

La aplicación del programa de multivariación GoldVarb en la investigación diacrónica

Miguel Gutiérrez Maté

Angaben zur Veröffentlichung / Publication details:

Gutiérrez Maté, Miguel. 2009. "La aplicación del programa de multivariación GoldVarb en la investigación diacrónica." In *Tendencias actuales en la investigación diacrónica de la lengua: actas del VIII Congreso Nacional de la Asociación de Jóvenes Investigadores de Historiografía e Historia de la Lengua Española Barcelona, del 2 al 4 de abril de 2008*, edited by Laura Romero Aguilera and Carolina Julià Luna, 133–46. Barcelona: Universitat de Barcelona.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>



LA APLICACIÓN DEL PROGRAMA DE MULTIVARIACIÓN *GOLDVARB* EN LA INVESTIGACIÓN DIACRÓNICA

MIGUEL GUTIÉRREZ MATÉ

Universidad de Valladolid

1. FUNDAMENTOS Y APLICACIONES DEL MÉTODO VARIACIONISTA

Si bien se insiste en el panorama lingüístico actual en las fases de preparación y presentación del corpus y de búsqueda y extracción de datos¹, no parece prestarse la misma atención a la fase de cuantificación de los datos, es decir, a qué hacer con estos una vez que se han obtenido, cómo ordenarlos y cuantificarlos para aprovechar al máximo los resultados. A menudo esta fase queda subsumida en las anteriores, por lo que la cuantificación se reduce, en el mejor de los casos, a cálculos, más o menos básicos, de frecuencias². La aplicación (o incluso el desarrollo) de técnicas estadísticas más complejas parece restringida a unas pocas áreas dentro de la Lingüística, una de las cuales es la Sociolingüística de corte laboviano o *variacionista* (denominación a la que se le incorpora, de forma más o menos tácita, la propiedad de cuantificadora).

No me detendré en reflexionar acerca de las relaciones entre la estadística y la (socio)lingüística, en parte porque escapa a los objetivos de este trabajo, en parte porque ya ha sido analizada por Moreno Fernández (1994: 95-154). Este artículo, que viene a constituir, aún en la actualidad, el mejor manual sobre el tema, aborda los orígenes del variacionismo y su relación de *amor-odio* con el generativismo³, diferencia claramente entre una estadística descriptiva y una estadística inferencial —la cual, por medio de un análisis probabilístico, permite extrapolar los resultados de una muestra a una población—, plantea los fundamentos del modelo logístico de regresión, que conduce al cálculo de probabilidades, y explica la evolución de los programas VARBRUL (desde su primera versión en 1971 hasta las de los años 90), así como las funciones principales de estos programas. Tan solo cabría añadir dos anotaciones a este ya clásico trabajo: por una parte, mencionar el volumen coordinado por Humberto López Morales que, bajo el título de *Corrientes actuales en la dialectología del Caribe hispánico*, reunió los trabajos de un simposio celebrado en la Universidad de Río Piedras en 1976 y que

¹ Estas son *grossa modo* las competencias de la llamada Lingüística de Corpus, que ha ido creciendo en las últimas décadas hasta convertirse en una disciplina autónoma, intrínsecamente ligada al manejo de herramientas informáticas, que en su aplicación a diacronía ha sido señalada también como disciplina metodológica con sus propios objetos de estudio (prueba de ello es la celebración del reciente *Col·loqui internacional de corpus diacrònics en llengües iberoromàniques*, Palma de Mallorca, 25-27 de octubre de 2007).

² Se suele considerar que estos cálculos constituyen una etapa de *pre-cuantificación* (Moreno Fernández 2005[1998]: 316).

³ Es significativo el hecho de que en aquellos momentos se considerara la «actuación» como un «reflejo estadístico de la competencia» (Cedergren y Sankoff 1974: 33-355).

constituye un buen ejemplo de los intereses iniciales en el ámbito hispánico por las labores de cuantificación (el trabajo de Cedergren sobre la aspiración de /s/ implosiva en Panamá, por ejemplo, recoge resultados elaborados ya por medio del programa VARBRUL); por otra parte, citar las últimas versiones (posteriores al texto de Moreno Fernández) del programa GoldVarb, continuador del VARBRUL: *GoldVarb 2001* (Robinson, Lawrence y Tagliamonte 2001a) y *GoldVarb X* (Sankoff, Tagliamonte y Smith 2005).

Aun a riesgo de caer en una peligrosa generalización, el método de análisis variacionista puede reducirse a cuatro etapas o pasos fundamentales, una vez constituida la muestra: la identificación de la variable dependiente (o variable objeto de estudio) y sus variantes, la acotación del dominio de la variación (esto es, la eliminación de los usos obligatorios —los contextos que no permiten variación o alternancia entre las variantes—), la identificación de las variables independientes (o *grupos de factores* explicativos) y el análisis estadístico de los datos, que permite en última instancia observar en qué medida los factores explicativos influyen sobre la variable dependiente (hallando índices de probabilidad con valores que oscilan entre 0 y 1, de tal modo que por encima de 0,5 consideramos que un factor favorece el uso de una variante determinada y por debajo de 0,5 lo desfavorece). Para esta última etapa contamos con la ayuda de los programas de regla variable que se han diseñado y perfeccionado específicamente para el estudio de la variación lingüística (no obstante, la mayoría de sus funciones forman parte de programas estadísticos más generales para el estudio de las Ciencias Sociales, como SPSS —*Statistical Package for the Social Sciences*—).

Este método de análisis ha superado los límites de la variación fonética, de la que dio cuenta de manera exclusiva en los primeros años, y se ha convertido en una valiosa herramienta para el estudio de la variación sintáctica⁴ e incluso se ha desvinculado a veces de la variación social para estudiar los factores intralingüísticos que condicionan la variación⁵. Además, hay otro salto que el método variacionista, al menos desde el punto de vista conceptual, puede dar fácilmente: de su aplicación en la investigación sincrónica a la diacrónica⁶, lo que supone abandonar los corpus de grabaciones orales en favor de los corpus escritos y analizar datos de estadios de lengua anteriores que no debemos interpretar conforme a nuestro conocimiento de la lengua actual. Sin embargo, hay que reconocer que su aplicación en el estudio de la historia del español ha sido hasta la fecha muy escasa⁷.

⁴ Un estado de la cuestión reciente de la polémica en torno a la conveniencia o no de considerar variables sintácticas, que arranca de las críticas de Lavandera (1978), puede leerse en López Morales (2006).

⁵ Es evidente que el método variacionista nace en el seno de la Sociolingüística porque otorga una especial atención, además de a factores estructurales, a factores sociales o extralingüísticos como la clase social, el sexo, la edad, etc. Sin embargo, en lo referente al nivel sintáctico, se ha dicho que la falta de datos que permite evaluar la conducta de un grupo social y compararla con la de otros provoca que el trabajo del investigador se convierta a menudo en «un estudio de variación sintáctica que culmina con la identificación de los factores internos que condicionan y explican la variación» (Silva-Corvalán 2001: 137).

⁶ Desde nuestra perspectiva diacrónica, en la que reconocemos la lengua como un sistema dinámico, el variacionismo habrá de interesarnos en la medida en que todo cambio lingüístico presupone variación en su origen, si bien la variación no implica necesariamente un cambio lingüístico: «la variación estable y a largo plazo que persiste por muchos siglos básicamente en la misma forma es quizá más común incluso que los cambios que avanzan hacia su conclusión» (Labov 2006[1994]: 128).

⁷ Encontramos, no obstante, excepciones en las que se llegan a utilizar los programas de regla variable: por ejemplo, el análisis multivariado de Gimeno Menéndez (1998: 123-133), que propone como variable dependiente la realización sorda o sonora del elemento dentoalveolar africado en posición implosiva en documentos del siglo XIII y como variables independientes, su posición en la palabra, el contexto fónico y el lugar de procedencia del documento en combinación con su grado de formalidad; huelga decir que en

2. FASES EN LA ELABORACIÓN DE UN ANÁLISIS MULTIVARIADO

Corresponde en este apartado mostrar cómo se lleva a cabo un análisis cuantitativo por medio del programa GoldVarb, después de seleccionar y ordenar el corpus y de decidir cuáles son las variables explicativas. Sin ánimo de explicar cada componente del programa (existen manuales para ello), me detendré en indicar los pasos que ha de seguir el investigador —pensando más en el historiador de la lengua que en el sociolingüista— para obtener el mayor beneficio de acuerdo con sus intereses. Para ello estableceré como objeto de estudio (o variable dependiente) la presencia/ausencia del pronombre *yo* en un pequeño corpus de documentos caribeños de finales del siglo XVII, fenómeno del que ya me ocupé, sobre todo de la parte cualitativa, en un trabajo anterior (Gutiérrez Maté, en prensa)⁸. A modo de pequeña muestra, aunque suficiente para los objetivos de esta exposición, tomaré del corpus documental con el que trabajo en mi Tesis doctoral solamente cinco documentos, si bien algunos son «documentos complejos» (Carrera de la Red y Gutiérrez Maté 2007), es decir, encierran a su vez varios textos —a menudo de distinta tipología desde el punto de vista diplomático y/o lingüístico⁹—, que tienen en común el estar fechados en la última década del siglo XVII en la ciudad de Cartagena de Indias¹⁰.

Es fundamental definir la muestra y clasificarla en función de criterios que nos puedan ayudar desde el principio a fijar algunas variables. Así, el oficio o cargo que ocupa el autor del texto constituye un criterio objetivo para la determinación de la variable *clase social*: son considerados de clase alta las autoridades gubernamentales y eclesiásticas (en menor medida, civiles) que aparecen en la documentación y de clase baja los sirvientes y los distintos oficios (albañiles, zapateros, pulperos, etc.); en medio queda una parte muy amplia de la sociedad indiana, donde incluimos a las autoridades intermedias, los militares, diversas personas ligadas a la administración —también secretarios de juzgado y escribanos—, etc. Por otra parte, también el tipo de texto, dadas las diferencias de tipo diplomático y discursivo entre cartas, autos, documentos dispositivos, etc., habrá de tener consecuencias importantes en la variación lingüística:

este tipo de trabajos lo que se analiza son variantes gráficas y no propiamente fonéticas (en el caso señalado, la alternancia <Z>/<C>). Esta situación contrasta con la que se presenta en el ámbito de nuestra lengua hermana, el portugués, donde empiezan a proliferar trabajos variacionistas, realizados por lo general en universidades brasileñas, que se ocupan bien de etapas pasadas de la historia de la lengua portuguesa (generalmente, de sus variedades ultramarinas), bien de los cambios que se han experimentado en ella a lo largo del tiempo —Santos Lopes (2003) estudia la variable *nós/a gente* como pronombre de primera persona del plural en el portugués de Brasil, Mendes Mothé (2006) la alternancia entre el gerundio y *a* + infinitivo en los dos últimos siglos, Barcia (2006) los tratamientos *tu/você* en cartas brasileñas del siglo XIX, etc.—.

⁸ He utilizado GoldVarb 2001 para este análisis, dado que la principal ventaja de GoldVarb X, es decir, poder trabajar con una variable dependiente de más de dos variantes, no es aplicable en nuestro caso.

⁹ El último documento de la muestra, unos autos criminales, constituye un universo discursivo propio, puesto que, aun habiendo excluido del presente análisis su parte final —más repetitiva formalmente y en cuanto al contenido (intervienen declarantes que ya lo habían hecho antes)—, encontramos textos muy diferentes (nueve declaraciones, más un *dicho* o declaración por escrito, dos cartas, cuatro autos —las partes propiamente dispositivas del documento—, cuatro certificaciones notariales, una notificación, una diligencia, una confesión y una petición —en este caso, no de un demandante sino del reo—). A su vez las declaraciones presentan fragmentos en estilo indirecto y otros relativamente extensos en estilo directo. Todo ello conforma una heterogeneidad de discursos que fijará las variables extralingüísticas con unos valores u otros a lo largo del documento.

¹⁰ Los textos fueron encontrados, bien por la Prof.^a Micaela Carrera bien por mí mismo, en el Archivo General de Indias de Sevilla y, al igual que otros documentos que transcribo y estudio en mi tesis, permanecen aún inéditos.

Ref. archivo	Descripción del contenido	Fecha	Tipo	Autor	Cargo/oficio del autor	Destinatario	Ext. (fs.)
A.G.I., Santa Fe 212	Traslado de tres cartas escritas desde Timiriguaco por Martín de Cevallos a Pedro Martínez de Montoya mostrando su conformidad y agradeciendo la dureza en el trato con los negros de los palenques y de Cartagena	1693 julio 18	Cartas (traslado)	Martín de Cevallos (Diego de Benasar)	Gobernador de Cartagena (Escribano público)	Pedro Martínez de Montoya	4 fs.
A.G.I., Santa Fe 212	Autos en los que se toma declaración a tres negros por su supuesta conspiración con los negros de los palenques	1693 mayo 1	Autos judiciales	José Blanco García	Escribano público	(Audiencia de Santa Fe/Consejo de Indias)	14 fs.
A.G.I., Santa Fe 213	Copia de una carta del Teniente General Pedro Martínez de Montoya al alcalde Marcos de Vega pidiéndole un documento en que se justifique la recusación hecha contra él, para presentarlo a Su Magestad y copia de la carta de respuesta de dicho alcalde	1693 junio 2	Cartas (copia)	Pedro Martínez de Montoya Marcos de Vega (José Blanco)	Teniente general de Gobernación Alcalde de Cartagena (Escribano público)	Marcos de Vega Pedro Martínez de Montoya	4 fs.
A.G.I., Santa Fe 212	Carta del mulato Diego de Frias pidiendo su liberación, incluida en la contracubierta de los autos criminales contra Pacho de Vera	1694 ¿?	Carta	Diego de Frias	Oficial de albañil	Gobernador de Cartagena ¿?	1 fol.
A.G.I., Santa Fe 212	Autos criminales contra Francisco de Vera por la supuesta conspiración con los negros esclavos de la ciudad	1694 junio 16	Autos judiciales	José Blanco García	Escribano público	(Audiencia de Santa Fe/Consejo de Indias)	110 fs.

Tabla 1: Documentos que constituyen la muestra

Otro paso previo a la introducción de los datos en el programa es la determinación de las variables independientes y la asignación de un valor (un número o una letra) a cada uno de los factores. Para esta prueba hemos distinguido grupos de tipo morfosintáctico (números 2, 6, 7 y 8), semántico-discursivo (3 y 4), semántico-oracional (5), pragmático (9 y 10), textual (11), discursivo (12) y sociolingüístico (13 y 14), además de un grupo referido al número del documento en el corpus (15), que cumple con una función similar a la de la variable *número de informante* en muchos estudios sociolingüísticos. Así mismo, según el carácter más o menos subjetivo de la realidad que pretendemos describir con las variables, algunas de ellas (4, 5, 10, etc.) requieren mayor grado de interpretación que otras (2, 6, 7, etc.) para asignarles un valor. La variable dependiente siempre ha de aparecer en primer lugar:

1. Omisión de *yo* = 0; Expresión de *yo* = 1.
2. *Ambigüedad de la forma verbal entre la primera y la tercera persona*: Forma verbal ambigua = A; Forma verbal no ambigua = B.
3. *Mantenimiento o cambio de referencia*: Misma referencia que el sujeto del verbo anterior = R; Cambio de referencia = S; Nueva referencia = N.
4. *Topicalidad*¹¹: Continuidad de tópico = c; Mantenimiento de tópico = m; Reanudación de tópico = r; Vuelta momentánea al tópico = v; Cambio de tópico = t; Establecimiento de nuevo tópico = n.

¹¹ Los factores establecidos para esta variable resultan muy intuitivos, tanto en la denominación como en el concepto que encierran; sin embargo, es lícito preguntarse cómo afecta a la estructura informativa del texto la introducción de la primera persona y distinguir grados de topicalidad en cada caso, siempre que el criterio seguido esté fundamentado lingüísticamente y, ante todo, sea mantenido rigurosamente durante el análisis.

5. *Papel temático del sujeto*: Sujeto agente = a; Tema o paciente = t; Experimentante = e; Poseedor = p; Beneficiario = b; Causante = c; Meta = m; Estado = s; Sujetos semi-agentivos = o.
6. *Transitividad verbal*¹²: Verbo transitivo = T; Verbo intransitivo = I.
7. *Presencia de un morfema reflexivo*: Verbo pronominal = p; no pronominal = q.
8. *Tiempo verbal*: Presente = a; Imperfecto = b; Indefinido = c; Futuro = d; Condicional = e; Pretérito perfecto = f; Pluscuamperfecto = g; Pretérito anterior = h; Futuro perfecto = i; Condicional perfecto = j; Presente subjuntivo = k; Imperfecto subjuntivo = l; Futuro subjuntivo = m; Perfecto subjuntivo = n; Pluscuamperfecto subjuntivo = o; Futuro perfecto subjuntivo = p.
9. *Capacidad del acto de habla de transformar la realidad*: Acto no declarativo = 1; Acto declarativo = 2.
10. *Illocución del acto de habla*: Acto asertivo = a; directivo (de mandato) = d; directivo (petición) = p; comisivo = c; expresivo = e; agradecimiento = g.
11. *Tipo textual*: Auto = A; Declaración = D; Carta informe/petición = C; Texto legislativo/dispositivo = L; Testimonio/Certificación = T; Petición = P.
12. *Mímesis de lo oral*: Fragmentos en estilo directo = d; Partes que no están en estilo directo = i.
13. *Clase social*: alta = a; media = m; baja = b.
14. *Etnia*: Blancos = b; Negros = n; Mulatos = m.
15. *Número de documento en la muestra*: 1; 2; 3; 4; 5.

Una vez fijados los factores con los que vamos a trabajar, se deben introducir los datos en el programa GoldVarb. La base sobre la que descansan todas las operaciones que realiza el programa es el archivo de *tokens* (en buena medida, un archivo de texto), donde a cada ejemplo le habrá de preceder un código en el que cada una de las variables, dependiente e independientes, tiene asignado un valor concreto:

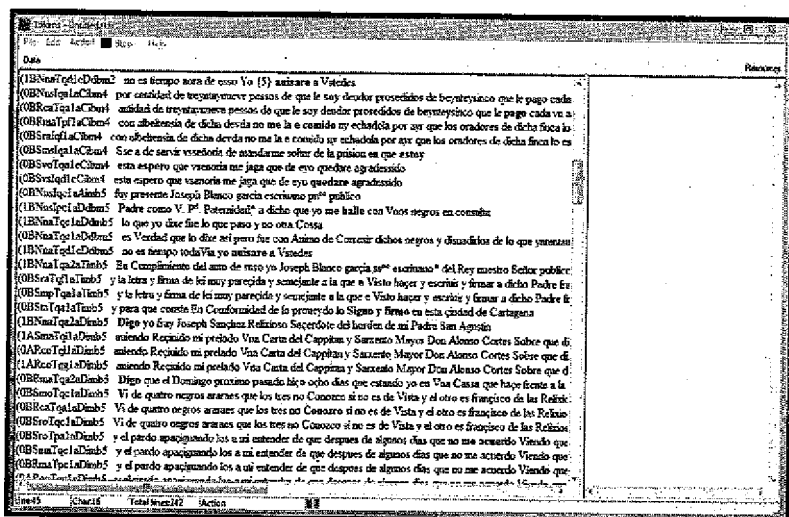


Figura 1: Codificación de los ejemplos en el archivo de tokens

¹² Se podrían haber señalado varios tipos de transitividad, en línea con la hipótesis de la transitividad gradual (Hopper y Thompson 1980), o haber distinguido entre intransitivos puros e inacusativos, pero se ha optado por la clasificación dicotómica tradicional, si bien el concepto de 'transitivo' resulta más lato de lo habitual (por ejemplo, muchos verbos con complemento de régimen se consideraron transitivos).

En la ventana de *tokens* se disponen los ejemplos y sus respectivos códigos en líneas sucesivas: escribimos en primer lugar el código, introducido obligatoriamente por un paréntesis de apertura para que el programa lo reconozca como tal código, y después, separado por tres espacios, que sirven para indicar al programa el final del código, el correspondiente ejemplo, que podemos copiar y pegar desde los archivos que contienen los textos digitalizados de nuestro corpus¹³.

En la primera línea (en realidad, la n.º 45 del archivo de *tokens*) observamos que el ejemplo «no es tiempo aora de esso Yo ausare a Vsteden» se ha codificado como «1BNnatqdlcDdbm2» —con 15 dígitos que resultan de indicar un valor para cada variable—, que debe leerse del siguiente modo: el pronombre sujeto *yo* está expreso (1); la forma verbal no presenta ambigüedad morfológica entre la primera y la tercera persona (B); el sujeto no tiene la misma referencia que el sujeto del verbo anterior —de hecho, es la primera vez que se hace referencia a este sujeto— (N); el empleo del pronombre inaugura un nuevo tema o tópic o en la estructura informativa del texto o párrafo (n); el papel temático del sujeto es agente (a); el verbo es transitivo (T) y no pronominal (q); el tiempo verbal es el futuro de indicativo (d); el acto de habla es *no declarativo* en cuanto a su capacidad de transformación de la realidad (1) y comisivo o compromisorio en cuanto a su carga ilocutiva (c); el ejemplo está inserto en la declaración de un testigo durante unos autos (D), concretamente en un fragmento en estilo directo (d), proferido por una persona de clase social baja (b) y de raza mulata (m), que hallamos en el documento n.º 2 de la muestra (2). La mayor parte del tiempo empleado en el manejo del programa se ocupa en esta fase de codificación manual de los ejemplos; si bien la tarea resulta ardua, el investigador ha de estar seguro de que la precisión de los resultados estadísticos finales y su potencial de interpretación habrán compensado el esfuerzo.

El siguiente paso consiste en indicar al programa qué y cuántos factores y *grupos de factores* (las variables) hemos tenido en cuenta durante la fase de codificación (*Factor specification*)¹⁴:

Group	Factor	No. factors	Grid
1	B	2	1 1
2	A	2	A B
3	R	3	R S H
4	C	5	C D E F G
5	A	8	A B C D E F G H

Figura 2: Especificación de factores

¹³ Además, podríamos añadir cuantos comentarios lingüísticos juzgásemos oportunos, siempre que fueran escritos bajo los ejemplos a los que se refieren e introducidos por punto y coma (;).

¹⁴ Utilizamos los recuadros de la parte superior de la ventana para pasar de un grupo de factores a otro (*New group*) e introducir manualmente los valores correspondientes a cada factor (*New factor*), que se escriben automáticamente en la hoja cuadrículada de la parte inferior. En esta el número que corresponde a cada grupo aparece en la columna de la izquierda; en la segunda columna se indica un valor por defecto que podemos asignar a cada variable (en caso de que decidamos no utilizar esta función bastará con asignarle el valor de uno cualquiera de los factores de ese grupo —por ejemplo, el primero—); en la tercera columna aparece el número de factores por grupo y, en las siguientes, sus respectivos valores.

Después el programa aplica los factores especificados al archivo de *tokens* y comprueba que no se hayan cometido errores en la codificación de los ejemplos (*check tokens*). Los errores, como los propios creadores del programa advierten (Robinson, Lawrence y Tagliamonte 2001b: 7), son prácticamente inevitables (figura 3), pero se pueden corregir fácilmente¹⁵:

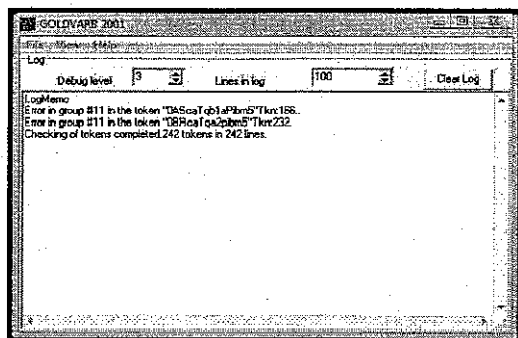


Figura 3: Comprobación de tokens

Una vez corregidos los errores, el programa está listo para llevar a cabo los análisis de frecuencias: es la fase de *creación de celdas*. Se nos preguntará primero por el *valor de aplicación* (*application value*), que equivale a pedir de nuevo los valores que hemos concretado para la variable dependiente (en nuestro caso, introduciremos 01), y a continuación se crea un archivo de resultados donde se presenta el análisis de frecuencias absolutas y relativas (porcentajes) (por motivos de espacio muestro solo algunas partes de la ventana *Results*):

The screenshot shows the 'Results' window of the GOLDVARB 2001 software, displaying absolute and relative frequencies for two groups. The data is organized into three main sections: 'Cell creation', 'Group 1 (11)', and 'Group 2 (12)'. Each section contains a table of frequencies for different factors.

Cell creation		Group 1 (11)		Group 2 (12)	
Factor	Value	Factor	Value	Factor	Value
Y	170	Y	167	Y	167
X	35	X	33	X	33
Z	35	Z	35	Z	35
W	31	W	31	W	31
Total	242	Total	242	Total	242

Figura 4: Ventana de resultados (frecuencias absolutas y relativas)

¹⁵ El programa localiza automáticamente estos errores. El error en el grupo 11 del ejemplo n.º 166 se debe a que en la especificación de factores original se olvidó introducir el valor *P* para este grupo y por tanto no es reconocido como tal factor en ningún *token*, de los que el n.º 166 es el primero en aparecer: la solución pasa por volver a la ventana de especificación de factores e introducir *P* como un valor posible. El error en el grupo 11 (y, en realidad, en todos los grupos que le siguen) del *token* n.º 232 advierte de que en la codificación del ejemplo no se introdujo ningún valor para este grupo, de modo que los siguientes ocupan una columna que no les corresponde: la solución, lógicamente, es buscar ese ejemplo en el archivo de *tokens* e insertar el valor oportuno en la columna 11.

11291	Total N				205	37	242
1151					%	84	15
0	Number of cells: 204						
	Application values: 00						
	Total no. of factors: 59						
Group	0	1	Total %				
1 (2)							
B	N	179	21	200	82		
	%	89	10				
A	N	26	16	42	17		
	%	61	38				
Total N		205	37	242			
	%	84	15				
2 (3)							
B	N	32	17	49	19		
	%	62	38				
A	N	128	12	140	58		
	%	57	12				
Line 26	Char 1	Total lines: 55		Line 26	Char 2	Total lines: 55	
15 (13)							
B	N	47	1	48	19		
	%	57	2				
A	N	40	24	64	26		
	%	52	30				
B	N	15	22	37	15		
	%	61	38				
Total N		205	37	242			
	%	84	15				
15 (14)							
B	N	110	15	125	51		
	%	58	22				
A	N	6	8	14	6		
	%	100	5	"Stockout"			
B	N	21	22	43	18		
	%	58	22				
Total N		205	37	242			
	%	84	15				
Line 26	Char 1	Total lines: 55		Line 26	Char 2	Total lines: 55	

Figura 4: Ventana de resultados (frecuencias absolutas y relativas) (continuación)

En la parte superior aparecen las condiciones que, sobre la base de los factores que hemos especificado, genera automáticamente el programa (y presenta también en un archivo independiente), las cuales, en este momento del análisis, consisten solamente en la distribución ordenada entre paréntesis del número de cada grupo de factores. A continuación se recogen los datos sobre el número total de celdas creadas (204), el valor de aplicación (01), el número total de factores contando todos los grupos (59) y, de ahí en adelante, se presentan los datos sobre la frecuencia de aparición de cada uno de los factores explicativos con respecto a la variable dependiente. Para el primer grupo de factores explicativos, por ejemplo, observamos esta tabla de resultados:

Group		0	1	Total %
1 (2)				
B	N	179	21	200 82
	%	89	10	
A	N	26	16	42 17
	%	61	38	
Total N		205	37	242
	%	84	15	

Estos números se interpretan del siguiente modo: del total de ejemplos con formas verbales no ambiguas (200), presentan la forma nula del pronombre sujeto 179 (el 89% de ellas), mientras que la forma expresa *yo* se da en 21 ejemplos (es decir, en el 10% de los casos); por el contrario, si la forma verbal se caracteriza por la ambigüedad morfológica entre la primera y la tercera persona (como en el imperfecto de indicativo y de subjuntivo, el condicional, el presente de subjuntivo, etc.), lo que ocurre en 42 ejemplos, hay un considerable 38% de los casos (16 ejemplos) donde se expresa el pronombre sujeto, frente a un 61% (26 ejemplos) donde se omite.

Algunas veces, sobre todo cuando partimos de un número de *tokens* relativamente escaso, podemos encontrar factores que se presentan solamente con una de las variantes (por ejemplo, solo en casos de omisión del pronombre). Nos detenemos en los resultados obtenidos para el grupo n.º 14, que representa la variable *raza*:

13	(14)				
b	N	110	15	125	51
	%	88	12		
n	N	4	0	4	1
	%	100	0	* KnockOut *	
m	N	91	22	113	46

Pronto nos percatamos de que el factor *n* solo se presenta en cuatro ejemplos (solo hay cuatro formas verbales de primera persona en los textos o fragmentos atribuidos a personas de raza negra) y de que, además, esos cuatro ejemplos se dan siempre con el valor de aplicación 0, lo que significa que los hablantes de raza negra omiten el pronombre sujeto en el 100% de los casos. El problema surge porque, lógicamente, el programa interpreta que el factor *n* no permite la alternancia entre los valores de aplicación 0 y 1, de modo que no constituye una verdadera variable estadística, y lo marca con un *KnockOut* que impide seguir con los cálculos necesarios para realizar el análisis probabilístico.

Existen dos formas fundamentales de resolver los *KnockOuts*. La primera es la recodificación de los factores especificados originalmente: para ello abrimos la ventana *Recode definition*, que nos permite, entre otras funciones, agrupar dos o más factores en uno solo:

Groups in token file						
Group	a	p	d	c	e	g
10	a	p	d	c	e	g
11	A	D	C	L	T	P
12	d	i				
13	a	m	b			
14	b	n	m			
15	1	2	3	4	5	

Groups after recoding				
Type	Old			
Exclude	11	A	D	C
Recode	14	b	m	m
Recode	14	b	n	m
Recode	15	1	2	1
Recode	15	1	2	3

Current operation: Idle

Click OK to write new conditions file from these recode instructions:

OK - write new conditions from recode instructions Cancel Clear all recoded groups

Figura 5: Recodificación de factores

De la parte izquierda de la ventana seleccionamos aquellos factores que queremos modificar, que se copian automáticamente en la parte derecha, repetidos en dos filas: en la inferior se muestran los valores originales y en la superior introducimos los nuevos, de modo que cada factor es la recodificación del que tiene debajo. El grupo n.º 14 consta originalmente de los factores *b*, *n* y *m* —fila inferior—, que se sustituyen por los factores nuevos *b*, *m* y *m* —fila superior—, de modo que *n* se recodifica como *m*. La recodificación ha de ser coherente con los objetivos de la investigación, dado que la interpretación de los resultados será distinta de la original, aunque utiliza los mismos datos (el archivo de *tokens* no se ha modificado): por ejemplo, al oponer el factor *b* al

factor *m* —que recoge los antiguos *n* y *m*— estamos intentando averiguar el uso del pronombre sujeto que hacen los hablantes de raza blanca frente a los demás (tanto mulatos como negros, indistintamente).

La otra opción posible para resolver un *KnockOut* se ha de llevar a cabo directamente en el archivo de condiciones. Después de aceptar las recodificaciones necesarias (*Ok-write new conditions from recode instructions*) el programa genera un archivo con las nuevas condiciones en el que podemos escribir manualmente la condición de excluir del cálculo probabilístico los resultados que ofrece un determinado factor —aquel que provoca el *KnockOut*—, lo cual, en el lenguaje de programación que emplea *GoldVarb*, se expresa por medio del signo «/»¹⁶. Así, ante los dos *KnockOuts* que, según informa la ventana de resultados, se produjeron en el grupo n.º 8 (*tiempo verbal*), correspondientes a los factores *e* y *o* —esto es, el condicional y el pluscuamperfecto de subjuntivo, a los que siempre acompaña la forma nula del pronombre sujeto—, se optó por suprimirlos del análisis¹⁷. El archivo de condiciones, tras las recodificaciones necesarias¹⁸ y la eliminación de los dos factores del grupo 8, queda como sigue:

```

(1)
(2)
(2)
(4) (a (COL 4 a))
      (m (COL 4 m))
      (w (COL 4 w))
      (x (COL 4 x))
      (t (COL 4 t))
      (n (COL 4 n))
(5) (a (COL 5 a))
      (p (COL 5 b))
      (p (COL 5 m))
      (p (COL 5 v))
      (p (COL 5 e))
      (p (COL 5 p))
      (a (COL 5 c))
      (s (COL 5 s))
      (a (COL 5 a))
(6)
(7)
(8) (/ (COL 8 e))
      (/ (COL 8 o))
(9)
(10) (a (COL 10 a))
      (a (COL 10 p))
      (a (COL 10 d))
      (a (COL 10 c))
      (a (COL 10 e))
      (a (COL 10 g))
(11) (T (COL 11 A))
      (D (COL 11 D))
      (C (COL 11 C))
      (L (COL 11 L))
      (T (COL 11 T))
      (P (COL 11 P))
(12)
(12)
(13) (b (COL 14 B))
      (m (COL 14 M))
      (m (COL 14 M))
      (1 (COL 15 1))
      (2 (COL 15 2))
Line=30 Total lines=9
  
```

Figura 6: Archivo de condiciones

¹⁶ Esta condición, por la cual el programa no tiene en cuenta un factor presente en algunos de los códigos del archivo de *tokens*, es muy diferente de la de realizar el análisis probabilístico eliminando los *tokens* que presentan ese factor en concreto —función que se podría llevar a cabo desde la ventana de recodificación—, porque de este modo eliminaríamos los códigos completos de una serie de ejemplos y podríamos alterar los resultados obtenidos para los factores de todos los grupos.

¹⁷ Dado que al establecer esta variable explicativa nuestra pretensión era atender a un criterio estructural objetivo, no pareció conveniente agrupar unos tiempos verbales con otros según ciertos criterios (el modo verbal, el aspecto, la anterioridad, etc.), por más que en fases posteriores de la investigación (reanálisis) sería interesante la agrupación.

¹⁸ Se pueden observar varias recodificaciones, aparte de la explicada arriba: en el grupo 4 se recodificó *v* como *r*; en el grupo 5 se creó un grupo general para los factores *b*, *m*, *t*, *e* y *p*, opuesto a *a*, *s* y *o*; en el grupo 11, *A* se recodificó como *T*, etc.

```

C:\Program\Veritas\CVS\src\bin\Documents\workspace\src\bin\known\bin\17200
File Edit Action ☒ Stop Close Help
Date
Run # 160, 30 cells:
Convergence at Iteration 13
Input 0.937
Group # 3 -- n: 0.189, z: 0.497, m: 0.713, c: 0.673, t: 0.157
Group # 7 -- c: 0.605, a: 0.697, f: 0.625, l: 0.129, b: 0.029, d: 0.264, g: 0.010, k: 0.265
Group #11 -- l: 0.659, d: 0.212
Log likelihood = -63.943 Significance = 0.025

Run # 161, 35 cells:
Convergence at Iteration 11
Input 0.924
Group # 3 -- n: 0.131, x: 0.532, m: 0.812, c: 0.642, t: 0.171
Group # 7 -- c: 0.615, a: 0.684, f: 0.495, l: 0.111, b: 0.049, d: 0.149, g: 0.823, k: 0.151
Group # 8 -- l: 0.514, t: 0.270
Log likelihood = -69.597 Significance = 0.000

All remaining groups significant

Groups eliminated while stepping down: 1 13 9 6 14 12 10 5 4 2
Best stepping up run: #46
Best stepping down run: #153

Line1302 Char28 TotalLines1894 Action

```

Figura 7: Final del análisis binomial de subida y bajada

En el mejor análisis de subida, por ejemplo, la probabilidad de que los resultados sean significativos (*Significance*) es de 0,037 (por debajo de la cifra de 0,05 que se ha fijado convencionalmente en las ciencias sociales como límite para que los resultados se

¹⁹ En nuestro caso, dada la variable dependiente que hemos tomado, con dos variantes, solo podemos llevar a cabo un análisis binomial (en caso de que fuera necesario realizar análisis multinomiales, habría que trabajar con la versión x de GoldVarb).

20. Primero se realiza el análisis de subida, que confronta el modelo del nivel 0, sin variables independientes, con el modelo del nivel 1 (con una variable independiente), luego el modelo del nivel 1 con el del nivel 2 (con dos variables independientes), y así sucesivamente hasta construir el modelo con todas las variables independientes codificadas. A continuación, el análisis de bajada, que efectúa un análisis por sustracción, que confronta el modelo con todas las variables independientes con el modelo del nivel inmediatamente anterior (con una variable menos) y así hasta llegar al nivel uno (Paolillo 2002: 57-70; Vietti 2005: 60-64). Los datos del mejor análisis de subida y el mejor de bajada deben coincidir en señalar como significativos los mismos grupos de factores; en caso de haber diferencias, se está indicando cierta incertidumbre con respecto a los grupos incluidos en un análisis pero excluidos en el otro (Rand y Sankoff 1990).

consideren plenamente fiables); por su parte, la función de verosimilitud (*Log likelihood*) refleja, en comparación con otros análisis, una buena adecuación de los datos al modelo. Se seleccionaron como significativos los siguientes grupos de factores:

```
Run # 46, 38 cells:
Convergence at Iteration 17
Input 0,946
Group # 2 -- N: 0,212, S: 0,464, R: 0,751
Group # 7 -- c: 0,565, a: 0,690, f: 0,644, l: 0,130, b: 0,013, d:
0,239, g: 0,004, k: 0,218
Group # 8 -- 1: 0,534, 2: 0,083
Group #11 -- i: 0,691, d: 0,168
Log likelihood = -63,271 Significance = 0,037
```

Los grupos de factores significativos son los números 2, 7, 8 y 11, que corresponden a las variables 3, 8, 9 y 12, respectivamente; los índices de probabilidad se refieren en este caso al primer valor de aplicación (0), esto es, a la omisión del pronombre sujeto. El *Input* hace referencia a la probabilidad general de aparición de esta primera variante (0,946 indica que la muestra seleccionada la favorece claramente, pero no llega a exigirla).

No deben sorprender los resultados de la variable *cambio de referencia*²¹, cuya influencia en español actual sobre la variable sujeto ha merecido detallados estudios (Cameron 1995: 1-27) y ha sido señalada como el factor de mayor influencia sobre los pronombres de primera persona (Bentivoglio 1987): en efecto, el mantenimiento de la referencia con respecto al sujeto del verbo anterior favorece el uso del pronombre nulo ($R = 0,751$) mientras que el cambio de referencia lo desfavorece levemente ($S = 0,464$) o severamente si además es la primera mención en el texto a ese sujeto ($N = 0,212$)²². La variable *tiempo verbal*, señalada por Silva-Corvalán (2003: 849-860) como la más importante para el estudio de la variable sujeto, ofrece también resultados interesantes, aunque en un primer momento su interpretación presenta mayor dificultad (inversamente proporcional, podemos decir, a la que supuso asignar los valores en cada caso: *cf.* nota 17): el presente de indicativo favorece la forma nula ($a = 0,690$), al igual que el perfecto ($f = 0,644$) y, ligeramente, el indefinido ($c = 0,565$); por el contrario, parecen favorecer el pronombre expreso los tiempos de subjuntivo —presente ($k = 0,218$) e imperfecto ($l = 0,130$)— y otros dos tiempos de pasado del indicativo —imperfecto ($b = 0,013$) y pluscuamperfecto ($g = 0,004$)—, todos ellos caracterizados por la ambigüedad morfológica entre la primera y la tercera persona, además del futuro de indicativo ($d = 0,239$)²³. La capacidad del acto de habla de transformar la realidad favorece de forma contundente el uso del pronombre expreso ($2 = 0,083$), incluyendo en este grupo no solo actos declarativos prototípicos como un nombramiento o una

²¹ Si bien los resultados del mejor análisis de bajada coinciden, casi totalmente, con los del mejor análisis de subida, en cuanto a los grupos seleccionados y los pesos relativos de los factores, el grupo 2 constituye la excepción al ser descartado como grupo significativo en favor del grupo 3, correspondiente a la variable *grado de topicalidad*, de modo que, por ejemplo, el factor *mantenimiento de tópico* favorece el pronombre nulo ($m = 0,725$) y el factor *cambio de tópico* lo desfavorece ($t = 0,129$). Esto quiere decir que ambas variables deben tenerse muy en cuenta, pero no alcanzan el grado de significatividad de las variables que fueron seleccionadas en ambos análisis.

²² Este uso corresponde en buena medida al que observé en un trabajo anterior a propósito de algunas cartas-informe/petición con carácter oficial, al que denominé «yo presentador» (Gutiérrez Maté, en prensa).

²³ El futuro, cuando se refiere a la primera persona del discurso, supone un compromiso o una implicación especial por parte del hablante que puede explicar este mayor uso del pronombre expreso (Gutiérrez Maté en prensa).

delegación de poderes, sino todos los actos que transforman un cierto discurso en una acción diplomática o documental concreta: por ejemplo, «certifico y doy fe» escrito por un notario transforma el texto anterior en un documento con validez legal, «declaro» en boca de un testigo determina que estamos ante una declaración oficial, del mismo modo que «pido» en boca de un demandante transforma su discurso en el tipo documental de la petición, tan recurrente en autos, etc. Por último, la reproducción de fragmentos en estilo directo —«transcripciones de enunciados de la inmediatez» (Cano Aguilar 1998: 219-242; Oesterreicher 2005: 729-769)— favorece el uso del pronombre ($d = 0,168$), mientras que los textos o fragmentos que no tratan de mimetizar hasta ese punto la lengua oral favorecen el pronombre nulo ($i = 0,691$).

En suma, el análisis estadístico nos permite entender los motivos que llevan a la expresión u omisión del pronombre sujeto de primera persona en la variedad lingüística contenida en los documentos cartagenos de finales del siglo XVII. De las variables explicativas que considerábamos en la fase inicial solo tres resultan verdaderamente significativas, aunque dos más presentan un grado importante de significatividad (el cambio de referencia y la topicalidad). Dada la relativa escasez de ejemplos (apenas 250), los resultados no son definitivos, pero cumplen perfectamente con la función de *estudio piloto*, y sirven para distinguir una serie de factores que habrá que tener en cuenta en todo estudio ulterior sobre la variable presencia/ausencia del *yo*. A lo largo de las páginas anteriores he insistido también en que las intuiciones del lingüista no solo están presentes en la interpretación de los resultados sino que recorren todas las fases del análisis, pues no podemos separar el desarrollo de las operaciones computacionales de los hechos lingüísticos que intentamos comprender por medio de ellas.

REFERENCIAS BIBLIOGRÁFICAS

- BARCIA, Lucía Rosado (2006): *As Formas de Tratamento em Cartas de Leitores Oitocentistas: peculiaridades do gênero e reflexos da mudança pronominal*. Dissertação de Mestrado em Língua Portuguesa, Universidade Federal do Rio de Janeiro [en línea], <<http://www.letras.ufrj.br/posvernal/mestrado/BarciaL.R.pdf>> [Consulta: 15/09/2008].
- BENTIVOGLIO, Paola (1987): *Los sujetos pronominales de primera persona en el habla de Caracas*. Caracas: Universidad Central de Venezuela.
- CAMERON, Richard (1995): «The Scope and Limits of Switch Reference as a Constraint on Pronominal Subject Expression». *Hispanic Linguistics*, 6/7, 1-27.
- CANO AGUILAR, Rafael (1998): «Presencia de lo oral en lo escrito: la transcripción de las declaraciones en documentos indios del siglo XVI». Wulf Oesterreicher, Eva Stoll y Andreas Wesch (eds.), *Competencia escrita, tradiciones discursivas y variedades lingüísticas. Aspectos del español europeo y americano en los siglos XVI y XVII*. Tübingen: Gunter Narr, 219-242.
- CARRERA DE LA RED, Micaela y Miguel GUTIÉRREZ MATÉ (2007): «Los documentos americanos en la red CHARTA», comunicación presentada en el *Col·loqui internacional de corpus diacrònics en llengües iberoromàniques*. Palma de Mallorca, 25-27 de octubre de 2007.
- CEDERGREN, Henrietta Jonas y David SANKOFF (1974): «Variable Rules: Performance as a Statistical Reflection of Competence». *Linguistics*, 50/2, 333-355.
- GIMENO MENÉNDEZ, Francisco (1998): «Grafemática y sociolingüística histórica: a propósito del *Libro de los Primitivos Privilegios de Alicante*». José Manuel Blecua Perdices, Juan Gutiérrez Cuadrado y Lidia Sala Caja (eds.), *Estudios de grafemática en el dominio hispano*. Salamanca: Universidad de Salamanca/Instituto Caro y Cuervo, 123-133.
- GUTIÉRREZ MATÉ, Miguel (en prensa): «Análisis multivariado de la presencia/ausencia del *yo* en documentos coloniales del Caribe». Concepción Company Company et al. (eds.), *Actas del VII Congreso Internacional de Historia de la Lengua Española*. Madrid: Arco/Libros.
- HÖPPER, Paul y Sandra A. THOMPSON (1980): «Transitivity in Grammar and Discourse». *Language*, 56/3, 251-299.
- LABOV, William (2006[1994]): *Principios del cambio lingüístico*. Madrid: Gredos, II [Versión española de Pedro Martín Butragueño].

- LAVANDERA, Beatriz (1978): «Where Does the Linguistic Variable Stop?». *Language in Society*, 7, 171-182.
- LÓPEZ MORALES, Humberto (ed.) (1978): *Corrientes actuales en la dialectología del Caribe hispánico*. Río Piedras: Universidad de Puerto Rico.
- LÓPEZ MORALES, Humberto (2006): «El estudio de la variación sintáctica: precisiones metodológicas». Mercedes Sedano, Adriana Bolívar y Martha Shiro (comps.), *Haciendo Lingüística. Homenaje a Paola Bentivoglio*. Caracas: Universidad Central de Venezuela, 779-795.
- MENDES MOTHÉ, Núbia Graciella (2006): «Gerúndio versus Infinitivo Gerundivo: Brasil e Portugal em contraste nos séculos XIX e XX». *Estudos Lingüísticos*, XXXV, 1554-1563.
- MORENO FERNÁNDEZ, Francisco (1994): «Sociolingüística, estadística e informática». *Lingüística (Asociación de Lingüística y Filología de América Latina)*, 6, 95-154.
- MORENO FERNÁNDEZ, Francisco (2005[1998]): *Principios de sociolingüística y sociología del lenguaje*. Barcelona: Ariel.
- OESTERREICHER, Wulf (2005): «Textos entre la inmediatez y distancia comunicativas. El problema de lo hablado escrito en el Siglo de Oro». Rafael Cano Aguilar (ed.), *Historia de la lengua española*. Barcelona: Ariel, 729-769 [2.ª edición].
- PAOLILLO, John C. (2002): *Analyzing Linguistic Variation. Statistical Models and Methods*. Stanford: CSLI Publications.
- RAND, David y David SANKOFF (1990): *GoldVarb Version 2. A Variable Rule Application for Macintosh* [en línea], <http://individual.utoronto.ca/tagliamonte/Goldvarb/GoldVarb_MANUAL.htm> [Consulta: 15/09/2008].
- ROBINSON, John, Helen LAWRENCE y Sali TAGLIAMONTE (2001a): *GoldVarb 2001: A Multivariate Analysis Application for Windows* [en línea], <http://www.crm.umontreal.ca/~sankoff/GoldVarb_Eng.html> [Consulta: 15/09/2008].
- ROBINSON, John, Helen LAWRENCE y Sali TAGLIAMONTE (2001b): *GoldVarb 2001. User's Manual* [en línea], <<http://courses.essex.ac.uk/Ig/Ig654/GoldVarb2001forPCmanual.htm>> [Consulta: 15/09/2008].
- SANKOFF, David, Sali TAGLIAMONTE y Eric SMITH (2005): *GoldVarb X: A Multivariate Analysis Application for Macintosh and Windows* [en línea], <http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm> [Consulta: 15/09/2008].
- SANTOS LOPES, Célia Regina dos (2003): *A inserção de a gente no quadro pronominal do português*. Frankfurt/Madrid: Vervuert/Iberoamericana.
- SILVA-CORVALÁN, Carmen (2001): *Sociolingüística y pragmática del español*. Washington D.C.: Georgetown University Press.
- SILVA-CORVALÁN, Carmen (2003): «Otra mirada a la expresión del sujeto como variable sintáctica». Francisco Moreno Fernández et al. (coords.), *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*. Madrid: Arco/Libros, II, 849-860.
- VIETTI, Alessandro (2005): «Approcci quantitativi all'analisi della variazione linguistica: il caso di *GoldVarb 2001*». *Linguistica e Filologia*, 20, 31-69.