CORPUS FORENUCA: DISEÑO, OBJETIVOS Y ESTADO ACTUAL EN EL MARCO DEL INSTITUTO DE INVESTIGACIÓN EN LINGÜÍSTICA APLICADA

Mario Crespo Miguel
Universidad de Cádiz

RESUMEN

Una de las disciplinas lingüísticas más recientes en el ámbito hispánico es la Lingüística forense, caracterizada por el uso de técnicas lingüísticas para investigar delitos. Entre sus principales focos de investigación se encuentra la determinación del emisor de textos electrónicos como emails, redes sociales o mensajería móvil. El estudio de los componentes dialectales y sociolectales del habla es esencial para una caracterización del género, edad o nivel educativo del emisor de un texto determinado. En el ámbito hispánico existe escasez de corpus de textos electrónicos asociados a diferentes variables sociolingüísticas, y que sirva como soporte científico en el ámbito de la Lingüística forense. Este trabajo presenta el Corpus ForenUCA de actual desarrollo en el Instituto de Investigación en Lingüística Aplicada de la Universidad de Cádiz, que recopila textos procedentes de nuevos medios de comunicación social — mensajería corta, email y redes sociales —. Este trabajo presenta las directrices, diseño y objetivos finales de este corpus que actualmente cuenta con más de 200 mil palabras.

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 linguistic analysis to investigate crime. Among the main points of interest are the authorship attribution of electronic texts such as emails, social networks or mobile

messaging. The study of dialectal and sociolinguistic parameters of a text is essential when characterizing the gender, age or educational level of a certain text sender. There is a lack of corpus of Spanish electronic texts linked to different sociolinguistic variables able to provide scientific support to Forensic Linguistics. This works presents the ForenUCA Corpus, under development at the Applied Linguistics Research Institute of the University of Cádiz, aiming at collecting texts from new social media — short mobile messaging, email and social networks —. This paper presents the guidelines, design and objectives of this corpus that currently contains more than 200 thousand words.

Keywords: 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, 2012), caracterizada por el uso de técnicas lingüísticas para investigar delitos (Crystal, 1987). Entre sus principales focos de investigación se encuentra la detección de emisor de textos electrónicos como emails, redes sociales o mensajería móvil. El estudio de los componentes dialectales y sociolectales del habla es esencial para una caracterización del género, edad o nivel educativo del emisor de un texto determinado (Watt, 2010).

Actualmente se está prestando mucha atención científica a la determinación automática de variables sociolingüísticas, 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). Gran relevancia también posee el workshop internacional anual PAN[1] centrado en la evaluación de tecnologías forenses sobre textos electrónicos para la determinación de emisor, plagio, perfiles lingüísticos.

En el ámbito hispánico existe escasez de corpus disponibles de textos electrónicos asociados a variables sociolingüísticas. Entre los más relacionados se destaca el proyecto PRESEEA[2] (Moreno Fernández, 1996) cuyo objetivo principal es el estudio del habla de comunidades urbanas de todo el mundo hispánico a partir de variables sociales como el *género*, la *edad* y el *grado de instrucción*. Sin embargo, los textos representan conversaciones orales grabadas entre investigadores e informantes.

El workshop PAN mencionado anteriormente provee públicamente de un corpus para cada una de sus tareas forenses (Rosso *et al.*, 2016), pero estos pertenecen exclusivamente a Twitter y solo contienen información sobre género y procedencia. Por otro lado, también se encuentra el corpus TweetNorm_es (Alegría *et al.*, 2014) centrado también en Twitter pero sin motivación sociolingüística y sí en el estudio de la normalización ortográfica de este tipo de textos.

[1] http://pan.webis.de
[2] http://preseea.linguas.net/

Este trabajo presenta el Corpus ForenUCA de actual desarrollo en el Instituto de Investigación en Lingüística Aplicada de la Universidad de Cádiz[3] que recopila textos procedentes de nuevos medios de comunicación social (mensajería corta, email y redes sociales). Para esto se están teniendo en cuenta variables sociolingüísticas, psicológicas y clínicas que puedan afectar al uso lingüístico idiolectal en estos géneros discursivos.

2. METODOLOGÍA

A continuación se presentan las directrices, diseño y objetivos finales del corpus ForenUCA que actualmente cuenta con más de 200 mil palabras.

2.1. Técnica de muestreo

Nos encontramos con una población formada por textos electrónicos de cualquier género discursivo tomados de hablantes nativos de español. El principio de inclusión de informantes en la recopilación de textos está siendo muy genérico y poco excluyente, de tal manera que se permite la participación a cualquier individuo que cumpla ciertas condiciones. En primer lugar, ser nativo del español — monolingüe o bilingüe —, pertenecer a una población de un territorio de habla española y, por supuesto, usar algún tipo de medio de comunicación electrónica.

Dado que no se partía de ningún corpus previo, en este primer proceso de recopilación, la técnica de muestreo no ha sido aleatoria en su totalidad y se ha optado por recopilar información de cualquier voluntario. Se quería observar de esta manera los problemas tanto a la hora de adquirir el corpus como a la hora de responder al cuestionario de características sociolingüísticas, psicológicas y clínicas.

2.2. Formato y tamaño textual

La cantidad de texto tomado de cada informante ha seguido un enfoque estadístico de estimación muestral (Triola, 2012). Dado que en Lingüística predomina el análisis cualitativo, el enfoque más adecuado es el ajuste muestral basado en proporciones:

[3] ila.uca.es

$$n = \frac{Z_a^{\,2} \times p \times q}{d^2}$$

Zα = puntuación crítica z a un nivel de confianza 1-α
p = probabilidad de éxito, o proporción esperada
q = probabilidad de fracaso
d = precisión (error máximo admisible en términos de proporción)

Figura 1. Fórmula estadística para el cálculo del tamaño muestral

Asumimos un margen de error al 3% y nivel de confianza al 99% (α=0,01) de acuerdo con estándares estadísticos normales. Finalmente, al no contar con ningún parámetro proporcional previo de estudio, asumimos p=0.5 que maximiza el tamaño muestral. Con tal información se obtiene un tamaño textual de 1843 palabras, que ajustamos a perdidas con un 10% resultando un tamaño textual adecuado mínimo de unas 2169 palabras por informante. El promedio actual de palabras por individuo es de 2325.

Una vez recopilados, los textos se desproveen de otro tipo de información gráfica adicional al texto, permitiéndose emoticonos o representaciones de expresiones faciales para aludir al estado de ánimo, y se almacenan en texto plano formato UTF-8. La figura 2 ejemplifica el corpus:

Whatsapp	Twitter	Facebook	Email
no me veo con alguien que kiera aserme mu felix😊	Vive tu vida de tal manera q cuan2 tus pies tquen el suelo xla manana hasta el Diablo se estremezca y dga Hostia!! :D #estacabronayaselevantó	Tod@s los que estubieron y los que no pero se acordaron, sois los mejores amigos del mundo.	Algunos tienen años, así que dan asco y eso que los he seleccionado de los que no dan asco puto XD

Figura 2. Muestras de texto del corpus ForenUCA

2.3. Variables de estudio

Se puede definir estilo como la manera peculiar de escribir de un escritor. Es el resultado de las diferentes opciones que escoge de manera inconsciente entre todas las que hubieran sido posibles (McMenamin, 2010). Esta variación proviene del origen geográfico, edad, ámbito social, rasgos de su personalidad o hábitos de procedencia patológica o aprendida (Lucena Molina, 2005). El corpus ForenUCA complementa los textos recopilados con un cuestionario sobre características individuales. Estas se dividen en tres bloques:

1. Bloque sociolingüístico, con las siguientes variables (al modo de PRESEEA, 2004):

- Género
- Edad
- Lugar de nacimiento, lugares donde ha vivido y número de años.
- Escolaridad máxima alcanzada
- Ocupación (previa si estuviera parado o jubilado): Idiomas nativos.
- Idiomas estudiados (no nativos)

2. Bloque psicológico, donde el informante debe situar en una escala likert de 5 puntos sus características personales:

a. Estabilidad emocional:

1.- Inestable, ansioso, hostil, proclive a la depresión.

5.- Estable, relajado, no se molesta con facilidad.

b. Extraversión o introversión:

1.- Extrovertido, afable, seguro, entusiasta.

5.- Introvertido, inseguro, reservado.

c. Amplitud de miras:

1.- Imaginativo, sensitivo, abierto a las ideas nuevas.

5.- Conservador, resistente al cambio, realista.

d. Nivel de cooperación y disponibilidad a los demás:

1.- Compasivo, cooperativo, considerado, amigable.

5.- Precavido, Suspicaz, poco cooperativo con los demás.

e. Nivel de diligencia:

1.- Disciplinado, persistente, perfeccionista.

5.- Espontáneo, impulsivo, conseguir logros no es tan importante.

3. Bloque clínico, que recopila factores que puedan influir en su estado de salud.

a. Lateralidad (mano con la que escribe o trabaja)

b. Enfermedades actuales o crónicas, y patologías padecidas

c. Enfermedades familiares

d. Medicación

e. Posible enfermedad mental diagnosticada

f. Tabaquismo

g. Ingesta alcohol

h. Drogas

i. Ejercicio semanal

La figura 4 muestra parte del cuestionario presentado a cada informante.

ANTECEDENTES MÉDICOS PERSONALES

- **Lateralidad (mano con la que escribe o trabaja):** (izquierda / derecha)
- **Heredofamiliares (** enfermedades de la familia) :

  _____

- **Patológicos** (lista de enfermedades padecidas):

  _____

- **No patológicos:**
  - Tabaquismo (sí / no)

Nunca – casi nunca – de vez en cuando – ocasionalmente – muchas veces – siempre

  - Ejercicio físico (veces por semana): _____
  - Ingesta de alcohol:

Nunca – casi nunca – de vez en cuando – ocasionalmente – muchas veces – siempre

Figura 3. Cuestionario sobre aspectos clínicos

3. ESTADO ACTUAL DEL CORPUS

Actualmente tenemos un corpus de algo más de 200 mil palabras (209.786) proveniente de un total de 89 informantes en su mayoría de la provincia de Cádiz, formado por 61 mujeres y 28 hombres, con una media de edad de 26,3 años y desviación típica de 9,8. La distribución por edades se puede observar en la siguiente figura:

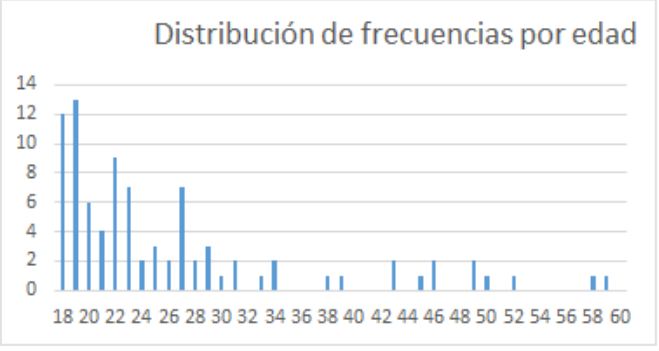

Figura 4. Distribución de frecuencias por edad

La gran masa de informantes se encuentra entre los 18 y 27 años. Esto es debido a que muchos de ellos eran voluntarios de la propia universidad y además, estos medios electrónicos gozan de gran aceptación en estas edades.

Los géneros discursivos electrónicos de nuestro corpus son Facebook, emails, Twitter, mensajería Whatsapp o Line, y blogs con las siguientes proporciones:

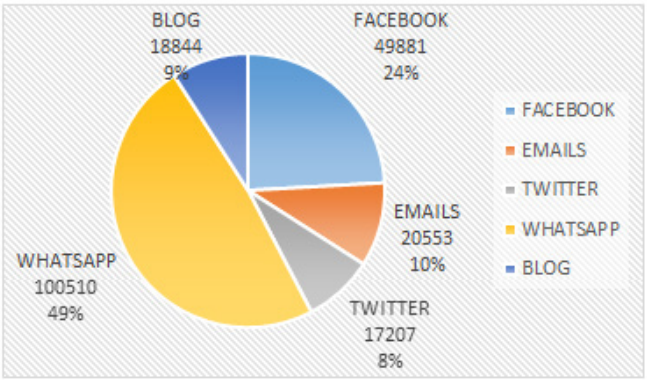

Figura 5. Distribución de géneros discursivos

Como se puede apreciar predomina la mensajería telefónica Whatsapp al estar más universalizada y poseer mayor productividad que el resto de géneros.

4. CONCLUSIONES Y TRABAJO FUTURO

El corpus ForenUCA recopila textos electrónicos a los que se les añade información sociocultural, geográfica, de personalidad y clínica del individuo. Se observa que uno de los principales problemas de nuestro corpus es la preponderancia de mujeres informantes de 18 a 27 años procedentes de la provincia de Cádiz. Nuestra siguiente labor será ampliar estas variables mediante un muestreo aleatorio estratificado para incluir progresivamente a informantes con nuevas características.

De mayor dificultad es encontrar a informantes que no provengan o estén asentados en la provincia de Cádiz. Partiendo de esta primera experiencia nuestra intención es aumentar el trabajo de recopilación con otras universidades e instituciones interesadas para hacer un estudio de mayor complejidad.

También está previsto el enriquecimiento del texto con información lingüística tales como clases de palabras o información sintáctica. Existen herramientas que pueden aligerar el proceso como la herramienta *TreeTagger* (Schmid, 1995) destinada al etiquetado automático de clases de palabras. Sin embargo, este tipo de herramientas, entrenadas sobre un corpus de referencia, producen un mayor error en textos que se desvían de la norma estándar, por lo que está tarea deberá ser eminentemente manual.

Finalmente queremos destacar que esperamos que estos materiales ayuden en la conformación de un corpus de textos electrónicos de referencia para el español, herramienta imprescindible si se quiere trabajar con este tipo de textos desde un punto de vista forense.

REFERENCIAS BIBLIOGRÁFICAS

Alegria, I., Aranberri, N., Comas, P.R., Fresno, V., Gamallo, P., Padró, L., San Vicente, I., Turmo, J., y Zubiaga, A. (2014). TweetNorm_es: an annotated corpus for Spanish microtext normalization. En *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14) European Language Resources Association (ELRA).

Argamon, S., Koppel, M., Pennebaker, J. W., y Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119-123

Bamman, D., Eisenstein, J., y Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135-160.

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

Dunn, J., Argamon, S., Rasooli, A. y Kumar, G. (2016). Profile-based authorship analysis. *Digital Scholarship in the Humanities*, *31*(4), 689-710

Jiménez Bernal, M., Reigosa Riveiros, M., y Garayzábal Heinze, E. (2012). La lingüística forense: licencia para investigar la lengua. En E. Garayzábal Heinze, M. Jiménez Benal y M. Reigosa Riveiros (Eds.), *Lingüística forense: la lingüística en el ámbito legal y policial* (pp. 28-50). Madrid: Euphonia.

Lucena Molina, J.J. (2005). *La acústica forense*. Instituto Universitario sobre seguridad interior. Universidad Nacional de Educación a distancia.. Sitio web: http://portal.uclm.es/descargas/idp_docs/doctrinas/referencia_autores_guardia_civil_i_u_i_ s_i.pdf

McMenamin, G.R. (2010.): Forensic stylistics: theory and practice of forensic stylistics. En M. Coulthard y A. Johnson (Eds.). *The Routledge Handbook of Forensic Linguistics* (473-486). London: Routledge.

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

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

Nguyen, D., Gravel, R., Trieschnigg, D., y Meder, T. (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 (pp. 439-448). Palo Alto, CA,: AAAI Press.

Paolo Rosso, F. R., Potthast, M., Stamatatos, E., Tschuggnall, M., y Stein, B. (2016). Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative* (CLEF 16). Berlin Heidelberg New York: Springer.

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

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

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

Schwartz, A.H., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., y Ungar, L.E. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach, *PloS one*, *8*(9), e73791.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538-556.

Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., y Stein, B. (2015). Overview of the PAN/CLEF 2015 Evaluation Lab. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer*

Science, 9283 (518-538). NY: Springer International Publishing.
Triola, M. F. (2012): *Estadística*. México: Pearson.
Watt, D. (2010). The identification of the individual through speech. En C. Llamas y D. Watt (Eds.), *Language and Identities* (pp. 76-85). Edinburgh.