Análisis estadístico de datos textuales. Aplicación al estudio de las declaraciones del Libertador Simón Bolívar

ZULAIMA OSUNA
MARÍA PURIFICACIÓN GALINDO VILLARDÓN
JAVIER MARTÍN VALLEJO
UNIVERSIDAD DE SALAMANCA

RESUMEN. En este trabajo se presenta una estrategia metodológica basada en el Análisis Estadístico de Datos Textuales de corpus cronológicos. Se emplean distintas herramientas tales como: los métodos estadísticos multidimensionales descriptivos y métodos derivados de la estadística léxica. La utilización del Análisis Estadístico de Datos Textuales permite, entre otras posibilidades, una lectura de los textos sustancialmente distinta y complementaria a las lecturas realizadas desde un enfoque más cualitativo. Consideramos que esta aplicación es novedosa ya que la misma ha sido escasamente empleada para tratar declaraciones e inferir las concepciones mentales de los sujetos y analizar cambios evolutivos.

PALABRAS CLAVE: análisis estadístico de datos textuales, estadística léxica, lematización.

RESUMO. Neste trabalho se apresenta uma estratégia metodológica baseada na Análise Estatística de Dados Textuais de corpus cronológicos. Utiliza-se distintas ferramentas, tais como: os métodos estatísticos multidimensionais descritivos e métodos derivados da estatística léxica. A utilização da Análise Estatística de Dados Textuais permite, entre outras possibilidades, uma leitura dos textos substancialmente distinta e complementaria às leituras realizadas a partir de um enfoque mais qualitativo. Consideramos que esta aplicação é uma novidade já que a mesma foi escasamente utilizada para tratar declarações e inferir as concepções mentais dos sujeitos e analisar mudanças evolutivas.

PALAVRAS CHAVE: análise estatístico de dados textuais, estatística léxica, lematização.

ABSTRACT. In this investigation a methodological strategy based on the analysis of textual data of chronological corpus is presented. Diverse instruments related to this analysis are used, such as methods of descriptive multidimensional statistics and methods derived from lexical statistics. The statistical analysis of textual data allows, among other things, a reading of the data substantially different and complementary to the readings carried out from most qualitative approaches. We consider that this application is original since it has been rarely applied to deal with declarations and to infer mental conceptions of individuals as well as to analyze changes progressively.

KEY WORDS: statistical analysis textual data, lexical statistics, lemmatisation

Introducción

El análisis de lo escrito ha motivado un interés muy particular en el seno de las diversas disciplinas de las ciencias humanas. Estos análisis se llevan a cabo desde distintas perspectivas y puntos de vistas. En este trabajo abogamos por el uso de las técnicas multivariantes de análisis de datos textuales que consideran la complejidad del objeto de estudio en sus múltiples determinaciones e interdependencias.

Las técnicas de Análisis Estadístico de Datos Textuales (AEDT), a partir de los modelos desarrollados por Lebart y Salem (1994), basadas en análisis cuantitativo cuyo punto de partida básico es el recuento de ciertas unidades textuales definidas previamente, y que se fundamentan principalmente en el análisis a través de la comparación, son aplicadas en este trabajo al procesamiento de datos textuales procedentes de las principales declaraciones del Libertador Simón Bolívar.

Objetivos y métodos

El objetivo principal es presentar una estrategia estadística para el análisis de datos textuales. Los objetivos específicos son los siguientes:

- Determinar las ideas "nucleares" o temáticas que se derivan o infieren del conjunto de las declaraciones del Libertador Simón Bolívar.
- Determinar cuáles declaraciones presentan un perfil léxico similar.
- Estudiar la evolución del vocabulario y organización de las declaraciones a lo largo de los años 1812-1826.

Se aplica la técnica fundamental del AEDT, la cual es el Análisis Factorial de Correspondencia de tablas léxicas. El Análisis Factorial de Correspondencias (AFC) fue propuesto por Benzècri en 1973 como un método para explorar tablas de contingencias multidimensionales (Tablas Léxicas), en el cual se enfatiza la potencialidad de las representacio-

nes gráficas. La clasificación del AFC como técnica exploratoria se debe a que, para su utilización no se requiere supuestos acerca de la distribución subyacente en los datos.

El AFC es una técnica apropiada para investigar la magnitud y naturaleza de las inter-asociaciones entre palabras y declaraciones. Para ello se construyen los subespacios de dimensión reducida que mejor se ajustan, en el sentido de los mínimos cuadrados, a la nube que se desea describir. En estos subespacios es posible interpretar las distancias entre las proyecciones de las palabras sobre los planos, como medida de su semejanza en cuanto a la frecuencia con que aparece en cada una de las declaraciones. Por su parte, las proximidades entre las declaraciones representan afinidades entre ellas, es decir, si dos declaraciones aparecen muy cercanas en proyección, probablemente se deba a que están caracterizadas por las mismas palabras. Las distancias entre palabras y declaraciones sólo deben interpretarse en términos baricéntricos.

Para el procesamiento de los datos textuales se utilizaron los programas computacionales LEXICO 3 y ADE-4.

Corpus analizado

El corpus analizado lo constituyó una base de datos textuales creada a partir de las algunas declaraciones pronunciadas después de la proclamación de la independencia de Venezuela (5 julio 1811) por el Libertador Simón Bolívar entre los años 1812 y 1826 (Tabla N° 1). El corpus tiene una extensión o longitud de 28870 ocurrencias y está formado por 8783 formas gráficas distintas o palabras propias. El porcentaje de diversidad es 30,42%.

Como fase previa a los análisis multivariantes, el corpus fue intervenido a objeto de eliminar palabras que, desde el punto de vista estadístico no aportan información adicional a los mismos. Las modificaciones fueron:

- Exclusión de las preposiciones y conjunciones
- Exclusión de pronombres personales y demostrativos
- Exclusión de grafías romanas, números y abreviaturas
- Exclusión de palabras que aparecen una sola vez (Hápax)
- Las letras mayúsculas, sólo se han mantenido en el caso de los nombres propios, en el resto de contextos tipográficos no se distingue entre letras mayúsculas y minúsculas.

A los palabras restantes se les aplicó un nivel moderado de lematización (Bolasco, 1993), es decir se han reagrupado algunas flexiones del mismo lema para «reducir el número de variaciones no significativas», llevando las palabras de alta frecuencia de ocurrencia y con significado relevante a sus formas canónicas (verbos al infinitivo, sustantivos en singular y adjetivos en masculino singular).

Tabla N° 1. Declaraciones analizadas

Año	*Declaración*	*Extensión*	*P. Propias*	*Hápax*
1812	Memoria dirigida a los ciudadanos de la Nueva Granada por un caraqueño	5765	2006	1453
1813	Decreto de Guerra Muerte	667	326	253
1815	Carta de Jamaica	7932	2411	1678
1819	Discurso de Angostura	10304	2636	1731
1826	Mensaje al Congreso Constituy. de Bolivia	4202	1404	991

Resultados

A fin de ver si el léxico de las declaraciones del Libertador es diferenciable, se trabajó con la Tabla Léxica formada por las cinco declaraciones y 476 palabras distintas, ya que el Análisis Factorial de Correspondencias permite establecer tipologías en términos de semejanza y disimilitud, entre las partes de un corpus en función de su frecuencia. El umbral de frecuencia seleccionado fue seis, es decir en la tabla léxica sólo esta constituida por aquellas palabras cuya frecuencia de ocurrencia en todo el corpus es mayor o igual a seis. En el análisis de esta tabla se observó que la información se distribuye en cuatro ejes, en los que pueden observarse las tendencias en el comportamiento de las palabras. Los valores propios, los porcentajes de inercia y la inercia acumulada de cada eje principal asociado a cada valor propio correspondiente al AFC de la Tabla Léxica TL (476x5) se muestran en la Tabla N° 2 y estos miden la inercia captada o varianza sobre cada uno de los ejes principales.

Tabla N° 2. Valores propios y porcentaje de inercia

Valor Propio	*% Inercia*	*Inercia acumulada*
0,17	35,75	35,75
0,12	25,00	60,75
0,10	21,98	82,73
0,08	17,27	100,00

En este caso se analizará sólo el plano factorial generado por lo dos primeros valores propios el cual recoge o explica el 60,75% de la variabilidad de los datos, los elementos de partida serán la información proporcionada por las contribuciones absolutas y relativas (del Elemento al Factor).

Este análisis nos permitió corroborar la existencia de un léxico diferenciado por declaraciones. Más específicamente, se observó que el

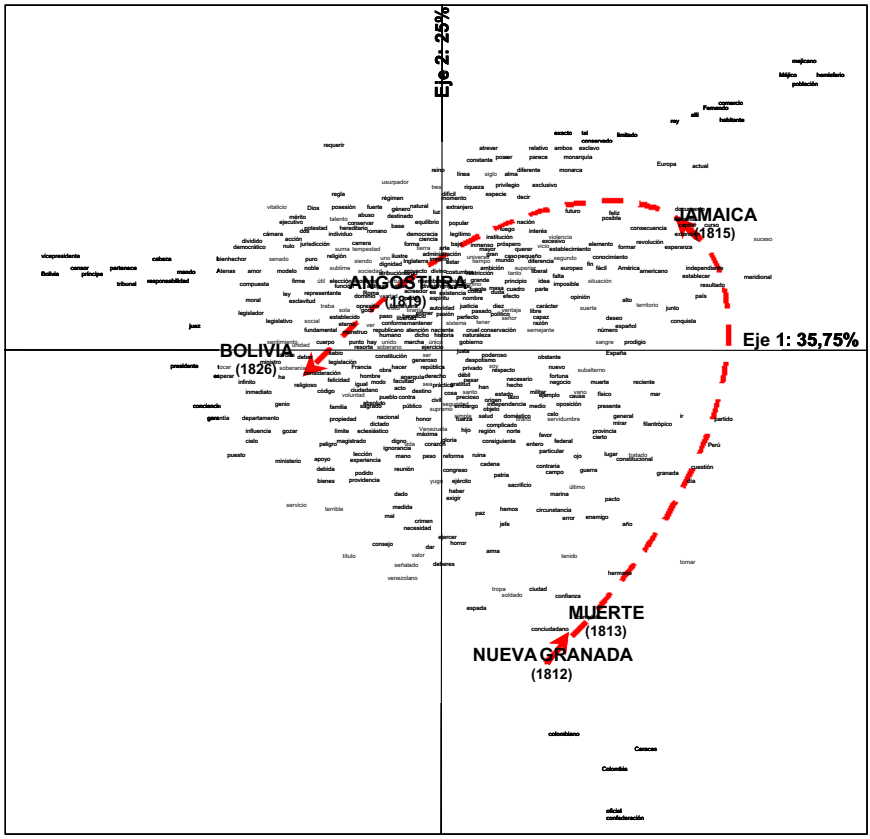

Figura 1. Análisis Factorial de Correspondencia

léxico de la Carta de Jamaica se contrapone con el léxico utilizado en el Decreto de Guerra a Muerte y la Memoria dirigida a los ciudadanos de la Nueva Granada por un caraqueño (Ver Figura 1). De igual forma se observa en esta figura el ordenamiento cronológico de las cinco declaraciones, lo cual pone de manifiesto la estructura evolutiva del vocabulario empleado; es decir confirma la existencia de un tiempo léxico en la práctica discursiva.

Este análisis nos permitió corroborar la existencia de un léxico diferenciado por declaraciones. Más específicamente, se observó que el léxico de la Carta de Jamaica se contrapone con el léxico utilizado en el Decreto de Guerra a Muerte y la Memoria dirigida a los ciudadanos de la Nueva Granada por un caraqueño (Ver Figura 1). De igual forma se observa en esta figura el ordenamiento cronológico de las cinco declaraciones, lo cual pone de manifiesto la estructura evolutiva del vocabulario empleado; es decir confirma la existencia de un tiempo léxico en la práctica discursiva.

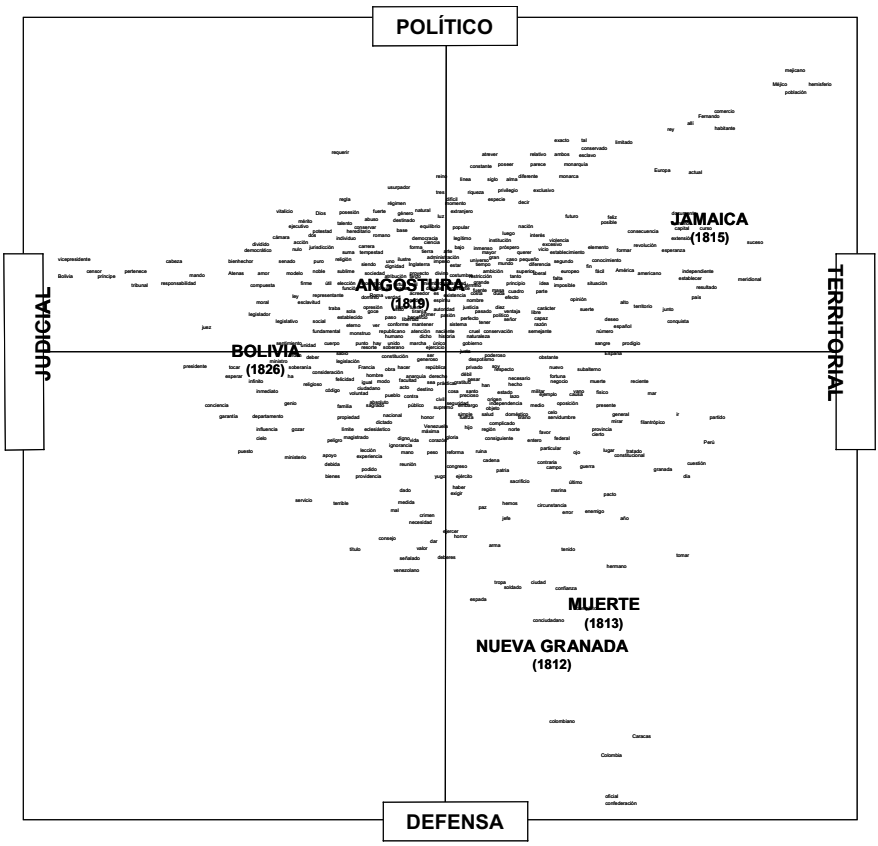

Figura 2. Plano Factorial. Temáticas asociadas a las declaraciones

El reconocimiento de las temáticas asociadas a cada declaración (Ver Figura 2), permite establecer los significados de los ejes y reconstruir los polos semánticos. De la lectura del plano factorial conjuntamente con las contribuciones relativas determinamos que las palabras exclusivas del Factor 1 o Eje 1 son: *país, independencia, capital, extensión, conquista, territorio, mando, juez, garantía, moral legislador, ministro, cámara, legislativo, ley, jurisdicción* entre otras, lo cual pone de manifiesto dos polos semánticos asociados con las declaraciones, los cuales son: al lado derecho o positivo del eje, aspectos relacionados con lo territorial, y la lado izquierdo o negativo aspectos relacionados con el poder judicial. De forma análoga para el Factor 2 o Eje 2 las palabras exclusivas son: *monarquía, reino, régimen, monarca, democracia, soldado, arma, honor, jefe, hermano*, entre otras cuyos polos semánticos asociados son los aspectos relacionados con lo político y la defensa correspondiente a la parte superior e inferior del plano factorial respectivamente.

Conclusiones

1. La segmentación del corpus en períodos no solamente permite dar cuenta de las temáticas abordadas, sino también encontrar el ritmo del discurso dentro de un contexto definido.

2. El análisis directo de las declaraciones del Libertador Simón Bolívar funciona mejor a partir del corpus lematizado. El reagrupamiento de las palabras en lemas reduce el número de unidades y como consecuencia aumenta sus frecuencias. Por esta razón, el número de palabras compartidas por dos a más declaraciones aumenta, lo que permite dar mayor sentido a las distancias calculadas entre ellos.

3. La lematización influye en la selección de las palabras conservadas. El caso particular de los verbos, cuyas inflexiones son numerosas y mucho más dispersas que la de los sustantivos, adjetivos y adverbios, es eliminada en mayor cantidad por el proceso de filtrado mediante el umbral de frecuencia. En efecto una vez agrupadas las diferentes inflexiones de un mismo verbo bajo la forma infinitiva, aumenta la posibilidad de que la frecuencia global supere el umbral seleccionado.

REFERENCIAS BIBLIOGRAFICAS

BOLASCO, S. (1993). Choix de lemmatisation en vue des reconstructions syntagmatiques du texye par l'analyses des correspondence. *Secondes Journées Internationales d'Analyses des Données Textuelles*. Ottobre 21-22 Montpellier.

BENZÉCRI, J. P. (1973). *L'Analyse Des Dones. II L'Analyse Des Correspondances.* París: Dunod

LEBART, L. & SALEM, A. (1994). *Statistique Textuelle*. París: Dunod.

ZULAIMA OSUNA. Doctoranda de Estadística Multivariante Aplicada en la Universidad de Salamanca. Licenciada en Ciencias Estadística por la Universidad Central de Venezuela (1996). Estadístico de la Oficina de planificación del Sector Universitario. Línea de investigación Análisis de Datos Textuales.
Correo-e: zulosu@cantv.net

MARIA PURIFICACION GALINDO. Doctora en Estadística Multivariante por la Universidad de Salamanca (1985), es profesora titular e investigadora de la Universidad de Salamanca desde 1986. Directora del Departamento desde 1991. Coordinadora del Programa de Doctorado en Estadística Multivariante Aplicada desde 1991. Profesora visitante en 12 universidades Europeas y Americanas. Directora de 15 tesis doctorales, 8 tesinas de licenciatura. Ha publicado 15 libros y/o monográficas científicas, más de 70 artículos en revistas, más de 150 comunicaciones en congresos nacionales e internacionales, ha impartido 49 cursos de postgrado

invitada por universidades europeas y americanas, ha coordinado varias reuniones científicas nacionales e internacionales.
Correo-e: pgalindo@usal.es

JAVIER MARTIN VALLEJO. Doctor en Estadística Multivariante por la Universidad de Salamanca (1995), Master en Psicología por la Oxford Brookes University (1997). Profesor titular e investigador de la Universidad de Salamanca. Su investigación esta dirigida a la aplicación de las a diferentes técnicas multivariantes (especialmente los métodos Biplot y Análisis de Correspondencias) a distintos campos así como la utilización de estas técnicas en Meta-análisis.
Correo-e: jmv@usal.es