LAS PERÍFRASIS VERBALES EN ESPAÑOL COMO MARCAS DISTINTIVAS EN LA ATRIBUCIÓN FORENSE DE AUTORÍA[1]

MARIA S. SPASSOVA
Laboratori de Lingüística Forense
Institut Universitari de Llingüística Aplicada
Universitat Pompeu Fabra

RESUMEN. *El objetivo de este estudio, que se enmarca dentro de los trabajos pioneros en determinación de autoría a partir de marcas sintácticas en lengua española, es evaluar el potencial discriminatorio de las perífrasis verbales en español empleadas como marcas distintivas y las posibilidades de su implementación en una técnica de identificación de autor. Siguiendo la metodología recomendada por Coulthard (1994) el estudio se basa, por un lado, en los resultados preliminares de un análisis cuantitativo de los datos de un corpus general del español y por otro, en los resultados del análisis estadístico de las ocurrencias de perífrasis verbales extraídas de nuestro corpus. Dicho corpus está constituido por 15 textos narrativos producidos por 3 escritores hispanohablantes nativos. Este estudio nos permite afirmar que las perífrasis verbales son de uso idiosincrásico y manifiestan un potencial discriminatorio elevado como marca identificativa.*
PALABRAS CLAVES*: determinación de autoría***,** *marcas identificativas, perífrasis verbal*

ABSTRACT. *The main objective of this study, which is set among the pioneer works in language-based author identification by means of syntactic markers in Spanish, is to evaluate the discriminatory potential of verbal periphrases in Spanish as markers of authorship identification. More specifically, it investigates the possibilities of verbal periphrases implementation in an author identification technique. In accordance with the methodology recommended by Coulthard(1994) the study was based both on the results from prior quantitative analysis of the data obtained from general language corpus queries and the data derived from the statistical analysis of verbal periphrases occurrences in our corpus. The corpus consists of 15 narrative texts produced by 3 native Spanish-speaking writers. From the analysis it can be concluded that the verbal periphrases are idiosyncratic in their use and reveal high discriminatory potential when used as identification markers.*
KEYWORDS: *language-based* author *identification, identification markers, verbal periphrases*

1. INTRODUCCIÓN

Los seres humanos somos seres complejos con peculiaridades que, a la vez que nos diferencian, nos unifican. La ciencia nos estudia incansablemente con el fin de poder aislar aquellas idiosincrasias físicas, genéticas, psicológicas o de cualquier otra índole, que luego plasmar en técnicas de análisis. En el caso de las disciplinas forenses, son un buen ejemplo el dactilograma (huella dactilar) en dactiloscopia y el ácido desoxirribonucleico en la prueba de ADN, entre otras, que permiten identificar a cada individuo.

Para elaborar sus técnicas de análisis, la Lingüística Forense a su vez tiene como materia viva de sus experimentos las producciones lingüísticas. Mediante un escrutinio detenido, se dedica a buscar aquellos rasgos idiosincrásicos distintivos únicos e irrepetibles del habla o de la escritura de cada individuo que lo pueden distinguir de otros usuarios de la misma lengua (Coulthard 2004). En la presente fase de desarrollo de la disciplina, en su rama dedicada a la identificación de autor de textos escritos, es de primordial importancia la detección de dichas idiosincrasias, denominadas *marcas identificativas*, y la evaluación de su fiabilidad y relevancia en casos reales para su posible posterior incorporación en nuevas técnicas de identificación de autor (Chaski 1997; Grant y Baker 2001).

La atribución de autoría de textos debitados[2] a partir de marcas de identificación sintácticas es un tema incipiente y apenas explorado en otras lenguas, sobre todo el inglés, y sin precedentes en el español. Este artículo refleja la primera aproximación al problema en un

estudio preliminar que servirá para verificar si los objetivos del trabajo de tesis son factibles y merecen futura investigación.

2. Objetivos e hipótesis

2.1. *Objetivos*

A diferencia de otras lenguas como el inglés que obedece a unas reglas estrictas del orden de las palabras, el español destaca por la libertad estructural de sus enunciados. Por lo tanto, lo que se pretende alcanzar en la investigación es seleccionar las estructuras sintácticas en español cuya variabilidad interna sea relativamente limitada y así permita considerarlas como posibles candidatas a marcas de identificación. Un grupo de estructuras con estas características por excelencia es el grupo de las perífrasis verbales.

El objetivo del estudio que se presenta en este artículo ha sido la evaluación empírica del potencial discriminatorio de las perífrasis verbales en español como marca sintáctica de Atribución de Autoría en textos de narrativa.

2.2. *Hipótesis*

La Lingüística Forense, en todas sus ramas ligadas a la identificación de un individuo concreto, parte del concepto de la unicidad de las realizaciones lingüísticas de los usuarios de una lengua. La hipótesis general que esta disciplina defiende, y a la que nos atenemos sobre todo en Identificación de Autor, establece que cada individuo, como usuario de una variedad de una lengua concreta, dispone de su propia gramática interna que forma parte de su idiolecto y que viene modificada por varios factores externos (geográficos, sociales, psicológicos etc.). Estos cambios quedan reflejados en todos los niveles lingüísticos, y por lo tanto, podemos considerar únicas las realizaciones fonéticas, léxicas y sintácticas de cada autor.

Encontrar marcas sintácticas que posibiliten la atribución de autoría de textos escritos en español constituye una verdadera rémora. Y esto es así debido a la variación sintáctica que caracteriza a esta lengua. A pesar de ello, en el español se observan algunas regularidades en el orden de las palabras dentro de la cláusula y en la organización de las cláusulas en la oración, cosa que, por lo menos en el primer caso, podemos constatar de forma sencilla realizando una consulta en cualquier corpus general de la lengua. Este hecho nos permite admitir como hipótesis que la lengua española dispone de construcciones y estructuras sintácticas que presentan poca variabilidad, lo que por su parte permite que éstas puedan ser aplicadas como marcas para los fines de la Lingüística Forense en el campo de la Identificación de Autoría. Son de especial interés para la investigación, las estructuras que tienen un número delimitado de variantes, a fin de clasificarlas y estudiar su aplicabilidad como marcas identificativas. Las que nos ocupan en este estudio son las unidades verbales formadas por dos verbos, llamadas perífrasis verbales. En virtud de varias características que las perífrasis verbales poseen, y a las que se prestará mayor atención en los apartados siguientes, se espera poder demostrar con este estudio que las perífrasis verbales en español pueden ser marcas de identificación de autor.

3. LAS PERÍFRASIS VERBALES: SELECCIÓN Y CLASIFICACIÓN

Las propiedades estructurales y funcionales que caracterizan las perífrasis verbales (en lo sucesivo PV) y que favorecen su aplicación como marcas de identificación en Autoría son, a grandes rasgos, su constancia combinatoria de los componentes integrantes y su amplio uso en la lengua objeto de análisis.

La clasificación de PV que se sigue en este estudio se basa en la bibliografía pertinente sobre el uso actual de esta clase de unidades verbales (Gómez 1988; Fernández de Castro 1990) y en los resultados de las consultas previas de corpus realizadas para determinar cuáles son las perífrasis más frecuentes en español, ya que éstas supuestamente serían las que tendrían mayor posibilidad de aparecer en cualquier tipo de texto. Entre los componentes de cada grupo de perífrasis seleccionadas en base a su frecuencia de uso en un corpus de referencia se han incluido también otras en cuyo uso se han observado ciertas idiosincrasias en cada autor. Se han excluido, en cambio, las construcciones que, aunque se suelen manifestar con el mismo esquema combinatorio que los elementos de las perífrasis, no tienen las mismas características sintácticas. Por ejemplo, en las PV el segundo verbo en forma no personal determina la semántica y selecciona los argumentos de la construcción (ej.1), mientras que en la pseudoperífrasis (PP) dichas funciones las desempeñan el primer verbo (ej.2).

Ej. (1) La pelea empezó cuando el otro <u>se puso a decir</u> aquellas cosas tan feas. – PV
(2) La mirada atónita de la mujer le <u>dio a entender</u> que allí pasaba algo raro. – PP

Finalmente, en este estudio se ha trabajado con las ocurrencias de 56 tipos de PV. En la Tabla 1 se detalla su adjudicación en los tres grupos generales de infinitivo, gerundio y participio.

Se ha de notar aquí que en la fase del análisis estadístico algunos de los grupos de PV han quedado excluidos para evitar *el ruido* que se produce cuando hay una distribución desigual con pocas o nulas ocurrencias de las variables.

PV de Infinitivo		PV de Participio		PV de Gerundio	
Temporales	ir + a + inf.	**Pasivas**	ser + part. estar + part.	**Progresivas**	estar + ger.
Inminentes	estar + por + inf. estar + para + inf.	**Resultativas**	tener + part. quedar(se) + part. dejar + part. verse + part. hallarse + part. ponerse + part. volverse + part.	**Continuativas**	seguir + ger. continuar + ger. andar + ger. quedar(se) + ger. ir + ger.
Incoativas	ponerse + a + inf. echarse + a + inf. romper + a + inf. pasar + a + inf. empezar + por + inf. comenzar + a + inf. empezar + a + inf.				
Reiterativas	soler + inf. volver + a + inf.			**Durativas**	venir + ger. llevar + ger. tener + ger.
Terminativas	acabar + de + inf. acabar + por + inf. dejar + de + inf. terminar + de + inf. terminar + por + inf.	**Durativas**	seguir + part. continuar + part. permanecer + part.	**Terminativas**	acabar + ger. terminar + ger. salir + ger.
Resultativas	llegar + a + inf. alcanzar + a + inf. venir + a + inf.				
Voluntativas	querer + inf. poder + inf.	**Frecuentativas**	ir + part. venir + part. llevar + part. andar + part.	**Incoativas**	empezar + ger.
Obligativas	haber + de + inf. haber + que + inf. deber + inf. tener + que + inf.				
Dubitativas	deber + de + inf.				

Tabla 1. *Clasificación de las PV empleada en el estudio*

4. CORPUS

4.1. *Selección del corpus*

Con el propósito de elaborar conclusiones sobre el comportamiento de la variable de análisis y comprobar la validez de las hipótesis planteadas en este estudio se ha recurrido a la consulta del Corpus Técnico (CT) del Instituto Universitario de Lingüística Aplicada (IULA)[3], Universitat Pompeu Fabra, por un lado, y por otro, a la creación de un Corpus de Análisis (CA) constituido por textos narrativos de tres escritores de habla hispana, a saber: Cela, Vargas Llosa y Mendoza. La extensión media de los textos para su inclusión en el corpus ha sido de 3000 palabras. Para que los datos sobre la ocurrencia de las variables sean representativos del estilo del autor el número de documentos se fijó en 5 por escritor. En definitiva, se incorporaron 15 textos narrativos alcanzando a un corpus de 45000 palabras.

Con el objetivo de facilitar el manejo del gran tamaño de información recopilada y agilizar la búsqueda de las PV en los textos de cada autor, los documentos de CA se procesaron e incorporaron al CT que permite su interrogación en línea.

4.2. *Extracción de datos*

Las PV objeto de análisis fueron extraídas del CT y CA por medio de expresiones regulares, mediante la opción de consulta compleja del Bwananet[4].

Se formularon una serie de expresiones regulares con la ayuda de las cuales fue posible interrogar los corpus sobre varias perífrasis a la vez, tal que:

ej. [lemma="acabar|terminar|dejar"& pos="V.*"] [lemma="de|por"] [pos="VI----"]

Con la expresión del ejemplo se realiza la búsqueda de una secuencia de palabras en la que la primera puede ser *acabar*, *terminar* o bien *dejar* en función de verbo, seguida por la preposición *de* o *por*, seguida por un verbo en infinitivo.

5. ANÁLISIS DE LOS DATOS

En el tratamiento final de los datos se han llevado a cabo dos tipos de análisis: comparativo y estadístico. El análisis comparativo ha consistido en realizar una estimación de la frecuencia de uso de las PV en un corpus de la lengua general. La finalidad de esta estimación, a partir del concepto de *prominencia* en Autoría, es la de saber cuáles son los tipos de perífrasis verbales de valor más alto, y por consiguiente, de mayor uso en la lengua castellana y de comprobar si la tendencia se mantiene con valores aproximados en el Subcorpus General del Corpus Técnico del IULA (en adelante SCTG) y en el CA. Esto podría indicar que las PV son de uso idiosincrásico. Pero también señalar qué grupos de PV habría que eliminar en una fase posterior de análisis.

Con el análisis estadístico de las ocurrencias de PV en el CA, constituido por los textos narrativos producidos por tres escritores diferentes y estudiados en su conjunto, se evaluó la capacidad de las PV de discriminar entre autores y su fiabilidad como marca identificativa. En el análisis se utilizaron como variables sólo aquellas PV que han manifestado un uso elevado en el análisis comparativo.

5.1. *Análisis comparativo*

En el análisis comparativo mediante consultas consecutivas del SGCT y el CA, se recuperaron las ocurrencias para cada grupo de perífrasis (de infinitivo (PVI), de participio (PVC) y de gerundio (PVG)). Los Gráficos 1, 2 y 3 muestran el número de ocurrencias en porcentajes de PVI, PVC y PVG en CA en comparación con los valores de los datos del SGCT.

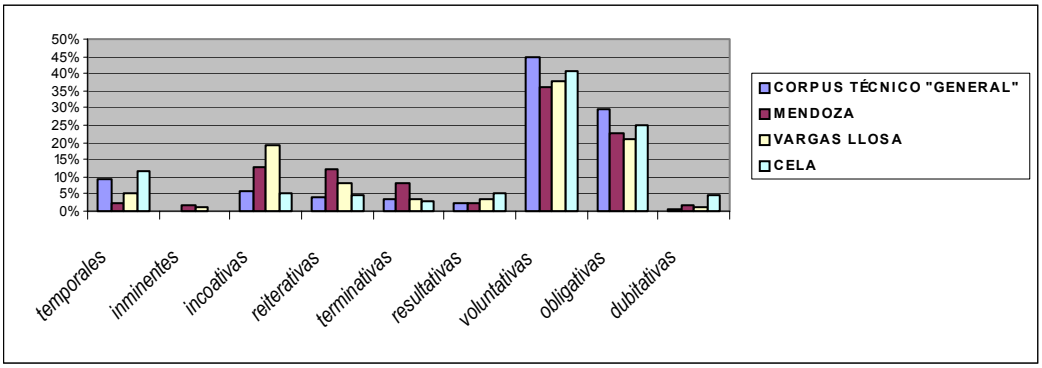

Gráfico 1. *Representación gráfica de las ocurrencias de PVI en el SGCT y CA*

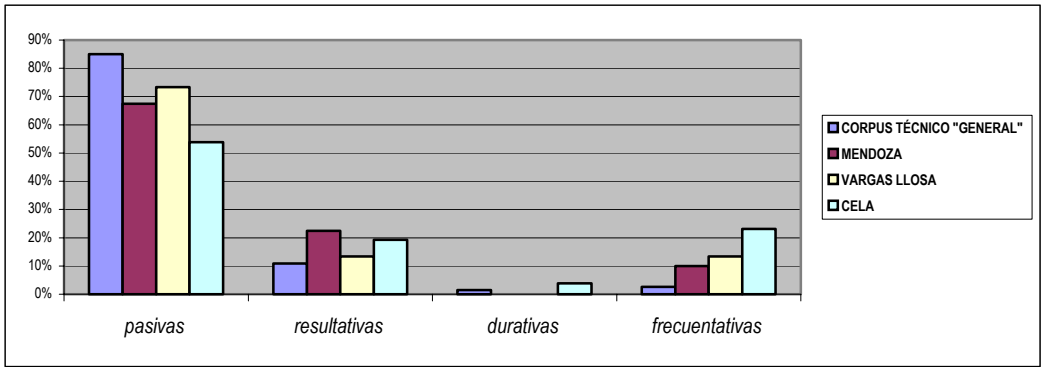

Gráfico 2. *Representación gráfica de las ocurrencias de PVC en el SGCT y CA*

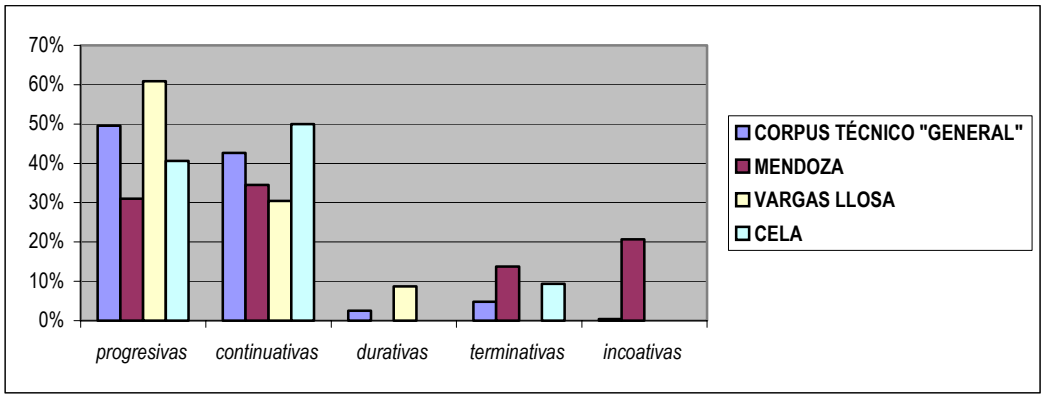

Gráfico 3. *Representación gráfica de las ocurrencias de PVG en el SGCT y CA*

El tipo de perífrasis que predomina en ambos corpus coincide. Además, según los datos del SGCT las PV que más se usan son las de participio y las que menos, las de infinitivo. Dato que no debe resultar sorprendente teniendo en cuenta que las construcciones que expresan la voz pasiva en la lengua española están en el grupo de las PVC.

Por otra parte, en todos los gráficos destaca una cierta distancia entre los valores de algunas de las variables menos frecuentes, lo que apunta a un uso idiosincrásico por parte de los autores. La observación gráfica permite considerar las perífrasis inminentes y dubitativas en el caso de las PVI, las durativas en el de las PVC, y por último las durativas entre las PVG, cuantitativamente irrelevantes y descartarlas a la hora del análisis estadístico.

5.2. *Análisis estadístico*

Para el análisis estadístico se ha aplicado la técnica de análisis discriminante. Las variables retenidas en el análisis son las siguientes:

PVI	PVC	PVG
• *Temporales* • *Incoativas* • *Reiterativas* • *Terminativas* • *Resultativas* • *Voluntativas* • *Obligativas*	• *Pasivas* • *Resultativas* • *Frecuentativas*	• *Progresivas* • *Continuativas* • *Terminativas* • *Incoativas*

Tabla 2. *PV empleadas como variables en el análisis estadístico*

Los resultados de las pruebas estadísticas realizadas pueden verse en los Gráficos 4, 5 y 3. Se han clasificado los textos de autor 1 (Mendoza), 2 (Vargas Llosa), y 3 (Cela) aplicando como marca de identificación las PVI.

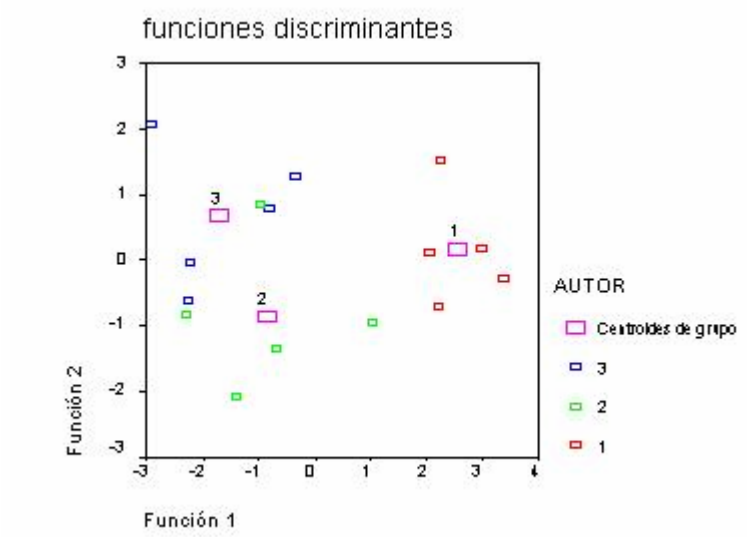

Gráfico 4. *Resultados de la clasificación de los textos según autor mediante PVI*

En el Gráfico 4 se observa que las PVI agrupan correctamente las muestras de escritura de los tres autores analizados en un 93% de los casos. Además se contempla una ligera dispersión en la agrupación de las muestras de autor 2 (Vargas Llosa) y autor 3 (Cela), probablemente debida a la mayor heterogeneidad en su forma de escribir en comparación con el autor 1(Mendoza) que se muestra más homogéneo en el uso de PVI.

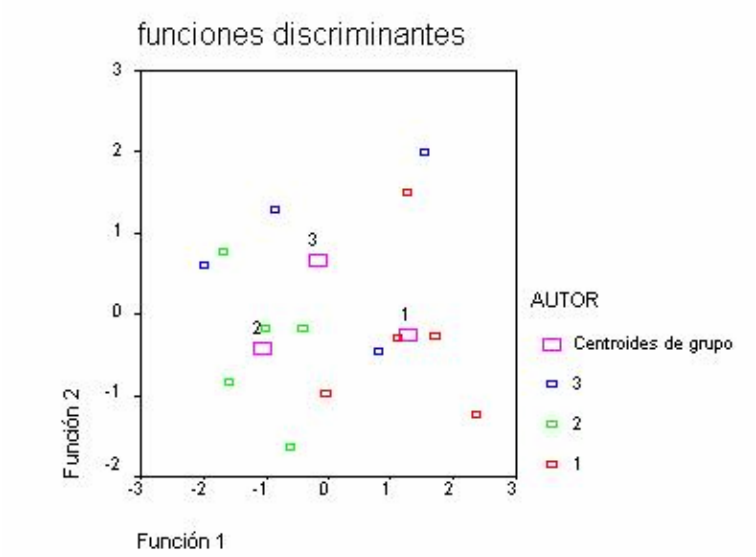

Gráfico 5. *Resultados de la clasificación de los textos según el autor mediante PVC*

La prueba estadística basada en las PVC, en cambio, no ha dado resultados tan relevantes, ya que según señala el Gráfico 5, la agrupación de los textos es correcta sólo en un 66% de los casos, lo que demuestra que las PVC no son suficientemente fiables para ser aplicadas por si solas como marca de identificación.

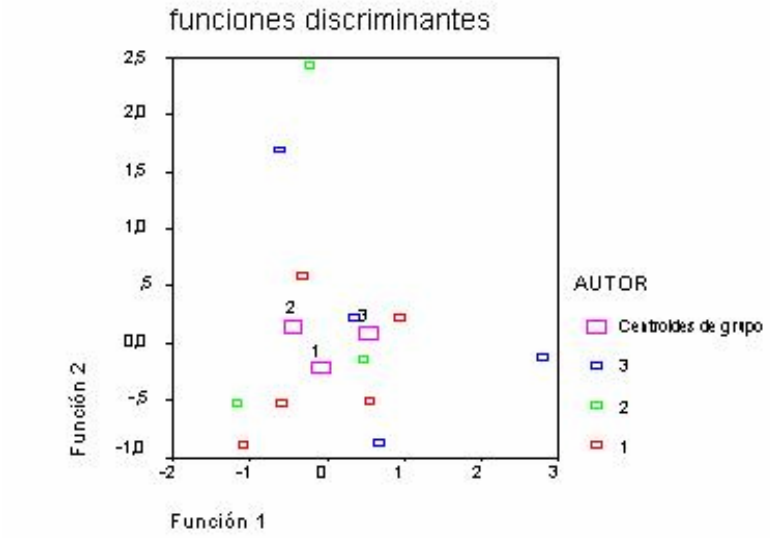

Gráfico 6. *Resultados de la clasificación de los textos según autor mediante PVG*

En el caso de las PVG, nos encontramos ante una situación similar. Los resultados de la clasificación descartan esta variable como rasgo idiosincrásico del estilo de los tres autores estudiados por el alto nivel de dispersión de los marcadores y por la proximidad de los centroides de cada grupo (véase el Gráfico 6).

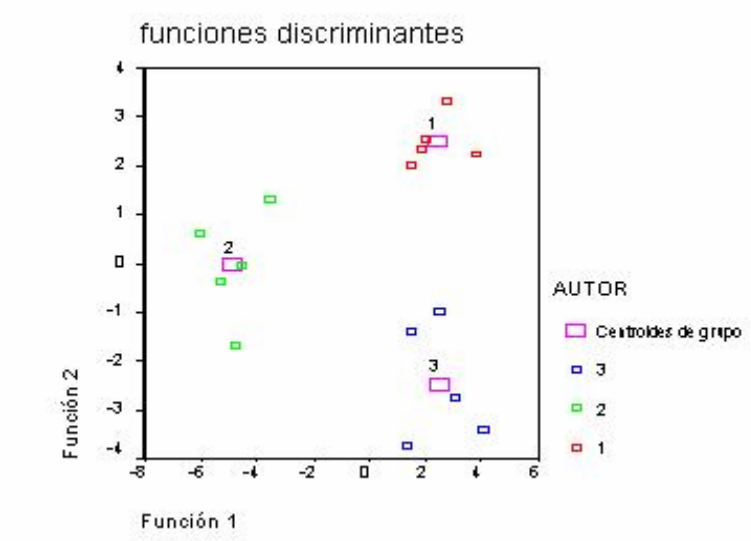

Gráfico 7. *Resultados de la clasificación de los textos según autor mediante PV*

En la prueba final se han empleado como descriptores todos los tipos de PV seleccionadas con anterioridad (Gráfico 7). Con una clasificación de los textos correcta en un 100% de los casos, el resultado del último análisis apunta a que las PV empleadas en su conjunto en vez de por grupos separados son más susceptibles de ser marca de identificación de autor. De las variables usadas en las pruebas estadísticas comentadas el *p valor* significativo, es decir inferior o igual a 0.05, se manifiesta sólo en dos: las incoativas y las reiterativas de infinitivo. De disponer de documentos de casos reales de autoría, el potencial discriminatorio de estos dos tipos de perífrasis merecería ser estudiado en un futuro

6. CONCLUSIONES

Tomando como referencia los resultados obtenidos, mediante el análisis estadístico discriminante de los datos extraídos de CA, se ha llegado a dos conclusiones generales en cuanto al potencial discriminatorio de las PV en la lengua castellana.

En primer lugar, que el uso de las PV es de carácter idiosincrásico y que está relacionado con el idiolecto de cada individuo, especialmente el de las perífrasis de infinitivo. De lo que se desprende que dichas perífrasis pueden discriminar entre autores con un grado de fiabilidad considerable.

En segundo lugar, detectar este tipo de idiosincrasia idiolectal implica aumentar el corpus analizado con más textos. Además, para poder evaluar la fiabilidad de las PV como marcas identificativas, debería estudiarse el comportamiento de las mismas en combinación con otras marcas en casos reales.

NOTAS

1. La presente comunicación informa sobre el trabajo experimental llevado a cabo para la obtención del DEA en Lingüística del Institut Universitari de Lingüística Aplicada (UPF), dentro del marco del estudio de Autoría del ForensicLab de este instituto.
2. Documentos escritos cuya autoría se cuestiona o desconoce
3. A la hora de realizar las consultas (julio de 2005) el Corpus Técnico del IULA contenía 911 documentos y un total de 16 912 597 palabras
4. Programa de explotación del Corpus Técnico del IULA: http://bwananet.iula.upf.edu/

BIBLIOGRAFÍA

Biber, D. 1990. "Methodological issues regarding corpus-based analyses of linguistic variation". *Literary and Linguistic Computing* 5: 257-269

Chaski, C.E. 1997. "Who wrote it? Steps towards a science of Authorship identification". *National Institute of Justice Journal*, September:15-21

Coulthard, M. 1994. "On the use of corpora in the analysis of forensic texts". *Forensic Linguistics*, 1: 26-43

Coulthard, M. 2004. "Author identification, idiolect, and linguistic uniqueness". *Applied Linguistics* 25: 431-447

Fernández de Castro, F. 1990. *Las perífrasis verbales en el español actual*. Madrid: Gredos

Gómez, L. 1988. *Perífrasis verbales: sintaxis, semántica y estilística*. Madrid: Arco libros

Grant, T., Baker, K. 2001. "Identifying reliable, valid markers of authorship: a response to Chaski". *Forensic Linguistics* 8: 66-79

Seco, M. 1989. Grámatica esencial del español. Introducción al estudio de la lengua. Madrid: Espasa Calpe