Sprachverarbeitung*, Universität Stuttgart, 43: 28.

Schwartz, Andrew H., Eichstaedt, Johannes C., Kern, Margaret L., Dziurzynski, Lukasz, Ramones, Stephanie M., Agrawal, Megha, Shah, Achal, Kosinski, Michal, Stillwell, David, Seligman, Martin E. P. y Ungar, Lyle H. (2013): "Personality, gender, and age in the language of social media: The open-vocabulary approach", *PloS one*, 8.9, e73791.

Stamatatos, Efstathios (2009): "A survey of modern authorship attribution methods", *Journal of the American Society for information Science and Technology*, 60.3, pp. 538-556.

Stamatatos, Efstathios, Potthast, Martin, Rangel, Francisco, Rosso, Paolo y Stein, Benno (2015): "Overview of the PAN/CLEF 2015 Evaluation Lab", *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science*, 9283, NY: Springer International Publishing, pp. 518-538.

Tamboli, Mubin Shaukat y Prasad, Rajesh S. (2013): "Authorship analysis and identification techniques: a review". *International Journal of Computer Applications*, 77.16, pp. 11-15.

Watt, Dominic (2010): "The Identification of the Individual through Speech", *Language and Identities*, Edinburgh: Edinburgh University, pp. 76-85.