UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Departament d'Estadística i Investigació Operativa

ANÁLISIS ESTADÍSTICO DE CORPUS CRONOLÓGICOS.

Aplicación al estudio de bases bibliográficas y textos retóricos

Tesis para optar por el grado de
Doctor en Estadística e Investigación Operativa

Autora:
Daría Micaela Hernández Ramírez

Directora:
Mónica Bécue-Bertaut

Barcelona, España - 19 de Diciembre 2016

A mi padre (†).

Agradecimientos

A todas las personas que hicieron posible la realización de esta tesis.

A mi directora de tesis, Dr. Mónica Bécue-Bertaut, por el tiempo dedicado en la dirección y supervisión de la tesis. Mi gratitud y estima para el Dr. Belchin Kostov porque a lo largo de estos años su paciencia, consejos y apoyo en todo momento, fueron imprescindibles para la culminación de este proceso.

Mi más sentido agradecimiento a mis compañeros y maestros de AC, en especial al Ing. Aquiles Córdova Morán y al Dr. Abel Pérez Zamorano porque sin su ejemplo e impulso no hubiera llegado a este momento. A todos mis compañeros y amigos del cemees, principalmente a Vania y Guy, por su exigencia, comprensión y cariño.

Aprecio y valoro el apoyo recibido de mis amigos el Dr. Carlos Pérez Santos y la Dra. Gabriela Zayas De Lille, con quienes compartí gratos momentos, me brindaron siempre ayuda y compartieron conmigo sus experiencias.

Gracias a mis compañeros del Departamento de Estadística e Investigación Operativa de la UPC: Diana, Nihan, Cristina, Jesús, Hajar, Jessica y Vicky. Por su compañía y agradables momentos que compartido juntos.

Agradezco infinitamente a los que siempre estuvieron presentes y son una razón muy importante de mi vida, mis padres Raquel y Camerino. A mis hermanos por todo su apoyo.

Resumen

Debido a la gran cantidad de datos textuales que se generan constantemente, los investigadores se enfrentan con la necesidad de clasificarlos y analizarlos, aunque existen diferentes técnicas y herramientas computacionales para facilitar su estudio. En esta tesis se proporciona un procedimiento metodológico, así como su herramienta computacional para el análisis de corpus cronológicos. Nuestro interés se centra en modelizar la estructura del corpus y clarificar el flujo de su vocabulario.

La metodología propuesta continúa con la linea metodológica desarrollada por Bécue-Bertaut (2014) la cual combina los métodos multidimensionales clásicos para el análisis de datos con los métodos para el estudio de la estructura y la evolución de los corpus. Para modelizar la estructura del corpus y clarificar el flujo de su vocabulario, el corpus se segmenta en tres partes, de acuerdo a las funciones que desempeñan las palabras: vocabulario especializado o local, que es inducido por el tema tratado, pero que también marca la estrategia evolutiva del corpus; vocabulario estable, conformado por las palabras utilizadas de forma regular a lo largo del corpus y, vocabulario aleatorio, formado por las palabras herramientas en general, como preposiciones y determinantes. En la descomposición del vocabulario según las funciones de las palabras, proponemos una metodología que combina el índice de reparto del vocabulario, las palabras características cronológicas y una prueba de bondad de ajuste para la distribución de Poisson. Después se analiza el vocabulario especializado y la metodología propuesta sigue la idea desarrollada por Benzécri (1973, 1981), implementada en el método de Análsis de una Matriz de Datos (AMADO), e incorpora un nuevo procedimiento que consiste en: primero, ordenar todas las palabras especializadas: a) por sus coordenadas sobre la primera dimensión de un Análisis de Correspondencia (AC) y b) de acuerdo con el documento o segmento de documentos caracterizados por las palabras características cronológicas; segundo, visualizar el vocabulario que determina la evolución a través de los gráficos de Bertin y, tercero, mostrar la estructura del modelo o esquema de evolución cronológica mediante AC. Los resultados que se obtienen muestran las ventajas que ofrece el análisis de los datos a través de un enfoque cronológico al responder a preguntas como: ¿Cuáles son los temas más relevantes? ¿Existe evolución en el vocabulario? ¿Qué es lo que determina su evolución? ¿El corpus está bien organizado? ¿Existe diversidad temática? ¿Qué papel desempeña cada una de las palabras según su función? ¿Cuáles son las palabras que permiten evolucionar al corpus?

Los resultados se muestran mediante el análisis de una base bibliográfica y de un texto retórico. La metodología fue implementada en un conjunto de funciones programadas en R y puede ser aplicada a cualquier tipo de corpus.

Abstract

Due to the huge amount of textual data that is persintently generated, researchers are obliged to classify and analize them, even though there are different ways as well as computing tools to facilitate their study. In this thesis a procedure method as well as its computing tools are given in order to analyze chronological corpus. Our goal is focused mainly in analysing the corpus structure and clarifying the vocabulary flow.

The proposed statistical methodology follows the one described by Bécue-Bertaut (2014) which allows the combination of classic multidimensional methods for data analysis with those that study the structure and evolution of corpus. In order to achieve this goal, the corpus is divided in three parts, according to the function of the words: specialized or local vocabulary, according to the addressed issue, which marks the evolutive corpus strategy; steady vocabulary, including those words used on a regular basis throughout the corpus and, random vocabulary, including those tool words like prepositions, conjunctions and so on. According to the words functions we suggest a methodology that combines the vocabulary index, the chronological characteristic words and a goodness of fit test for a Poisson distribution. After specialized vocabulary is analyzed and proposed methodology follows the idea developed by Benzécri (1973, 1981), implemented in the method of Analysis of a Data Matrix (AMADO), and incorporates a new procedure consisting of: first, order all specialized words: a) by theirs coordinates on the first dimension of a Correspondence Analysis (CA) and b) according to the document or segment documents characterized by the chronological characteristics words; second, display the vocabulary that determines the evolution through Bertin's Grafics and third, show the model structure or chronological evolution scheme by AC. The results obtained show the advantages of the analysis of data through a chronological approach to answer questions such as: What are the most important issues? Is there evolution in the vocabulary? What determines its evolution? The corpus is well organized? Is there thematic diversity? What role do each of the words according to their function? What are the words that allow evolve the corpus?

The results are shown by analyzing a bibliographic base and a rhetorical text. The methodology was implemented in a set of functions programmed in R and can be applied to any type of corpus.

Índice

Resumen v

Abstract vii

1. Introducción 1

1.1. Motivación 1

1.2. Estructura de la tesis 4

2. Métodos multidimensionales clásicos en análisis textual 5

2.1. Análisis Factorial General 5

2.1.1. Notación y principales aspectos 5

2.1.2. Esquema de dualidad del Análisis Factorial General 6

2.1.3. Herramientas e indicadores para la interpretación 7

2.2. Análisis de correspondencias 7

2.2.1. Equivalencia distribucional y distancia chi-cuadrado 8

2.2.2. Relaciones de transición 9

2.2.3. Herramientas para la interpretación 9

2.3. Análisis Factorial Múltiple 11

2.3.1. Tabla múltiple 11

2.3.2. El AFM como un AFG 11

2.3.3. Equilibrar los conjuntos de variables 11

2.3.4. Representación superpuesta de las l nubes de individuos 12

2.4. Clasificación 12

2.4.1. Clasificación jerárquica 12

3. Métodos para el análisis de corpus cronológicos **15**

3.1. Análisis factorial múltiple de tablas de contingencias 15

3.1.1. Notación 16

3.1.2. Algoritmo 16

3.1.3. Tipos de resultados 17

3.1.4. Validación 17

3.2. Clasificación 18

3.2.1. Clasificación jerárquica con restricción de contigüidad 18

3.2.2. Clasificación cronológica 18

3.3. Caracterización léxica de los períodos 19

3.3.1. Palabras características 20

3.3.2. Incrementos específicos 20

4. Métodos para determinar las funciones de las palabras **23**

4.1. Funciones de las palabras 23

4.1.1. Índice del reparto del vocabulario 24

4.1.2. Palabras características cronológicas 25

4.1.3. Criterios para dividir las palabras según su función 26

4.2. Modelo de evolución cronológica 26

4.2.1. Matriz reordenada de Bertin 26

4.2.2. Análisis de una matriz de datos (AMADO) 27

4.2.3. AC para el estudio de trayectorias 27

4.2.4. Matriz ordenada: implementación en R 29

5. Análisis cronológico de una base bibliográfica **31**

5.1. Introducción 31

5.2. Corpus Bibliográfico 31

5.2.1. Obtención de los datos 31

5.2.2. Características del corpus 34

5.2.3. Identificación de temas mediante el glosario 35

5.3. AC: Metallaves y Metadocumentos 36

5.3.1. Identificación de temas mediante Metallaves 36
5.4. Evolución cronológica del vocabulario 39
5.4.1. Importancia del AFMTC en la cronología 39
5.4.2. Relación con la cronología: AC y AFMTC 39
5.4.3. AFMTC : análisis globlal y parcial de la tabla múltiple 39
5.4.4. Clasificación cronológica: períodos homogéneos 41
5.4.5. Clasificación con restricción de contigüidad: Estructura y Palabras jerárquicas 43
5.5. Artículos pioneros 46
5.6. Conclusiones 47

6. Análisis cronológico de un texto retórico 49
6.1. Introducción 49
6.2. Texto no estructurado 50
6.2.1. División del texto en frases 51
6.2.2. Agrupación de las frases en partes homogéneas 51
6.3. Análisis cronológico 51
6.3.1. Forma y trayectoria de las partes 51
6.3.2. Estructura jerárquica de las partes 52
6.3.3. Segmentación cronológica y flujo de la argumentación 53
6.4. Conclusiones 56

7. Funciones de las palabras 57
7.1. Introducción 57
7.2. Funciones de las palabras 57
7.2.1. Reparto del vocabulario 57
7.2.2. Palabras cronólogicas 59
7.2.3. Funciones de las palabras 60
7.3. Representación y trayectoria del vocabulario cronológico 64
7.3.1. Diagonal ordenada 64
7.3.2. Trayectoria: efecto Guttman 64

7.4. Funciones de las palabras en una base bibliográfica 64
7.5. Conclusiones 68

8. Funciones en R para el análisis estadístico de textos 71

8.1. Introducción 71
8.2. Funciones en R 72
8.2.1. Relaciones entre funciones 73
8.3. MacroBiblio 75
8.3.1. Argumentos y valores 75
8.3.2. Aplicación 76
8.3.3. Resultados 77
8.3.4. Gráficas 81
8.4. MacroTxChrono 85
8.4.1. Argumentos y valores 85
8.4.2. Aplicación 87
8.4.3. Resultados 87
8.4.4. Gráficas 90
8.5. Conclusiones 91

9. Conclusiones 93

Manual de funciones en R 101

Lista de figuras

1.1. Tabla múltiple 2

2.1. Metallaves/Metadocumentos 10

3.1. Tabla de proporciones múltiple **P** y márgenes 16

4.1. Descomposición del corpus 23

4.2. Matriz reordenada de Bertin 28

4.3. El efecto Guttman 29

5.1. Diagrama de flujo 32

5.2. Artículo por año 34

5.3. Artículos por revista 34

5.4. Matriz abstracts×palabras 34

5.5. Representación de las palabras y abstracts con mayor contribución en el primer plano principal de AC, asociados a un tema 37

5.6. Representación global de los abstracts y de un extracto de las palabras con mayor contribución en el primer plano principal del MFACT. 40

5.7. Representación global de las palabras y años-categórica según sus coordenadas en el primer eje de AFMTC. 41

5.8. La representación parcial de la columna suplementaria años-categórica sólo desde el punto de vista del vocabulario en el primer plano principal del AFMTC 42

5.9. Palabras características e incrementos léxicos específicos de los períodos homogéneos 43

5.10. Evolución del vocabulario a través del árbol etiquetado 44

5.11. Representacion global de un extracto de palabras (investigación y medicamentos) y períodos homogéneos 45

5.12. Representación de los abstracts de 2005-2010 adelantados a su tiempo en el primer plano principal de AFMTC. 46

6.1. Diagrama de flujo 50

6.2. Trayectoria del discurso en el primer plano factorial de AC 53

6.3. Estructura jerárquica de las partes del discurso a través de clasificación con restricción de contigüidad temporal 54

6.4. Seguimiento del flujo del vocabulario a través del árbol jerárquico. Cada palabra es localizada en la parte que mejor caracteriza. 55

7.1. Descomposición del discurso de Badinter 60

7.2. Descomposición del vocabulario especializado del discurso de Badinter . . . 64

7.3. Modelo de evolución cronológica del discurso de Badinter 65

7.4. Descomposición del vocabulario especializado 66

7.5. Modelo de evolución cronológica de la base de LES 68

8.1. Diagramas de relación 73

8.2. Diagramas de relación 74

8.3. Resultados en el primer plano de AC 82

8.4. Clasificación jerárquica con restricción de contigüidad 83

8.5. Resultados de un AFMTC 84

8.6. Resultados en el primer plano de AC 90

8.7. Clasificación jerárquica con restricción de contigüidad 91

Lista de tablas

2.1. Esquema general de AFG 6

4.1. Criterios para dividir el corpus 26

5.1. Criterios de selección 33

5.2. Formato de un abstract bajado de MEDLINE 33

5.3. Glosario 35

5.4. Temas definidos mediante el glosario 36

5.5. Metallaves/Metadocumentos (palabras y abstracts con mayor contribución) 38

5.6. Proporción de la inercia de la nube de palabras explicada por los dos primeros ejes del AC y AFMTC 40

5.7. Extracto de los abstracts "adelantados a la fecha" de 2005 a 2010 47

6.1. Porcentaje de inercia explicada por los ejes. 52

6.2. Palabras características e incrementos específicos de los períodos 54

7.1. Índice de reparto del vocabulario del discurso de Badinter 58

7.2. Palabras cronológicas de las partes 59

7.3. Principales características del reparto del vocabulario del discurso de Badinter en función de sus categorías gramaticales 61

7.4. Vocabulario Regular o estable del discurso de Badinter 62

7.5. Vocabulario aleatorio o usual del discurso de Badinter 63

7.6. Vocabulario local o especializado del discurso de Badinter 63

7.7. Índice de reparto del vocabulario de la base de LES 66

7.8. Vocabulario cronológico de la base bibliográfica LES 67

8.1. Funciones en R 72

1. Índice 101

Capítulo 1

Introducción

1.1 Motivación

Los investigadores se enfrentan constantemente con la necesidad de estudiar, analizar y clasificar grandes volúmenes de datos textuales. De una manera más o menos artesanal, seleccionan la información que consideran relevante para su investigación. Este proceso, a medida que va generando más información sobre diversos temas, se vuelve más complejo y conlleva un esfuerzo considerable por parte de los investigadores.

Los métodos estadísticos multidimensionales constituyen una herramienta importante para el tratamiento de textos. Estos métodos han surgido como resultado del estudio cuantitativo de los textos literarios, por una parte, y de las aportaciones metodológicas desarrolladas por la Escuela Francesa de Análisis de Datos, por otra. Además, los avances en la ciencia de la computación y el desarrollo de sistemas informáticos, desarrollados con fines específicos, hacen posible la manipulación y el análisis estadístico de grandes volúmenes de información textual.

Desde los años cincuenta se han propuesto metodologías y herramientas computacionales que incluyen técnicas estadísticas para agilizar y potenciar el estudio de textos mediante procesos automatizados. Esta tesis tiene como finalidad proporcionar un procedimiento metodológico y su herramienta computacional para facilitar el análisis de corpus cronológicos. El objetivo principal es modelizar la estructura de un corpus y clarificar el flujo de su vocabulario.

Las razones principales que motivaron el desarrollo de esta tesis fueron:

- Proporcionar una herramienta más completa para el análisis de textos que incorpore métodos estadísticos ya existentes y de gran utilidad en el análisis de corpus cronológicos tales como:
  - Análisis Factorial Múltiple de Tablas de Contingencia (AFMTC; Bécue-Bertaut y Pagès, 2004, 2008)
  - Clasificación Cronológica (CC; Legendre y Legendre, 1998)
  - Clasificación con Restricción de Contigüidad (CCC; Legendre y Legendre, 1998)
  - Palabras Jerárquicas (PJ; Bécue-Bertaut et al., 2014)
  - Palabras Características Cronológicas (PCC; Lebart et al., 2000, 1998)
  - Crecimiento Específico del Vocabulario (CEV; Lebart et al., 1998)

- Indice de Reparto del Vocabulario (IV; Hubert y Labbé, 1990a,b),

que no se han implementado en ninguno de los paquetes que facilitan el anásis de datos textuales y que están disponibles en el CRAN de R, tales como: *Text Mining* (Feinerer, 2008; Feinerer y Hornik, 2012), *koRpus*(Michalke, 2014) y *textometry* (Loiseau et al., 2014) y *RcmdrPlugin.temis* (Bouchet-Valat y Bastin, 2013).

- Implementar un procedimiento metodológico para el análisis de textos, desde un punto de vista cronológico, mediante el desarrollo de macro funciones que combinen diferentes métodos, faciliten el análisis y proporcionen información relevante del corpus estudiado.

  - MacroBiblio: análisis de bibliografia científica
  - MacroCaChcpc: análisis de encuestas con preguntas abiertas
  - MacroTxChrono: análisis de textos no estructurados

La metodología que proponemos parte de la metodología propuesta por Bécue-Bertaut (2014) que combina los métodos multidimensionales clásicos para el análisis de datos y los métodos para el estudio de la estructura y evolución de los corpus cronológicos, de tal manera que se ofrece la posibilidad de introducir el factor tiempo (cronología) al análisis, para establecer criterios de medición que faciliten la búsqueda de los temas más relevantes del campo de estudio, la definición de las funciones de las palabras según su uso, el flujo del vocabulario y su relevancia a través del tiempo.

Para el desarrollo de la metodología propuesta, el primer factor a considerar en el análisis de corpus cronológicos es la codificación de los datos. El corpus se codifica en una tabla de frecuencias documentos×palabras, comúnmente conocida como tabla léxica para el análisis textual (Lebart et al., 1998). Las variables contextuales relacionadas con la tabla léxica se codifican en una tabla Documentos×variables-contextuales. Estas últimas juegan un papel esencial porque permiten hacer análisis agregados de la tabla léxica. La integración de ambas tablas constituye una tabla múltiple (Figura 1.1).

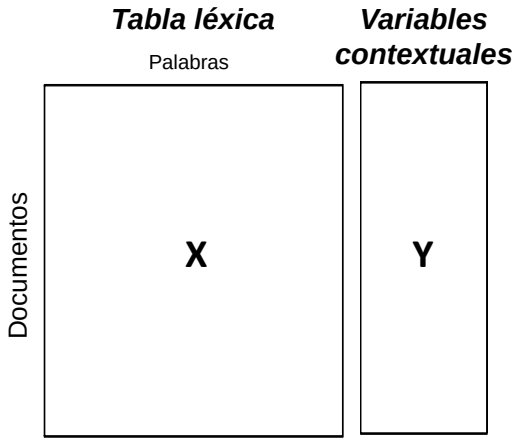


Figura 1.1: Tabla múltiple

La estructura de la tabla múltiple permite la implementación de los diferentes métodos, de acuerdo con el objetivo que se persigue. Por ejemplo, para la obtención de los temas más relevantes, el Análisis de Correspondencias (AC; Benzécri, 1973, 1981; Lebart et al., 1998; Murtagh, 2005) es una herramienta esencial en la organización de la tabla de frecuencias documentos×palabras; proporciona un resumen de las similitudes entre los documentos y una descripción de la asociación entre palabras, de tal manera que, las palabras que están íntimamente relacionadas pueden usarse, frecuentemente, en los mismos documentos y, todos estos documentos juntos, dado su grado de homogeneidad en el uso del vocabulario, abordan un tema específico (Benzécri, 1981). En este sentido, el hecho de identificar las palabras que definen cada tema y los documentos relacionados con cada uno de ellos, hace del AC una herramienta útil en el análisis de corpus con una estructura similar a la expuesta anteriormente (Bansard et al., 2007; Kerbaol et al., 2006; Kerbaol y Bansard, 2000; Morin, 2004, 2006; Rouillier et al., 2002; Šilić et al., 2012).

Para seguir la evolución del vocabulario en el tiempo, se utiliza el método llamado Análisis Factorial Múltiple de Tablas de Contingencia (AFMTC; Bécue-Bertaut y Pagès, 2004, 2008). El AFMTC ofrece la posibilidad de introducir la cronología como una columna activa en el análisis y describir los documentos tanto por su vocabulario, parecido a AC, como por su cronología. La estructura de la tabla múltiple que se va a analizar yuxtapone una tabla léxica con tantas columnas como palabras y la cronología (una tabla cuantitativa reducida a una sola columna, correspondiente a los años de publicación). AFMTC da un papel activo a ambos conjuntos, pero equilibrando su influencia en el análisis global. Si existe una relación entre la cronología y el vocabulario, se proporciona un primer eje altamente correlacionado con la cronología (Bécue-Bertaut, 2014; Bécue-Bertaut et al., 2008). Además la representación superpuesta de los documentos, ya sea desde el punto de vista de la cronología o del vocabulario, facilita encontrar los documentos que, desde el punto de vista de su vocabulario, están más avanzados a su cronología. Estos documentos los llamamos "artículos pioneros".

Para definir las funciones de las palabras es necesario dividir el corpus en: vocabulario local especializado, vocabulario regular o estable y vocabulario aleatorio. La metodología propuesta combina el índice de reparto del vocabulario propuesto por Hubert y Labbé (1990a,b), las palabras características cronológicas (Lebart et al., 2000, 1998) y una prueba de bondad de ajuste para la distribución de Poisson. Para modelizar la estructura del corpus y mostrar el flujo argumentativo del vocabulario, se trabaja sólo con el vocabulario especializado y se sigue la metodología propuesta por Benzécri (1973, 1981) e implementada en el método para el Análisis de una Matriz de Datos (AMADO) Chauchat y Risson (1995) y se incorpora un nuevo procedimiento que consiste en ordenar las palabras especializadas: primero, por sus coordenadas en la primera dimensión de un AC y segundo, de acuerdo con el documento o segmento de documentos caracterizados por las palabras. Después, mediante un AC, se pone de relieve la estructura del modelo de evolución cronológica y, a través de los gráficos de Bertin (1973, 1977, 1981) se hace una representación visual del vocabulario que le da evolución al corpus.

Aunque la metodología propuesta puede ser aplicada a cualquier tipo de corpus, la prioridad se asigna a corpus bibliográficos, compuestos de artículos científicos y a textos únicos no estructurados (como un discurso retórico y/o argumentativo). El objetivo es mostrar las ventajas que ofrece el análisis de los datos a través de un enfoque cronológico.

En el caso de una base bibliográfica, la metodología propuesta pretende responder a las siguientes preguntas: ¿Cuáles son los temas más relevantes de la investigación? ¿Los temas están relacionados con los años, autores, revistas, país, etc.? ¿Existe evolución en el vocabulario? ¿Qué es lo que determina su evolución? ¿Cuáles son los artículos pioneros?

Por otro parte, cuando se analiza un texto único, se busca poner de relieve su estructura y proporcionar información sobre su construcción. La metodología trata de solucionar los siguientes interrogantes: ¿El discurso está bien organizado? ¿En cuántas partes se puede dividir? ¿Existe diversidad temática? ¿Qué papel desempeña cada una de las palabras? ¿Cuáles son las palabras que permiten evolucionar al discurso?

1.2 Estructura de la tesis

En esta tesis se presenta la metodología para el análisis de textos cronológicos y las funciones programadas en R, donde fue implementada la metodología. En el capítulo 2, se resumen los métodos multidimensionales clásicos para análisis textual que respaldan nuestro enfoque: métodos de análisis factorial general (Análisis de Correspondencias (AC) y Análisis Factorial Múltiple (AFM)) y Clasificación Jerárquica (CJ). En el capítulo 3, se resumen los métodos para el estudio de corpus cronológicos tales como: Análisis Factorial Múltiple de Tablas de Contingencia (AFMTC), Clasificación Cronológica (CC), Palabras Características (PC), etc. En el capítulo 4, se propone la ´metodología para determinar las funciones de las palabras y modelizar la estructura del corpus. En el capítulo 5, se hace una aplicación de la metodología a una base de datos bibliográfica. En el capítulo 6, se aplica la metodología a un texto único no estructurado (discurso argumentativo). En el capítulo 7, se analiza las funciones de las palabras, mediante la aplicación a un discurso y a una base bibliográfica. En el capítulo 8, se describen las relaciones entre las funciones y se hace una demostración de las dos macro funciones utilizadas en los capítulos 5 y 6 (MacroBiblio y MacroTxCrhono) y, finalmente, en el capítulo 9 se exponen las conclusiones principales de esta tesis.

Capítulo 2

Métodos multidimensionales clásicos en análisis textual

El Análisis estadístico de datos textuales desarrollado con un enfoque multidimensional, a partir de las aportaciones de Jean Paul Benzécri (1973, 1977, 1981), ha permitido el estudio de textos de diversos tipos mediante la aplicación de métodos de Análisis Factorial General (AFG) y Clasificación a tablas de contingencia Documentos×Palabras.

La metodología presentada en esta tesis se basa en las técnicas estadísticas multidimensionales desarrolladas por la Escuela Francesa de Análisis de Datos. En este capítulo se presentan los métodos y sus propiedades, ya que serán utilizados en los siguientes capítulos.

2.1 Análisis Factorial General

2.1.1 Notación y principales aspectos

El Análisis Factorial General describe el proceso común a todos los métodos en ejes principales (Lebart et al., 1997); ofreciendo un enfoque geométrico para extraer y visualizar la información de la matriz de datos.

La notación y los principales aspectos se pueden resumir en los siguientes pasos:

- Dada una tabla $\mathbf{X}$ con I filas y J columnas, se consideran dos nubes: la nube de puntos-fila, N_I en $\mathbb{R}^J$, y la nube N_J en $\mathbb{R}^I$. Los pesos de las filas $\mathbf{D_I}$, también usados como métrica en el espacio de las columnas y los pesos de las columnas $\mathbf{D_J}$, también usados como la métrica en el espacio de las filas.

- Se buscan los ejes de mayor inercia, llamados ejes principales, de las nubes N_I y N_J. En N_I (resp. N_J), los vectores propios u_s (resp. z_s) son vectores en $\mathbb{R}^J$ (resp. $\mathbb{R}^I$) que satisfacen las siguientes ecuaciones:

$$\mathbf{X^T D_I X D_J u_s} = \lambda_s \mathbf{u_s} \tag{2.1}$$

con la restricción $||\mathbf{u_s}||_{\mathbf{D_J}} = \mathbf{u_s^T D_J u_s} = 1$.

$$\mathbf{X D_J X^T D_I z_s} = \lambda_s \mathbf{z_s} \tag{2.2}$$

con la restricción $||\mathbf{z_s}||_{\mathbf{D_I}} = \mathbf{z_s^T D_I z_s} = 1$.

- N_I y N_J se proyectan sobre los ejes de máxima inercia: las coordenadas de los puntos de N_I (resp. N_J) sobre el eje s constituyen el $I-factor$ de rango s (resp. $J-factor$), denotado F_s (resp. G_s):

$$\mathbf{F_s} = \mathbf{X D_J u_s}. \tag{2.3}$$

$$\mathbf{G_s} = \mathbf{X^T D_I z_s}. \tag{2.4}$$

- La proyección de la fila i (resp. columna j) sobre el eje de rango s en $\mathbb{R}^J$ (resp. $\mathbb{R}^I$) se puede calcular a partir de las coordenadas N_J (resp. N_I) sobre el eje de rango s en $\mathbb{R}^I$ (resp. $\mathbb{R}^J$) mediante las formulas de transición.

$$\mathbf{F_s} = \mathbf{X D_J G_s} \lambda_s^{-1/2} \tag{2.5}$$

$$\mathbf{G_s} = \mathbf{X^T D_I F_s} \lambda_s^{-1/2} \tag{2.6}$$

2.1.2 Esquema de dualidad del Análisis Factorial General

	Nube N_I	Nube N_J
Espacio	$\mathbb{R}^J$	$\mathbb{R}^I$
Métrica	$\mathbf{D_J}$	$\mathbf{D_I}$
Datos	$\mathbf{X}$	$\mathbf{X^T}$
Pesos	$\mathbf{D_I}$	$\mathbf{D_J}$
Ejes de inercia	$\mathbf{U}$	$\mathbf{Z}$
Ecuación	$\mathbf{X^T D_I X D_J U} = \mathbf{U}\Lambda$ (2.7)	$\mathbf{X D_J X^T D_I Z} = \mathbf{Z}\Lambda$ (2.8)
Ortonormalidad	$\mathbf{U^T D_J U} = \mathbf{Id}$	$\mathbf{Z^T D_I Z} = \mathbf{Id}$
Factores	$\mathbf{F} = \mathbf{X D_J U}$	$\mathbf{G} = \mathbf{X^T D_I Z}$
Ecuaciones	$\mathbf{X D_J X^T D_I F} = \mathbf{F}\Lambda$ (2.9)	$\mathbf{X^T D_I X D_J G} = \mathbf{G}\Lambda$ (2.10)
Ortogonalidad	$\mathbf{F^T D_I F} = \Lambda$	$\mathbf{G^T D_J G} = \Lambda$
Relaciones de transición	$\mathbf{F} = \mathbf{X D_J G}\Lambda^{-1/2}$	$\mathbf{G} = \mathbf{X^T D_I F}\Lambda^{-1/2}$

Tabla 2.1: Esquema general de AFG

El esquema general para AFG expuesto por Escofier y Pagès (1992) se resume en la Tabla 2.1. El Análisis de Correspondencias (AC) y el Análisis Factorial Múltiple (AFM) se pueden ver como casos particulares del AFG.

2.1.3 Herramientas e indicadores para la interpretación

Las herramientas más importantes que ayudan en la interpretación de los resultados de los métodos de análisis factorial son:

- **Varianza explicada por el factor**: es igual al cociente entre la inercia proyectada y la inercia total. Para el factor s es igual a

$$\frac{\lambda_s}{\sum_{s\in S}\lambda_s}. \tag{2.11}$$

  multiplicada por 100, este indicador da el porcentaje de inercia expresado por el factor de rango s.

- **Contribución de un elemento a la inercia del eje s**: las filas (resp. las columnas) contribuyen a la inercia del eje s mediante

$$\frac{\mathbf{D_I F_s}^2}{\lambda_s} \times 100 \tag{2.12}$$

$$\frac{\mathbf{D_J G_s}^2}{\lambda_s} \times 100. \tag{2.13}$$

- **Calidad de la representación**: la fila i en el factor s puede ser medida por la distancia entre el punto en el espacio y la proyección sobre el factor

$$qlt_s(i) = \frac{Inercia\ proyectada\ de\ i\ sobre\ \mathbf{u_s}}{Inercia\ total\ de\ i} = cos^2\theta_i^s \tag{2.14}$$

  donde θ_i^s es el ángulo entre **Oi** (vector que conecta el origen al punto i) y $\mathbf{u_s}$.

2.2 Análisis de correspondencias

El AC fue propuesto por Benzécri (1973, 1981) como un método inductivo para trabajar con datos textuales. El punto de partida es codificar el corpus en una tabla de frecuencias documentos×palabras $\mathbf{X}$, llamada tabla léxica. Los documentos pueden ser textos, cuando el corpus está dividido en textos; capítulos, cuando se trata del análisis de un libro; abstracts, cuando se trata de un estudio bibliográfico; secuencias cortas, como frases, cuando se analiza un solo texto, etc. Entre los resultados, se privilegian las representaciones gráficas. El término general de la tabla léxica x_{ij} contiene la frecuencia de la palabra j en el documento i. $\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij} = n$, siendo n el total de ocurrencias. El margen de las columnas contiene la frecuencia total de las palabras j, $x_{.j} = \sum_{i=1}^{I} x_{ij}$; $j = 1, ..., J$. El margen de las filas contiene la longitud de los documentos i, $x_{i.} = \sum_{j=1}^{J} x_{ij}$; $i = 1, ..., I$.

La tabla de frecuencias $\mathbf{X}$ es transformada en una tabla de proporciones $\mathbf{P}$ con el término general

$$\mathbf{P} = [p_{ij}] = [\frac{x_{ij}}{n}] \tag{2.15}$$

$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Los márgenes de las filas y columnas de $\mathbf{P}$ están dados por $p_{i.} = \sum_{j=1}^{J} p_{ij}$ y $p_{.j} = \sum_{i=1}^{I} p_{ij}$ (matriz $\mathbf{D_I}$ y $\mathbf{D_J}$, respectivamente). Los perfiles-fila y perfiles-columna se expresan como $(p_{ij}/p_{.j}, j = 1, ...I)$ y $(p_{ij}/p_{i.}, i = 1, ...J)$. Cada perfil-fila i está ponderado en relación a la proporción de sus ocurrencias sobre el total de ocurrencias, esto es: $p_{i.} = \sum_{j=1}^{J} p_{ij}$, simétricamente el perfil-columna j está ponderado por $p_{.j} = \sum_{i=1}^{I} p_{ij}$.

2.2.1 Equivalencia distribucional y distancia chi-cuadrado

Una distancia entre perfiles sobre I (respectivamente, J) está definida de tal manera que la distancia entre dos filas (respectivamente, dos columnas), es cero cuando los perfiles asociados son iguales y pequeña cuando son similares. Se llama sinónimos distribucionales a las filas y columnas con perfiles idénticos. Las distancias deben también cumplir el *principio de equivalencia distribucional*. Este principio establece que la distancia entre dos filas i e $i^{'}$ no cambia si se fusionan dos columnas j y $j^{'}$ que son sinónimos distribucionales. Simétricamente, la distancia entre dos columnas j y $j^{'}$ no cambia si se fusionan dos filas i e $i^{'}$.

El principio de equivalencia distribucional condujo a Benzécri (1973, 1977, 1981) a escoger las siguientes distancias al cuadrado entre filas y entre columnas

$$d^2(i, i') = \sum_{j \in J} \frac{1}{p_{.j}} (\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}})^2 \tag{2.16}$$

$$d^2(j, j') = \sum_{i \in I} \frac{1}{p_{i.}} (\frac{p_{ij}}{p_{.j}} - \frac{p_{ij'}}{p_{.j'}})^2 \tag{2.17}$$

La distancia entre dos filas (respectivamente columnas), llamada por Benzécri distancia distribucional, es conocida como distancia chi-cuadrado. Otras distancias entre perfiles que obedecen al principio de equivalencia distribucional han sido tratadas por Escofier (2003) y Greenacre y Lewi (2009).

Los resultados clásicos de CA se pueden obtener mediante un AFG aplicado a

$$\mathbf{Q} = \mathbf{D_I^{-1}} \mathbf{P} \mathbf{D_J^{-1}} = [q_{ij}] = [\frac{p_{ij}}{p_{i.} p_{.j}}] \tag{2.18}$$

o, de forma equivalente, a la matriz con las dimensiones (IXJ)

$$\bar{\mathbf{Q}} = [\bar{q}_{ij}] = [\frac{p_{ij} - p_{i.} p_{.j}}{p_{i.} p_{.j}}] \tag{2.19}$$

con métricas/pesos $\mathbf{D_J}$ y $\mathbf{D_I}$, esto es, AFG($\bar{\mathbf{Q}}$, $\mathbf{D_J}$, $\mathbf{D_I}$) (Escofier y Pagès, 1988; Pagès y Bécue-Bertaut, 2006). Este cálculo ubica a CA en el esquema general de AFG, mostrando que CA analiza la desviación entre $\mathbf{P}$ con las dimensiones (IXJ) y la matriz del modelo independiente $[p_{i.} p_{.j}]$.

2.2.2 Relaciones de transición

Las representación simultánea de filas y columnas se basa en las relaciones de transición que une las coordenadas $F_s(i)$ de los puntos-fila i $(i = 1, ..., I)$ y las coordenada $G_s(j)$ de los puntos-columna j $(j = 1, ..., J)$ sobre los ejes de dispersión s $(s = 1, ..., Min(I-1, J-1))$.

Las relaciones de transición se escriben:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j \in J} \frac{p_{ij}}{p_{i.}} G_s(j) \tag{2.20}$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{p_{ij}}{p_{.j}} F_s(i) \tag{2.21}$$

- $F_s(i)$: proyección de la fila i sobre la dimensión s de N_I
- $G_s(j)$: proyección de la columna j sobre la dimensión s de N_J
- λ_s valor común de la inercia asociada a cada una de las dimensiones

2.2.3 Herramientas para la interpretación

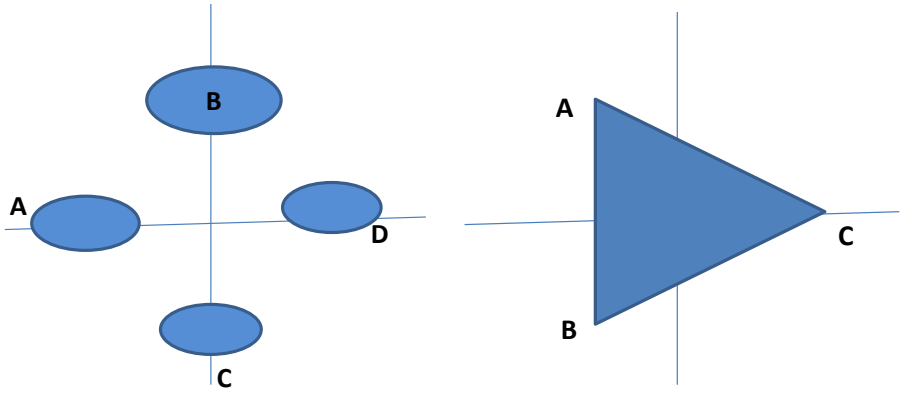
En las visualizaciones del AC, los documentos se presentan más cercanos cuando usan un vocabulario similar, y las palabras se presentan más cercanas cuando están presentes con más frecuencia en el mismo documento o asociadas con las mismas palabras. Este resultado demuestra la habilidad del AC para encontrar relaciones entre los documentos y las palabras. Es decir, los documentos que tienen un significado similar, aunque estén expresados con palabras diferentes, se asocian entre sí. Ambas presentaciones están relacionadas de tal forma que las relaciones de las palabras y los documentos se muestran, permitiendo así la interpretación de las similitudes entre los documentos en términos de su vocabulario y contenido (Lebart et al., 1998). Por tanto, el AC tiene en cuenta lo siguiente:

- Similitudes entre documentos basándose en su contenido.
- Similitudes entre palabras basándose en su distribución entre los documentos, teniendo en cuenta el contexto, es decir, las asociaciones entre las palabras.
- Asociaciones mutuas entre los documentos y las palabras mediante una representación simultánea de las filas (documentos) y las columnas (palabras) en la misma gráfica.

Herramientas visuales

Kerbaol et al. (2006) denomina *metallaves/metadocumentos* a los grupos de palabras/documentos cuyas contribuciones son muy altas en un eje. La interpretación de

los resultados de AC consiste en observar, eje por eje, los *"metallaves"* y los *"metadocumentos"* que caracterizan a cada eje. En un eje dado pueden existir dos *"metallaves"*(*"metallave+"/"metallave-"*), es decir, un conjunto de palabras que más contribuyen a su inercia y que se encuentran en su parte positiva/negativa. De forma similar, para un eje dado, pueden existir dos *"metadocumentos"* (*"metadocumento+"/"metadocumento-"*), conjunto de los documentos que más contribuyen a su inercia y que se encuentra en la parte positiva/negativa. Por tanto, uno o dos *metallaves/* uno o dos *metadocumentos* pueden caracterizar cada eje dependiendo de la configuración de las palabras/documentos. Las palabras que pertenecen al mismo *metallave* pueden usarse, frecuentemente, en los mismos documentos y, todas juntas corresponden a un tópico dado. Una palabra puede pertenecer a varios *"metallaves"*, pero asociada en cada uno de ellos con otras palabras diferentes y en contextos diferentes. Cada contexto corresponde a un significado diferente (Bécue-Bertaut, 2014) Para facilitar la interpretación gráfica se presenta el esquema de Morin (2006).



(a) Esquema ideal (b) Esquema común

Figura 2.1: Metallaves/Metadocumentos

La Figura 2.1a muestra lo que se puede obtener en el primer plano factorial cuando los documentos son monotemáticos. Sea A, B, C y D los grupos de palabras/documentos (*metallaves/metadocumentos*) que conforman temas. En este caso, cada tema tiene su proyección sobre un eje. La interpretación a partir de esta figura permite identificar sin ambigüedad los distintos grupos temáticos y los documentos relacionados con ellos. Lo ideal es encontrar esquemas como este, donde la interpretación es evidente.

La Figura 2.1b corresponde a la situación más frecuente. Algunos de los tópicos están bien representados (por ejemplo C) sobre el primer eje positivo y están en oposición con otros (C se opone a A y B). Las proyecciones de los temas A y B en la parte izquierda del primer eje se mezclan. Cuando se presenta este tipo de esquema, donde los temas se mezclan, los resultados son difíciles de interpretar. Para poder hacerlo, se seleccionan las palabras y los documentos cuyas contribuciones a la inercia son grandes (por lo general, tres veces la contribución media por palabras o por documento); esto se hace en cada parte de los ejes que fueron conservados (positiva y negativa).

2.3 Análisis Factorial Múltiple

El Análisis Factorial Múltiple (AFM; Escofier y Pagès, 1988) analiza tablas en las cuales un conjunto de individuos está descrito por varios grupos de variables. El AFM puede verse como un AFG ponderado, de tal forma que se equilibra la influencia de los distintos grupos de variables. En un mismo grupo, las variables deben de ser del mismo tipo (cuantitativas o categóricas), pero los grupos de variables pueden ser de diferentes tipos. Para simplificar el resumen de AFM, se consideran los pesos de las filas y los pesos de las columnas iguales a la unidad.

2.3.1 Tabla múltiple

En una tabla múltiple, el conjunto de los individuos está descrito por varios grupos de variables.

- *Tabla global:* los I individuos i $(i = 1, ..., I)$ constituyen la nube N_I situada en $\mathbb{R}^J$; las J variables j $(j = 1, ..., J)$ constituyen la nube N_J situado en $\mathbb{R}^I$
- *Subtabla l:* considerando solo la tabla l $(l = 1, ..., L)$, los I individuos son denotados por i^l $(i = 1, ..., I)$ y constituyen la nube N_I^l situada en $\mathbb{R}^{J_l}$; las J_l variables constituyen la nube N_J^l en $\mathbb{R}^I$.

2.3.2 El AFM como un AFG

El AFM trata las variables continuas como un AFG y las variables categóricas como un Análisis de Correspondencias Múltiple, pero teniendo en cuenta la ponderación. Los resultados proporcionados por AFM son los resultados clásicos del análisis factorial general:

- las coordenadas, contribuciones y calidad de representación de los individuos;
- los coeficientes de correlación entre los factores y las variables continuas;
- las coordenadas de las categorías, como centros de gravedad de los individuos que pertenecen a esa categoría. A cada coordenada se asocia un indicador, llamado *valor test* que permite seleccionar las categorías asociadas con más fuerza a cada eje.

2.3.3 Equilibrar los conjuntos de variables

Si se consideran todos los grupos de variables como elementos activos, pero sin equilibrar su influencia, es posible que un único grupo domine la construcción de los primeros ejes. La influencia de un grupo de variables deriva de su estructura, en el sentido de la distribución de la inercia en las nubes N_I^l y N_J^l inducidas. Este fenómeno sugiere normalizar la mayor inercia axial de cada conjunto. Esto se obtiene mediante la ponderación de cada variable del conjunto l por $1/\lambda_1^l$ donde λ_1^l es el primer valor propio del análisis factorial general aplicado al conjunto l. Dicha ponderación se puede interpretar fácilmente: normaliza cada una de las dos nubes inducidas por el grupo l de variables N_I^l y N_J^l dándoles una inercia axial máxima igual a 1. Sin embargo, no se equilibra la inercia total de los

diferentes grupos; un grupo de mayor dimensionalidad tendrá una mayor influencia global en el sentido de que contribuirá a un mayor número de ejes.

2.3.4 Representación superpuesta de las l nubes de individuos

A cada conjunto l, se asocia la nube N_I^l de individuos en el espacio $\mathbb{R}^{J_l}$. Dicha nube, llamada "parcial", contiene individuos "parciales" i^l (individuo i según el conjunto l).

Para poner de relieve las similitudes entre las diferentes nubes N_I^l, es decir, representar los puntos homólogos tan próximos como sea posible, se proyectan las nubes N_I^l sobre los ejes del análisis global como elementos ilustrativos. Las coordenadas de i^l sobre el eje s se expresan como $F_s(i^l)$ y pueden calcularse a partir de $G_s(j)$, $j \in J_l$, mediante la siguiente relación:

$$F_s(i^l) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\lambda_1^l}} \sum_{j \in J_l} x_{ij} G_s(j) \tag{2.22}$$

Esta es la fórmula de transición usual, pero restringida a las variables del grupo J_l.

2.4 Clasificación

Los métodos de clasificación constituyen, junto con los métodos factoriales, una segunda familia importante de técnicas de análisis multidimensional de datos. Estos métodos nos permiten representar las proximidades entre las filas o las columnas de una tabla léxica mediante la formación de clases (Lebart et al., 2000, 1998). Los resultados proporcionados por los métodos de clasificación revelan ser, en la práctica, unos complementos indispensables de los resultados proporcionados por el análisis de correspondencias (Lebart et al., 2000, 1998; Murtagh, 2005; Murtagh et al., 2009, 2011).

2.4.1 Clasificación jerárquica

El punto de partida de la clasificación jerárquica es una matriz de disimilitudes $\mathbf{X}$ (estas disimilitudes pueden ser distancias euclidianas), entre individuos donde el término general $d(i, i')$ es la disimilitud entre los individuos i y i'

En primer lugar, se agregan los individuos i y i' más próximos. Este par de individuos agregado constituye un nuevo elemento, (i, i'), cuyo peso es la suma de los pesos de los individuos agregados. Después se actualiza la matriz $\mathbf{X}$, calculando las distancias entre este nuevo elemento y cada uno de los individuos que quedan por clasificar. A la salida de esta etapa, el problema se reduce a clasificar $I - 1$ individuos. Se agregan de nuevo los dos individuos más próximos y se repite el proceso ($I - 1$ veces en total) hasta agotar el conjunto de los individuos. La última $(I - 1)$ *ésima* operación reagrupa el conjunto de los individuos en el seno de una única clase (Husson et al., 2010; Lebart et al., 2000, 1998).

Cada uno de los reagrupamientos efectuados recibe el nombre de *nodo*. Al conjunto de individuos reunidos en un nodo se le llama clase .

Existen diferentes métodos de agregación que son divididos de acuerdo a la forma en que operan y a los resultados que brindan. El método de agregación utilizado es el de ligamiento completo o vecino más lejano, porque a diferencia de otros métodos, este define la distancia entre los nodos por el más distante de los dos individuos comparados.

La clasificación obtenida se puede representar en forma de árbol jerárquico o dendograma. Esta representación muestra, de manera clara, que las clases formadas a lo largo del proceso de clasificación constituyen una jerarquía indexada de clases parcialmente anidadas unas en otras, que puede ser vista como una continuación de particiones.

Inercia y partición

La partición en el árbol jerárquico debe buscar que:

- en el interior de las clases, definidas por el corte, los individuos sean homogéneos.
- de una clase a otra, los individuos sean diferentes.

Si los documentos están en un espacio euclidiano, el teorema de Huygens descompone la inercia total (de la nube de los individuos) en dos partes (Husson et al., 2010):

Inercia total = Inercia interclases + Inercia intraclase.

$$\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\sum_{j=1}^{J}(x_{iqj}-\bar{x_j})^2=\sum_{q=1}^{Q}\sum_{j=1}^{J}I_q(\bar{x_{qj}}-\bar{x_j})^2+\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\sum_{j=1}^{J}(x_{iqj}-\bar{x_{qj}})^2 \tag{2.23}$$

- x_{iqj} el valor para la variable j del individuo i en la clase q
- $\bar{x_{qj}}$ la media de la variable j en la clase q
- I_q el número de individuos en la clase q
- $\bar{x_j}$ la media general de la variable j

- **inercia intraclase**: es un indicador de la homogeneídad de los individuos. Establecida sobre la diferencia entre cada punto y el centro de gravedad de la clase a la cual pertenece;
- **inercia interclases**: fundada sobre la diferencia entre cada centro de gravedad de una clase y el centro de gravedad general.

Si se considera esta descomposición como marco de análisis para la búsqueda de una buena partición, es indiferente minimizar la variabilidad intraclase o maximizar la variabilidad interclases (Husson et al., 2010). De esto depende que la calidad de una partición puede ser medida por:

$$\frac{Inercia\ interclases}{Inercia\ total} \tag{2.24}$$

Este cociente indica la parte de variabilidad total expresada por la partición.

Capítulo 3

Métodos para el análisis de corpus cronológicos

El objetivo de este capítulo es presentar los métodos cronológicos que facilitan el entendimiento de los cambios presentados en el corpus. Estos métodos están organizados en tres categorías: Análisis Factorial Múltiple de Tablas de Contingencias, métodos de clasificación (Clasificación con Restricción de Contigüidad y Clasificación Cronológica) y métodos para caracterizar léxicamente los períodos o partes (Palabras Características, Palabras Cronológicas, Incrementos Léxicos Específicos).

3.1 Análisis factorial múltiple de tablas de contingencias

Análisis Factorial Múltiple de Tablas de Contingencias (AFMTC) permite un estudio cronológico más completo al incorporar la cronología como una variable activa, y describir las partes o períodos del corpus tanto por su vocabulario como por su cronología (Bécue-Bertaut, 2014)

El AFMTC fue propuesto por Bécue-Bertaut y Pagès (2004, 2008). Este método es una extensión del AFM; parte de los principios del análisis de correspondencias binarias intra-tablas (Benzécri, 1973; Escofier, 1983), generalizado por Cazes y Moreau (1991) con el nombre de Análisis de Correspondencias Interno (ACI), y, así, toma en cuenta las diferencias entre los márgenes de las filas. Además, adopta el enfoque del AFM para equilibrar la influencia de las diferentes tablas y proporciona gráficos específicos de la estructura en grupos de las columnas.

El AFMTC puede manejar una tabla múltiple yuxtaponiendo varios conjuntos de columnas cuantitativas, categóricas y de frecuencias. En este trabajo, la estructura de los datos es muy específica; el corpus cronológico se organiza en una tabla múltiple. Los documentos se describen por el vocabulario (conjunto de frecuencias con tantas columnas como palabras diferentes) y la cronología (una columna cuyos diferentes valores son los años de publicación, considerados como valores cuantitativos). Por tanto, la tabla múltiple que se va a analizar yuxtapone por filas una tabla léxica con tantas columnas como palabras y una tabla cuantitativa reducida a una sola columna (años de publicación). AFMTC da un papel activo a ambos conjuntos pero equilibra su influencia en el análisis global. Si existe una relación entre la cronología y el vocabulario, se proporciona un primer eje altamente correlacionado con la cronología. Bécue-Bertaut (2014) detalla AFMTC aplicado a unos

datos con estructura similar.

3.1.1 NOTACIÓN

Varias tablas de frecuencias $\mathbf{X_1}, ..., \mathbf{X_l}, ..., \mathbf{X_L}$, de dimensión $(I \times J_l)$, se yuxtaponen por filas en la tabla de frecuencias múltiple $\mathbf{X}$ de dimensión $(I \times J)$.

$\mathbf{X}$ se transforma en una tabla de proporciones $\mathbf{P}$ (Figura 3.1). Entonces, p_{ijl} es la proporción asociada a la fila i $(i = 1, ..., I)$ en la columna j $(j = 1, ..., J_l)$ de la tabla l $(l = 1, ..., L)$; $\sum_{l \in L} \sum_{i \in I} \sum_{j \in J_l} p_{ijl} = 1$. Las filas y columnas marginales de la tabla $\mathbf{P}$ son $p_{i..} = \sum_{l \in L} \sum_{j \in J_l} p_{ijl}$ y $p_{\bullet jl} = \sum_{i \in I} p_{ijl}$, respectivamente.

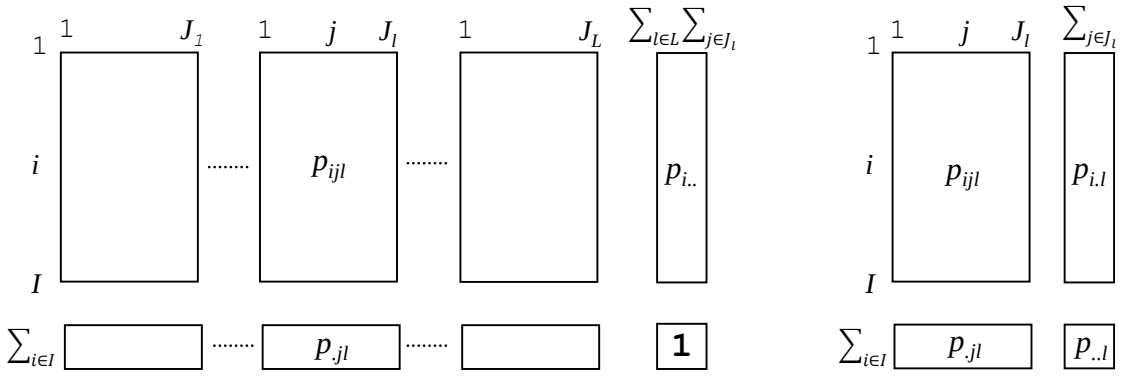


Figura 3.1: Tabla de proporciones múltiple $\mathbf{P}$ y márgenes

3.1.2 ALGORITMO

La introducción de las tablas de frecuencias como conjuntos de variables induce un problema específico. El Análisis Factorial Múltiple (AFM) requiere que los pesos unitarios sean idénticos en todas las tablas. En el caso del AFM clásico, al analizar cualquier conjunto de variables cuantitativas o categóricas, se adoptan pesos uniformes. En el caso de una tabla de frecuencias, CA impone los coeficientes de los márgenes de las filas como los pesos de las filas.

El AFMTC combina el Análisis de Correspondencia Interno (ACI; Cazes y Moreau, 1991; Escofier, 1983) y resuelve este problema centrando las subtablas en sus propios márgenes. Se puede ver como un CA que hace referencia al modelo de independencia de intra-tablas cuyo término genérico es

$$m_{ijl} = \left(\frac{p_{i.l}}{p_{..l}} \right) p_{.jl} \tag{3.1}$$

donde $p_{i.l} = \sum_{j \in J_l} p_{ijl}$ es el margen de la fila de la tabla l y $p_{..l} = \sum_{i \in I} \sum_{j \in J_l} p_{ijl}$ es la suma de los términos de la tabla l dentro de la tabla de proporciones $\mathbf{P}$.

Entonces la tabla $\mathbf{Y}$ es igual a

$$\mathbf{Y} = \frac{p_{ijl} - m_{ijl}}{p_{i..} \times p_{.jl}} = \frac{p_{ijl} - \left(\frac{p_{i.l}}{p_{..l}}\right) p_{.jl}}{p_{i..} \times p_{.jl}} = \frac{1}{p_{i..}} \left(\frac{p_{ijl}}{p_{.jl}} - \frac{p_{i.l}}{p_{..l}}\right). \tag{3.2}$$

El término genérico de la tabla $\mathbf{Y}$ es el peso residual respecto al modelo de independencia de las intra-tablas. Este modelo neutraliza las diferencias entre los perfiles de columnas medios separados.

AFMTC realiza un AFG no estandarizado de la tabla global $\mathbf{Y}$ dando el peso $p_{i..}$ a la fila i y el peso $p_{.jl}/\lambda_1^l$ a la columna j de la tabla l con λ_1^l el primer valor propio del AFG separado de la subtabla $\mathbf{Y_l}$.

3.1.3 Tipos de resultados

En el caso de la estructura particular descrita anteriormente, AFMTC ofrece los siguientes resultados (Bécue-Bertaut, 2014):

- Resultados globales en las filas y columnas activas; valores propios, representaciones de la fila-documentos y columna-palabras de una forma similar al AC.
- Resultados parciales de las filas y columnas activas: contribución de cada grupo en la construcción de cada eje; representación superpuesta de los documentos, ya sea desde el punto de vista de la cronología o del vocabulario. Esto facilita encontrar los documentos que, desde el punto de vista de su vocabulario, están más avanzados a su cronología. Estos documentos son llamados "trabajos pioneros" porque su vocabulario se usa más a menudo que antes de su fecha de publicación.

Se pueden introducir y representar columnas ilustrativas en los gráficos principales. El año-categórica es una variable categórica ilustrativa. Ésta se construye a partir de la variable cuantitativa "año de publicación", con los diferentes años como las categorías de esta variable. Cada categoría (año) se representa como ilustrativa, ya sea en las representaciones globales o parciales. En este caso, vale la pena enfatizar lo siguiente:

- La trayectoria de los años-categórica (desde el punto de vista del vocabulario), traza el ritmo y la evolución del vocabulario con posibles cambios hacia adelante y hacia atrás y permite identifica períodos léxicamente homogéneos en el corpus, caracterizados por el vocabulario.

3.1.4 Validación

Un test de permutaciones evalúa el nivel de significación del primer valor propio global que da cuenta de la relación entre el vocabulario y la cronología. La hipótesis nula es la no existencia de una dimensión cronológica en la variabilidad del vocabulario. De esta manera, las filas de la columna-año son permutadas aleatoriamente y, para cada permutación se lleva a cabo un AFMTC, conduciendo a una distribución empírica del

primer valor propio bajo la hipótesis nula. Un gran número de réplicas permite un cálculo del valor p asociado con el valor observado del primer valor propio.

3.2 Clasificación

La clasificación con restricción de contigüidad temporal descubre la estructura del corpus mediante un esquema de árbol, y la clasificación cronológica permite segmentar partes o períodos léxico homogéneos.

3.2.1 Clasificación jerárquica con restricción de contigüidad

La clasificación jerárquica con restricción de contigüidad completa la vizualización de los métodos factoriales al tener en cuenta una mayor dimensionalidad, corrigiendo ciertas deformaciones inherentes a la representación sobre un espacio de dimensión reducida. Este método de clasificación reúne las partes o períodos a partir de sus coordenadas en los ejes seleccionados. El método de enlace que se utiliza es *Complete linkage*. Cualquier algoritmo de aglomeración puede ser utilizado; sin embargo, se recomienda el método de Complete linkage porque, al calcular las distancias entre los grupos de fechas/partes, cuya restricción es que al menos un miembro de los dos grupos que se unen debe ser adyacente, se evitan las inversiones o retrocesos (Bécue-Bertaut et al., 2014; Legendre y Legendre, 1998; Murtagh, 1985). En el caso de un corpus cronológico, permite tener en cuenta los distintos cambios que se pueden presentar en la estructura a través del tiempo. Se deben seguir los siguientes pasos:

- Calcular una matriz de similitud entre las partes o períodos.
- A partir de este esquema de conexión, se construye una matriz de contigüidad con 1 para períodos conectados y con 0 para los no conectados.
- Calcular el producto de Hadamard de estas dos matrices, es decir, el producto, elemento por elemento.
- Buscar el valor de máxima similitud y formar un nuevo nodo con el par de períodos o grupo de períodos.
- Actualizar la matriz de similitud.
- Actualizar la matriz de contigüidad.
- Volver al punto 3 y seguir hasta que todas las partes o períodos pertenezcan a un mismo grupo.

3.2.2 Clasificación cronológica

El corpus puede ser segmentado en partes o períodos léxico homogéneas. La decisión de segmentar o no el corpus depende del tipo de texto que se analice. En el caso de textos no estructurados, como discursos, Bécue-Bertaut et al. (2014) proponen segmentar el texto en partes lo suficientemente grandes de modo que la variabilidad local se suavice. Esta

segmentación se realiza a partir de una secuencia inicial en partes cortas equivalentes a frases. Estas pueden ya existir en el texto, o se deducen a partir de una lectura clásica del texto, o de manera automatizada se cortan pseudo-frases de igual tamaño conforme fluye el texto. Después, las diferentes palabras son identificadas y se calcula su frecuencia. El texto es considerado como una serie de frases (serie temporal multi-palabras), que se ajusta muy bien con el enfoque de AC, al que se aplica el método de clasificación cronológica (CC; Legendre y Legendre, 1998), desarrollado para identificar discontinuidad en series temporales multidimensionales, como el seguimiento del recuento de una serie de especies a lo largo del tiempo. CC utiliza el algoritmo de clasificación con restricción de contigüidad, pero agrega un nuevo paso al algoritmo: cada agregación se somete a una prueba estadística que autoriza, o no, la fusión entre las dos frases o grupos de frases. Es decir, dependiendo del p-value asociado con el test de permutaciones se permite o no, la agregación. Los pasos que se siguen son:

- Se calculan todas las distancias entre pares de frases o grupos adyacentes que son candidatos a fusionarse (n_1 frases en el primer grupo y n_2 en el segundo grupo). Estas distancias se dividen en dos grupos: 50 % de las distancias con los valores más altos son codificadas con 1 y 50 %; con los valores más bajos se codifican con 0.
- Las distancias con los valores más altos en la matriz entre el grupo (área compartida entre objetos de un grupo y de otro) se suman y se indican con una h.
- Para cada combinación distinta, se calcula el recuento de las distancias más altas entre los grupos permutados. Si el total de combinaciones es muy grande, las permutaciones pueden ser seleccionadas al azar para formar la distribución de referencia para las pruebas de significación.
- Con número de permutaciones que produce un resultado igual o superior a h, dividido por el número de permutaciones realizadas, se obtiene una estimación de la probabilidad p de la observación de los datos bajo la hipótesis nula (los objetos de los dos grupos son extraídos de la misma población estadística y , en consecuencia, es sólo un artefacto de aglomeración del algoritmo de agrupación que temporalmente forma dos grupos).
- La probabilidad p se compara con un nivel de significación preestablecido α. Si $p \leq \alpha$, la hipótesis nula es rechazada y se impide la unión de los dos grupos.

El nivel de significación α utilizado en la prueba, determina la facilidad en que la hipótesis nula puede ser rechazada. Aumentar el valor α hace más fácil rechazar la hipótesis nula y que más grupos se formen. Los grupos resultantes son, por tanto, más pequeños y sobresalen más discontinuidades en la serie de datos. Así, cambiando el valor de α, en realidad cambia la resolución de los resultados agrupados.

3.3 Caracterización léxica de los períodos

Las palabras características son las palabras cuya frecuencia en un período es significativamente mayor que lo que indicaría la aleatoriedad. Éstas proporcionan evidencia sobre el contenido del corpus en cada época.

Los incrementos léxicos específicos comparan el vocabulario de un período con el vocabulario de períodos anteriores para mostrar los nuevos temas que aparecen en el corpus como resultado del incremento del nuevo vocabulario.

3.3.1 Palabras características

Las palabras características identifican palabras con una frecuencia muy alta o con una frecuencia muy baja en cada período del corpus (Lebart et al., 2000, 1998). El procedimiento se describe a continuación.

Sea:

- n_{ij} el número de ocurrencias de la palabra i en el período j
- $n_{..}$ el número total de ocurrencias en todo el corpus
- $n_{.j}$ el número de ocurrencias en el período j
- $n_{i.}$ el número de ocurrencias de la palabra i en todo el corpus

La frecuencia de palabras i en el período j se compara con las frecuencias que se podrían obtener con todas las posibles muestras comprendidas de $n_{.j}$, ocurrencias extraídas de manera aleatoria sin reemplazo del total del corpus (que es la hipótesis nula). Si la palabra i es relativamente más frecuente en el período j que en toda la muestra, esto es, si $n_{ij}/n_{.j} > n_{i.}/n$ (respectivamente, menos frecuente en el período j que en toda la muestra), el *p-value* de la prueba, para el caso donde la palabra i es más frecuente, se calcula mediante expresión 3.3 y, para el caso donde la palabra i es menos frecuente, mediante la expresión 3.4.

$$p_{i,j} = \sum_{x=n_{ij}}^{n_{.j}} \frac{\binom{n_{i.}}{x}\binom{n_{..}-n_{i.}}{n_{.j}-x}}{\binom{n_{..}}{n_{.j}}} \qquad (3.3) \qquad p_{i,j} = \sum_{x=1}^{n_{ij}} \frac{\binom{n_{i.}}{x}\binom{n_{..}-n_{i.}}{n_{.j}-x}}{\binom{n_{..}}{n_{.j}}} \qquad (3.4)$$

3.3.2 Incrementos específicos

Los incrementos léxicos específicos permiten comparar el vocabulario de un período con el vocabulario de los períodos anteriores (Lebart et al., 1998). Supongamos que el corpus está dividido en P períodos o partes. El objetivo es comparar la frecuencia n_{ij} de la palabra en el período j, donde $2 \leq j \leq P$, con la frecuencia de esa misma palabra en el conjunto de los períodos 1 hasta j. Para llevar a cabo esta comparación hay que recurrir a cálculos probabilistas similares a los que se efectúan para la extracción de las palabras características, pero considerando que en este caso se sustituye la totalidad del corpus por el conjunto de los períodos de 1 hasta j.

Sea:

- n_{ij} frecuencia de la palabra i en el período j

- $n_{.j}$ el número de ocurrencias en el período j
- $n^j_{..}$ el número de ocurrencias de los primeros j períodos
- $n^j_{i.}$ el número de ocurrencias de la palabra i en el mismo sub-corpus (1 hasta j períodos)

Si n_{ij} es claramente superior a la moda de la distribución hipergeométrica, se calcula la probabilidad de observar un número de ocurrencias de la palabra i igual o superior a n_{ij} cuando las $n_{.j}$ ocurrencias se seleccionan al azar. esto es, si $n_{ij}/n_{.j} > n^j_{i.}/n^j_{..}$. Si, por el contrario, este valor n_{ij} es claramente inferior a la moda, entonces se calcula la probabilidad de observar un número de ocurrencias de la palabra i igual o inferior a n_{ij}. El *p-value* de la prueba, para el caso donde la palabra i es más frecuente, se calcula mediante la expresión 3.5 y para el caso donde la palabra i es menos frecuente, mediante la expresión 3.6.

Sea:

$$p_{i,j} = \sum_{x=n_{ij}}^{n_{.j}} \frac{\binom{n^j_{i.}}{x}\binom{n^j_{..}-n^j_{i.}}{n_{.j}-x}}{\binom{n^j_{..}}{n_{.j}}} \qquad (3.5) \qquad p_{i,j} = \sum_{x=1}^{n_{ij}} \frac{\binom{n^j_{i.}}{x}\binom{n^j_{..}-n^j_{i.}}{n_{.j}-x}}{\binom{n^j_{..}}{n_{.j}}} \qquad (3.6)$$

Capítulo 4

Métodos para determinar las funciones de las palabras

El objetivo de este capítulo es presentar los métodos para determinar las funciones de las palabras y modelizar la estructura de un corpus cronológico. Estos métodos son: el índice del reparto del vocabulario, las palabras características cronológicas y una prueba de bondad de ajuste para la distribución Poisson (ayudan a establecer los criterios para dividir las palabras según su función) y los métodos que permiten visualizar el vocabulario y mostrar la estructura del modelo o esquema de evolución cronológica (Matriz reordenada de Bertin, Análisis de una matriz de datos (AMADO) y AC para el estudio de trayectorias).

4.1 Funciones de las palabras

Para definir las funciones de las palabras y caracterizar el flujo del vocabulario, nuestra hipótesis es que necesitamos dividir el corpus en: vocabulario local o especializado, vocabulario regular o estable y vocabulario aleatorio (Figura 4.1)

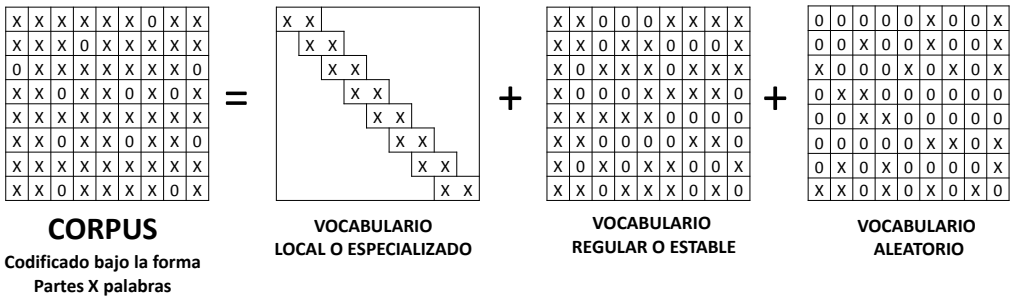


Figura 4.1: Descomposición del corpus

- Vocabulario local o especializado: palabras locales que indican ideas o temas más concretos, abordados en ciertas partes del corpus. Estas palabras son las que marcan la estrategia evolutiva del corpus.
- Vocabulario regular o estable: palabras utilizadas de forma regular a lo largo del corpus. Constituye, en cierta manera, la trama. En el caso de un texto argumentativo,

la mayoría de palabras que conforman este grupo son aquellas que reflejan la voluntad del orador de dar un tono de seguridad o de duda al discurso.

- Vocabulario aleatorio o usual: palabras herramientas como preposiciones, artículos, conjunciones, etc. Presentan una distribución de Poisson.

4.1.1 Índice del reparto del vocabulario

El índice de reparto del vocabulario divide el vocabulario del corpus según su uso, en regular o especializado (Hubert y Labbé, 1990a,b). Según los autores, en todo discurso, el orador dispone de un vocabulario general polivalente y de un vocabulario especializado.

A partir del momento en que una palabra aparece varias veces, nos podemos preguntar: ¿La palabra es regular? ¿De qué parte del texto es característica esa palabra? La respuesta dependerá del uso que se haga de esa palabra. La repartición permite saber si la palabra es de uso regular (o estable) o uso especializado en el corpus considerado.

Sea un texto compuesto de N palabras en las que hay V palabras diferentes. Consideramos una palabra teniendo F frecuencia en el texto. Se asocia a esta palabra una dimensión característica T igual a la inversa de su frecuencia relativa.

$$T = N/F \tag{4.1}$$

Después se calculan los intervalos que separan cada ocurrencia de la palabra. Sea d_i el número de palabras que separan las i ésima y $(i+1)$ ésima ocurrencias de esta palabra. Para la primera y la última ocurrencia del texto se suma el número de palabras separándolas respectivamente desde el comienzo hasta el final del texto. Los límites de de los intervalos estarán comprendidos en el cálculo, cuya suma será:

$$\sum_{i=1}^{i=F} d_i = N \tag{4.2}$$

Los F intervalos d_i están clasificados por tamaño creciente (i varia de 1 a F). El índice i está aumentado de 1 a $(k-1)$ mientras que d_i es inferior o igual a T. Sea k el valor de i cuando d_i es superior a T (generalmente T no es un entero. Entonces convine tomar la longitud del entero inmediatamente superior a T). El intervalo d_k contiene un cierto número de fragmentos de longitud igual a T en el que la palabra considerada no aparece. Su número es igual a $(d_k - T)$. Hay $(F-k)+1$ intervalos donde una situación tal es posible. El número total de fragmentos de longitud T de donde la palabra está ausente es igual a:

$$\sum_{i=k}^{i=F}(d_i - T) = \sum_{i=k}^{i=F} d_i - [((F-k)+1) * T] \tag{4.3}$$

Sea $N^{'}$ el número de segmentos de longitud T que contiene la palabra considerada.

Donde :

$$N^{'} = N - (\sum_{i=k}^{i=F} d_i - [((F-k)+1) * T]) \tag{4.4}$$

El índice de repartición se basa en la comparación de N y $N^{'}$ y variará entre 0 y 1:

- Si $F = k$ entonces $N^{'} = N$. Todos los segmentos posibles de longitud T contienen una ocurrencia de la palabra y ésta es conocida como una palabra regular. Cuando el índice toma su valor máximo, o sea, igual a 1, se dice que la palabra presenta una regularidad perfecta.
- Si $N^{'} = F$ entonces todas las ocurrencias de la palabra son contiguas y están contenidas en un intervalo de F palabras y el índice toma un valor cercano a cero; la palabra es de uso local.

Índice de repartición:

$$R = \frac{N^{'} - F}{N - F} \tag{4.5}$$

El índice es igual a 1 si $N^{'} = N$, es decir, en la hipótesis de una repartición de la palabra perfectamente regular. Es igual a 0 si $N^{'} = F$; es decir, cuando todas las ocurrencias de la palabra son contiguas. Por consiguiente, se define la repartición de una palabra en un texto como un conjunto de lugares en los que aparece esta palabra, o su "posición". Cuando la aparición es única, esta ubicación es en sí misma significativa y no hay necesidad de información adicional. Sin embargo, tan pronto como la palabra se presenta en varios lugares en el texto, surge una pregunta: ¿Es esta aparición regular y, de lo contrario, podemos considerar que la palabra o vocablo es característico de una parte en particular del texto? Es así como se propone el índice de repartición que cuantifica el grado de regularidad de una palabra en un corpus.

Procedimiento para repartir las palabras en función de la categoría gramatical

Labbé y Hubert (2016) proponen dividir el corpus en tantos sub-corpus como categorías gramaticales significativas existan (verbos, sustantivos, adjetivos nombres propios, pronombres, adverbios, determinantes, preposiciones y conjunciones). Dentro de cada una de las categorías gramaticales las palabras se ordenan en clases de frecuencia. Dentro de cada una de estas clases, se calcula el índice promedio de reparto del vocabulario (IPRVCF) y la desviación estándar, lo que permite identificar las palabras anormalmente repartidas (en los dos extremos de la distribución). Evidentemente, el cálculo sólo tiene sentido cuando las clases de frecuencia tienen suficientes efectivos.

4.1.2 Palabras características cronológicas

Para identificar las palabras cronológicas se buscan primero las palabras caracteristicas de cada período del corpus, esto es, las palabras cuya frecuencia en la parte o período

que corresponde es significativamente superior a lo que la aleatoriedad indicaría (primer nivel) y después, las de grupos de dos períodos consecutivos (palabras características del segundo nivel); después, las de tres períodos consecutivos (tercer nivel), y así sucesivamente. Al final del proceso, cada palabra es asignada como una palabra característica cronológica en el período o grupo de períodos que mejor caracteriza, es decir, para el cual presenta el más pequeño p-value de una distribución hipergeométrica (con la condición de que este p-value sea menor de 0.05).

4.1.3 Criterios para dividir las palabras según su función

Para dividir las palabras según su función se combinaron tres métodos: el índice del reparto del vocabulario, las palabras características cronológicas y una prueba de bondad de ajuste para la distribución Poisson. Esto es debido a que el índice de reparto propuesto por Hubert y Labbé (1990a,b); Labbé y Hubert (2016) y desarrollado en la sección anterior, sólo permite obtener una lista tanto del vocabulario estable como del vocabulario local. Las palabras están clasificadas por su categoría gramatical y, dentro de cada categoría, ordenadas en clases de frecuencia, pero no se conoce cómo están asociadas entre ellas, ni en qué momento intervienen. Para este caso, se recomienda regresar al texto, o por lo menos, al contexto de las palabras para resolver ambigüedades y captar mejor el significado de las palabras locales. Esta contextualización conlleva grandes dificultades en los análisis completos del flujo del vocabulario debido al tiempo y esfuerzo requeridos, pues el proceso tendría que realizarse de forma manual y, en un corpus de grandes dimensiones, el seguimiento del vocabulario se volvería prácticamente imposible. Dada esta problemática, nuestro planteamiento consiste en proponer una nueva alternativa que permita, no sólo separar el vocabulario en dos grupos (estable y local), sino también en separar las palabras de uso necesario, llamadas palabras herramientas, que se ajustan a una distribución de Poisson. Asimismo, conocer la asociación de las palabras y el momento en que éstas intervienen y la parte o secuencia que caracterizan. Esto último es posible mediante las palabras características cronológicas. En la Tabla 4.1 se detallan los criterios establecidos en cada uno de los grupos en los que se divide el corpus.

Vocabulario	Índice de reparto del vocabulario (IRV)	Característico	Distribución
Local o especializado	$IRV \leq IPRVCF - \sigma$	Sí	
Regular o estable	$IRV \geq IPRVCF + \sigma$	No	
Vocabulario aleatorio o usual	$IPRVCF - \sigma > IRV < IPRVCF + \sigma$	Sí o No	Poisson

[1] IPRVCF: Índice promedio de reparto del vocabulario en la clase de frecuencia

Tabla 4.1: Criterios para dividir el corpus

(Tabla 4.1).

4.2 Modelo de evolución cronológica

4.2.1 Matriz reordenada de Bertin

Bertin (1973, 1977, 1981) aporta una herramienta simple y eficaz para el análisis gráfico de matrices de datos. La matriz reordenada de Bertin consiste en transformar

mediante un proceso de permutación de las filas o las columnas de la matriz inicial, hasta lograr una estructura más homogénea y organizada que proporcione información relevante para la toma de decisiones. Su método ha sido implementado en un paquete en el software R, con el nombre de bertin (Falguerolles et al., 1997).

Para exponer sus ideas, Bertin utiliza un ejemplo sobre ocupación hotelera y, mediante una pequeña historia, muestra cómo el director de un hotel encuentra la forma de representar sus datos como una matriz mejor organizada, fácil de visualizar y con información de interés (Figura 4.2).

En los últimos años los criterios de Bertin definidos por permutaciones empíricas han dado lugar al desarrollo de nuevos métodos de ordenación como: los métodos heurísticos, métodos estadísticos multivariados (AC y ACP), clasificación, etc. Todos intentan encontrar la mejor solución al problema, con el fin de revelar la mayor cantidad de información de los datos.

Los criterios de ordenación definidos por permutaciones por Bertin son también los criterios de los métodos estadísticos multivariados. La idea de Benzécri (1973, 1981) sugiere que las coordenadas proporcionadas por un análisis de correspondencias pueden utilizarse para la clasificación de las filas y columnas en una matriz de Bertin. Por lo tanto, la combinación de ambos métodos ha permitido el desarrollo de nuevos sistemas computacionales como: GAP (Chen, 2002), Bertifier (Perin et al., 2014) y el método de análisis de una matriz de datos (AMADO) (Chauchat y Risson, 1995; Risson, 1994). En este trabajo nos interesamos por el último.

4.2.2 Análisis de una matriz de datos (AMADO)

Los criterios de ordenación propuestos por Chauchat y Risson (1995), para hacer visible la estructura de una matriz de datos combinan dos enfoques:

- Análisis estadístico de datos multidimensionales : permiten encontrar las similitudes o relaciones entre las filas o las columnas.
- Gráficos de Bertin: proporcionan un complemento visual a las soluciones encontradas mediante los métodos factoriales (AC, ACP y clasificación)

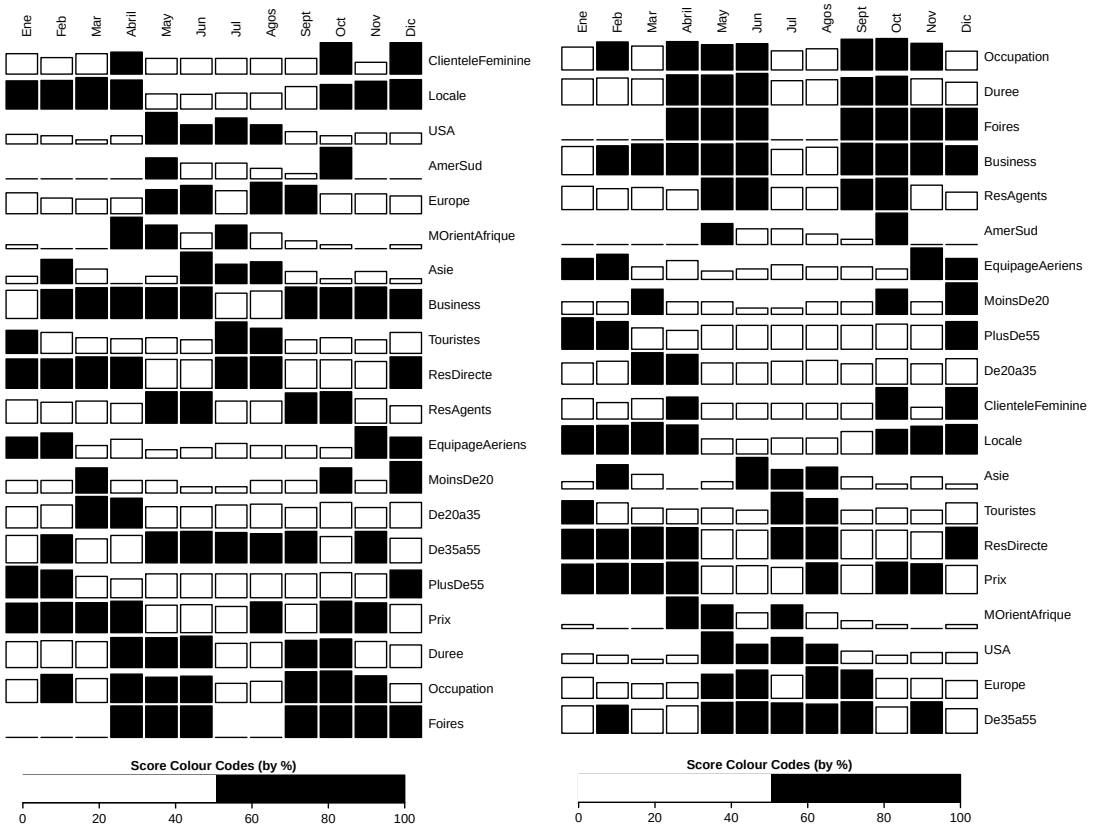
Es así como AMADO se convierte en un método potente de análisis para descubrir y mostrar similitudes y contrastes entre los elementos de una matriz. Este método fue implementado en el software SPAD (Lebart y Morineau, 1984).

4.2.3 AC para el estudio de trayectorias

AC se presenta como la herramienta esencial para poner de relieve la estructura evolutiva del corpus (Lebart et al., 2000, 1998; Murtagh, 2005; Murtagh et al., 2009, 2011). En el caso de un corpus dividido en partes o períodos homogéneos, el AC permite visualizar su estructura si éste está organizado en una tabla léxica agregada documentos $\times$ palabras. Los documentos representan las partes o períodos en que fue dividido el corpus y sus posiciones sobre el primer plano factorial conducen al efecto Guttman, lo cual implica una

	Ene	Feb	Mar	Abril	May	Jun	Jul	Agos	Sept	Oct	Nov	Dic
ClienteleFeminine	26	21	26	28	20	20	20	20	20	40	15	40
Locale	69	70	77	71	37	36	39	39	55	60	68	72
USA	7	6	3	6	23	14	19	14	9	6	8	8
AmerSud	0	0	0	0	8	6	6	4	2	12	0	0
Europe	20	15	14	15	23	27	22	30	27	19	19	17
MOrientAfrique	1	0	0	8	6	4	6	4	2	1	0	1
Asie	3	10	6	0	3	13	8	9	5	2	5	2
Business	78	80	85	86	85	87	70	76	87	85	87	80
Touristes	22	20	15	14	15	13	30	24	13	15	13	20
ResDirecte	70	70	75	74	69	68	74	75	68	68	64	75
ResAgents	20	18	19	17	27	27	19	19	26	27	21	15
EquipageAeriens	10	12	6	9	4	5	7	6	6	5	15	10
MoinsDe20	2	2	4	2	2	1	1	2	2	4	2	5
De20a35	25	27	37	35	25	25	27	28	24	30	24	30
De35a55	48	49	42	48	54	55	53	51	55	46	55	43
PlusDe55	25	22	17	15	19	19	19	19	19	20	19	22
Prix	163	167	166	174	152	155	145	170	157	174	165	156
Duree	2	2	2	2	2	2	2	2	2	2	2	1
Occupation	67	82	70	83	74	77	56	62	90	92	78	55
Foires	0	0	0	1	1	1	0	0	1	1	1	1

(a) matriz de datos



(b) Representación de los datos

(c) Matriz reordenada

Figura 4.2: Matriz reordenada de Bertin

trayectoria que forma un patrón de herradura en el primer plano principal del AC aplicado a la tabla documentos×palabras (Bécue-Bertaut et al., 2014). Este esquema evolutivo es el resultado típico cuando se analiza una tabla de datos agregados, caracterizada por una suave graduación en filas y columnas (Lebart et al., 2000, 1998). Esto debido a que dos partes consecutivas tienden a estar más próximas entre sí, en la medida en que contengan la misma frecuencia de las palabras, y éstas estarán más cerca entre ellas en la medida en que se asocian.

El fenómeno se presenta de manera más clara, cuando la tabla documentos×palabras es una tabla modelo con todos sus elementos nulos fuera de la banda diagonal y presenta los siguientes resultados: el segundo factor es una función polinómica de segundo grado del primer factor y los puntos se sitúan en el plano 1-2, exactamente sobre una parábola o patrón de herradura. Igualmente, el tercer factor es una función de tercer grado del primero y los puntos se sitúan en el plano 1-3 en una curva que corta tres veces al eje 1. (Figura 4.3). Todos los factores son funciones polinómicas del primer factor, de un grado cada vez más elevado. Este efecto persiste posiblemente en varios ejes y conduce a trayectorias específicas, detalladas en (Benzécri, 1973).

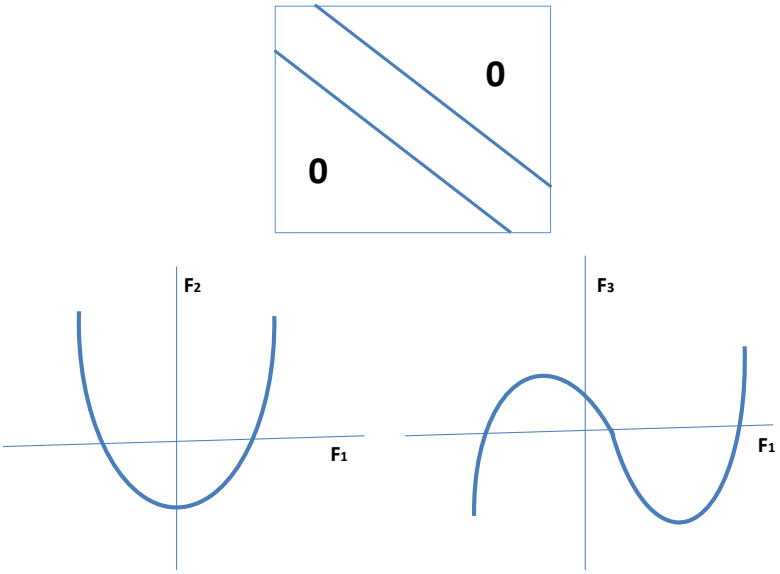


Figura 4.3: El efecto Guttman

En el caso de un efecto Guttman, cuando se presenta el patrón de herradura sobre el primer plano, el primer factor da una buena representación del conjunto de las distancias entre todas las parejas de puntos, pero no puede recoger correctamente las distancias entre perfiles; por lo tanto, es el segundo factor el que corrige y afina la aproximación de las distancias recogidas por el primer factor. Es decir, el que permite ver si algunos puntos se desvían del modelo (Escofier y Pagès, 1992)

4.2.4 Matriz ordenada: implementación en R

En este trabajo partimos de una tabla DocumentosXpalabras. Nuestro interés es visualizar de manera ordenada las palabras que caracterizan a cada documento para poner

de relieve una estructura que permita abordar de manera clara el flujo del vocabulario. La metodología propuesta sigue la estrategia implementada en AMADO e incorpora un nuevo procedimiento a través de las palabras características cronológicas. Es decir, cada palabra será ordenada por su coordenada en las primeras dimensiones del AC y también por el orden de acuerdo con el documento o segmento de documentos que caracteriza. El procedimiento consiste en:

- Aplicar un AC a la tabla DocumentosXPalabras.
- Buscar las palabras características cronológicas
- Representar con Gráficas de Bertin la tabla ordenada

Capítulo 5

Análisis cronológico de una base bibliográfica

5.1 Introducción

En este capítulo se ofrece una herramienta bibliométrica que permite seguir los cambios y las novedades que se producen en un determinado campo de investigación. La metodología estadística propuesta sigue la línea marcada por Bécue-Bertaut (2014), la cual combina varios métodos del análisis textual, tales como: análisis de correspondencias, clasificación jerárquica, palabras características, análisis factorial múltiple de tablas de contingencia, clasificación jerárquica con restricción de contigüidad y palabras jerárquicas e incrementos y decrementos específicos.

Como ejemplo, la metodología se aplica a una base de 506 resúmenes científicos, descargados de la base Medline, relativos al Lupus Eritematoso Sistémico (LES) y publicados en 115 revistas diferentes con alto factor de impacto durante un período de 18 años.

Existen métodos que permiten el estudio de textos. Sin embargo, en esta tesis proponemos un procedimiento metodológico desde un punto de vista cronológico, es decir, nuestro interés es no sólo mostrar las ventajas que ofrece la incorporación del tiempo en el análisis de textos sino también desarrollar una herramienta que incorpore la metodología propuesta.

Los resultados muestran los temas más relevantes en el conjunto de las publicaciones, la evolución temporal de los temas, los momentos en los cuales se producen cambios marcados y, finalmente, se identifican los artículos pioneros. Esta metodología, que se aplica a cualquier colección de artículos o abstracts, ha sido implementada en una función en R, MacroBiblio.En la Figura 5.1 se presenta el diagrama de flujo de la función y en en el capítulo 7 su aplicación.

5.2 Corpus Bibliográfico

5.2.1 Obtención de los datos

Los datos se descargaron de la base médica MEDLINE (for Biotechnology Information, 2012), a través de PUBMED. Los criterios que se usaron para seleccionar los abstracts fueron especificados por un grupo especializado en LES del hospital Clínico de Barcelona,

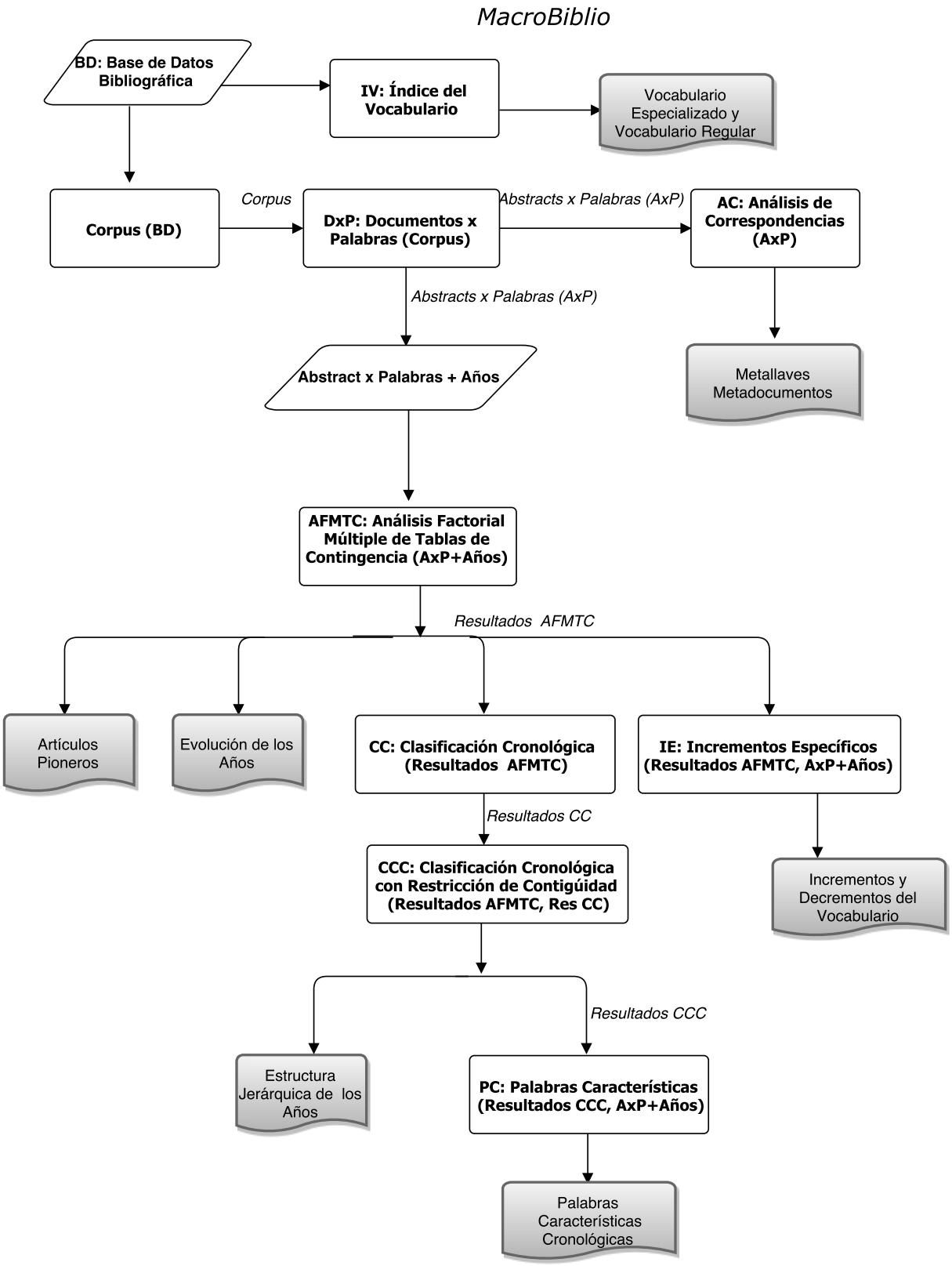


Figura 5.1: Diagrama de flujo

España. Estos son: que "LES" aparezca en el título, que esté basado en un ensayo clínico y que el abstract esté escrito en inglés (Tabla 5.1). Cumpliendo estos criterios fueron encontrados 506 abstracts, publicados entre enero de 1994 y diciembre de 2012, con el formato que se muestra en la Tabla 5.2.

Topic=(,(SLE[Title] OR systemiclupus erythematosus[Title])
AND "clinical trial"[Filter] AND "hasabstract"[Filter] AND "English"[Filter]
Refined by: Subject Areas=(systemic lupus erythematosus)
Refined by: Subject Areas=(systemic lupus erythematosus)
Publications = (Journals)
Languages= (English))

Tabla 5.1: Criterios de selección

PMID- 22466290
OWN - NLM
STAT- MEDLINE
DA - 20120403
DCOM- 20120724
LR - 20120831
IS - 1952-4005 (Electronic)
IS - 1148-5493 (Linking)
VI - 23
IP - 1
DP - 2012 Mar 1
TI - The -2518 A/G polymorphism in the monocyte chemoattractant protein 1 gene is associated with the risk of developing systemic lupus erythematosus in Argentinean patients a multicenter study.
PG - 7-11
AB - Systemic lupus erythematosus (SLE) is a systemic, autoimmune disorder. Monocyte chemoattractant protein 1 (MCP-1), a chemokine involved in the recruitment and migration of monocytes/macrophages, has been shown to be increased in the plasma of SLE patients. The aim of our study was to evaluate the possible association of the polymorphism -2518 of the MCP-1 gene with the risk of developing SLE, manifesting lupus nephritis (LN) and with other clinical features of SLE in an Argentinean population. A group of 171 SLE patients and 120 control subjects were examined. Genotypic and allelic frequencies of the MCP-1 -2518 A/G polymorphism showed significant differences between the SLE and the control groups (p=0.001 and p=0.01, respectively). However, the polymorphism showed no association with LN or with the other clinical variables studied. Our results suggest that the presence of the MCP-1 -2518 A/G polymorphism might be a risk factor for developing SLE in genetically predisposed individuals, but it does not seem to have a role in the evolution of the disease in the Argentinean population.
AD - Laboratorio de Hemostasia y Trombosis, Hospital de Infecciosas Dr. F. J. Muniz, Buenos Aires, Argentina. nanoaranda@hotmail.com
FAU - Aranda, Federico
AU - Aranda F
FAU - Wingeyer, Silvia Peres
AU - Wingeyer SP
FAU - Munoz, Sebastian Andres

Tabla 5.2: Formato de un abstract bajado de MEDLINE

Todos los abstracts utilizados en el estudio tienen título, autor, nombre de la revista y año de publicación. En la Figura 5.2 se muestra el número de abstracts, que se publicaron en cada año. En la Figura 5.3, el número de trabajos publicados por revista. Nótese que el

año con mayor número de publicaciones es 2007, con 42 artículos. Sólo 8 revistas publicaron más de 10 artículos relacionados con el tema que nos ocupa y, 5 de ellas, *Arthritis and Rheumatism, Lupus, The Journal of Rheumatology, Rheumatology Oxford, England and Annals of the Rheumatic Diseases*, concentran el 60 % de los artículos.

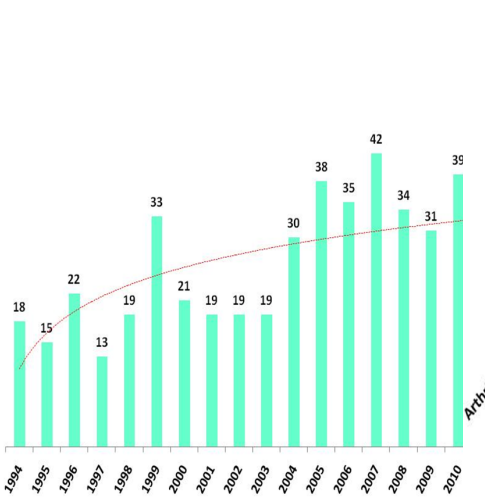


Figura 5.2: Artículo por año

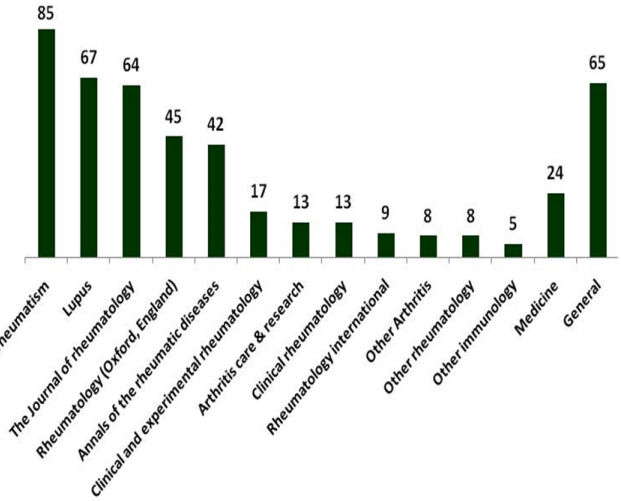


Figura 5.3: Artículos por revista

5.2.2 Características del corpus

El corpus inicial está formado por 89,155 ocurrencias, que corresponden a 6,276 palabras diferentes. Esto representa una longitud media de 176.20 ocurrencias por abstract. Las palabras con una frecuencia menor de 10 y/o, las que no se encontraron en, al menos, 5 abstracts, se eliminaron. Las preposiciones, las conjunciones, así como los pronombres personales y pronombres demostrativos también fueron eliminados. Este tratamiento permitió conseguir un corpus constituido por 51,904 ocurrencias, correspondientes a 1,120 palabras distintas, con el que se creó la matriz abstracts×palabras. En la Figura 5.4 se muestra la estructura de la tabla, compuesta de 506 filas y 1,120 columnas. Las filas representan los abstracts y las columnas, las palabras.

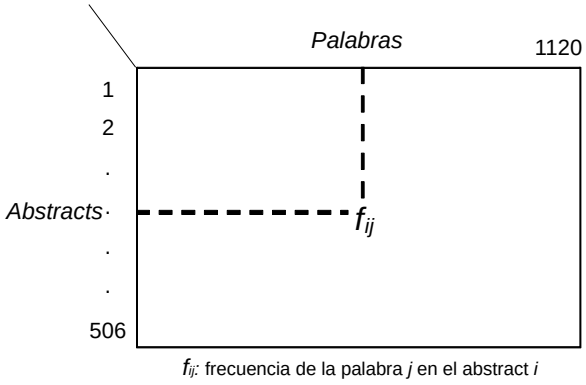


Figura 5.4: Matriz abstracts×palabras

5.2.3 Identificación de temas mediante el glosario

Glosario

La Tabla 5.3 muestra el glosario con las palabras más frecuentes. Para evitar duplicidad, se hicieron las equivalencias correspondientes, es decir, se agruparon las palabras cuyo significado es el mismo, las palabras que aparecen tanto en singular como en plural y las palabras que terminan en -mente. Por ejemplo: *female* con *women*; *male* con *men*; *patients* con *patient*; *diseases* con *disease*, entre otras.

Al analizar el glosario, del total de ocurrencias (51,904), 10 % se acumula en tan sólo diez palabras. No es extraño encontrar a *patient* como la palabra más frecuente, con 2,444 ocurrencias. Después se encuentran las palabras que definen al LES: *sle* (2,248), *lupus* (321) y *disease* (904). También *women*(242) se encuentra entre las palabras más frecuentes. Al comparar esta palabra con *men*, observamos que el número de veces que se menciona a las mujeres es cuatro veces mayor que la de los hombres. Esto coincide con investigaciones realizadas anteriormente: el LES es mucho más frecuente en mujeres que en hombres.

Palabra	Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia
patient	2444	methotrexate	67	blys	30
sle	2248	plasma	64	atherosclerosis	28
disease	904	organ	61	hypertension	27
lupus	321	immunosuppressive	60	creatinine	27
antibody	310	belimumab	57	cutaneous	27
placebo	248	neuropsychiatric	56	pulmonary	27
women	242	bmd	56	anticardiolipin	26
sledai	204	blind	54	azathioprine	25
damage	178	bcells	54	hydroxychloroquine	25
renal	159	depletion	54	lipoprotein	24
cohort	150	fatigue	53	snps	24
dhea	124	cardiovascular	48	protocol	23
manifestation	120	cholesterol	47	depression	23
prednisone	109	cell	46	interferon	23
cyclophosphamide	101	vascular	46	polymerase	22
gene	99	complement	45	thrombocytopenia	22
allele	96	tumornecrosisfactor	45	urinary	22
dna	93	proteinuria	43	platelet	22
blood	92	lymphocytes	42	april	22
steroid	90	tcells	41	clinicaltrial	21
polymorphism	90	genetic	40	interleukin	21
flares	84	abnormalities	40	cytotoxic	21
bilag	84	bone	39	fever	21
nephritis	81	intervention	39	antigen	20
corticosteroid	80	completed	34	anticoagulant	20
systemic	79	rheumatoidarthritis	32	antimalarial	20
symptom	77	estrogen	32	illness	20
rituximab	72	thrombosis	31	pga	20
rheumatology	69	lung	31	pressure	20
genotype	68	arthritis	30	chloroquine	20

Tabla 5.3: Glosario

Temas

Las palabras referentes a LES pueden ser clasificadas según seis grupos (Tabla 5.4): síntomas, etiología, diagnóstico, tratamiento, prognosis y epidemiología. Considerando esta clasificación, una palabra puede pertenecer a más de un grupo.

Síntomas		Dianóstico/ Prognóstico	Etiología	Tratamiento	Epidemiología
brain	**renal**	biopsy	antiphospholipid	*methotrexate*	*women*
pulmonary	**nervous**	prognosis	dsdna	*cytotoxic*	*race*
thrombocytopenia	**hypertension**	prognostic	dna	*dietary*	*ethnic*
chest	**urinary**	resonance	hormonal	*cyclophosphamide*	*cohort*
articular	**cholesterol**	magnetic	atidna	*prednisone*	*african*
vasculitis	**cardiovascular**	disease	lipoprotein	*dose*	*age*
blind	**carciac**	illness	hormone	*antibodies*	*male*
joint	**glomerulonephritis**	bilag	genotype	*drug*	*hispanic*
defect	**chronic**	april	cell	*intravenous*	*hrqlo*
symptoms	**systolic**	plama	B-Cell	*hydroxychloroquine*	*ethnicity*
nephropath	**arthritis**	vascular	estrogen	*anticardiolipin*	*caucasian*
abnormalities	**depletion**	rheumatology	antibody	*therapeutic*	*gender*
lung	**inflammation**	blood	allele	*immunosuppressive*	*children*
cutaneous	**neuropsychiatric**	bmd	pathogenesis	*corticosteroid*	*multiethnic*
platelet	**bone**	calcium	snps	*chloroquine*	*ethnically*
pressure	**damage**	sledai	gene	*azathioprine*	*population*
toxicity	**rheum**	died	polymorphism	*antigens*	*person*
urine	**coronary**	evaluation	genetic	*immunoregularors*	*human*
manifestations	**thrombosis**	infection	genotiped	*rituximab*	*patient*
anaemia	**atherosclerosis**	chronic	polymerase	*belimumab*	*candidate*
		mortality	haplotype	*blys*	*adult*
		predictive		*pharmacokinetics*	*case-control*

Tabla 5.4: Temas definidos mediante el glosario

5.3 AC: Metallaves y Metadocumentos

5.3.1 Identificación de temas mediante Metallaves

Uno de los objetivos que se persigue en este estudio es poner de relieve los diferentes temas encontrados en una bibliografía científica. Mediante la asociación de las palabras es posible detectar cambios, avances y novedades dentro de una base de datos textual de gran tamaño.

El método de análisis de correspondencias (Benzécri, 1973, 1981; Lebart et al., 1998; Murtagh, 2005) busca similitudes entre documentos mediante la relación de las palabras, es decir, los documentos que usan las mismas palabras están íntimamente relacionados.

Aplicando un AC, a la tabla abstract×palabras (506 abstracts y 1120 palabras), se encontraron los metallaves y los metadocumentos asociados a los ejes principales. Los dos primeros ejes, con valores propios iguales a 0.31 y 0.28, explican una porción pequeña de la inercia global (en conjunto, 2.84 %). Este bajo porcentaje, típico cuando se trabaja con matrices muy grandes, está frecuentemente asociado con una estructura satisfactoria de los datos, como fue comentado por Lebart et al. (1998).

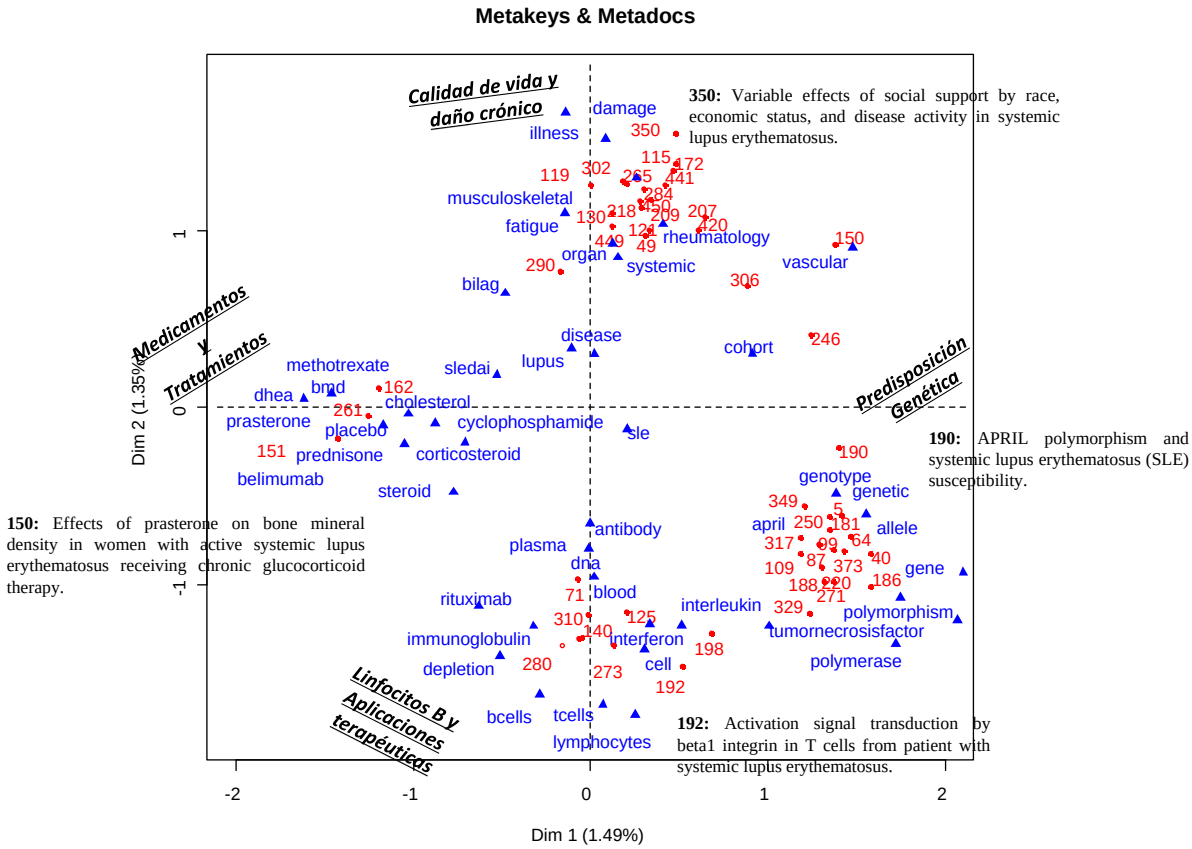


Figura 5.5: Representación de las palabras y abstracts con mayor contribución en el primer plano principal de AC, asociados a un tema

En la Figura 5.5 se muestran los *"metallaves"* y los *"metadocumentos"* que caracterizan el primer plano factorial del AC. Los *metallaves* y los *metadocumentos* reúnen las palabras y abstracts cuya contribución es más de 6 veces la contribución media de su respectivo grupo en los dos primeros ejes.

En la Figura 5.5 se pueden apreciar cuatro nubes de palabras y de abstracts. En el primer eje, parte negativa, un conjunto de palabras (o *"metallave1-"*) relacionadas con medicamentos-tratamientos, se contrapone con *metallave1+*, parte positiva del primer eje, predisposición genética. Estos dos *metallaves* (medicamentos-tratamientos y predisposición génetica) están, a su vez, enfrentados con el segundo eje donde se localizan los *metallaves*: linfocitos B y aplicaciones terapéuticas (*metallave2-*), parte negativa del segundo eje y calidad de vida y daño crónico (*metallave2-*), parte positiva. Estas oposiciones también se encuentran en los *metadocumentos*, o conjuntos de abstracts con mayor contribución. Estos abstracts se identifican por sus números en la gráfica.

Los abstracts pertenecientes a un *metadocumento* usan palabras pertenecientes al *metallave* asociado. Esto permite localizar fácilmente los abstracts relacionados con los temas identificados por los *metallaves*

En los sucesivos ejes se pueden identificar otros temas tales como: problemas de

DIM	Tema	Metallave	Metadocumento
1+	Predisposición genética	ASSOCIATION ALLELE GENE SUSCEPTIBILITY POLYMORPHISM POLYMORPHISMS HLA ASSOCIATED CONTROLS SNPS RISK SLE GENETIC ALLELES GENOTYPES GENES expression genotype genotyped cohort associations african european beta case acl april healthy found tnf vascular population odds	181 190 186 87 99
1-	Medicamentos y tratamientos	PLACEBO GROUP TREATMENT BMD PRASTERONE DHEA MONTHS DAY DOSE weeks randomized calcium prednisone spine therapy trial blind methotrexate efficacy week flare received double cholesterol belimumab month	151 162 261
2+	Calidad de vida, actividad de la enfermedad y daño crónico	DAMAGE HEALTH DISEASE PHYSICAL HRQOL SDI SOCIAL FACTORS quality status scores self activity mental american csle psychosocial canada life costs rheumatology slicc duration international fatigue outcomes race support college diagnosis clinics acr variables organ socioeconomic visits	350 441 172 119 115 130 450 284 290 209 193 302
2-	Linfocito B: etiopatogenia y aplicaciones terapéuticas	CELLS CELL EXPRESSION BETA DEPLETION RITUXIMAB LYMPHOCYTES ANTI anti gene blood activation serum dna levels treatment peripheral proliferation antibody binding normal	192 280 310
3+	Metabolismo óseo,	CALCIUM BMD SPINE ALLELE PRASTERONE POLYMORPHISM CHOLESTEROL GENE SNPS POLYMORPHISMS SUSCEPTIBILITY GROUP DENSITY HIP hdl association functional control bone intervention premenopausal genotypes controls genotype dietary groups dhea tnf genotyped significant april genes placebo program	162 343 151 261 181
3-	Afectación clínica de los principales órganos	SPECT CNS MRI PERFUSION SYNDROME ANTI ANTIBODIES diagnostic dsdna abnormalities positive patients acl involvement abnormal cyclophosphamide renal manifestations	278 83 480 113 318 364
4+	Terapias biológicas	ANTI CELLS CELL ACTIVITY DSDNA DEPLETION BILAG RITUXIMAB BELIMUMAB antibody dna damage csle disease	310 119 275 13 277
5-	Riesgo cardiovascular	VASCULAR EVENTS SEIZURES SPINE ACL OCCURRENCE RISK HLA ANTIBODIES PRASTERONE PREMENOPAUSAL THROMBOSIS alone renal mortality cohort taking cerebrovascular anti african bone pregnancy	162 150 306 151

Tabla 5.5: Metallaves/Metadocumentos (palabras y abstracts con mayor contribución)

metabolismo óseo, consecuencias clínicas, epidemiología, terapias biológicas, etcétera, que son identificados en otros ejes. En la Tabla 5.5 se muestran: todos los temas que fueron identificados en los primeros 5 ejes. Para cada uno aparece la lista de las palabras que los conforman. Las que cuentan con una contribución mayor a seis veces, la contribución

media, aparecen en mayúscula y, las de contribución menor que seis y mayor que tres veces la contribución media, con letra normal. A cada grupo se asocia un tema. En la parte de la derecha está el identificador de los artículos que están asociados a cada tema, con una contribución mayor a seis veces la contribución media.

5.4 Evolución cronológica del vocabulario

5.4.1 Importancia del AFMTC en la cronología

La incorporación del Análisis Factorial Múltiple de Tablas de Contingencia (MFACT; (Bécue-Bertaut y Pagès, 2004, 2008)) permite identificar tendencias, cambios bruscos y períodos homogéneos, caracterizados por el vocabulario. Para hacer el seguimiento de la evolución del vocabulario, se realiza una representación de las palabras y los abstracts, como en el AC, pero teniendo en cuenta tanto la cronología como la ocurrencia de las palabras.

Dadas las bondades que ofrece AFMTC con respecto a AC en el estudio de corpus cronológicos, se realiza un comparativo entre ambos métodos, con el fin de resaltar la importancia del AFMTC en estudios de esta naturaleza.

5.4.2 Relación con la cronología: AC y AFMTC

En el análisis anterior, donde se realizó un AC a la tabla abstracts×palabras, se yuxtapone la columna año, como una columna cuantitativa suplementaria. La cronología (años) tiene una correlación débil con los primeros ejes: 0.06 con el primer eje; 0.03 y 0.07 con el segundo y tercer eje, respectivamente. Considerando tan sólo 10 ejes, la calidad de la proyección de los años es de apenas 0.46. Este análisis, aunque tiene importancia en la búsqueda de los principales temas, como se mostró en la sección anterior 5.3, no explica la evolución del vocabulario en el tiempo, porque la influencia de la cronología se disemina en muchos ejes.

Para seguir la evolución o los cambios en la investigación, es necesario introducir la cronología, para lo cual el método adecuado es el AFMTC, porque incorpora la cronología como una variable activa, y describir las partes o períodos del corpus tanto por su vocabulario como por su cronología.

5.4.3 AFMTC : análisis globlal y parcial de la tabla múltiple

Características principales del análisis global en el primer plano

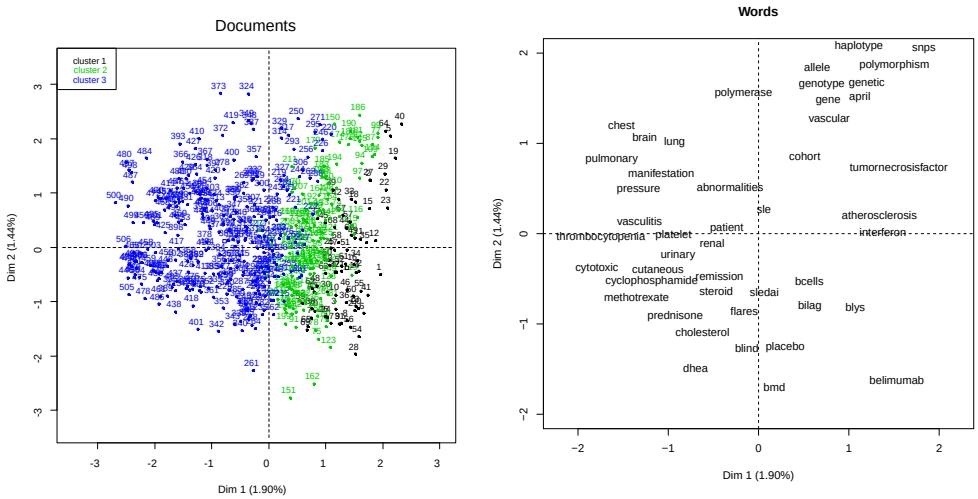
AFMTC se aplica a la tabla múltiple formada por la tabla abstracts×palabras (de dimensión 506 x 1120) y dos columnas años-cuantitativa y años-categórica. Esto permite representar los abstracts de acuerdo con el vocabulario y la cronología. AFMTC proporciona una representación global de las palabras en el primer plano. Éste cuenta con casi tanta variabilidad en la representación de las palabras como un CA el cual, por construcción, ofrece la representación óptima del plano de dicha variabilidad (Tabla 5.6). El primer valor propio de AFMTC es igual a 1.34. De acuerdo con la prueba de permutación (p-valor=

0.000 obtenido de 1000 replicaciones), el máximo valor es igual a 2. Esto significa que el primer valor propio aún está lejos de la aleatoriedad (Escofier y Pagès, 1992).

	Eje 1 (%)	*Eje 2 (%)*	*Acumulada (%)*
AC	1.49	1.35	2.84
AFMTC	0.67	1.65	2.32

Tabla 5.6: Proporción de la inercia de la nube de palabras explicada por los dos primeros ejes del AC y AFMTC

Respecto a la cronología, la alta correlación del primer eje de MFACT con el año de publicación *(corr= 0.94)* sugiere que este eje representa la variabilidad del vocabulario relacionado con el tiempo. Por otro lado, en el primer eje, las palabras contribuyen en un 30.1 %, y el año contribuye en un 69.9 % de la inercia.



(a) Abstracts (b) Extracto de palabras

Figura 5.6: Representación global de los abstracts y de un extracto de las palabras con mayor contribución en el primer plano principal del MFACT.

La Figura 5.6 muestra la representación global de las palabras y los abstracts en el primer plano principal. Éste enfrenta palabras relacionadas con síntomas y medicamentos (en el lado negativo) contra palabras utilizadas en el estudio de las causas de la enfermedad (desde dos puntos de vista: aspectos genéticos y aspectos hormonales), proyectadas en el lado positivo. Por otra parte, los años (columna años-categórica) son proyectados en el primer eje desde un punto de vista global. Como era de esperar, aunque estén separadas por intervalos de diferente longitud, se encuentran ordenados en el primer eje, como se puede ver en la Figura 5.7.

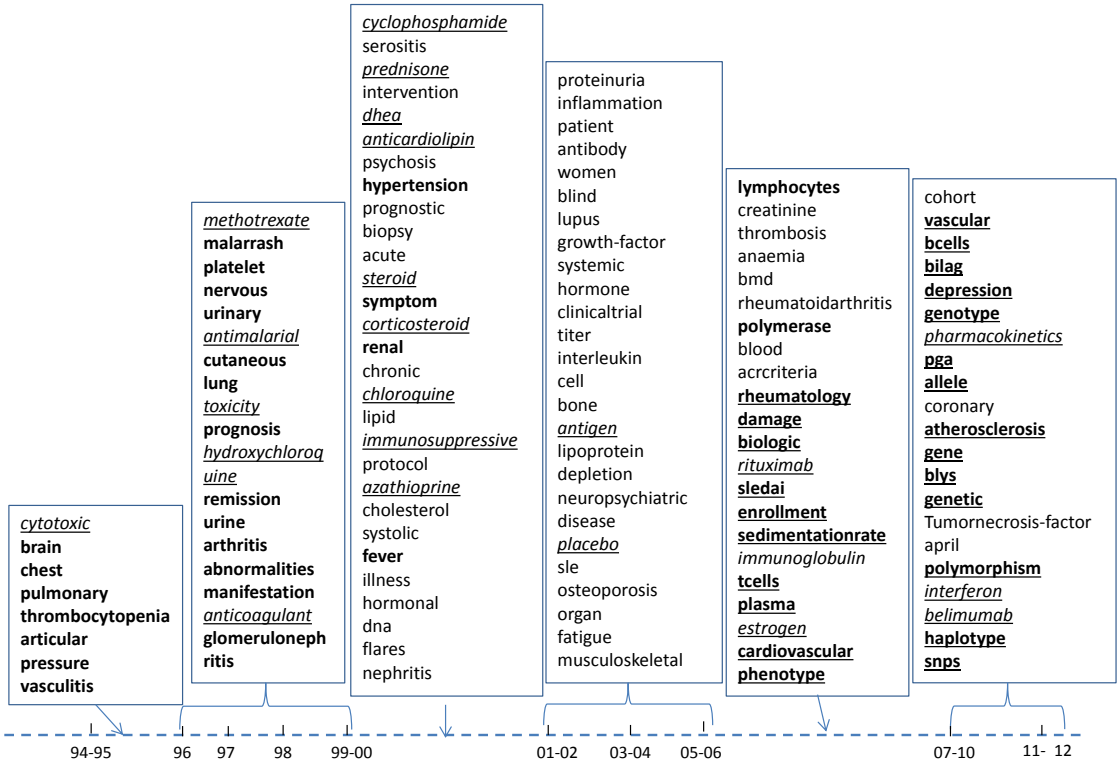


Figura 5.7: Representación global de las palabras y años-categórica según sus coordenadas en el primer eje de AFMTC.

Representación parcial de año-categórica desde un punto de vista del vocabulario

La representación parcial de año-categórica, desde el punto de vista del vocabulario (Figura 5.8), es interesante porque muestra la trayectoria de los abstracts agrupados por años. Esta trayectoria presenta cambios notables que corresponden a la publicación de trabajos novedosos que han incorporado palabras nuevas en los años que muestran cambios.

Aunque el corpus es relativamente pequeño para este tipo de estudio, se pueden identificar tendencias generales. De acuerdo a los cambios presentados en la trayectoria es de interés encontrar si los años están agrupados en períodos de tiempo. Esto indica que la investigación se puede precisar en períodos concretamente definidos.

5.4.4 Clasificación cronológica: períodos homogéneos

A través del algoritmo de agregación CC se identificaron discontinuidades entre los años. Este algoritmo garantiza, mediante un test de permutaciones, que sólo los años que comparten un vocabulario homogéneo se agrupen en clases. Con un nivel de significación de $\propto = 0.15$, los puntos de cambio identificaron en este análisis se dividen en tres períodos homogéneos (1994-2000, 2001-2006 y 2007-2012) y se pueden apreciar en la Figura 5.8.

De 1994 a 2000, se observan pocos cambios en el vocabulario. Una transición se produce en 2001, marcada por la introducción de nuevo vocabulario que se mantiene hasta

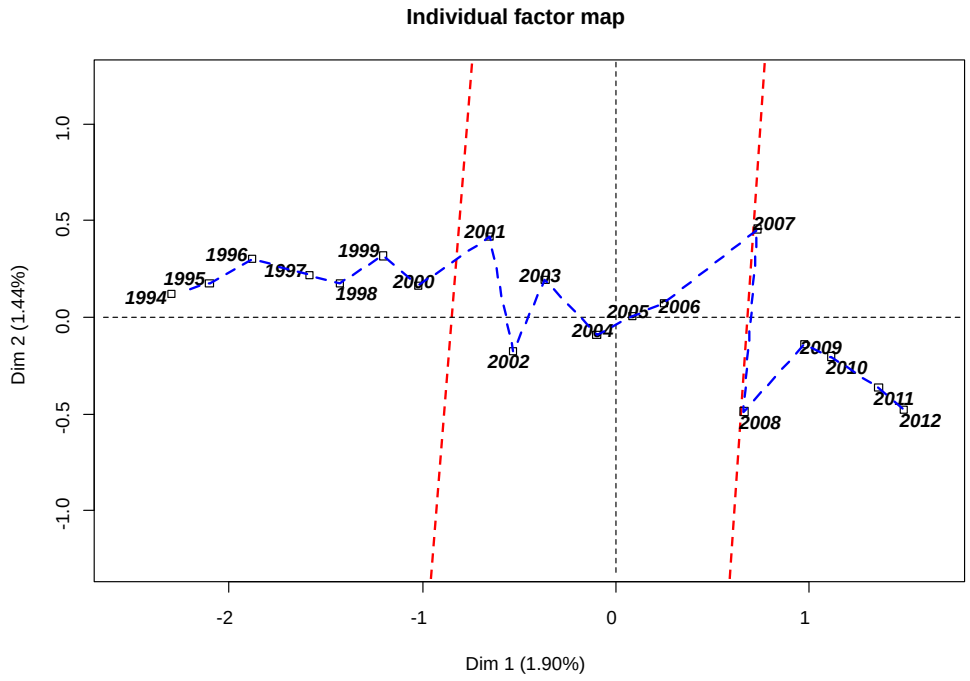


Figura 5.8: La representación parcial de la columna suplementaria años-categórica sólo desde el punto de vista del vocabulario en el primer plano principal del AFMTC

2002, como se muestra en la siguiente sección. Este vocabulario se utilizó hasta 2006. Otro cambio aparece entre 2006 y 2007. Este último año sobresale como un año marcado por la renovación de nuevo vocabulario, que es utilizado hasta que finaliza el período (2007-2012).

Palabras características e incrementos específicos de los períodos homogéneos

La segmentación obtenida con CC (Figura 5.8) y comentada anteriormente se debe conservar y los períodos 1994-2000, 2001-2006 y 2007-2012 se describen por sus características léxicas, permitiendo la caracterización de la evolución del vocabulario.

La Figura 5.9 muestra cómo los estudios médicos relacionados con el lupus se han diversificado desde 1994 a 2012. El primer período (1994-2000) está caracterizado por las palabras que engloban el estudio de la sintomatología u órganos afectados por la enfermedad (*manifestations* (manifestaciones), *brain* (cerebro), *hypertension* (hipertensión), *cutaneous* (cutáneo), *lung* (pulmón), *nervous* (nervioso), etc), así como por los médicamentos paliativos de estos síntomas (*prednisone*, *cytotoxic*, *dhea*, *cyclophosphamide*, *antimalarials*). En 2001-2006 aparecen nuevos medicamentos y hay un cambio en la investigación. Esta parece focalizar su interés principalmente en las mujeres (*women*). Hoy en día sabemos que es la mujer la que más padece de lupus.

Las palabras introducidas en 2007 hacen que este año se aleje un poco del resto de su grupo porque presenta una peculiaridad importante: es el año donde se llevaron a cabo estudios clínicos de cohortes y casos-control y se considera el inicio del estudio de nuevos tópicos relacionados con la etiología. Lo anterior se refuerza con la segmentación en dos

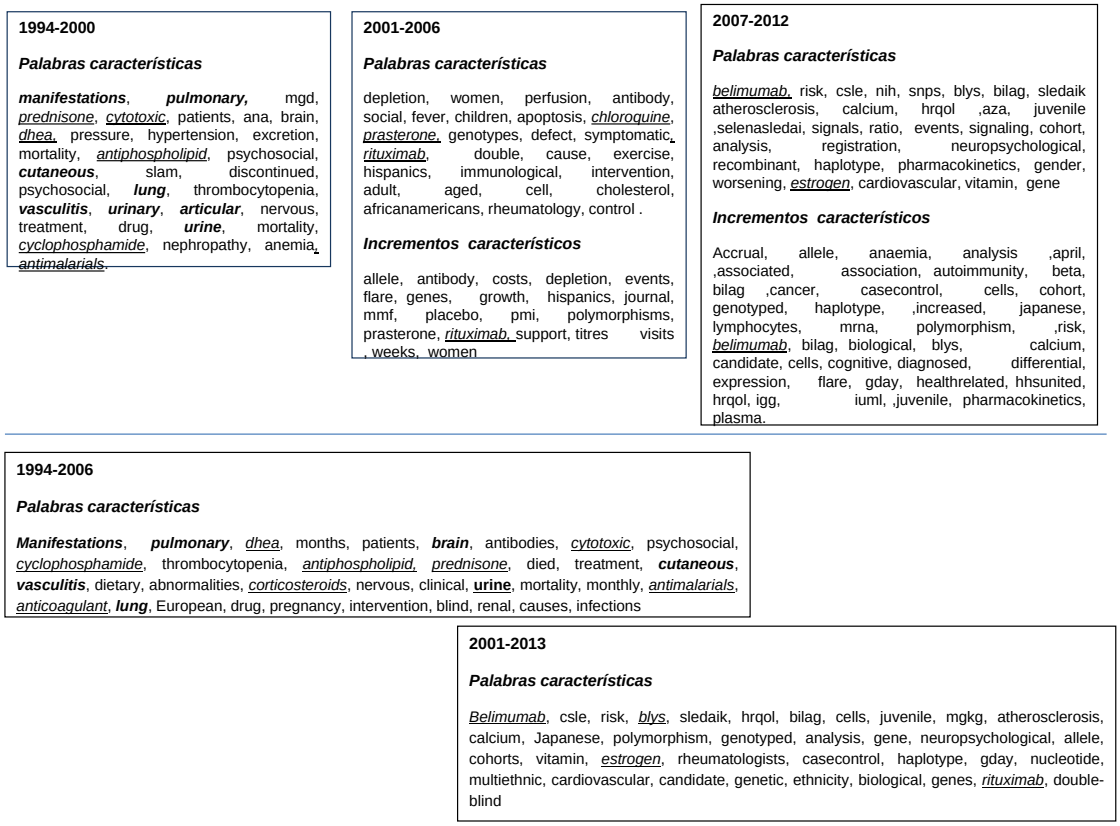


Figura 5.9: Palabras características e incrementos léxicos específicos de los períodos homogéneos

bloques (1994-2006 y 2001-2012).

A pesar de que algunos tópicos de interés anteriores a 2001 se mantuvieron en los años siguientes, hay más homogeneidad en el vocabulario observado de 2001 a 2012 que el de 1994 y 2006, como se puede deducir por el mayor número de palabras características entre 2001 y 2012 que el de 1994 y 2006 (Figura 5.8). Esta observación se ve reforzada por la gran similitud entre los incrementos léxicos observados en 2001-2006 y 2007-2012. Por ejemplo, *Belimumab* (medicamento que pertenece a la familia de los medicamentos llamados terapias biológicas) aparece entre 2001 y 2006 (2 citas) y su uso aumenta entre 2007 y 2012 (55 citas). Las palabras *gene* (gen) y genes se usan 4 veces entre 1994-2000, 42 veces entre 2001-2006 y 70 veces entre 2007-2012; la palabra *women* (mujeres), en algunos casos citada como *female* (hembra), son usadas 55 veces en 1994-2000, 93 entre 2001-2006 y 127 entre 2007-2012.

5.4.5 Clasificación con restricción de contigüidad: Estructura y Palabras jerárquicas

La clasificación es congruente con la información mostrada en secciones anteriores. Sin embargo, la lectura, por una parte del árbol y por otra, de las palabras cronológicas jerárquicas, hacen difícil el seguimiento del desarrollo de investigación. Para solucionar este problema Bécue-Bertaut et al. (2014) propuso un árbol etiquetado.

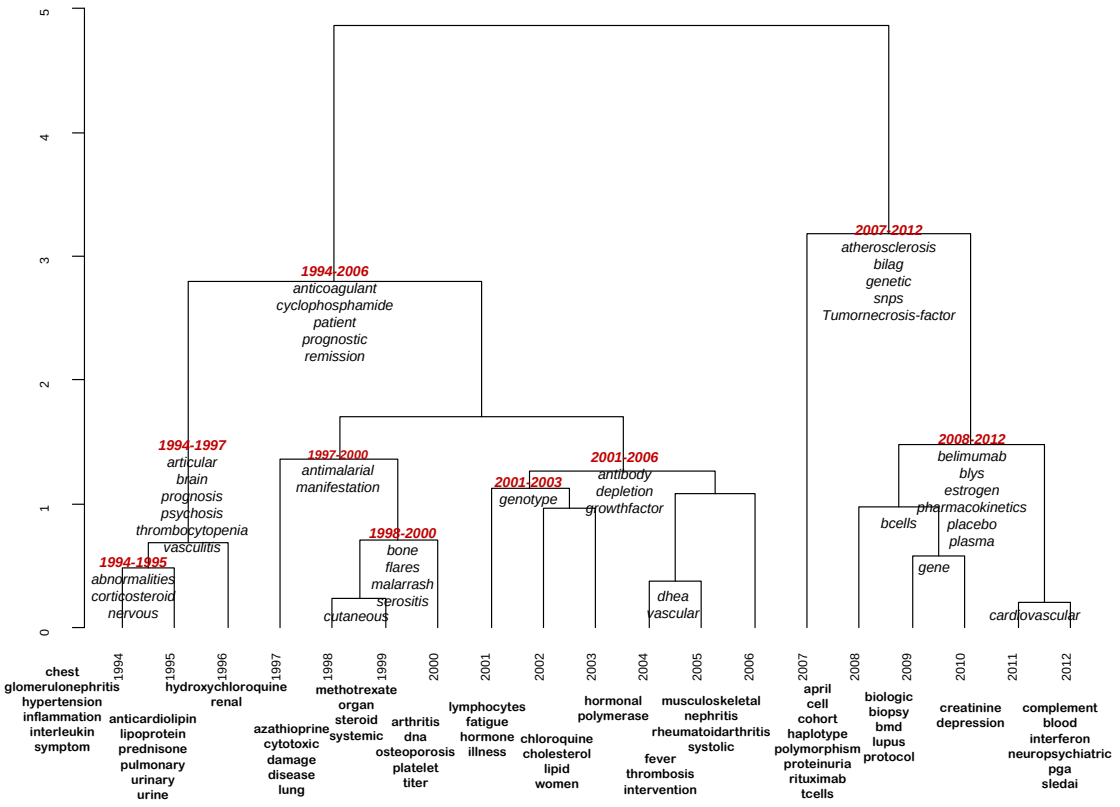


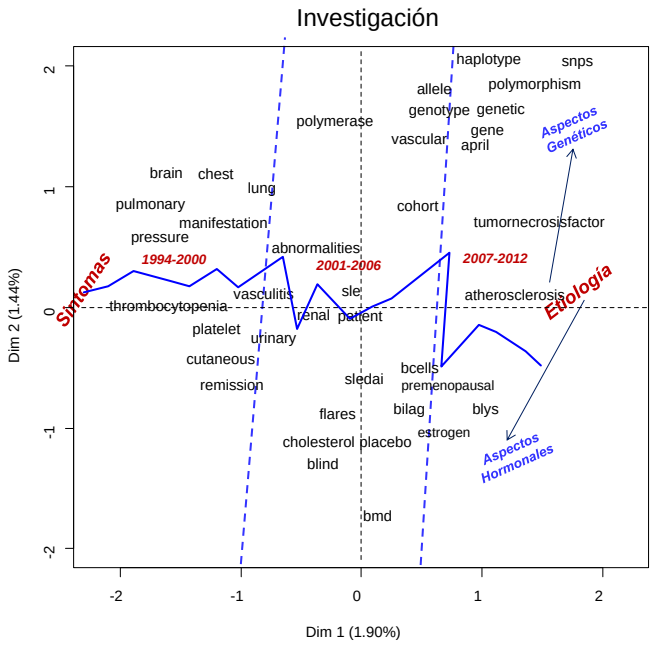
Figura 5.10: Evolución del vocabulario a través del árbol etiquetado

La representación en forma de árbol etiquetado (Figura 5.10) proporciona información de gran valor sobre la evolución del vocabulario a través del flujo de las palabras en cada período de tiempo.

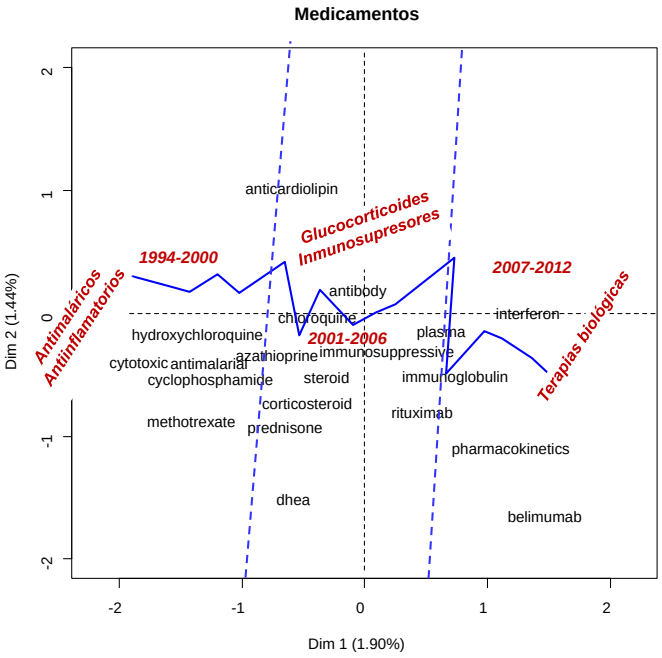
De acuerdo con la Figura 5.10 entre 1994 y 2000, el tema más tratado fue la sintomatología y, los tratamientos se basaron casi exclusivamente en medicamentos anti-maláricos (cloroquina e hidrocloroquina) y anti-inflamatorios (metotrexato y ciclofosmamida); de 2001 a 2006 se introdujeron nuevas palabras que reflejaban novedades en investigación. Se inició con el estudio de las causas, donde la mujer jugó un papel preponderante en la investigación; también en este período, se hizo uso de glucocorticoides (prednisona) e inmuno-supresores (azatropina y praesterona) como tratamiento parcial. Entre 2007 y 2012, la investigación se centró en aspectos genéticos y hormonales y se utilizaron terapias biológicas (rituximab y belimumab).

Actualmente, la mayor parte de las investigaciones se han centrado en las hormonas y su acción sobre el metabolismo óseo, especialmente en mujeres post-menopáusicas. Los últimos avances se han relacionado con un nuevo medicamento (*Belimumab*) el cual tiene una actividad inmuno supresora e inmuno reguladora sobre los linfocitos B. Parece ser que las investigaciones futuras se centrarán en medicamentos con una actividad similar.

Como conclusión del estudio, se considera que los cambios son impulsados por múltiples factores, en este caso, de acuerdo con los resultados analizados en cada una de las secciones y en conjunto, se determinó que los cambios en el léxico son debidos, principal-



(a) Evolución en la investigación



(b) Evolución en los medicamentos

Figura 5.11: Representacion global de un extracto de palabras (investigación y medicamentos) y períodos homogéneos

mente, al desarrollo en la investigación y a la presencia de nuevos medicamentos (Figura 5.11).

5.5 Artículos pioneros

La representación superpuesta de los puntos parciales de los abstracts, representa aquellos que desde el punto de vista de su vocabulario, están adelantados a su fecha de publicación (Figura 5.12).

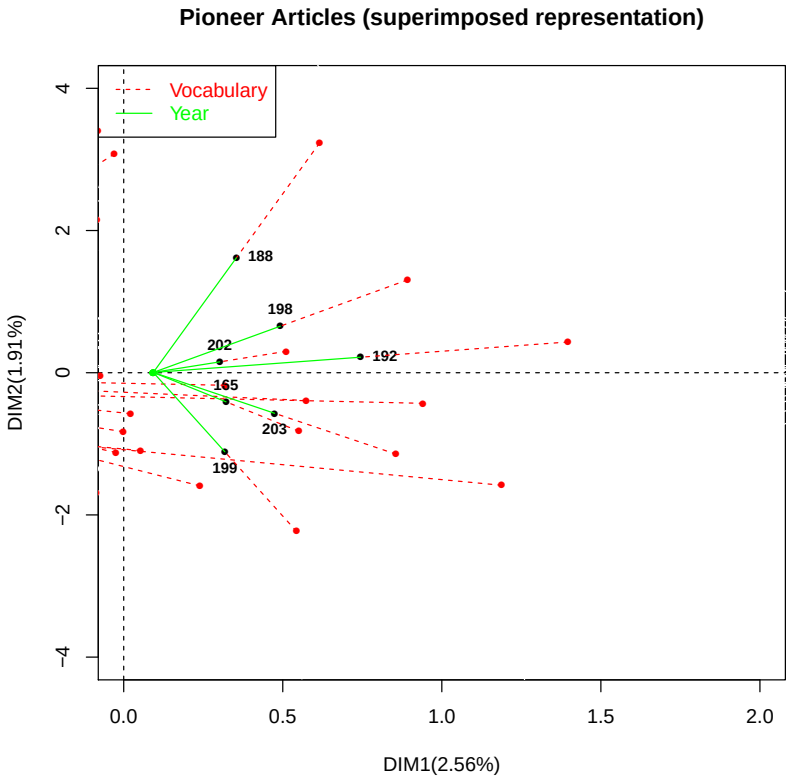


Figura 5.12: Representación de los abstracts de 2005-2010 adelantados a su tiempo en el primer plano principal de AFMTC.

Mientras que la columna de los años contribuyen un 69.9 % de la inercia del primer eje, el conjunto de palabras contribuye solamente en un 30.1 %. Esta desigualdad en la contribución de ambos conjuntos hace que los puntos parciales de la "cronología" estén más alejados del centroide en el primer eje que los puntos parciales del "vocabulario". Por lo tanto, esto es lo que permite que los abstracts tiendan a presentar un punto parcial del "vocabulario" por delante del punto parcial de la"cronología". Por el contrario, este efecto funciona de manera opuesta en el caso de los abstracts recientes.

La Tabla 5.7 proporciona el título, revista y año de los abstracts más "adelantados a la fecha" entre 2005 y 2010. Examinar su contenido facilita la comprensión de los avances en la investigación. Los abstracts "adelantados a la fecha", corresponden a innovaciones. Los abstracts más recientes no pueden ser seleccionados con este criterio (no se ha encontrado ningún trabajo pionero en 2011 o 2012)

Nº	Título	Año	Revista
188	Structural insertion/deletion variation in IRF5 is associated with a risk haplotype and defines the precise IRF5 isoforms expressed in systemic lupus erythematosus (Kozyrev et al., 2007).	2007	Arthritis and rheumatism
192	Activation signal transduction by beta1 integrin in T cells from patient with systemic lupus erythematosus (Nakayamada et al., 2007).	2007	Arthritis and rheumatism
198	A multicenter phase I/II trial of rituximab for refractory systemic lupus erythematosus (Tanaka et al., 2007).	2007	Modern rheumatology
202	Deficient CD4+CD25high T regulatory cell function in patient with active systemic lupus erythematosus (Valencia et al., 2007).	2007	Journal of immunology

Tabla 5.7: Extracto de los abstracts "adelantados a la fecha" de 2005 a 2010

5.6 Conclusiones

La metodología propuesta presenta avances importantes en el estudio de bibliografía científica. Nuestra contribución innovadora es la incorporación de AFMTC, clasificación cronológica, clasificación jerárquica con restricción de contigüidad, incrementos característicos y palabras jerárquicas. La aplicación de estos métodos en estudios bibliográficos es útil porque:

- permite identificar rápidamente los temas principales.
- se encontró que cambios en el vocabulario indican cambios en la investigación
- se identifican períodos homogéneos
- se identifican artículos pioneros

Esta metodología está disponible para los profesionales, investigadores y académicos, que están interesados en sacar información relevante de grandes conjuntos de datos.

Con el fin de obtener los resultados, MacroBiblio (capítulo 7) fue programada para ayudar en el análisis.

Capítulo 6

Análisis cronológico de un texto retórico

6.1 Introducción

Los objetivos y naturaleza del corpus a analizar varían de un estudio a otro. En el capítulo anterior se analizó una base de artículos científicos, pero los métodos expuestos son también adecuados para el estudio de corpus no estructurados como: discursos políticos, críticas literarias, ensayos, entrevistas, artículos de períodicos, cuentos, etc. Este tipo de corpus textuales son el objeto de estudio en este análisis.

El objetivo de este capítulo es proporcionar una herramienta metodológica para poner de relieve la estructura de un texto no estructurado y aportar información relevante sobre su construcción y el uso de su vocabulario.

La metodología propuesta combina varios métodos y se ilustra mediante la aplicación a un discurso retórico, con el fin de mostrar que la estadística textual, mediante el uso de métodos multidimensionales, ofrece grandes beneficios en el análisis de corpus no estructurados.

Bécue-Bertaut et al. (2014) presentan una secuencia de pasos. Continuando en la línea marcada en este trabajo, se propone lo siguiente:

- Dividir el discurso en frases artificialmente homogéneas
- Agrupar las frases en partes homogéneas a través de clasificación cronológica (sección 3.2.2).
- Visualizar la trayectoria de las partes homogéneas mediante análisis de correspondencias (sección 4.2.3).
- Descubrir la estructura jerárquica del discurso a través de clasificación con restricción de contigüidad temporal (sección 3.2.1).
- Seguir el flujo de la argumentación mediante la extracción de las palabras características, tanto en las partes como en los nodos de la jerarquía (sección 4.1.2).

Para facilitar el análisis de este tipo de textos, llamados no estructurados, se programó la función MacroTxChrono. En la Figura 6.1 se presenta el diagrama de flujo de la función y en el capítulo 7 se realiza su implementación.

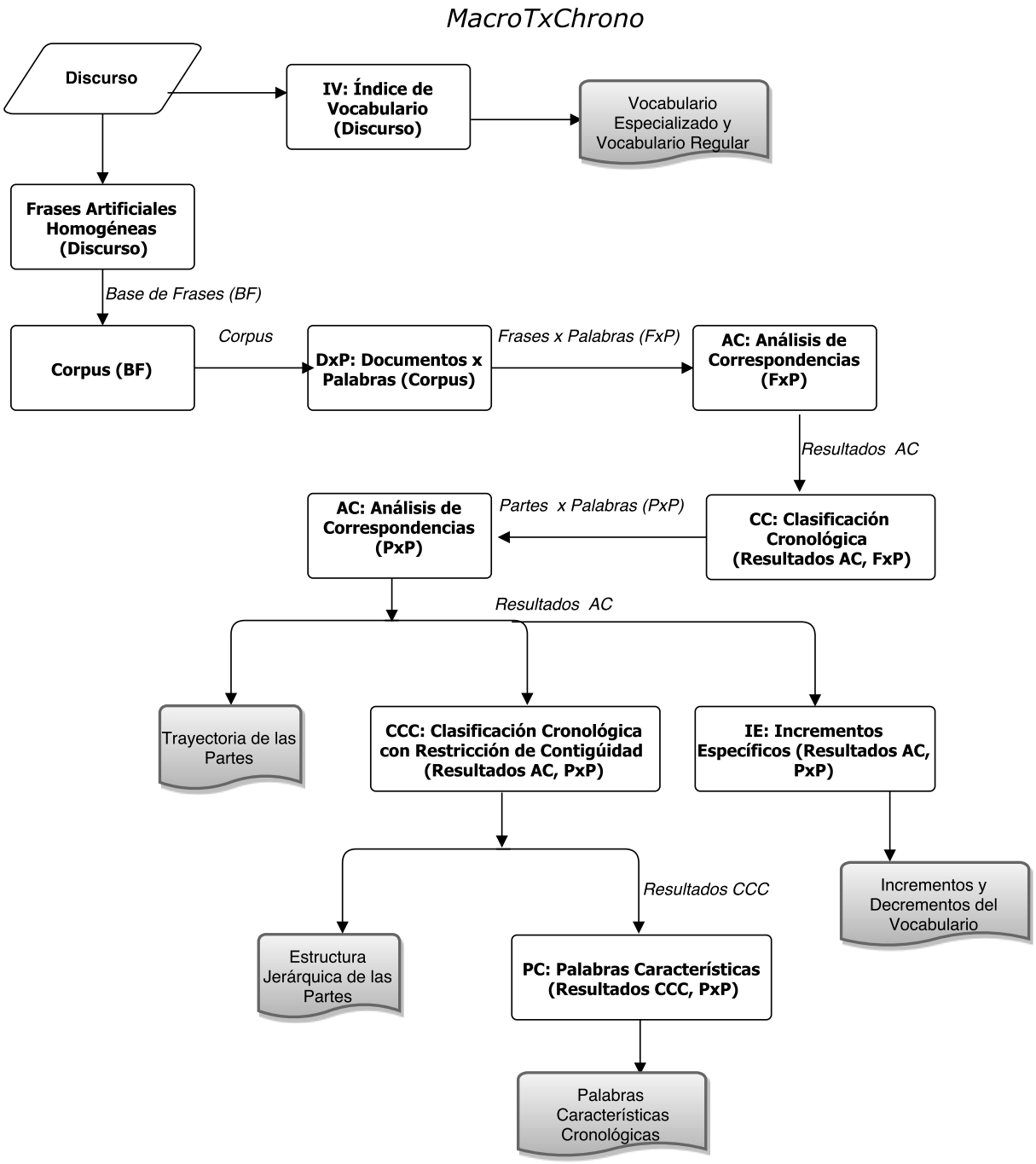


Figura 6.1: Diagrama de flujo

6.2 Texto no estructurado

El texto analizado es un discurso jurídico pronunciado el 17 de septiembre 1981 frente a la Asamblea Nacional Francesa por el Ministro de Justicia de François Mitterrand, el abogado y político francés Robert Badinter. La importancia e impacto de este discurso ha dado a conocer a Badinter como uno de los principales defensores de la abolición de la

pena de muerte en Francia y como defensor de los derechos humanos.

El discurso está formado por 7,940 ocurrencias, que corresponden a 1,737 palabras diferentes. Las palabras con una frecuencia menor a 5 y que no se encontraron, en al menos 3 frases, fueron eliminadas. Las preposiciones, conjunciones, pronombres y determinantes *à, au, l, la , le, les, un, une, d, de, des, du, qu, que, y, en* fueron descartados por considerarlos de poca relevancia en el estudio. Así, el corpus a analizar cuenta con 3,367 ocurrencias, que corresponden a 210 palabras seleccionadas.

6.2.1 División del texto en frases

El discurso es dividido en frases cortas. Éstas pueden ser definidas a partir de una lectura clásica del discurso, o de manera automatizada, se cortan pseudo-frases de igual tamaño. Después, las diferentes palabras que conforman las frases se identifican y se calcula su frecuencia.

El discurso se divide de manera automatizada en 74 frases léxicamente homogéneas (100 ocurrencias por frase).

6.2.2 Agrupación de las frases en partes homogéneas

Una vez que el discurso fue dividido en frases se aplicó el algoritmo de clasificación cronológica (CC; Legendre y Legendre, 1998), que fue desarrollado para identificar discontinuidad en series temporales multidimensionales, para las frases en partes homogéneas. En este caso, el corpus es considerado como una serie de frases (se contabiliza la ocurrencia de las palabras de cada frase y se crea un vector de frecuencias. Entonces, el algoritmo toma las frases y va considerando la distancia entre cada una de ellas y agregándolas en nuevos grupos. La distancia entre las frases es la distancia Chi-cuadrado, y el método de agregación, el de ligamiento completo). Para que se realice esta agregación, se requiere la validación de un test de permutación para asegurar que sólo frases léxicamente homogéneas se agrupan. El algoritmo se detiene cuando ya no se pueden realizar más agregaciones entre las frases.

Como resultado final, considerando un nivel de significación de $\alpha = 0.15$, el algoritmo de CC agregó las frases del discurso en 11 partes homogéneas, denotadas por Pj, $j = 1, ..11$. Estas partes se componen desde 2 frases (P1 y P8) hasta 10 frases (P2). Su longitud varía entre 78 y 440 ocurrencias.

6.3 Análisis cronológico

6.3.1 Forma y trayectoria de las partes

Para estudiar la trayectoria del discurso se construyó la tabla partes×palabras (11 partes, 210 palabras) y se realizó un AC. Los conceptos básicos en la aplicación de este método se exponen en la sección 4.2.3 y se basan en: primero, las partes del discurso que utilizan las mismas palabras con frecuencias similares están íntimamente relacionadas; segundo, las palabras estarán más cerca entre ellas en la medida en que se asocian y, tercero,

las partes estarán menos distantes a medida en que contengan la misma ocurrencia de la palabra.

Los resultados del AC muestran que el primer plano factorial conserva el 31 % de inercia total. Cada uno de los primeros cuatro ejes explica una inercia superior a la media y, juntos, explican un 53.55 % de la inercia total (Tabla 6.1).

	Valores propios	% de inercia	% inercia acumulada
dim 1	0,17	16,89	16,89
dim 2	0,15	14,21	31,10
dim 3	0,12	11,99	43,09
dim 4	0,11	10,46	53,55
dim 5	0,10	9,94	63,49
dim 6	0,09	8,46	71,95
dim 7	0,08	8,03	79,98
dim 8	0,08	7,35	87,33
dim 9	0,07	6,87	94,19
dim 10	0,06	5,81	100,00

Tabla 6.1: Porcentaje de inercia explicada por los ejes.

Este análisis se centra en la representación que ofrece el primer plano factorial de AC (Figura 6.2) y se analiza la estructura del discurso. Aquí no se hace referencia al vocabulario empleado en cada una de las partes para poder captar mejor la organización del discurso, que se representa a través de su forma, que es parte fundamental de nuestro interés.

Siguiendo la trayectoria de las partes, desde P1 a P8 se presenta una evolución regular. Ésta es interrumpida por el retroceso de P9 y, a partir de ésta, se inicia una progresión ordenada del discurso, resultado de la aparición de un nuevo argumento, que se sigue hasta el cierre del discurso. La forma observada en la trayectoria dibuja de manera significativa una parábola muy estable, con sólo dos pequeños retrocesos en su evolución (P3 y P9).

6.3.2 Estructura jerárquica de las partes

Para decidir el número de ejes en la agrupación de las partes en la jerarquía, la variable cuantitativa que represente la cronología de las partes incorporada en el AC, como una variable cuantitativa suplementaria, proporciona información de interés sobre la relación con las primeras dimensiones (Dim1: 0.35, Dim2: 0.20, Dim3: 0.06, Dim4: 0.04, Dim5: 0.05). Con un umbral de significación $\propto = 0.10$ para la correlación entre la cronología y la dimensión se consideraron los dos primeros ejes de AC. Estos, en conjunto, proporcionan una representación de alta calidad (31 %) y, recogen la mayor parte de la variabilidad asociada con el tiempo, Así, el árbol presentado en la Figura 6.3 conserva las características expuestas anteriormente.

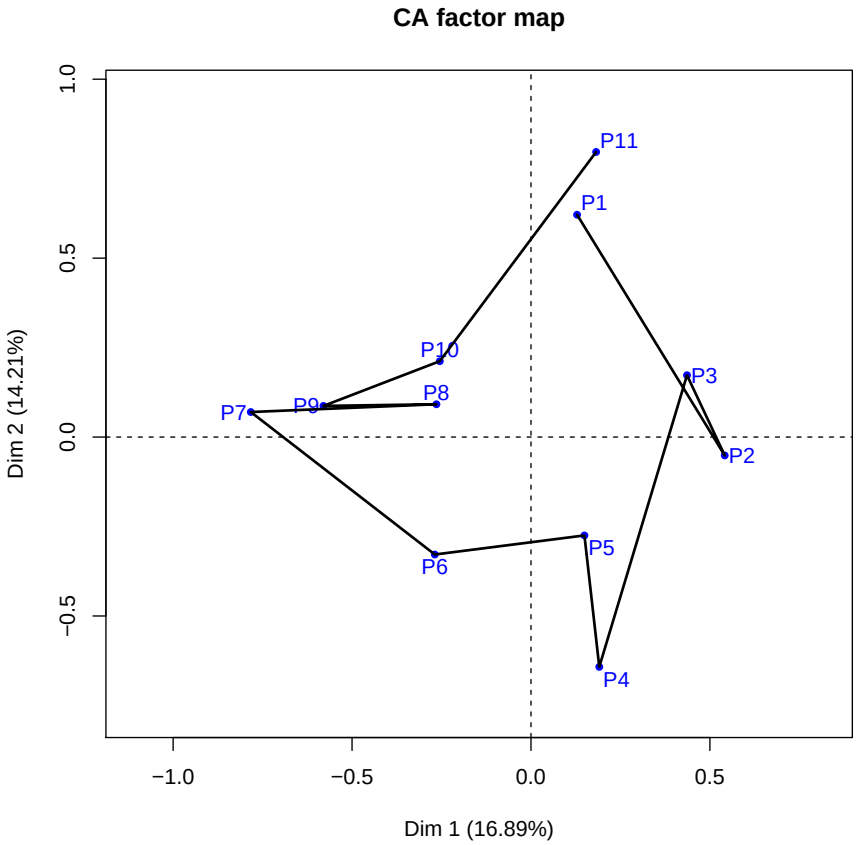


Figura 6.2: Trayectoria del discurso en el primer plano factorial de AC

6.3.3 Segmentación cronológica y flujo de la argumentación

Las partes se agruparon en tres principales períodos [P1-P3], [P4-P6] y [P7-P11]. Esta división homogénea se obtuvo a través del algoritmo de CC, aplicado a la tabla partes×palabras, con un nivel de significación de $\alpha = 0.15$.

Una vez determinados los períodos de cambio en el discurso, para estudiar el flujo del vocabulario, se encontraron las palabras que caracterizan a cada período así como los incrementos específicos de estos (Tabla 6.2). Las palabras asociadas con cada uno de los períodos reflejan perfectamente la diversidad temática entre las partes. De [P1-P3] el orador expone el objetivo del proyecto de ley resaltando el papel de la izquierda y de algunos líderes en abolir la pena de muerte en Francia; de [P4- P6] se consolida la información y se busca persuadir al auditorio; el orador insta a los miembros de la asamblea a seguir el ejemplo de países occidentales, diciendo que en: "todos los países democráticos europeos occidentales, no en Estados Unidos, la pena de muerte ha sido abolida. ¿Qué le ha pasado a Francia? " Argumenta que aplicar la pena de muerte a un terrorista es utilizar sus mismas armas por parte de un estado democrático. Da pruebas irrefutables de cómo en una sociedad democrática, el estado no puede decidir quién debe vivir y quién debe morir. En [P7-P11], el ministro desarrolla una serie de pruebas y, mediante su elocuencia, busca conseguir el apoyo total de la asamblea. Trata de penetrar en la conciencia de los asistentes, haciéndoles

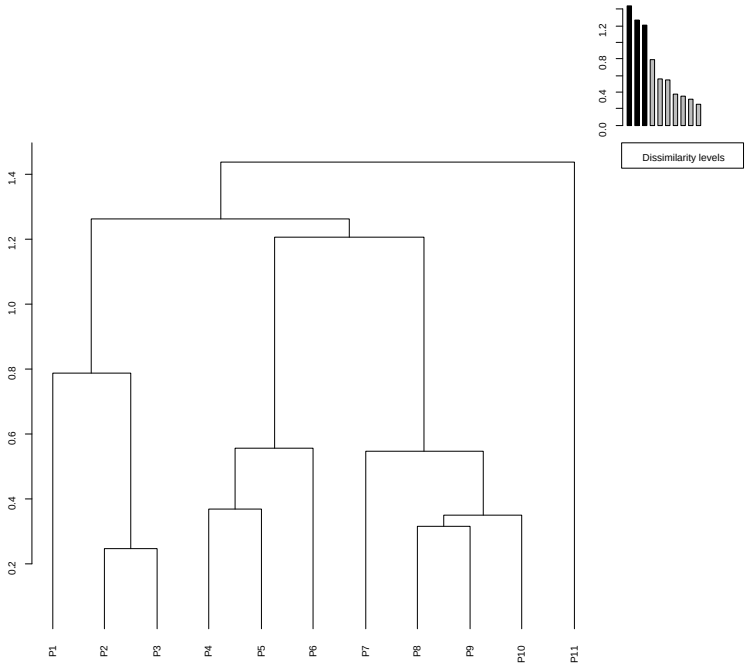


Figura 6.3: Estructura jerárquica de las partes del discurso a través de clasificación con restricción de contigüidad temporal

ver que la justicia francesa no debe ser una justicia que mate; obliga a pensar sobre las repercusiones futuras y termina invitando a todos los legisladores a votar por la abolición de la pena de muerte.

	Período [P1-P3]	Período [P4-P6]	Período [P7-P11]
Palabras Características	Gauche (izquierda), éloquence (elocuencia), question (pregunta), abolition (abolición), france (Francia), Jaurès, attendre (esperar), non (no), assemblée (asamblea), première (primero), cause (causa), seulement (solo), pourquoi (porque), grande (grande), esprit (espíritu), pratique (práctico), messieurs (caballeros)	Jamais (nunca), on (nosotros), mort (muerte), contre (contra), terrorismo (terrorismo), sanglante (sangriento), démocraties (democracias), criminalité (criminalidad), jusqu (arriba), droit (derecho), démocratie (democracia), liberté (libertad), celle (que)	Justice (justicia), malheur (desgracia), élimination (eliminación), victimes (víctimas), gouvernement (gobierno), criminel (criminal), doit (debe), parce (porque), société (sociedad), loi (ley), vie (vida).
Incrementos Específicos		Crime (crimen), criminalité (criminalidad), démocraties(democracias), homme (hombre), jamais (nunca), mort (muerte), on (nosotros), réalité (realidad), sanglante (sangriento), terrorismo (terrorismo), violence (violencia)	Aucune (ninguna), condamnés (condenado) ,criminel (criminal), doit (debe), donc (por lo tanto), elimination (eliminación), et (y), être (ser), gouvernement (gobierno), justice(justicia), loi (ley), malheur (desgracia), même (incluso), mesure (medir), será (ser), société (sociedad), soit (sí), tout (todos), tue(muertes), victimes(victimas), vie (vida).

Tabla 6.2: Palabras características e incrementos específicos de los períodos

La jerarquía etiquetada resulta ser una herramienta de gran importancia para la organización de los argumentos. Aquí se calcula la intensidad con la cual cada palabra caracteriza cada período de la jerarquía, incluyendo las terminales que corresponden a las partes. La Figura 6.4 muestra de manera detallada el flujo del vocabulario.

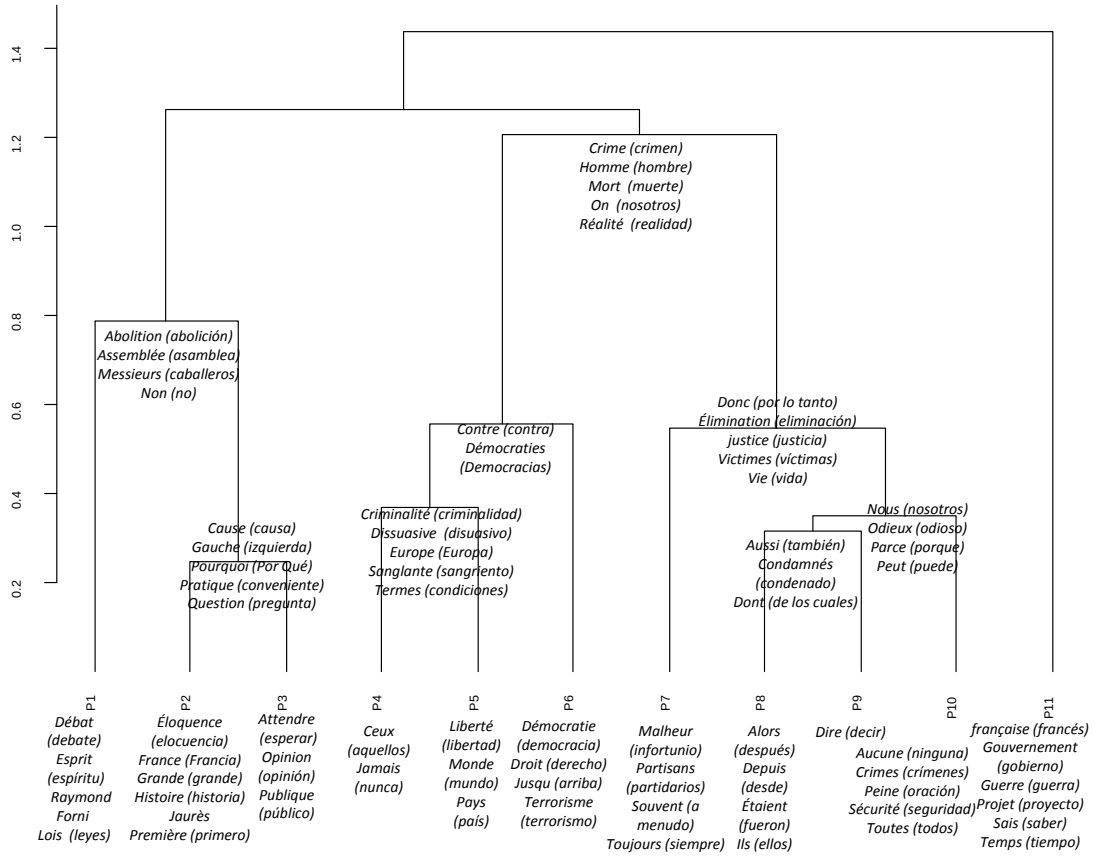


Figura 6.4: Seguimiento del flujo del vocabulario a través del árbol jerárquico. Cada palabra es localizada en la parte que mejor caracteriza.

[P1-P3] muestra una evolución regular del vocabulario. Aquí se establecen las bases del discurso. El orador solicita a la Asamblea Nacional abolir la pena de muerte en Francia.

- Reconoce la labor de Raymond Forni (político socialista francés) en la lucha contra la abolición de la pena [P1]
- Hace un recuento histórico sobre los derechos humanos en Francia y exalta su grandeza y valentía; recuerda con orgullo que Francia fue uno de los primeros países del mundo en abolir la esclavitud; sin embargo, se pregunta ¿Por qué la demora en la abolición de la pena de muerte? [P2]
- Busca una respuesta a su pregunta y la encuentra en la política. Considera que la supresión ha sido siempre una de las principales causas de la izquierda francesa y de las fuerzas progresistas que buscan el cambio. Recuerda a Jean Jaurès (político socialista pacifista francés) como un político elocuente y de gran corazón que debe ser recordado en la historia de Francia como un defensor de la libertad y la abolición [P2-P3].

[P4-P6] A través de ejemplos y datos relevantes, invita a la audiencia a seguir el ejemplo de los gobernantes y grandes hombres de otros países occidentales que han luchado por la abolición de la pena de muerte en sus países. Afirma que, en todos los países democráticos del mundo, la pena de muerte debe ser abolida. Sólo en regímenes totalitarios se puede entender que el estado decida quién debe vivir y quién debe morir, pero en una república democrática, nadie puede tener ese derecho. El terrorrismo no se combate con la muerte.

[P7-P11] Persuade a los oyentes con el argumento de que un hombre no es totalmente culpable ni la justicia es totalmente infalible y, dado que la abolición es una decisión moral, hay que pronunciarse con toda claridad a favor de la abolición.

- Insiste en que, a los adversarios del proyecto de abolición, lo que les parece insoportable no es la vida del criminal en prisión sino la posibilidad de que lo vuelva a hacer en otra ocasión y piensan que, como medida de precaución, es mejor que el criminal muera [P7-P8].
- Invita a votar por la abolición de la pena de muerte sin restricción alguna, sin tener en cuenta las categorías de los crímenes. Porque la muerte de un niño o de una persona de edad puede suscitar mayor compasión que la muerte de una mujer de treinta años o de un hombre maduro. Aunque cualquier discriminación es una injusticia [P9-P10]
- Finaliza su discurso no sólo invitando a votar por la propuesta de ley sino haciendo un matiz para explicar que la ley no se aplicará en tiempos de guerra [P11]

6.4 Conclusiones

Con la metodología propuesta, en la que se combinan varios métodos estadísticos, se obtienen buenos resultados cuando se aplica a textos organizativos, además de que el procedimiento puede ser aplicado a cualquier tipo de texto oral o escrito.

Mediante AC se descubre la organización evolutiva del texto. La segmentación homogénea de las partes, así como su estructura jerárquica, logradas a través de los métodos de clasificación cronológica y clasificación con restricción de contigüidad, facilitan el flujo argumentativo del discurso dado que cada parte desempeña un papel específico en la evolución del discurso.

Los resultados en esta aplicación son muy safisfactorios. Las 11 partes léxicamente homogéneas en las que fue dividido el discurso de Badinter, muestran una evolución regular; las partes son muy especializadas desde el punto de vista del vocabulario. La temática es fácil de identificar tanto en las partes individuales como en segmentos de ellas. Podemos decir que la estructura y organización del discurso son buenas, de acuerdo con la metodología utilizada para su análisis. Esto no se podría afirmar a través de una lectura convencional.

Capítulo 7

Funciones de las palabras

7.1 Introducción

En este capítulo se implementa la metodología desarrollada en el Capítulo 4, la cual tiene como finalidad mostrar las funciones que desempeñan las palabras según el uso que de ellas se hace en un corpus cronológico dado; según se trate de un corpus no estructurado, como un discurso, un ensayo, un artículo, un capítulo de un libro, etc. o de un corpus estructurado.

La metodología propuesta se sigue aplicando al disurso sobre la abolición de la pena de muerte en Francia. Discurso que ya se ha utilizado en un capítulo precedente; con ello se quiere demostrar que esta metodología es aplicable también a una amplia base bibliográfica, formada por 506 artículos científicos sobre Lupus Eritematoso Sistémico (LES), ya utilizada en el capítulo 5.

Los pasos que se siguen son los siguientes:

- Distribuir el vocabulario de acuerdo con el índice de reparto del vocabulario (sección 4.1.1).
- Determinar las funciones del vocabulario de acuerdo al índice de reparto y a las palabras cronológicas (sección 4.1.3).
- Visualizar el vocabulario que determina la evolución a través de los gráficos de Bertin (sección 4.2.1).
- Mostrar la trayectoria o esquema evolutivo mediante análisis de correspondencias (sección 4.2.3).

7.2 Funciones de las palabras

7.2.1 Reparto del vocabulario

El índice de reparto del vocabulario (Hubert y Labbé, 1990a,b) intenta medir la regularidad del uso de las palabras. Un valor próximo a 1 indica uso regular, mientras que un valor próximo a 0, marca un uso circunstancial o local de la palabra (sección 4.1.1).

En el discurso de Badinter, con un umbral de frecuencia igual a cinco, el índice de reparto varía entre los valores más bajos 0.14 y 0.24 - para *malheur* (infortunio) y *opinion* (opinión) - y los más altos 0. 88 y 0.79 - para *sûr* (seguro) y *abolitionnistes* (abolicionistas), respectivamente (Tabla 7.1). La estabilidad de un número de palabras se revela por

Vocabulario local o especializado			Vocabulario regular o estable					
Palabra	Índice	Frec.	Palabra	Índice	Frec.	Palabra	Índice	Frec.
malheur (infortunio)	0,14	9	sûr (seguro)	0,88	5	assemblée (asamblea)	0,64	10
opinion (opinión)	0,24	7	abolitionnistes (abolicionistas)	0,79	6	cette (esta)	0,64	36
terrorisme (terrorismo)	0,26	5	humaine (humano)	0,76	7	mais (pero)	0,64	44
jaurcs (jaurcs)	0,27	5	simplement (simplemente)	0,76	8	nom (nombre)	0,64	9
élimination (eliminación)	0,28	7	égard (respeto)	0,75	5	point (punto)	0,64	6
éloquence (elocuencia)	0,28	6	encore (de nuevo)	0,75	6	pour (para)	0,64	65
publique (Público)	0,28	7	simple (sencillo)	0,75	5	celles (aquellas)	0,63	5
attendre (esperar)	0,30	5	toute (todos)	0,74	7	cet (este)	0,63	14
gauche (izquierda)	0,30	7	enfin (finalmente)	0,73	5	elle (eso)	0,63	25
temps (tiempo)	0,31	11	république (república)	0,73	8	forni (forni)	0,63	8
projet (proyecto)	0,32	5	aprcs (después)	0,72	7	mort (muerte)	0,63	104
criminalité (criminalidad)	0,33	5	devant (delante de)	0,71	5	non (no)	0,63	12
démocratie (democracia)	0,33	14	avoir (tener)	0,70	5	sans (sin)	0,63	13
sanglante (sangriento)	0,33	5	esprit (espíritu)	0,70	7	sûr (seguro)	0,63	15
criminel (criminal)	0,35	5	majorité (mayoría)	0,70	5	abolir (abolir)	0,62	6
gouvernement (gobierno)	0,36	15	prononcer (pronunciar)	0,70	5	ainsi (así)	0,62	7
droit (derecho)	0,37	9	raison (razón)	0,70	10	cela (eso)	0,62	6
guerre (guerra)	0,37	11	raymond (raymond)	0,70	6	fois (tiempo)	0,62	7
française (Francés)	0,38	7	bien (bueno)	0,69	22	il (eso)	0,62	68
europe (Europa)	0,39	11	effet (efecto)	0,69	5	qui (cual)	0,62	98
termes (condiciones)	0,39	5	faire (hacer)	0,69	8	seulement (solamente)	0,62	9
victimes (víctimas)	0,39	6	moins (menos)	0,68	5	trois (tres)	0,62	5
justice (justicia)	0,40	39	valeurs (valores)	0,68	6	politiques (políticas)	0,61	6
pratique (conveniente)	0,42	5	ait (tiene)	0,67	5	soit (si)	0,61	28
débat (debate)	0,43	8	parlement (parlamento)	0,67	7	tue (muertes)	0,61	6
sureté (seguridad)	0,43	5	peine (pena)	0,67	77	abord (a bordo)	0,60	7
dissuasive (disuasivo)	0,44	5	dont (de los cuales)	0,66	13	alors (después)	0,60	10
libération (liberación)	0,44	5	et (y)	0,66	157	crime (crimen)	0,60	15
souvent (a menudo)	0,44	8	grandes (grande)	0,66	6	crimes (crímenes)	0,60	11
grande (grande)	0,45	9	quelle (qué)	0,66	8	dans (en)	0,60	80
odieux (odioso)	0,45	5	serait (sería)	0,66	9	loin (lejos)	0,60	6
sécurité (seguridad)	0,45	5	années (años)	0,65	9	même (incluso)	0,60	18

Tabla 7.1: Índice de reparto del vocabulario del discurso de Badinter

un alto índice y por un uso frecuente. El vocabulario regular no sólo está compuesto por palabras herramienta, sino también por un lenguaje que define la personalidad y el estado de ánimo del orador. El orador da un tono seguro al discurso sirviéndose de adverbios como: *simplement* (simplemente), *seulement* (solamente), *totalement* (totalmente), *vraiment*(verdaderamente), *même* (incluso), *enfin* (finalmente), etc. También palabras como: *abolir* (abolir), *mort* (muerte), *politiques* (políticas), *abolitionnistes* (abolicionistas),*égard* (respeto) y *humaine* (humano) aparecen regularmente para mostrar que el razonamiento del orador se establece sobre elementos sólidos. Asimismo, puede apreciarse que ciertos conectores son también estables, por ejemplo: *cette* (esta), *celles* (aquellas), *outre* (además), *donc* (por lo tanto), etc.

7.2.2 PALABRAS CRONÓLOGICAS

Ciertas palabras pueden caracterizar no sólo una parte, sino un grupo de partes consecutivas. La Tabla 7.2 muestra la lista de palabras que caracterizan a cada parte o grupo de partes en que fue dividido el discurso de Badinter. Para detectarlas, se determinan primero las palabras características de cada parte (es decir, el primer nivel) y después, sucesivamente, las de grupo de dos partes consecutivas (palabras características del segundo nivel); después, grupos de tres partes consecutivas (palabras características del tercer nivel), y así sucesivamente. Al final del proceso, cada palabra se asocia a la parte o grupo de partes/años que mejor caracteriza: es decir, aquel para el cual la probabilidad asociada al test es mayor.

Palabra	Parte	Frecuencia		P. value	V.test	Palabra	Parte	Frecuencia		P. value	V.test
		Interna	Global					Interna	Global		
débat (debate)	P1	2	8	0,05	2,00	gouvernement (gobierno)	P11	7	15	0,00	3,53
esprit (espíritu)	P1	2	7	0,03	2,11	guerre (guerra)	P11	5	11	0,00	2,86
forni (forni)	P1	2	8	0,05	2,00	projet (proyecto)	P11	3	5	0,01	2,44
lois (leyes)	P1	2	7	0,03	2,11	sais (saber)	P11	3	6	0,03	2,21
raymond (raymond)	P1	2	6	0,03	2,24	temps (tiempo)	P11	5	11	0,00	2,86
éloquence (elocuencia)	P2	5	6	0,00	3,54	abolition (abolición)	P1-P3	23	46	0,00	3,00
france (francia)	P2	10	19	0,00	3,94	assemblée (asamblea)	P1-P3	7	10	0,01	2,44
grande (grande)	P2	5	9	0,01	2,75	messieurs (caballeros)	P1-P3	4	5	0,05	1,97
histoire (historia)	P2	5	8	0,00	2,97	cause (causa)	P2-P3	6	8	0,01	2,62
jaurcs (jaurcs)	P2	5	5	0,00	3,96	gauche (izquierda)	P2-P3	7	7	0,00	3,83
premicre (primero)	P2	6	8	0,00	3,70	pourquoi (por qué)	P2-P3	6	9	0,02	2,31
opinion (opinión)	P3	4	7	0,01	2,54	pratique (conveniente)	P2-P3	4	5	0,03	2,14
publique (público)	P3	4	7	0,01	2,54	question (pregunta)	P2-P3	9	12	0,00	3,35
seulement (solamente)	P3	4	9	0,03	2,13	criminalité (criminalidad)	P4-P5	5	5	0,00	3,27
liberté (libertad)	P5	5	8	0,00	3,16	europe (europa)	P4-P5	7	11	0,01	2,67
monde (mundo)	P5	3	5	0,02	2,25	sanglante (sangriento)	P4-P5	5	5	0,00	3,27
pays (país)	P5	7	21	0,01	2,48	termes (condiciones)	P4-P5	4	5	0,02	2,32
démocratie (democracia)	P6	9	14	0,00	4,40	démocraties (democracias)	P4-P6	5	5	0,01	2,60
droit (derecho)	P6	7	9	0,00	4,27	crime (crimen)	P4-P10	14	15	0,02	2,40
terrorisme (terrorismo)	P6	5	5	0,00	4,06	homme (hombre)	P4-P10	16	17	0,01	2,69
criminel (criminal)	P7	4	5	0,00	3,37	mort (muerte)	P4-P10	82	104	0,00	3,57
malheur (infortunio)	P7	9	9	0,00	6,08	réalité (realidad)	P4-P10	9	9	0,03	2,20
partisans (partidarios)	P7	4	5	0,00	3,37	élimination (eliminación)	P7-P10	7	7	0,00	3,65
dire (decir)	P9	3	7	0,04	2,10	justice (justicia)	P7-P10	31	40	0,00	6,38
crimes (crímenes)	P10	4	11	0,01	2,50	victimes (víctimas)	P7-P10	6	6	0,00	3,30
peine (oración)	P10	11	77	0,05	1,98	vie (vida)	P7-P10	8	13	0,02	2,27
sécurité (seguridad)	P10	3	5	0,01	2,69	condamnés (condenado)	P8-P9	4	8	0,01	2,45
française (francés)	P11	3	7	0,04	2,01	odieux (odioso)	P8-P10	4	5	0,01	2,58

Tabla 7.2: Palabras cronológicas de las partes

En la Tabla 7.2 se puede apreciar que en el discurso de Badinter, unas pocas palabras caracterizan grupos de partes; las partes del discurso presentan una gran homogeneidad temática y están bien diferenciadas. Por ejemplo, *lois* (leyes), *abolition* (abolición), *France* (Francia), *assemblée* (asamblea), *gauche* (izquierda), caracterizan al conjunto de las partes P1, P2 y P3. En estas partes [P1-P3] se expone toda la información sobre el proyecto de abolición de la pena de muerte liderado por la izquierda francesa. De la misma manera, las partes P4, P5 y P6 están caracterizadas por *liberté* (libertad), *droit* (derecho), *pays* (país), *démocratie* (democracia), *criminalité* (criminalidad) y *terrorisme* (terrorismo). En este bloque [P4-P6], el orador se apoya en las palabras "libertad" y "democracia" para generar conciencia de la necesidad de un cambio en las políticas de los países que se denominan democráticos, tales como Francia.

Organizando la lista completa de las palabras cronológicas en cada parte, es posible seguir la evolución del corpus, pero sin conocer qué papel desempeñan o si son de uso local o regular.

7.2.3 Funciones de las palabras

Para definir las funciones de las palabras, considerando los criterios expuestos en la sección 4, el corpus fue dividido en 3 grupos. La Figura 7.1 muestra la división del discurso de Badinter.

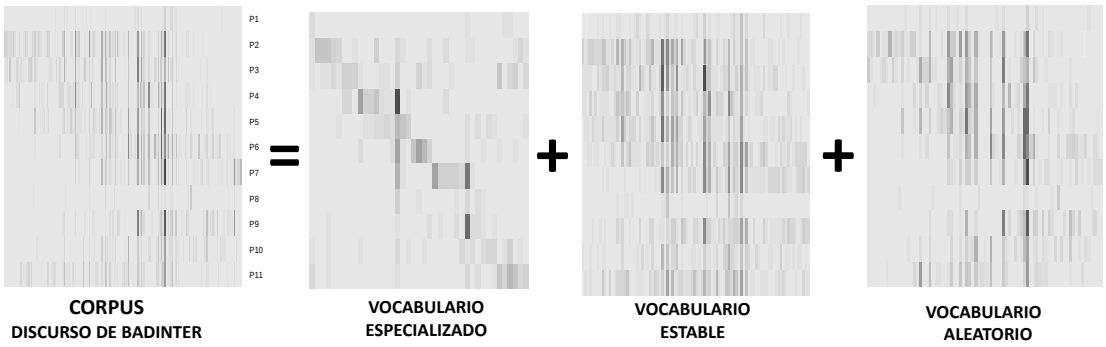


Figura 7.1: Descomposición del discurso de Badinter

Categorías gramaticales

El discurso fue subdividido en categorías gramaticales (verbos, sustantivos, adjetivos nombres propios, pronombres, adverbios, determinantes, preposiciones y conjunciones). Dentro de cada una de las categorías gramaticales las palabras están ordenadas en clases de frecuencia.

Las Tabla 7.3 permite visualizar las diferencias de cada uno de los grupos de acuerdo con su categoría gramatical. Los índices de los verbos son en promedio más regulares que los de los sustantivos y se aproximan de manera significativa alrededor de la media (0.56). El perfil de los adjetivos se acerca mucho al de las palabras herramienta (adverbio, conjunción, preposición y pronombre), los cuales no son muy diferentes de los verbos. Por último, los nombres propios y los sustantivos son los que más se aleja de la media, es decir, que pueden ser considerados como los grupos con las palabras más localizadas. Por ejemplo, existen 69 sustantivos cuya frecuencia está entre 5 y 110 ocurrencias. Esta categoría fue organizada en tres clases de frecuencia (5-10, 11-30, 31-110). La primera clase [5-10],cuenta con 50 palabras, tiene un índice de reparto medio de 0.51, con una desviación estándar de 0.14. En esta clase de frecuencia, el sustantivo más localizado es *malheur* (infortunio) con un índice de reparto de 0.14.

Vocabulario regular o estable

La regularidad del vocabulario en el discurso puede tener varias causas. Por un lado, tiene que ver con el tema que se está tratando, con el lenguaje propio del orador y por

Categoría	Clase de frecuencia	Efectivos	Frecuencia media	Reparto medio	Desv. estandar
Adjetivo	[5-10]	20	6,40	0,58	0,17
	[11-30]	4	16,75	0,65	0,06
Adverbio	[5-10]	15	6,93	0,63	0,09
	[11-50]	10	26,30	0,57	0,05
Conjunción	[5-10]	4	7,50	0,56	0,13
	[11-200]	5	56,00	0,61	0,08
Nombre propio	[5-20]	5	9,80	0,50	0,18
Preposición	[5-20]	6	9,17	0,63	0,06
	[21-100]	4	65,25	0,60	0,05
Pronombre	[5-10]	14	6,93	0,60	0,11
	[11-30]	13	16,00	0,60	0,05
	[31-100]	8	53,75	0,59	0,06
Sustantivo	[11-30]	15	13,60	0,50	0,10
	[5-10]	50	6,64	0,51	0,14
	[31-110]	4	66,50	0,55	0,12
Verbo	[5-10]	19	6,37	0,59	0,13
	[11-110]	10	26,20	0,60	0,05

Tabla 7.3: Principales características del reparto del vocabulario del discurso de Badinter en función de sus categorías gramaticales

otro lado, con la estrategia que éste utilice para convencer y persuadir a su auditorio.

Las palabras más regulares son las que imponen un tono al discurso, o un sentimiento de convicción constante a través de la regularidad de su aparición como: *sûr* (seguro), *simplement* (simplemente), *enfin* (finalmente), *seulement* (solamente), etc. Estas palabras se localizan a lo largo de todo el discurso y, por lo tanto, no caracterizan ninguna de las partes.

Vocabulario aleatorio

Gran número de las palabras del discurso son palabras gramaticales como: determinantes, preposiciones y auxiliares que deben estar presentes en el vocabulario habitual, impuesto por la lengua.

Las palabras herramientas no son de las que destacan por un notable índice alto; sin embargo, de acuerdo al valor test, pueden ser características de una o más partes. Por ejemplo, *et* (y), palabra con mayor frecuencia (157), es característica en el segmento [P7-P10] porque en esta parte su uso es mayor (36 %, 57 veces); *dans* (en), con una frecuencia de 80, caracteriza a P6 (20 %, 16 veces); *mort* (muerte), con un uso regular a lo largo del discurso; sin embargo, es más representativa en el segmento [P4-P10]. Las palabras que conforman este grupo presentan una distribución Poisson.

Vocabulario especializado o local

El estudio detallado de cada uno de los grupos de vocabulario (regular, especializado y aleatorio) puede ser de gran utilidad, dependiendo del interés del investigador. En este

Categoría	Clase de frec.	Palabra	Índice	Frec.
Adjetivo	[5-10]	loin (lejos)	0,6	6
		grandes (grande)	0,66	6
		toute (todos)	0,74	7
		simple (sencillo)	0,75	5
		humaine (humano)	0,76	7
		abolitionnistes (abolicionistas)	0,79	6
		autre (otro)	0,83	8
		sur (seguro)	0,88	5
	[11-30]	bien (bueno)	0,69	22
		deux (dos)	0,72	12
Adverbio	[5-10]	seulement (solamente)	0,63	9
		ainsi (así)	0,62	7
		fois (tiempo)	0,62	7
		moins (menos)	0,68	5
		tant (así)	0,68	5
		aprcs (después)	0,72	7
		enfin (finalmente)	0,73	5
		encore (de nuevo)	0,75	6
		simplement (simplemente)	0,76	8
	[11-50]	quand (cuándo)	0,57	11
		ne (no)	0,61	54
		sur (sobre)	0,63	15
Conjunción	[5-10]	comme (como)	0,74	9
	[11-200]	mais (pero)	0,64	44
		si (si)	0,7	24
Preposición	[5-20]	avec (con)	0,58	13
		aux (contracción a les)	0,65	10
		devant (delante de)	0,71	5
	[21-100]	pour (para)	0,64	65
Verbo	[5-10]	tue (morir)	0,61	6
		abolir (abolir)	0,62	6
		avait (tenía)	0,65	6
		serait (sería)	0,66	9
		ait (tiene)	0,67	5
		faire (hacer)	0,69	8
		avoir (tener)	0,7	5
		prononcer (pronunciar)	0,7	5
		soient (son)	0,71	5
		soit (ser)	0,61	28
	[11-110]	sont (son)	0,64	20
Pronombre	[5-10]	celles (aquellos)	0,63	5
		quelle (qué, cuál)	0,66	8
		aucun (no)	0,73	5
		moi (yo)	0,78	5
		ma (mi)	0,81	5
	[11-30]	notre (nuestro)	0,57	15
		tous (todos)	0,57	19
		cet (este)	0,63	14
		elle (eso)	0,63	25
		sa (su)	0,64	18
		lui (él)	0,65	14
		son (su)	0,65	18
	[31-100]	je (yo)	0,61	52
		il (eso)	0,62	68
		cette (este)	0,64	36
		ce (este)	0,65	41
Sustantivo	[5-10]	paix (paz)	0,53	7
		coeur (corazón)	0,54	6
		peur (miedo)	0,54	7
		demande (pregunta)	0,55	5
		grâce (gracia)	0,57	5
		passion (pasión)	0,57	6
		conscience (conciencia)	0,57	8
		vérité (verdad)	0,59	7
		instant (momento)	0,59	9
		violence (violencia)	0,6	7
		politiques (políticas)	0,61	6
		point (punto)	0,64	6
		nom (nombre)	0,64	9
		années (anos)	0,65	9
		parlement (parlamento)	0,67	7
		valeurs (valores)	0,68	6
		effet (efecto)	0,69	5
		majorité (mayoría)	0,7	5
		raison (razón)	0,7	10
		république (república)	0,73	8
		égard (respeto)	0,75	5
	[11-30]	choix (elección)	0,56	11
		politique (político)	0,56	11
	[31-100]	société (sociedad)	0,56	11

Tabla 7.4: Vocabulario Regular o estable del discurso de Badinter

caso nos enfocamos en el vocabulario especializado de los dos ejemplos, porque aquí se localizan las palabras que le dan evolución al corpus. Estas palabras son las que se encuentran exclusivamente en ciertas partes del corpus y abordan ideas o temas específicos.

Es importante analizar la función de cada una de estas palabras en el discurso para entender la razón de su uso especializado o local. La localización depende del tema específico que se está desarrollando. Por ejemplo, analizando tres palabras, éloquence (elocuencia), Jaurès (político francés), y opinion (opinión). estas palabras hacen referencia a un momento específico en el que Badinter recuerda las cualidades de Jean Jaurès y su opinión; sobre la abolición de la pena de muerte. Este grupo de palabras caracteriza una parte o un segmento de partes del discurso.

Para mostrar más claramente las palabras que marcan la evolución del discurso, dado que se conoce de qué parte o segmento de partes son características, la tabla de frecuencias del vocabulario especializado se separa en dos: una tabla con sólo las ocurrencias de las

Categoría	Clase de frec.	Palabra	Índice	Frec.	P.Value	Categoría	Clase de frec.	Palabra	Índice	Frec.	P.Value
Adjetivo	[5-10]	ailleurs (en otra parte)	0,48	5	0,57	Sustantivo	[5-10]	ans (años)	0,48	7	0,30
		judiciaire (judicial)	0,56	7	0,77			réalité (realidad)	0,53	9	0,40
		trois (tres)	0,62	5	0,72			histoire (historia)	0,53	8	0,71
	[11-30]	même (incluso)	0,6	18	0,73			monde (mundo)	0,54	5	0,56
Adverbio	[5-10]	toujours (siempre)	0,53	5	0,35			messieurs (señores)	0,54	5	0,56
		jusqu (hasta)	0,53	7	0,63			cause (causa)	0,56	8	0,74
		contre (contra)	0,57	8	0,39			abord (trato)	0,6	7	0,93
		là (la)	0,58	7	0,95			esprit (espíritu)	0,7	7	0,58
		alors (después)	0,6	10	0,54		[11-30]	loi (ley)	0,52	11	0,41
	[11-50]	aussi (también)	0,56	21	0,44			vie (vida)	0,56	13	0,44
		non (no)	0,63	12	0,76			homme (hombre)	0,57	17	0,80
Preposición	[5-20]	depuis (desde)	0,57	7	0,72			crime (crimen)	0,6	15	0,71
		sans (sin)	0,63	13	0,95			crimes (crímenes)	0,6	11	0,52
	[21-100]	dans (en)	0,6	80	0,77		[31-110]	mort (muerte)	0,63	104	0,28
		a (a)	0,64	67	0,89	Verbo	[5-10]	doit (debe)	0,49	9	0,42
Pronombre	[5-10]	leurs (su)	0,51	7	0,29			aurait (tendría)	0,56	5	0,55
		ses (su)	0,53	8	0,43			faut (debe)	0,58	6	0,80
		celle (ésta)	0,55	7	0,85			assemblée (juntar)	0,64	10	0,91
		nos (nuestro)	0,55	10	0,46			fait (hecho)	0,5	13	0,31
		leur (su)	0,58	10	0,99		[11-110]	ai (tener)	0,57	16	0,43
	[11-30]	toutes (todos)	0,51	12	0,44			ont (tener)	0,57	17	0,71
		ils (ellos)	0,53	14	0,32			est (es)	0,61	108	0,82
		ces (estas)	0,56	11	0,44			été (verano)	0,65	18	0,88
		ceux (aquellos)	0,57	17	0,58	Conjunción	[5-10]	ni (o)	0,47	6	0,53
		nous (nosotros)	0,58	18	0,74			donc (por lo tanto)	0,48	6	0,51
		dont (de los cuales)	0,66	13	0,95			pourquoi (por qué)	0,54	9	0,49
	[31-100]	se (se)	0,57	32	0,67		[11-200]	ou (o)	0,57	30	0,34
		qui (cual)	0,62	98	0,90	Nombre propio	[5-20]	forni (forni)	0,63	8	0,94

Tabla 7.5: Vocabulario aleatorio o usual del discurso de Badinter

Categoría	Clase de frec.	Palabra	Índice	Frec.	Parte	Categoría	Clase de frec.	Palabra	Índice	Frec.	Parte
Adjetivo	[5-10]	gauche (izquierda)	0,3	7	P2-P3	Sustantivo	[5-10]	projet (proyecto)	0,32	5	P11
		sanglante (sangriento)	0,33	5	P4-P5			criminalité (criminalidad)	0,33	5	P4-P5
		française (Francés)	0,38	7	P11			criminel (criminal)	0,35	5	P7
		pratique (práctica)	0,42	5	P2-P3			droit (derecho)	0,37	9	P6
		dissuasive (disuasivo)	0,44	5	P4-P5			victimes (víctimas)	0,39	6	P7-P10
		odieux (odioso)	0,45	5	P8-P10			termes (condiciones)	0,39	5	P4-P5
		grande (grande)	0,45	9	P2			débat (debate)	0,43	8	P1
		premicre (primero)	0,5	8	P2			sécurité (seguridad)	0,45	5	P10
	[11-30]	autres (otro)	0,59	15	P5			liberté (libertad)	0,46	8	P5
Nombre propio	[5-20]	jaurcs (jaurcs)	0,27	5	P2			partisans (partidarios)	0,47	5	P7
		europe (Europa)	0,39	11	P4-P5			condamnés (condenado)	0,48	8	P8-P9
Verbo	[5-10]	élimination (eliminación)	0,28	7	P7-P10			démocraties (democracias)	0,49	5	P4-P6
		attendre (esperar)	0,3	5	P3			lois (leyes)	0,5	7	P1
		sais (saber)	0,55	6	P11		[11-30]	temps (tiempo)	0,31	11	P11
		sera (sera)	0,55	6	P11			démocratie (democracia)	0,33	14	P6
	[11-110]	peut (puede)	0,59	11	P8-P10			gouvernement (gobierno)	0,36	15	P11
Sustantivo	[5-10]	malheur (infortunio)	0,14	9	P7			guerre (guerra)	0,37	11	P11
		opinion (opinión)	0,24	7	P3			pays (país)	0,51	21	P5
		terrorisme (terrorismo)	0,26	5	P6			question (pregunta)	0,51	12	P2-P3
		publique (Público)	0,28	7	P3		[31-110]	justice (justicia)	0,4	39	P7-P10
		éloquence (elocuencia)	0,28	6	P2			abolition (abolición)	0,51	46	P2-P3

Tabla 7.6: Vocabulario local o especializado del discurso de Badinter

palabras en la parte o segmentos de partes que caracterizan, y la otra con los residuos (Figura 7.2).

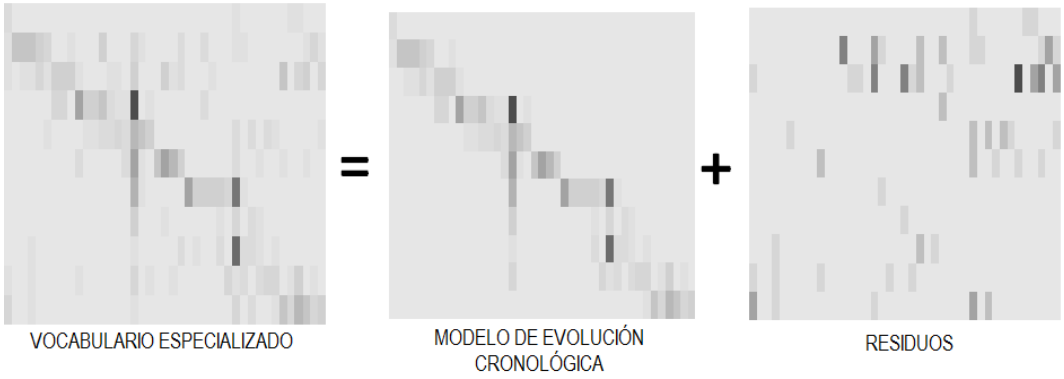


Figura 7.2: Descomposición del vocabulario especializado del discurso de Badinter

7.3 Representación y trayectoria del vocabulario cronológico

7.3.1 Diagonal ordenada

Para la representación visual del las palabras que marcan la evolución del discurso fue implementado el procedimiento de reordenamiento de matrices Documentos×Palabras (Partes×Palabras en el caso del discurso). Las palabras se permutaron, primero, con base en las coordenadas proporcionadas por el análisis de correspondencias y, segundo, por la caracterización cronológica de las palabras en las partes o segmentos de partes. Una vez reordenada la matriz, se realizó la representación mediante los gráficos de Bertin. Si una tabla de datos, con filas o columnas ordenadas por los valores de las coordenadas de AC, proporciona una diagonal cargada, como puede observarse en este caso, se reconoce como el efecto Guttman.

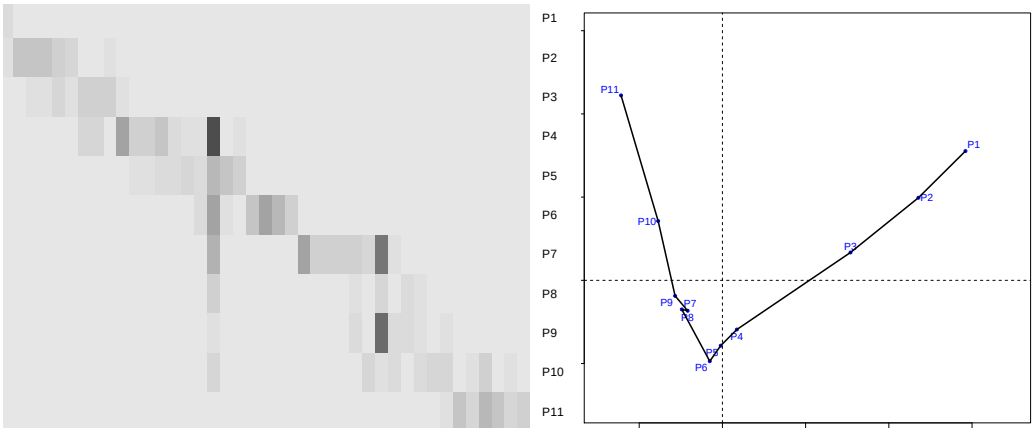
7.3.2 Trayectoria: efecto Guttman

Analizando la tabla diagonal mediante AC, las posiciones de las partes del discurso sobre el primer plano factorial presentan una forma de arco o herradura. Con este patrón evolutivo y la integración de las palabras cronológicas, se consigue visualizar la progresión argumentativa y narrativa del discurso. Se puede observar que el discurso de Badinter está bien organizado porque presenta una renovación gradual del vocabulario y, la trayectoria de las partes, facilita el flujo argumentativo del discurso, dado que cada parte desempeña un papel específico en el desarrollo, al introducir una nueva temática (7.3).

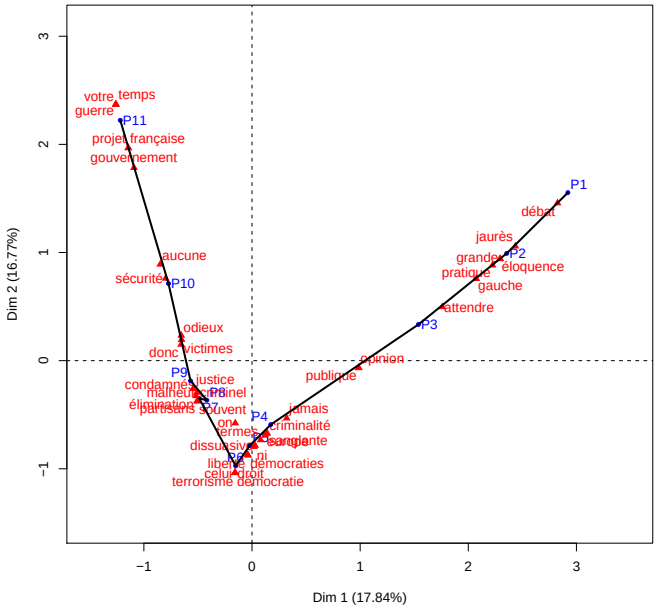
7.4 Funciones de las palabras en una base bibliográfica

En la base bibliográfica de LES, los valores más bajos del índice de reparto se localizan en la categoría de nombres propios y determinan un uso especializado de las palabras (Tabla 7.7). Estos valores corresponden a cuatro medicamentos: *belimumab* (0.14),*dhea* (0.15), *rituximab* (0.17) y *methotrexate* (0.21).

Los valores más altos (0.79 y 0.78) que determinan un uso estable, son *treating*



(a) Representación con gráficos de Bertin (b) Trayectoria: efecto Guttman



(c) Trayectoria de las partes y palabras cronológicas en el primer plano factorial de AC

Figura 7.3: Modelo de evolución cronológica del discurso de Badinter

(tratar) y *recently* (recientemente) y están clasificadas dentro de la categoría gramatical de verbo y adverbio.

Al igual que en el discurso, en este caso también se puede apreciar que la estabilidad del lenguaje depende en gran medida del uso de adverbios como: *mainly* (principalmente), *frequently* (frecuentemente), *especially* (especialmente), *strongly* (fuertemente), *potentially* (potencialmente), etc; las palabras que definen el tema como: *erythematosus* (eritematoso), *systemic* (sistémico), *SLE*, *patient* (paciente); además de las palabras herramienta como: *furthermore* (además), *moreover* (por otra parte), *even* (incluso), *however* (sin embargo), *also* (también), *and* (y), entre otras.

En la Figura 7.4 se muestra la descomposición del vocabulario especializado. Al

Vocabulario local o especializado				Vocabulario regular o estable							
Categoría	Palabra	Clase de Frec.	Índice	Categoría	Palabra	Clase de Frec.	Índice	Categoría	Palabra	Clase de Frec.	Índice
Adjetivo	articular	[10-20]	0,29	Adjetivo	complex	[10-20]	0,61	Sustantivo	documented	[10-20]	0,62
	cytotoxic		0,34		estimated		0,61		complications		0,61
	biologic		0,37		decreasing		0,62		classification		0,73
	systolic		0,37		dependent		0,62		majority		0,69
	vascular	[21-50]	0,29		even		0,67		influence		0,65
Nombre propio	pga	[10-20]	0,25		experienced	[21-50]	0,62		medication	[21-50]	0,66
	chloroquine		0,35		possible		0,68		evaluation		0,65
	vasculitis		0,37		consecutive		0,64		development		0,63
	snps	[21-50]	0,31		erythematosus		0,64		evidence		0,7
	blys		0,32		systemic	[51-200]	0,62		findings	[51-200]	0,65
	belimumab	[51-200]	0,14	Adverbio	mainly	[10-20]	0,74		conclusion	[>200]	0,73
	dhea		0,15		approximately		0,62		methods		0,77
	bmd		0,17		potentially		0,61		results		0,74
	rituximab		0,19		recent		0,78		patient		0,61
	methotrexate		0,21		recently		0,63	Verbo	treating	[10-20]	0,79
	bilag		0,26		moreover		0,7		applied		0,71
Sustantivo	interferon	[10-20]	0,26		furthermore		0,72		conclude		0,71
	brain		0,26		especially		0,67		monitored		0,63
	anaemia		0,34		commonly		0,61		occur		0,75
	malarrash		0,35		therefore	[21-50]	0,7		predict		0,75
	chest		0,37		strongly		0,65		define		0,66
	april	[21-50]	0,19		frequently		0,71		established		0,61
	fever		0,27		clinically		0,61		decreases		0,63
	cholesterol		0,29		however		0,66		described		0,75
	intervention		0,33		also	[51-200]	0,63		confirmed		0,71
	lung		0,34	Conjunción	and	[>200]	0,65		demonstrate		0,71
	bone		0,34	Nombre propio	sle	[>200]	0,6		confirm		0,63
	fatigue		0,36	Preposición	against	[10-20]	0,69		affected		0,65
	tcells		0,37		with	[>200]	0,61		indicate		0,67
	genotype	[51-200]	0,28	Sustantivo	reactions	[10-20]	0,76		analysed	[21-50]	0,61
	bcells		0,29		hypothesis		0,61		analyzed		0,66
	depletion		0,34		observation		0,68		controlled		0,63
	damage		0,35		show		0,74		identified		0,64
	allele		0,36		report		0,76		suggest		0,67
	polymorphism		0,37		improvements		0,73		evaluated		0,61
	placebo	[>200]	0,32		retrospective		0,69		determine		0,65

Tabla 7.7: Índice de reparto del vocabulario de la base de LES

realizar esta descomposición se visualiza el modelo de evolución mediante los gráficos de Bertin y mediante AC se proyectan las palabras de evolución cronológica.

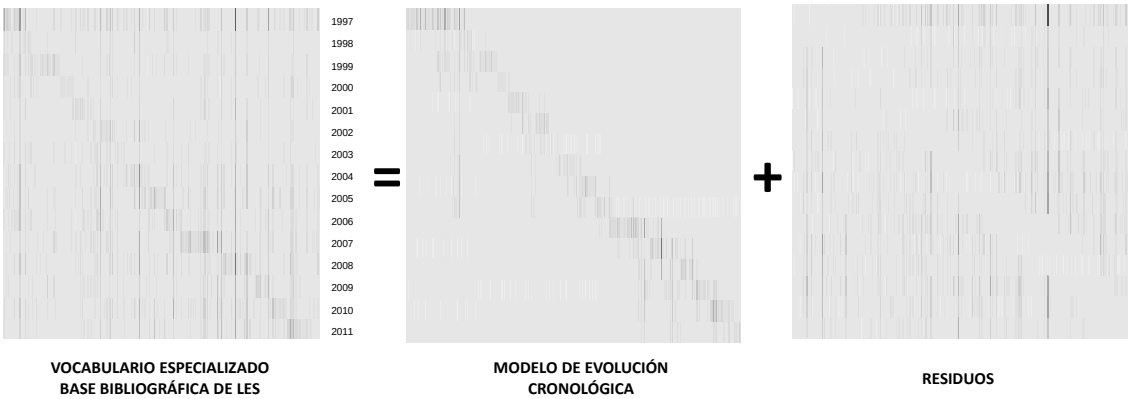


Figura 7.4: Descomposición del vocabulario especializado

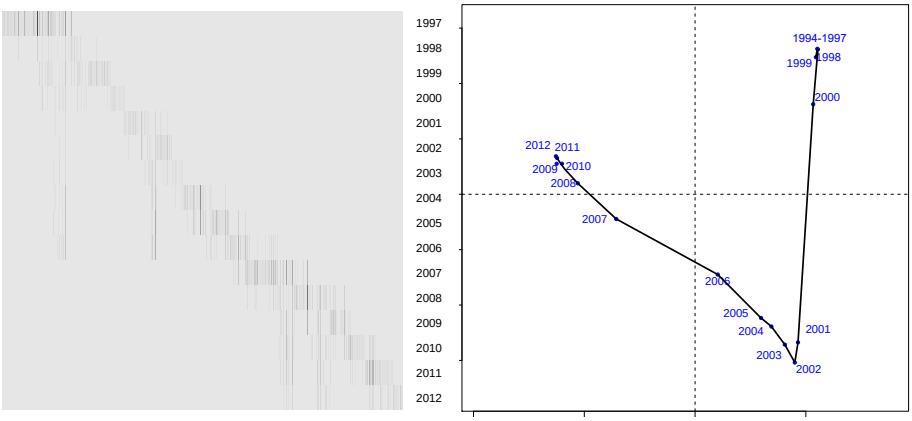
En el Capítulo 5, al estudiar cómo ha evolucionado el vocabulario de LES, se encontraron tres períodos de desarrollo (1994-2000, 2001-2006 y 2007-2012). En la Tabla 7.8 se confirman los períodos encontrados en ese capítulo, en el que las palabras cronológicas que

Palabra	Índice	Frecuencia	P.value	Año(s)	Palabra	Índice	Frecuencia	P.value	Año(s)
chest	0,37	11	0,00	1994	fever	0,27	21	0,00	2005
corticosteroid	0,48	71	0,00	1994	thrombosis	0,4	30	0,00	2005
lipoprotein	0,4	23	0,00	1995	nephritis	0,44	73	0,00	2006
prednisone	0,38	109	0,00	1995	remission	0,45	52	0,00	2006
pressure	0,46	19	0,01	1995	systolic	0,37	18	0,00	2006
pulmonary	0,39	25	0,00	1995	antibody	0,42	298	0,00	2002-2006
urinary	0,49	22	0,01	1995	depletion	0,34	53	0,00	2002-2006
anticoagulant	0,37	21	0,00	1996	women	0,46	239	0,00	2002-2006
cyclophosphamide	0,38	99	0,00	1996	dhea	0,15	121	0,00	2004-2005
manifestation	0,45	116	0,00	1996	vascular	0,29	45	0,00	2004-2005
renal	0,45	154	0,00	1996	prognostic	0,44	11	0,50	2004-2006
azathioprine	0,48	25	0,02	1997	allele	0,36	91	0,00	2007
cytotoxic	0,34	19	0,00	1997	anaemia	0,34	15	0,00	2007
lung	0,34	31	0,00	1997	april	0,19	22	0,00	2007
methotrexate	0,21	67	0,00	1997	cell	0,47	38	0,00	2007
organ	0,48	52	0,00	1997	cohort	0,45	147	0,00	2007
steroid	0,39	75	0,00	1997	haplotype	0,38	17	0,00	2007
dna	0,47	76	0,00	1999	polymorphism	0,37	89	0,00	2007
osteoporosis	0,44	10	0,30	1999	proteinuria	0,43	41	0,00	2007
platelet	0,44	20	0,00	1999	rituximab	0,19	67	0,00	2007
fatigue	0,36	52	0,00	2000	tcells	0,37	41	0,00	2007
illness	0,38	12	0,01	2000	biologic	0,37	10	0,01	2008
abnormalities	0,46	40	0,00	1994-1996	bmd	0,17	53	0,00	2008
articular	0,29	11	0,00	1994-1996	gene	0,43	95	0,00	2009
thrombocytopenia	0,38	21	0,01	1994-1996	depression	0,38	22	0,00	2010
vasculitis	0,37	17	0,01	1994-1996	interferon	0,26	12	0,00	2011
brain	0,26	16	0,00	1995-1996	neuropsychiatric	0,39	56	0,00	2011
psychosis	0,49	13	0,03	1995-1996	pga	0,25	17	0,00	2011
urine	0,46	17	0,15	1995-1996	sledai	0,39	170	0,00	2011
bone	0,34	37	0,00	1998-1999	atherosclerosis	0,44	26	0,00	2007-2012
cutaneous	0,42	26	0,00	1998-1999	bilag	0,26	82	0,00	2007-2012
flares	0,48	78	0,00	1998-1999	snps	0,31	23	0,00	2007-2012
malarrash	0,35	12	0,14	1998-1999	bcells	0,29	52	0,00	2008-2010
genotype	0,28	66	0,00	2001	belimumab	0,14	54	0,00	2008-2012
lymphocytes	0,39	38	0,00	2001	blys	0,32	24	0,00	2008-2012
chloroquine	0,35	18	0,00	2002	estrogen	0,41	28	0,00	2008-2012
cholesterol	0,29	44	0,00	2002	pharmacokinetics	0,44	11	0,42	2008-2012
lipid	0,43	10	0,03	2002	placebo	0,32	205	0,00	2008-2012
hormonal	0,44	11	0,02	2003	plasma	0,43	62	0,00	2008-2012
intervention	0,33	36	0,00	2004	cardiovascular	0,44	48	0,00	2011-2012

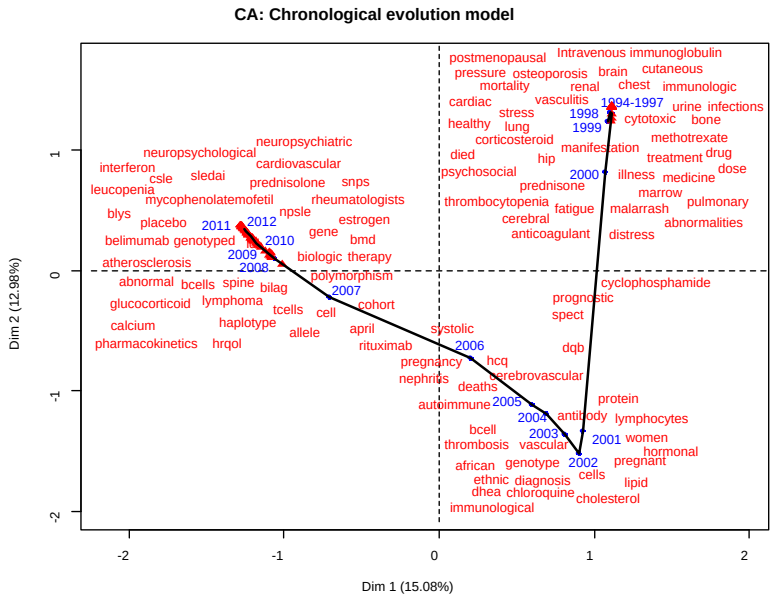
Tabla 7.8: Vocabulario cronológico de la base bibliográfica LES

caracterizan el primer período (1994-2000) son las relacionadas con los síntomas de LES y los tratamientos que se utilizaron fueron antimaláricos. De 2001-2006 hay un crecimiento en el vocabulario; aparecen nuevos medicamentos y la investigación se centra en el estudio de las causas y de 2007-2012, se realizan estudios de cohorte y aparecen los medicamentos conocidos como terapias biológicas (Belimumab y Rituximab).

La separación del corpus bibliográfico de LES, de acuerdo con la función que desempeñan las palabras, permite claramente seguir el flujo de desarrollo del tema, encontrar los momentos o períodos en que se producen cambios importantes en la investigación y



(a) Representación: gráficos de Bertin (b) Trayectoria: efecto Guttman



(c) Trayectoria de las partes y palabras cronológicas en el primer plano factorial de AC

Figura 7.5: Modelo de evolución cronológica de la base de LES

determinar las causas o factores que los provocaron (Figura 7.5).

7.5 CONCLUSIONES

La distinción sobre el uso del vocabulario, ya sea de forma regular o especializada, y la caracterización de este uso en cada una de las partes, así como la representación gráfica con las matrices de reorden de Bertin, permiten definir los diversos temas abordados, también encontrar el ritmo del corpus y detectar los giros o retrocesos mal empleados en el desarrollo del tema.

Cuando un corpus cronológico, ya sea estucturado o no estructurado, muestra una renovación sistemática del vocabulario y cada parte o período define una temática o desem-

peña un papel en el proceso de desarrollo, garantiza que el corpus analizado tiene una buena organización y que en la representación de su trayectoria, a través de un AC, las posiciones de las partes en que fue dividido el discurso, o los años en que fueron publicados los abstracts sobre el primer plano factorial, presentan una forma de arco o modelo de evolución cronológico.

Con el modelo de evolución se garantiza una mejor interpretación de la organización y estructura del corpus, permite hacer un comparativo y observar cuánta diferencia se presenta entre la trayectoria del corpus y la definida sólo por el vocabulario de evolución cronológica. Porque en un análisis completo se pueden presentar partes atípicas o bucles que rompen con la regularidad del corpus y dificultan su interpretación.

Capítulo 8

Funciones en R para el análisis estadístico de textos

8.1 Introducción

El desarrollo de programas informáticos enfocados al análisis de datos textuales surgen a la par del desarrollo de los métodos multidimensionales y sus aplicaciones en este campo. Aguirre (2003) hace un recuento sistemático del surgimiento e implicaciones de los primeros sistemas de información especializados en datos textuales. Primero, Lebart y Morineau (1984) desarrollaron una aplicación de tratamiento de textos en el programa SPAD; después, en 1989, Bécue-Bertaut desarrolla un sistema informático para el análisis de datos textuales Bécue-Bertaut (1989), el mismo que se incorpora al programa SPAD en versión SPAD.T (Bécue-Bertaut, 1991; Lebart et al., 1989); paralelamente al trabajo de Bécue-Bertaut, A. Salem desarrolla el programa Le Lexicloud (Salem, 1987), que es el inicio del software hoy conocido como *Lexico*.

Hoy en día, en el comercio se encuentran una gran variedad de herramientas especializadas en minería de textos; un estudio comparativo de ellas se puede consultar en (Spinakis y Chatzimakri, 2005; van Gemert, 2000). Sin embargo, dado su carácter comercial, en este trabajo no se pone énfasis en sus características. Aquí sólo interesan los que fueron diseñados con el mismo propósito, pero con la ventaja de que la comunidad académica puede acceder libremente a ellos. Estos son, principalmente, los desarrollados en el entorno del software R (Core, 2014) tales como: *Text Mining* (Feinerer, 2008; Feinerer y Hornik, 2012), *textometry* (Loiseau et al., 2014), *koRpus*(Michalke, 2014), *FactoMineR* (Husson et al., 2015; Lê et al., 2008), *RcmdrPlugin.temis* (Bouchet-Valat y Bastin, 2013) y *IRaMuTeQ* (Ratinaud, 2009). Cada uno de estos paquetes ofrecen herramientas para el análisis de datos textuales.

A pesar de la existencia de programas especializados en datos textuales, el conjunto de funciones programadas en esta tesis, es una herramienta versátil y con aplicación en grandes conjuntos de datos. Se hace uso de los métodos factoriales implementados en *FactorMineR* (análisis de correspondencias, análisis factorial múltiple y los métodos de clasificación); se aprovechan las ventajas que ofrece *Text Mining* en la construcción del corpus, la matriz documentos×palabras y las listas de los *stop-word*; se incorpora nueva metodología para el estudio de textos desde un punto de vista cronológico (Clasificación cronológica, clasificación con restricción de contigüidad, crecimiento específico del vocabulario, índice de reparto del vocabulario, palabras cronológicas, etc.)

8.2 Funciones en R

El número de funciones programadas son 30 (Tabla 8.1). Estas se dividen en tres tipos:

- funciones simples, programadas para implementar un método específico.
- funciones macros, (MacroBiblio, MacroTxChrono y MacroCaHcpc), subrutinas que ayudan en la organización y análisis de textos y hacen uso de varias funciones simples de acuerdo con un procedimiento metodológico que garantice el análisis completo de determinado tipo de textos.
- funciones métodos que muestran de manera ordenada los resultados de salida de las funciones simples o macros; ya sea a través de una impresión en pantalla (print) o mediante un resumen (summary).

Herramientas para el análisis estadístico de textos	
Función	Descripción
CharDocWord	Characteristic Documents and Words (CharDocWord)
DocVarTable	Documents by Variables Table (DocVarTable)
DocWordTable	Documents by Words Table (DocWordTable)
HierarchWords	Hierarchical Words (HierarchWords)
MacroBiblio	Analysis of Bibliography (MacroBiblio)
MacroCaHcpc	Correspondence Analysis and Hierarchical Clustering (MacroCaHcpc)
MacroTxChrono	Chronological Corpus (MacroTxChrono)
MDocWordTable	Multiple Document by Words Table (MDocWordTable)
META.CA	Metakeys-Metadocs (META.CA)
print.DocWordTable	Prints DocWordTable results
print.MacroBiblio	Prints MacroBiblio results
print.MacroCaHcpc	Prints MacroCaHcpc results
print.MacroTxChrono	Prints MacroTxChrono results.
print.TxCA	Prints TxCA results
print.TxCHCPC	Prints TxCHCPC results
print.TxMFACT	Prints TxMFACT results
print.VocIndex	Prints VocIndex results
summary.DocWordTable	Summary DocWordTable objects
summary.MacroBiblio	Summary MacroBiblio objects
summary.MacroTxChrono	Summary MacroTxChrono objects
summary.TxCA	Summary TxCA object
summary.TxMFACT	Summary TxMFACT objects
summary.VocIndex	Summary VocIndex objects
TxCA	Correspondence Analysis of Lexical Tables (TxCA)
TxCharClust	Characteristic Documents and Words of the Clusters (TxCharClust)
TxCHCPC	Constrained Hierarchical Clustering (TxCHCPC)
TxMFACT	Multiple Factor Analysis Contingency Tables for Textual Data (TxMFACT)
uCutDoc	Cut the sentences in homogeneous group (uCutDoc)
uHomo.groups	Homogeneous groups (uHomo.groups)
uSentences	Arbitrary sentences (uSentences)
VocIndex	Index of Vocabulary (VocIndex)

Tabla 8.1: Funciones en R

En el anexo 9 se pueden consultar de forma detallada cada una de las funciones. También en este capítulo se presentan dos aplicaciones con el fin de mostrar el uso de MacroBiblio y MacroTxChrono. Macro funciones utilizadas en los capítulos 4 y 5.

8.2.1 Relaciones entre funciones

En los siguientes diagramas se muestran las relaciones entre las funciones.

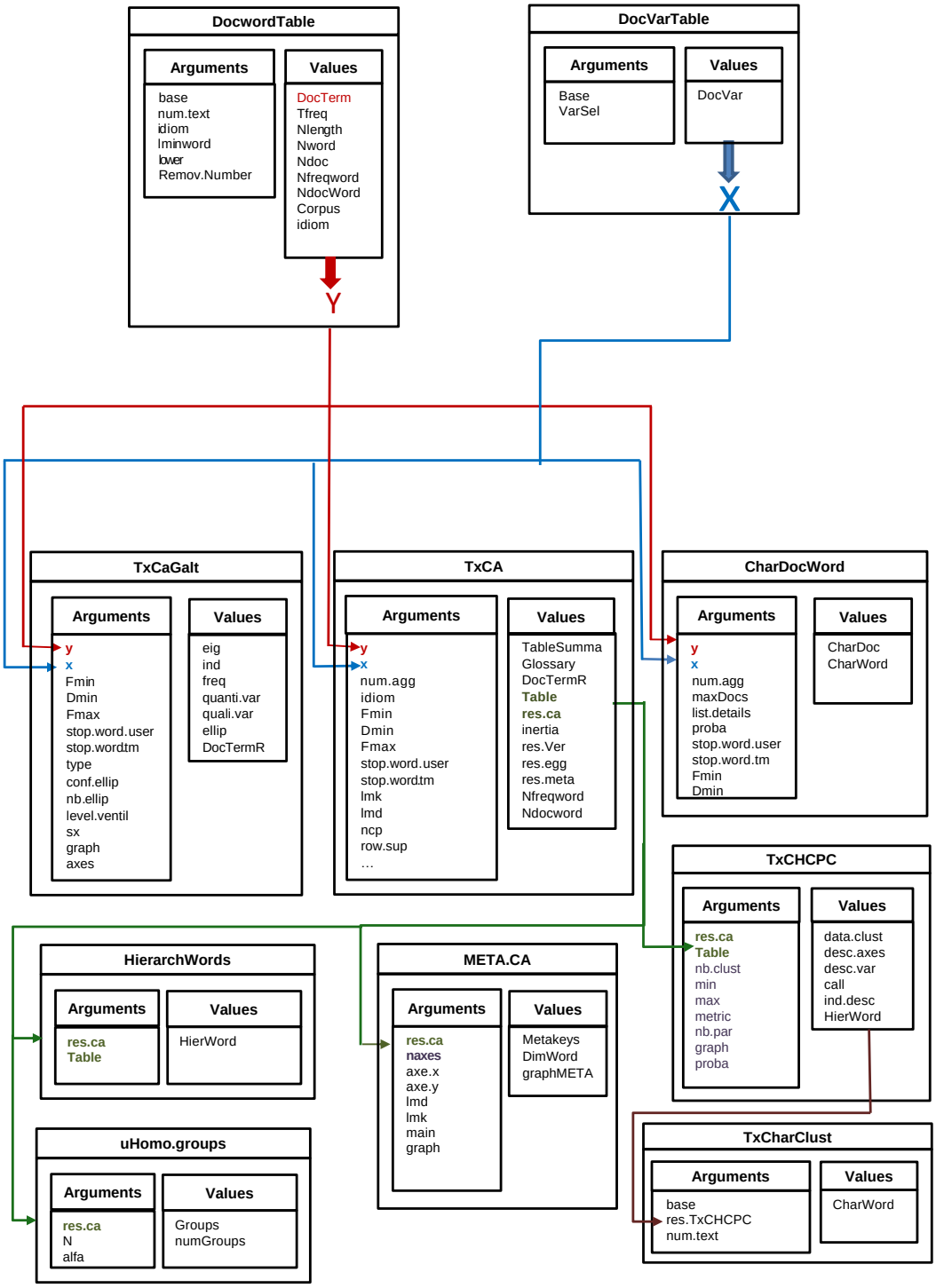


Figura 8.1: Diagramas de relación

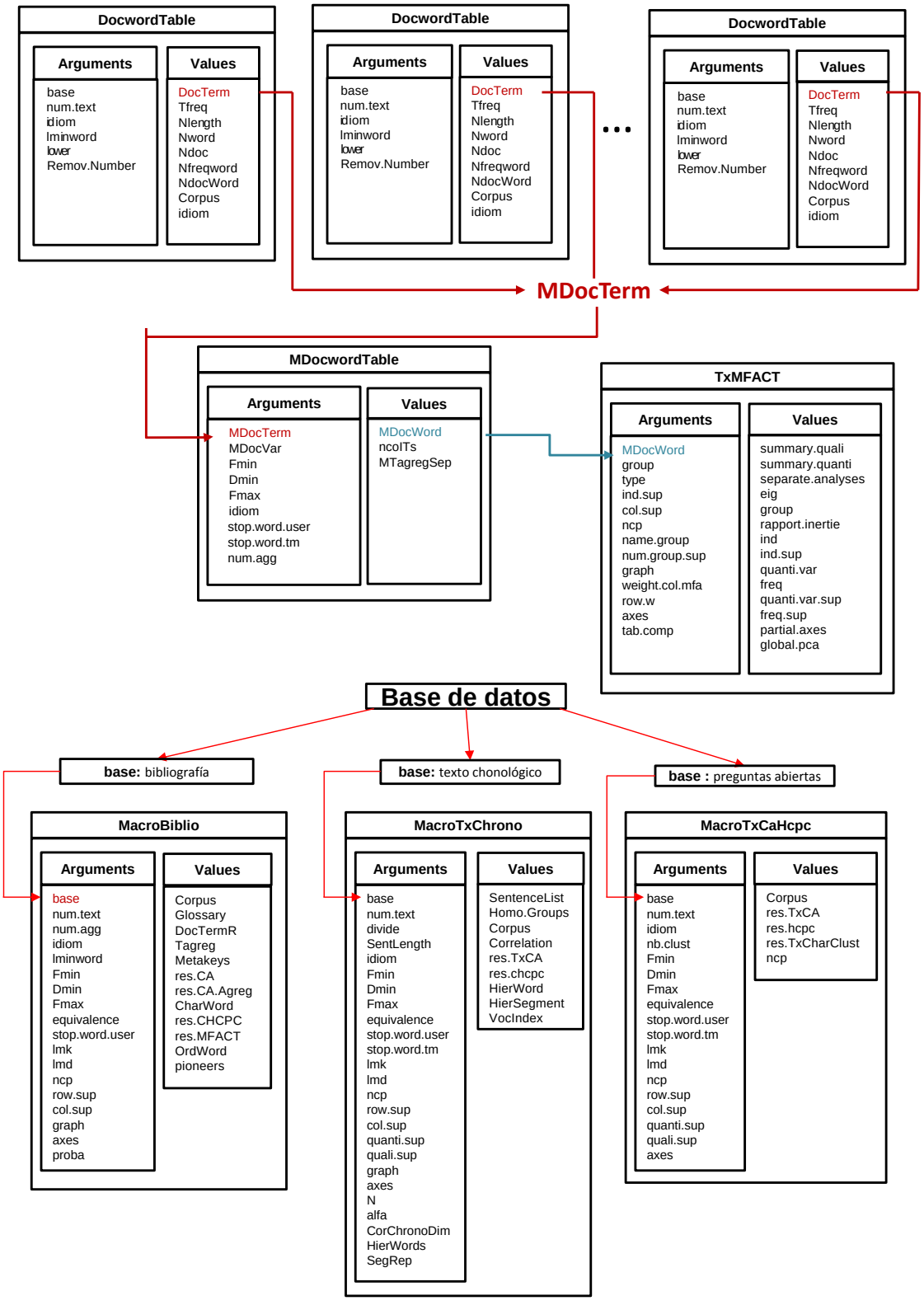


Figura 8.2: Diagramas de relación

8.3 MacroBiblio

Esta función es una herramienta de análisis bibliográfico relativamente simple que permite comparar, analizar y extraer información importante de una base bibliográfica de artículos científicos. La metodología, que se aplica a cualquier colección de artículos científicos, ha sido implementada en una función llamada *MacroBiblio*.

La entrada de la macro función *MacroBiblio* en R es:

```
MacroBiblio (base,num.text="Abstract", num.agg="Year", idiom ="en",lminword=3,
Fmin=10, Dmin=5,Fmax=NULL,equivalence=NULL,stop.word.user=NULL, lmd=3, lmk=3,
ncp=10, row.sup=NULL, col.sup=NULL,graph=TRUE, axes=c(1,2), proba=0.01)
```

8.3.1 Argumentos y valores

Argumentos

Los argumentos de *MacroBiblio* son:

- *base*: base de datos con I filas (abstracts/artículos) y J columnas. El nombre de las principales columnas deben ser: Title, Year, Abstract, Journal; La base de datos también puede tener otro tipo de variables, ya sean cuantitativas o cualitativas tales como: autor, país, etc.
- *num.text*: índice o nombre de la variable textual (por defecto "Abstract´´)
- *num.agg*: índice o nombre de la variable de agregación (por defecto "Year´´)
- *idiom*: idioma de la variable textual (por defecto inglés)
- *lminword*: longitud mínima de la palabra (por defecto 3)
- *Fmin*: umbral mínimo de frecuencia de la palabra (por defecto 10)
- *Dmin*: umbral mínimo de aparición de la palabra en al menos Dmin documentos (por defecto 5)
- *Fmax*: umbral máximo de frecuencia de la palabra
- *equivalence*: tabla o matriz con n filas y dos columnas (palabra original y palabra nueva)
- *stop.word.user*: vector con los stopwords seleccionadas por el usuario
- *lmd*: umbral mínimo de contribución para seleccionar los metadocumentos (por defecto tres veces mayor a la contribución media)
- *lmk*: umbral mínimo de contribución para seleccionar los metallaves (por defecto tres veces mayor a la contribución media)
- *ncp*: número de dimensiones consideradas en los resultados (por defecto 10)

- *row.sup*: vector con los índices o nombres de las filas suplementarias
- *col.sup*: vector con los índices o nombres de las columnas suplementarias
- *graph*: valor booleano, si es TRUE se muestran los gráficos
- *axes*: un vector de longitud 2 especificando las dimensiones de la gráfica
- *proba*: umbral de significación para caracterízar las palabras dentro de las categorías (por defecto 0.01)

Valores

Los valores de salida de *MacroBiblio* son:

- *Corpus*: resumen del corpus
- *Glossary*: glossario con la frecuencia de las palabras
- *DocTermR*: tabla documentos×palabras
- *Tagreg*: tabla léxica agregada
- *Metakeys.Metadocs*: representación de las palabras y documentos con mayor contribución
- *res.CA*: resultados del AC directo
- *res.CA.Agreg*: resultados del AC agregado
- *CharWord*: palabras características en cada categoría de la variable agregada
- *res.CHCPC*: resultados de la clasificación jerárquica con restricción de contigüidad
- *spec.growth*: crecimiento específico del vocabulario
- *res.MFACT*: resultados del AFMTC
- *OrdWord*: palabras ordenadas de acuerdo a sus coordenadas en la primera dimensión de AFMTC
- *pioneers*: artículos pioneros

8.3.2 Aplicación

Para ilustrar los resultados obtenidos con *MacroBiblio*, utilizamos la misma base bibliográfica presentada en el capítulo dos, relativa a Lupus Eritematoso Sistémico entre 1994 a 2012. La dimensión de la base de datos es de 506 filas (abstract) y 5 columnas (Título, año, revista, autor y abstract)

```
res.dataBiblio<-MacroBiblio(base, Fmin=20, Dmin=5, lmd=6, lmk=3, ncp =5)
```

8.3.3 Resultados

Los resultados son muchos, pero están organizados y etiquetados de acuerdo con el método implementado y pueden ser escudriñados haciendo print o una llamada.

```
> names(res.dataBiblio)
```

```
 [1] "Corpus"            "Glossary"          "DocTermR"          "Tagreg"
 [5] "Metakeys.Metadocs" "res.CA"            "res.CA.Agreg"      "CharWord"
 [9] "res.CHCPC"         "spec.growth"       "res.MFACT"         "OrdWord"
[13] "pioneers"
```

```
> print(res.dataBiblio)
```

```
**Results of Analysis of bibliography (MacroBiblio)**
*The results are available in the following objects:

   name                 description
1  "$Corpus"            "Summary of the information about corpus"
2  "$Glossary"          "Glossary of words"
3  "$DocTermR"          "Documents by words table"
4  "$Tagreg"            "Lexical aggregate table"
5  "$Metakeys.Metadocs" "Representation of words/documents with higher contribution "
6  "$res.CA"            "Results of  correspondence analysis direct"
7  "$res.CA.Agreg"      "Result of Correspondence analysis aggregate"
8  "$CharWord"          "characteristic words in each group of the aggregation variable"
9  "$res.CHCPC"         "Result of constrained hierarchical clustering"
10 "$res.MFACT"         "Result of multiple factor analysis for contingency tables"
11 "$OrdWord"           "words and their coordinates in the first dimension"
12 "$pioneers"          "pioneers articles"
```

La visualización e intepretación de los resultados es más fácil utilizando la función *summary.MacroBiblio*, la cual imprime los resultados más importantes en el análisis bibliográfico.

```
summary.MacroBiblio(object, nword=50, nEig=5, ...)
```

Los argumentos de la función summary pueden ser modificados.

```
> summary.MacroBiblio(res.dataBiblio,nword=10)
```

```
CORPUS

DocWordTable summary

Number of documents
  506
Corpus size
 91142
Vocabulary size
 6222
Glossary of the  10  most frequent  words
              Frequency N.Documents
the                4792         497
```

```
and                4317         499
with               2384         489
patients           2273         462
sle                2126         466
were               1649         448
was                1501         447
for                1074         385
disease             864         310
lupus               832         471

CORRESPONDENCE ANALYSIS CA

Summary table of information
                Occurrences Documents Dif-Words Doc-Average-Length
Before-selection       91142       506      6222             180.12
After-selection        47473       506       669              93.82

Call:
MacroBiblio(base, Fmin = 20, Dmin = 5, lmd = 6, lmk = 3, ncp = 5,
     num.agg = "Year_class")

Eigenvalues
                         dim 1     dim 2     dim 3     dim 4     dim 5
Variance               0.29023   0.25546   0.20718   0.19878   0.18370
% of var.              1.96564   1.73016   1.40318   1.34626   1.24410
Cumulative % of var.   1.96564   3.69580   5.09898   6.44524   7.68935

Total Inertia    Cramer V
  14.7654         0.171

Rows (the 10 first)
                   Dim.1      ctr     cos2      Dim.2      ctr     cos2      Dim.3      ctr     cos2
1              |  0.05535  0.00193  0.00011 |  0.99801  0.71452  0.03563 |  0.28719  0.07295  0.00295 |
2              |  0.89137  0.35753  0.06407 | -0.40129  0.08233  0.01299 |  0.62905  0.24943  0.03191 |
3              | -0.47185  0.16320  0.01143 |  0.04797  0.00192  0.00012 |  0.45149  0.20932  0.01047 |
4              | -0.08137  0.00553  0.00100 |  0.16261  0.02507  0.00398 | -0.20144  0.04744  0.00610 |
5              |  1.38086  1.07944  0.09096 | -0.30269  0.05893  0.00437 |  0.87479  0.60687  0.03651 |
6              | -0.40795  0.18118  0.01738 |  0.05062  0.00317  0.00027 | -0.13313  0.02703  0.00185 |
7              | -0.58641  0.30698  0.03008 |  0.07798  0.00617  0.00053 | -0.07265  0.00660  0.00046 |
8              | -0.61398  0.29549  0.02656 | -0.13508  0.01625  0.00129 |  0.95379  0.99890  0.06409 |
9              | -0.59787  0.32947  0.02628 |  0.00110  0.00000  0.00000 |  0.60028  0.46528  0.02650 |
10             | -0.18089  0.01686  0.00325 | -0.33100  0.06414  0.01088 |  0.04769  0.00164  0.00023 |
Columns (the 10 first)
                   Dim.1      ctr     cos2      Dim.2      ctr     cos2      Dim.3      ctr     cos2
abnormal       |  0.74388  0.12852  0.01395 |  0.33444  0.02951  0.00282 | -1.24278  0.50250  0.03893 |
abnormalities  |  0.35059  0.03568  0.00411 | -0.60002  0.11874  0.01205 | -1.25440  0.63992  0.05267 |
according      |  0.04892  0.00052  0.00013 |  0.00841  0.00002  0.00000 |  0.14422  0.00634  0.00111 |
achieved       | -0.57869  0.07778  0.01409 | -0.30703  0.02487  0.00397 | -0.56833  0.10509  0.01359 |
acid           |  0.10875  0.00197  0.00012 | -0.37459  0.02661  0.00144 | -0.16050  0.00602  0.00026 |
acl            |  1.24610  0.52967  0.02324 | -0.71496  0.19810  0.00765 | -1.00695  0.48452  0.01517 |
acr            |  0.41667  0.06174  0.01116 |  1.16220  0.54573  0.08686 |  0.05739  0.00164  0.00021 |
across         |  0.01508  0.00003  0.00001 |  0.65414  0.07057  0.01170 |  0.41169  0.03446  0.00463 |
activation     |  0.35889  0.05328  0.00291 | -1.18723  0.66247  0.03182 | -0.52219  0.15802  0.00616 |
active         | -0.29883  0.12314  0.01458 | -0.34984  0.19174  0.01998 | -0.22230  0.09546  0.00807 |

Glossary of the  10  most frequent words after selection
             Frequency N.Documents
patients          2273         462
sle               2126         466
disease            864         310
lupus              832         471
activity           592         224
systemic           588         453
erythematosus      542         448
treatment          505         209
study              501         307
group              477         160
treatment          400         167
```

```
METAKEYS-METADOCS

$DIM1
$DIM1$`Metakeys+`
 [1] "association"    "allele"         "gene"           "susceptibility" "polymorphism"   "polymorphisms"
 [7] "hla"            "associated"     "controls"       "risk"           "sle"            "snps"
[13] "genetic"        "alleles"        "genotypes"      "genes"          "expression"     "genotype"
[19] "genotyped"      "cohort"         "associations"   "beta"           "african"        "acl"
$DIM1$`Metadocs+`
[1] "181" "190" "87"  "99"

$DIM1$`Metakeys-`
 [1] "placebo"      "group"        "treatment"    "bmd"          "prasterone"   "dhea"
 [7] "months"       "day"          "dose"         "weeks"        "calcium"      "randomized"
[13] "spine"        "prednisone"   "therapy"      "trial"        "blind"        "methotrexate"
[19] "belimumab"    "month"
$DIM1$`Metadocs-`
[1] "151" "162" "261"

$DIM2
$DIM2$`Metakeys+`
 [1] "damage"        "health"        "disease"       "physical"      "hrqol"         "sdi"
 [7] "social"        "factors"       "quality"       "scores"        "status"        "self"
[13] "activity"      "mental"        "csle"          "american"      "psychosocial"  "canada"
[19] "life"          "costs"         "rheumatology"  "slicc"         "duration"      "fatigue"
$DIM2$`Metadocs+`
 [1] "350" "441" "172" "119" "115" "130" "450" "290" "284" "209" "302" "193"

$DIM2$`Metakeys-`
 [1] "cells"         "cell"          "beta"          "expression"    "depletion"     "rituximab"
 [7] "anti"          "lymphocytes"   "gene"          "blood"         "activation"    "serum"
[13] "dna"           "levels"        "peripheral"    "treatment"     "proliferation" "antibody"
$DIM2$`Metadocs-`
[1] "192" "280" "310"
...
DIMENSION OF WORD IN METAKEYS-METADOCS
                Dim Total.Dim %Dim
anti              4         5   80
bmd               4         5   80
calcium           4         5   80
prasterone        4         5   80
spine             4         5   80
abnormal          3         5   60
abnormalities     3         5   60
acl               3         5   60
alone             3         5   60
...

AGGREGATE CA

Correspondence analysis summary
                Occurrences Documents  Dif-Words    Mean-length
Before-selection       91142       506       6222      180.12
After-selection        47473       506        669       93.82

Call:
MacroBiblio(base, Fmin = 20, Dmin = 5, lmd = 6, lmk = 3, ncp = 5,
     num.agg = "Year_class")

Eigenvalues
                         dim 1     dim 2     dim 3     dim 4     dim 5
Variance               0.05949   0.03913   0.03148   0.03012   0.02847
% of var.             27.64018  18.18113  14.62764  13.99554  13.23048
Cumulative % of var.  27.64018  45.82131  60.44895  74.44450  87.67498

Total Inertia   Cramer V
  0.2152          0.1894
```

```
Rows
                Dim.1      ctr     cos2      Dim.2      ctr     cos2      Dim.3      ctr     cos2
<1997         | -0.41905 35.06992  0.58488 | -0.08708  2.30223  0.02526 | -0.11880  5.32607  0.04701 |
1998-2000     | -0.33009 23.71672  0.43969 |  0.01452  0.06981  0.00085 |  0.18545 14.14611  0.13879 |
2001-2003     | -0.04855  0.45476  0.00957 |  0.21243 13.23534  0.18317 | -0.39098 55.72434  0.62047 |
2004-2006     |  0.00066  0.00016  0.00000 |  0.09981  5.48267  0.09096 |  0.17904 21.92760  0.29267 |
2007          |  0.35143 17.13011  0.28739 |  0.42482 38.05463  0.41995 |  0.05413  0.76798  0.00682 |
2008-2009     |  0.19145  8.42049  0.17251 | -0.16681  9.71891  0.13097 |  0.02168  0.20397  0.00221 |
2010-2012     |  0.21140 15.20784  0.29081 | -0.24533 31.13642  0.39164 | -0.05441  1.90393  0.01927 |
Columns (the 10 first)
                Dim.1      ctr     cos2      Dim.2      ctr     cos2      Dim.3      ctr     cos2
abnormal      | -0.10934  0.01355  0.04747 | -0.11614  0.02324  0.05355 |  0.35172  0.26488  0.49112 |
abnormalities | -0.55594  0.43777  0.76583 | -0.12525  0.03378  0.03887 | -0.14383  0.05537  0.05126 |
according     | -0.12127  0.01562  0.08730 |  0.01185  0.00023  0.00083 | -0.23079  0.10692  0.31616 |
achieved      | -0.04176  0.00198  0.00559 |  0.01563  0.00042  0.00078 |  0.05886  0.00742  0.01111 |
acid          | -0.67591  0.37208  0.62385 | -0.32116  0.12771  0.14085 |  0.14576  0.03270  0.02901 |
acl           | -0.65056  0.70436  0.26240 |  0.43468  0.47806  0.11714 |  0.36352  0.41559  0.08193 |
acr           |  0.17762  0.05474  0.10307 |  0.41882  0.46271  0.57309 | -0.11010  0.03974  0.03960 |
across        |  0.09388  0.00624  0.02347 | -0.17852  0.03431  0.08488 |  0.33741  0.15235  0.30323 |
activation    |  0.23844  0.11475  0.13541 | -0.11861  0.04317  0.03351 |  0.01385  0.00073  0.00046 |
active        |  0.00631  0.00027  0.00147 |  0.05338  0.02915  0.10505 |  0.00624  0.00049  0.00143 |

Aggregation of documents according to the categorical variable
          Documents Non-empty-before Non-empty-after
<1997            68               68              68
1998-2000        73               73              73
2001-2003        57               57              57
2004-2006       103              103             103
2007             42               42              42
2008-2009        65               65              65
2010-2012        98               98              98
overall         506              506             506

Distribution of the words in the groups
          Occurrences-before % Of Total Occurrences-after Mean-length Words-before Words-after
<1997                  11195      12.28             5640      164.63         2167         601
1998-2000              12183      13.37             6147      166.89         2260         613
2001-2003              10147      11.13             5448      178.02         1887         606
2004-2006              19727      21.64            10223      191.52         2888         659
2007                    7754       8.51             3917      184.62         1746         564
2008-2009              12138      13.32             6488      186.74         2226         633
2010-2012              17998      19.75             9610      183.65         2709         648
overall                91142     100.00            47473      180.12         6222         669

Glossary of the  10  most frequent words after selection
             Frequency N.Documents
patients          2273         462
sle               2126         466
disease            864         310
lupus              832         471
activity           592         224
systemic           588         453
erythematosus      542         448
treatment          505         209
study              501         307
group              477         160
treatment          400         167

CHARACTERISTIC WORDS OF AGGREGATED VARIABLE

$`<1997`
$`<1997`$Over_represented_word
 [1] "manifestations"   "methotrexate"     "management"       "dietary"          "lung"
 [6] "corticosteroids"  "pulse"            "drug"             "transient"        "side"
[11] "patients"         "pulmonary"        "hypertension"     "renal"            "urinary"
$`<1997`$Infra_represented_word
 [1] "standard"       "background"     "response"       "african"        "association"    "tnf"
```

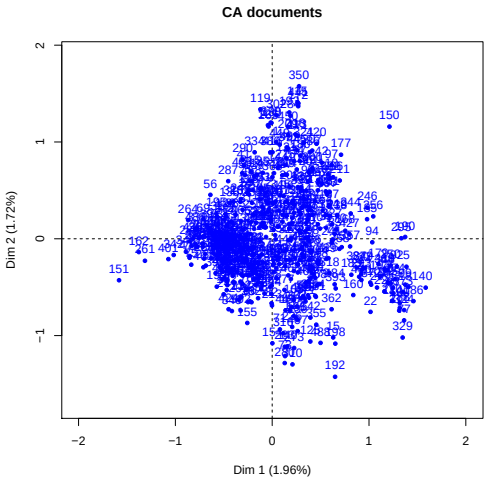
```
 [7] "levels"         "polymorphisms"  "time"           "polymorphism"   "year"           "acl"
[13] "expression"     "beta"           "cardiovascular" "mmf"            "gene"           "scores"
[19] "belimumab"      "juvenile"       "bmd"            "antibody"       "allele"         "versus"
...
$‘2010-2012‘
$‘2010-2012‘$Over_represented_word
 [1] "belimumab"        "csle"             "cognitive"        "events"           "flare"
 [6] "hdl"              "assessment"       "juvenile"         "sledai"           "performance"
[11] "atherosclerosis"  "attributed"       "analysis"         "neuropsychiatric" "risk"
$‘2010-2012‘$Infra_represented_word
 [1] "selected"   "dna"        "anti"       "six"        "pulmonary"  "social"     "cell"
 [8] "pulse"      "perfusion"  "renal"      "ethnic"     "dsdna"      "hla"        "spect"

MOST CONTRIBUTIVE WORD IN THE DIM1

                Word    Coord.Dim1
582            spect     -1.323
493        pulmonary     -1.279
623 thrombocytopenia     -1.246
494            pulse     -1.073
443        perfusion     -1.031
382     methotrexate     -1.021
34               ana     -0.989
564             side     -0.987
185          dietary     -0.959
181             dhea     -0.816
866           selena      1.058
100        belimumab      1.248
...
PIONEER ARTICLES

N° Year  Journal
188 2007  Arthritis and rheumatism
192 2007  Arthritis and rheumatism
198 2007  Modern rheumatology
202 2007  Journal of immunology
```

Los resultados de salida del summary muestran la información más relevante como: la construcción del corpus (tamaño de corpus, número de documentos, número de palabras y glosario con las palabras ordenadas por su frecuencia), algunos resultados del AC (valores propios, total de inercia, V de Cramer, coordenadas y contribuciones de los documentos y las palabras), los Metallaves-Metadocumentos (palabras/documentos que más contribuyen en la formación de los ejes y que determinan temas), las palabras características e incrementos específicos de los años, las palabras más contributivas en la primera dimensión y los artículos pioneros.

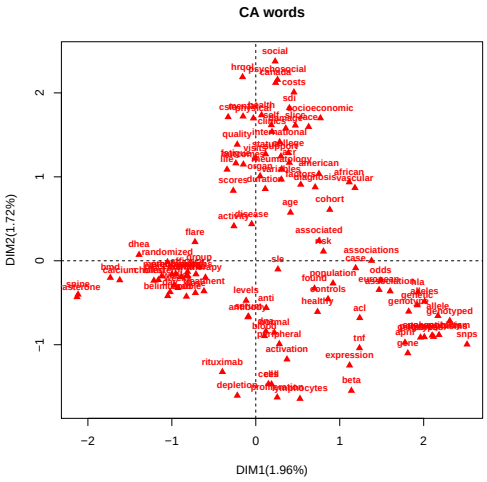
Aquí no se hace análisis de los resultados de salida proporcionados por el summary porque estos fueron analizados detalladamente en el capítulo 4.

8.3.4 Gráficas

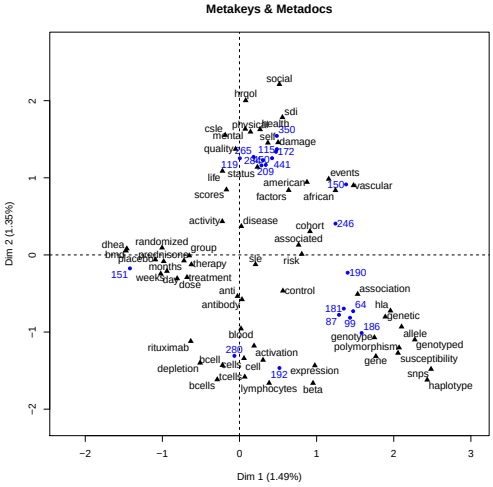
Por defecto la función de *MacroBiblio* devuelve 12 gráficas: 4 del análisis de correspondencias directo y agregado; 3 correspondientes a la clasificación jerárquica con restricción de contigüidad y 5 del análisis factorial múltiple con tablas de contingencia.



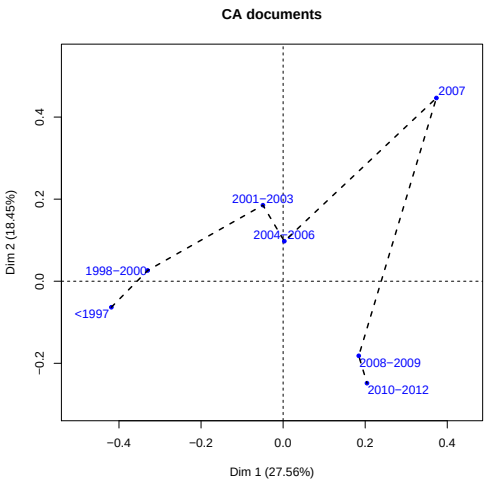
(a) Documentos (abstracts)



(b) Palabras

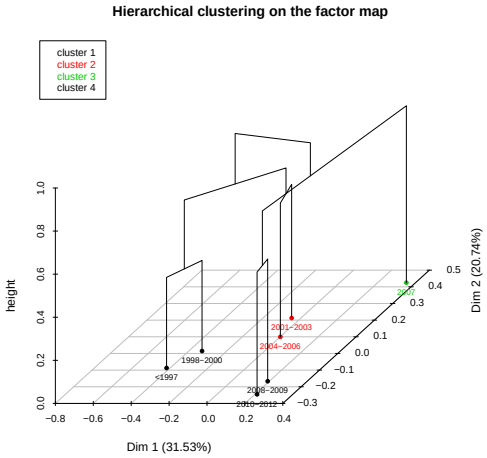


(c) Metakeys-Metadocs

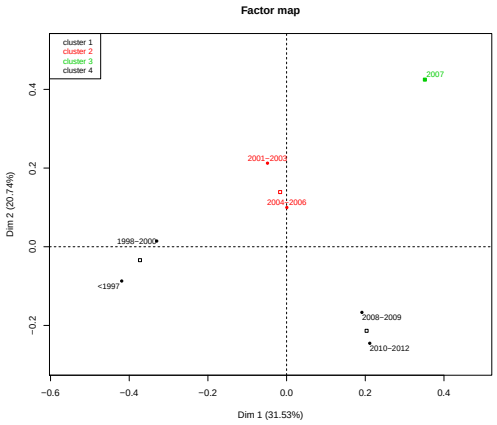


(d) Documentos (Años)

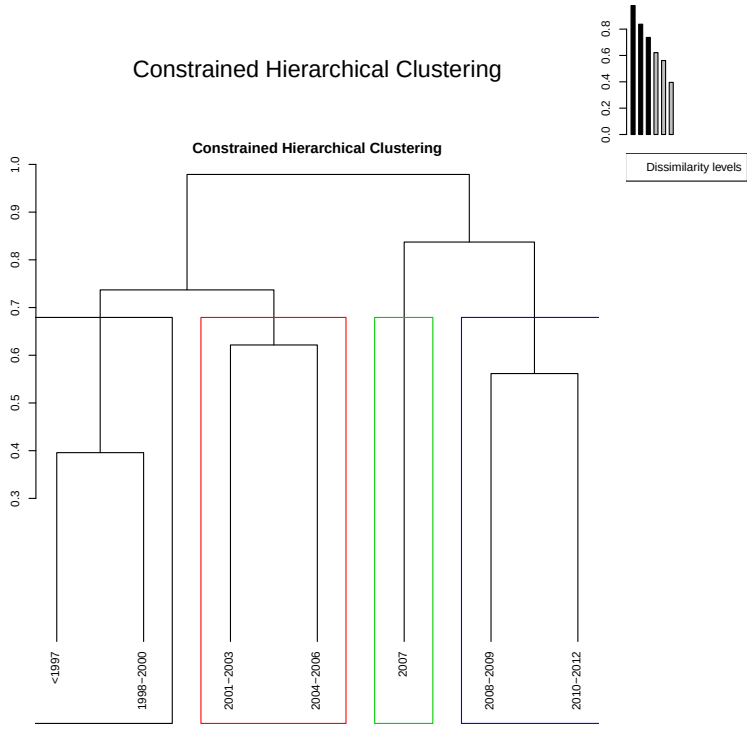
Figura 8.3: Resultados en el primer plano de AC



(a) Dendrograma en tres dimensiones

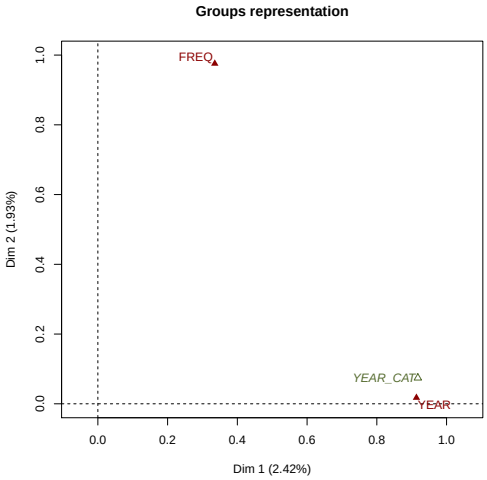


(b) Representación de la partición

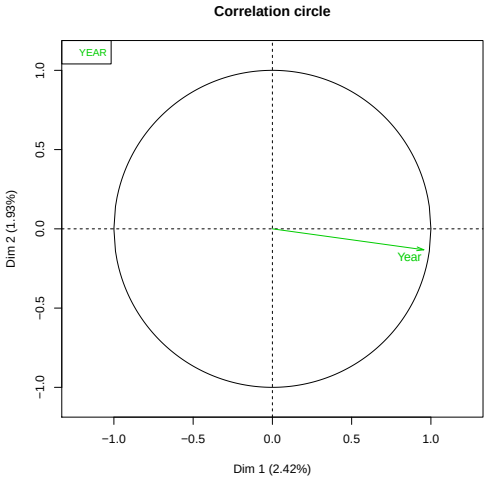


(c) Árbol jerárquico

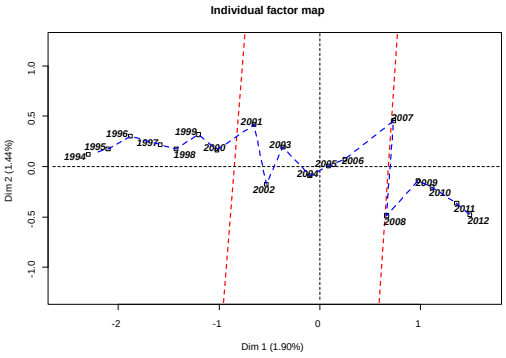
Figura 8.4: Clasificación jerárquica con restricción de contigüidad



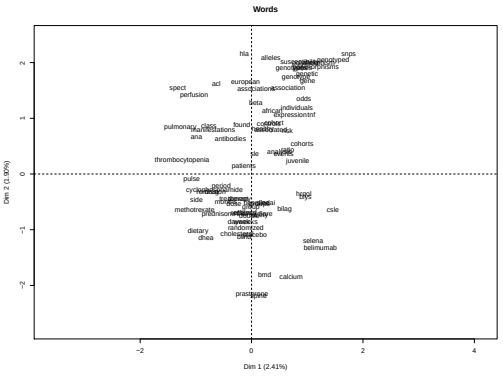
(a) Representación de los grupos



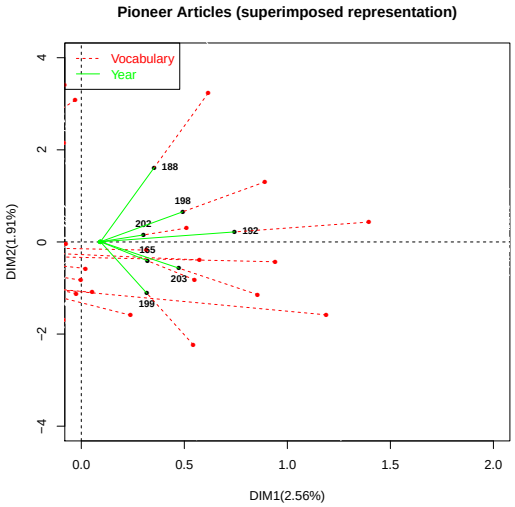
(b) Correlación de los años en el primer eje



(c) Representación parcial de los años



(d) Representación global de palabras más contributivas



(e) Artículo pioneros

Figura 8.5: Resultados de un AFMTC

8.4 MacroTxChrono

MacroTxChrono es una herramienta para el análisis de corpus cronológicos como: discursos políticos, textos literarios, artículos de períodicos sobre un tema, entrevistas, etc. Permite dividir el texto en frases o partes artificialmente homogéneas; crear el corpus y la matriz partesXpalabras; visualizar la trayectoria de las partes; descubrir la estructura jerárquica y analizar la evolución y uso del vocabulario a través de las palabras características e incrementos específicos del vocabulario.

La función de *MacroTxChrono* se presenta por defecto en R como:

```
MacroTxChrono(base, num.text, divide=TRUE, SentLength=100,  idiom ="en",
 Fmin=5, Dmin=1, Fmax=NULL, equivalence=NULL, stop.word.user=NULL,
stop.word.tm=FALSE, lmk=3,lmd=3, ncp=5,  row.sup = NULL, col.sup = NULL,
quanti.sup=NULL, quali.sup = NULL, graph = TRUE, axes = c(1,2), N=5000,
alfa=0.15, CorChornoDim=0.10, SegRep=FALSE)
```

8.4.1 Argumentos y valores

Argumentos

Los argumentos de *MacroTxChrono* son:

- *base*: data frame con al menos una variable textual
- *num.text*: índice o nombre de la variable textual
- *divide*: valor boleano, si es TRUE el corpus es dividido en sentencias arbitrarias de tamaño SentLength
- *SentLength*: longitud de las sentencias (por defecto 100 palabras)
- *idiom*: idioma de la variable textual (por defecto inglés)
- *lminword*: longitud mínima de la palabra (por defecto 3)
- *Fmin*: umbral mínimo de frecuencia de la palabra (por defecto 10)
- *Dmin*: umbral mínimo de aparición de la palabra en al menos Dmin documentos (por defecto 5)
- *Fmax*: umbral máximo de frecuencia de la palabra
- *equivalence*: tabla o matriz con n filas y dos columnas (palabra original y palabra nueva)
- *stop.word.user*: vector con los stopwords seleccionadas por el usuario
- *stop.word.tm*: valor booleano, si es TRUE las palabras que se encuentran en la lista de stopword de *tm* se eliminan

- *lmk*: umbral mínimo de contribución para seleccionar los metallaves (por defecto tres veces mayor a la contribución media)
- *lmd*: umbral mínimo de contribución para seleccionar los metadocumentos (por defecto tres veces mayor a la contribución media)
- *ncp*: número de dimensiones consideradas en los resultados (por defecto 5)
- *row.sup*: vector con los índices o nombres de las filas suplementarias
- *col.sup*: vector con los índices o nombres de las columnas suplementarias
- *quanti.sup*: vector con los índices o nombres de las variables cuantitativas suplementarias
- *quali.sup*: vector con los índices o nombres de las variables cualitativas suplementarias
- *graph*: valor boleano, si es TRUE se muestran los gráficos
- *axes*: un vector de longitud 2 especificando las dimensiones de la gráfica
- N: número de permutaciones (por defecto 5000)
- *CorChornoDim*: umbral para la correlación entre la cronología y las dimensiones (por defecto 0.10)
- *SegRep*: valor boleano, si es TRUE los segmentos repetidos son mostrados

Valores

TxChorono devuelve los siguientes valores:

- *SentenceList*: base de datos con las sentencias
- *Homo.Groups*: lista con la descripción de las partes homogéneas
- *Corpus*: descripción del corpus
- *Correlation*: correlación entre la cronología y las dimensiones
- *res.TxCA*: resultados del análisis de correspondencias
- *res.chcpc*: resultados de la clasificación jerárquica con restricción de contigüidad
- *HierWord*: palabras características en cada nodo de la jerarquía
- *HierSegment*: segmentos característicos en cada nodo de la jerarquía
- *VocIndex*: índice del vocabulario

8.4.2 Aplicación

El texto analizado es un discurso jurídico pronunciado por Robert Badinter, ex abogado designado por el Ministro de Justicia de François Mitterrand, para defender el proyecto de la ley para la abolición de la pena de muerte en Francia, pronunciado el 17 de septiembre 1981, frente a la Asamblea Nacional. Los resultados fueron analizados detalladamente en el capítulo 5.

Los datos se leen de un archivo externo

```
> base<-read.csv2("Badinter.csv", row.names=1)
```

Los los argumentos definidos para esta aplicacion en la función de *MacroTxChrono* son:

```
> stop=c("à","au","l","la","le","les","un","une","d","de",
"des","du","qu","que", "y", "en")
> res.MacroTxChrono<-MacroTxChrono(base, Fmin=5, Dmin=3, idiom="fr", num.text=1,
  stop.word.user=stop, SegRep=TRUE)
```

8.4.3 Resultados

Los resultado están organizados y etiquetados de acuerdo con el método implementado se pueden explorar haciendo print o una llamada.

```
> print(res.MacroTxChrono)
```

```
**Results for the MacroTxChrono**
  name            description
1 "$SentenceList" "dataset with  sentences"
2 "$Homo.Groups"  "Description homogeneous group"
3 "$Corpus"       "Description of corpus"
4 "$Correlation"  "Correlation between chronology and dimensions"
5 "$res.TxCA"     "Results of correspondence analysis"
6 "$res.chcpc"    "Results for the Constrained hierarchical clustering"
7 "$HierWord"     "Characteristic words for every node of the hierarchy"
8 "$HierSegment"  "Characteristic Segments for every node of the hierarchy"
9 "$VocIndex"     "Vocabulary index"
```

Para visualizar y hacer uso de los resultados de manera más eficiente utilizamos la función *summary.MacroTxChrono*. Ésta muestra los resultados relevantes para el análisis.

```
MacroTxChrono summary

DocWordTable summary

Number of documents
  74
Corpus size
 7940
Vocabulary size
```

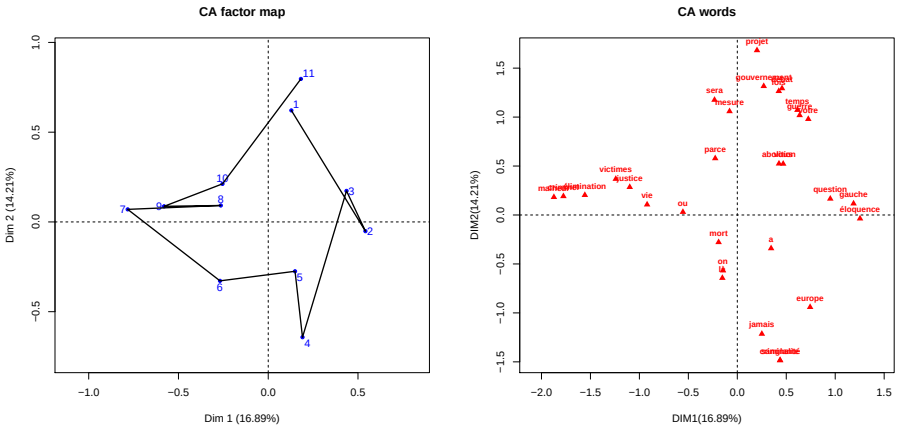
```
 1737
Glossary of the  10  most frequent  words
      Frequency N.Documents
de          469          74
la          380          73
l           200          70
le          164          68
que         160          64
et          157          66
les         142          61
à           139          61
est         113          58
mort        104          54


Correspondence analysis summary

                 Occurrences Documents  Dif-Words   Mean-length
Before-selection        7940        74       1737      107.3
After-selection         3367        74        210       45.5

Call:
MacroTxChrono(base, Fmin = 5, Dmin = 3, idiom = "fr", num.text = 1,
     stop.word.user = stop, SegRep = TRUE)

Eigenvalues
                        dim 1     dim 2     dim 3     dim 4     dim 5
Variance              0.17469   0.14690   0.12392   0.10818   0.10278
% of var.            16.89482  14.20780  11.98508  10.46267   9.93987
Cumulative % of var. 16.89482  31.10262  43.08771  53.55038  63.49025

Total Inertia  Cramer V
  0.3216          1.034
Rows (the 10 first)
                     Dim.1      ctr     cos2      Dim.2      ctr     cos2
1                |  0.12885  0.28791  0.00690 |  0.62124  7.95871  0.16048 |
2                |  0.54142 21.92914  0.35533 | -0.05140  0.23503  0.00320 |
3                |  0.43605 13.22178  0.27551 |  0.17285  2.47050  0.04329 |
4                |  0.19091  2.27420  0.03883 | -0.64218 30.59835  0.43935 |
5                |  0.14938  1.45681  0.03476 | -0.27491  5.86718  0.11773 |
6                | -0.26866  4.94557  0.08519 | -0.32823  8.77748  0.12716 |
7                | -0.78289 33.55461  0.49688 |  0.06994  0.31841  0.00397 |
8                | -0.26420  0.92566  0.01618 |  0.09162  0.13238  0.00195 |
9                | -0.57995 16.86924  0.32591 |  0.08703  0.45171  0.00734 |
10               | -0.25475  2.73640  0.05791 |  0.21204  2.25419  0.04012 |
Columns (the 10 first)
                     Dim.1      ctr     cos2      Dim.2      ctr     cos2
a                |  0.34524  1.39828  0.35720 | -0.34082  1.62037  0.34810 |
abolir           |  0.79613  0.64656  0.35734 | -0.17893  0.03884  0.01805 |
abolition        |  0.42573  1.41749  0.35720 |  0.52431  2.55658  0.54178 |
abolitionnistes | -0.15775  0.02538  0.02264 | -0.53462  0.34670  0.26004 |
abord            | -0.71308  0.60515  0.13286 |  0.55096  0.42959  0.07932 |
ai               |  0.49210  0.69992  0.38800 |  0.42408  0.61811  0.28815 |
ailleurs         |  0.90953  0.70323  0.40921 | -0.63364  0.40586  0.19861 |
ainsi            | -0.46491  0.25723  0.18886 |  0.38803  0.21309  0.13156 |
ait              |  0.66974  0.38131  0.36215 | -0.27817  0.07822  0.06247 |
alors            |  0.05788  0.00570  0.00086 |  0.44279  0.39638  0.05036 |

Aggregation of documents according to the categorical variable
        Documents non-empty-before non-empty-after
1               2                2               2
2              10               10              10
3               9                9               9
4               8                8               8
5               8                8               8
6               9                9               9
7               7                7               7
8               2                2               2
```

```
9              6              6              6
10             6              6              6
11             7              7              7
overall       74             74             74

Distribution of the words in the groups
        Occurrences-before % Of Total Occurrences-after Mean-length Words-before Words-after
1                      223       2.81               102      111.50          138          67
2                     1044      13.15               440      104.40          406         134
3                      965      12.15               409      107.22          381         141
4                      854      10.76               367      106.75          366         109
5                      857      10.79               384      107.12          368         129
6                      980      12.34               403      108.89          386         126
7                      744       9.37               322      106.29          298         109
8                      201       2.53                78      100.50          117          47
9                      659       8.30               295      109.83          285         108
10                     644       8.11               248      107.33          295         109
11                     769       9.69               319      109.86          320         116
overall               7940     100.00              3367      107.30         1737         210

Glossary of the  10  most frequent words after selection
      Frequency N.Documents
et          157          66
est         113          58
mort        104          54
qui          98          56
dans         80          44
peine        77          51
il           70          45
a            69          46
pour         65          42
pas          62          39

Correlation between chronology and dimensions
                Dim.1    Dim.2     Dim.3      Dim.4      Dim.5
Correlation 0.3572169 0.200985 0.0619215 0.03542674 0.05220848

  2  axes will be taken into account in the constrained hierarchical clustering

Hierarchical words
            Group Node  Intern %    glob % Intern freq Glob freq       p.value   v.test
crime        4-10   19 0.6676204 0.4455004          14         15 1.638319e-02 2.400265
homme        4-10   19 0.7629948 0.5049005          16         17 7.085092e-03 2.692818
mort         4-10   19 3.9103481 3.0888031          82        104 3.626355e-04 3.565881
on           4-10   19 2.2889843 1.6632017          48         56 1.889492e-04 3.733351
réalité      4-10   19 0.4291845 0.2673003           9          9 2.801621e-02 2.197059
ses          4-10   19 0.3814974 0.2376002           8          8 4.504857e-02 2.004201
abolition     1-3   18 2.4185068 1.3662014          23         46 2.728396e-03 2.996789
assemblée     1-3   18 0.7360673 0.2970003           7         10 1.460160e-02 2.442112
messieurs     1-3   18 0.4206099 0.1485001           4          5 4.908264e-02 1.967873
non           1-3   18 0.8412198 0.3564004           8         12 1.255237e-02 2.496223
contre        4-6   17 0.6065858 0.2376002           7          8 6.159655e-03 2.739159
...

VocIndex summary

    10 Regular Words                      10 specialized Words
                Word Index Freq                  Word Index Freq
505             sûr  0.88    5      1303      malheur  0.14    9
514          autre   0.83    8      573       opinion  0.24    7
281             ma   0.81    5      1149   terrorisme  0.26    5
484 abolitionnistes  0.79    6      251        jaurès  0.27    5
303             moi  0.78    5      1664       demain  0.28    5
229         humaine  0.76    7      232     éloquence  0.28    6
314      simplement  0.76    8      1367  élimination  0.28    7
194          encore  0.75    6      362      attendre  0.30    5
428          simple  0.75    5      299        gauche  0.30    7
```

Los resultados de salida proporcionados por summary.MacroTxChrono sintetizan la información más relevante del corpus tal como: el tamaño de corpus, el número de documentos y palabras que lo componen, el glosario con las palabras ordenadas por su frecuencia, los resultados del AC (valores propios, total de inercia, V de Cramer, coordenadas y contribuciones de las partes y las palabras), las correlaciones entre la cronología y las dimensiones, el número de ejes a considerar para la clasificación jerárquica con restricción de contigüidad y las palabras divididas según su uso (palabras regulares o especializadas).

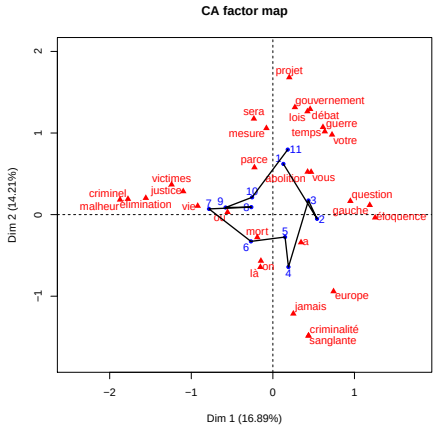
8.4.4 GRÁFICAS

MacroTxChrono, por defecto, devuelve 6 gráficas; 3 de la aplicación de AC y las otras de la clasificación con restricción de contigüidad.



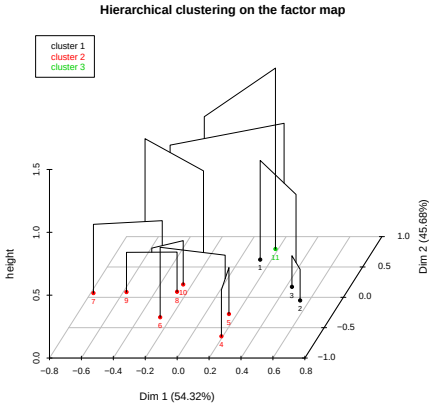
(a) Trayectoria de las partes

(b) Palabras más contributivas

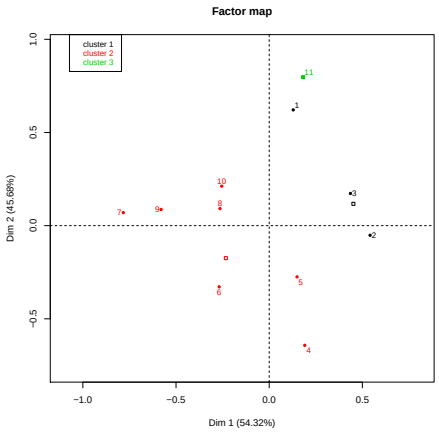


(c) Representación de las palabras y las partes

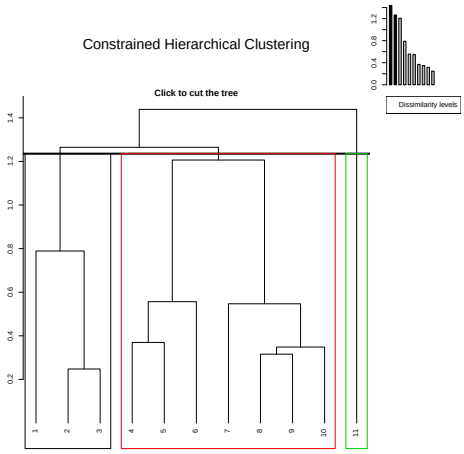
Figura 8.6: Resultados en el primer plano de AC



(a) Dendrograma en tres dimensiones



(b) Representación de las partes homogéneas



(c) Árbol jerárquico

Figura 8.7: Clasificación jerárquica con restricción de contigüidad

8.5 Conclusiones

Las características principales del conjunto de funciones han sido explicadas e ilustradas en este capítulo, a través de la implementación de dos macro funciones (MacroBiblio y MacroTxChrono), utilizando la base bibliográfica de LES (analizada en el capítulo 4) y el discurso de Badinter sobre la abolición de la pena de muerte (analizado en el capítulo 5).

En el anexo A se dan ejemplos de cada una de las funciones disponibles. También se pueden revisar detalladamente los argumentos y valores de las funciones.

Capítulo 9

Conclusiones

En esta tesis se presenta un nuevo procedimiento metodológico y un conjunto de funciones programadas en R para el análisis estadístico de textos. Mediante su aplicación se obtienen resultados interesantes cuando se estudian corpus que pueden ser organizados en una tabla documentos×palabras como son: bases de datos bibliográficos, discursos, críticas literarias, encuestas con preguntas abiertas, ensayos, opiniones en las redes sociales, textos narrativos, etc.

La metodología propuesta y las principales características del conjunto de funciones programadas en R se han presentado a través de dos ejemplos. El primer ejemplo corresponde a los datos de una base bibliográfica y el segundo, a un discurso retórico. Aunque se trata de dos conjuntos de datos con una estructura completamente distinta, la metodología propuesta permite analizarlos de forma similar; es decir, el análisis en ambos casos parte de una tabla múltiple compuesta por una tabla léxica documentos × palabras y una variable contextual correspondiente a la cronología. En el primer caso se analiza la tabla abstracts×palabras y se yuxtapone la fecha de publicación (variable cronológica); en el segundo caso, el discurso se divide en partes lexicalmente homogéneas; por lo tanto, la tabla múltiple se compone de una tabla léxica partes×palabras y la secuencia de las partes determina la variable cronológica. Los resultados proporcionados mediante estos dos ejemplos demuestran por una parte, que la implementación de un procedimiento para el análisis de corpus, desde un punto de vista cronológico, es de gran utilidad para los investigadores porque ofrece información de interés en el análisis de textos, la cual es difícil de obtener a través de un estudio convencional.

Por otra parte, al analizar los resultados por separado, se logra contestar a preguntas específicas del corpus estudiado. Por ejemplo, en el caso de una base bibliográfica ¿Cuáles son los temas más relevantes de la investigación? ¿Existen períodos homogéneos en la investigación? ¿Qué es lo que determina cada período?, etc. Cuando se analiza un discurso se responde a las siguiente preguntas: ¿El discurso está bien organizado? ¿En cuántas partes se puede dividir? ¿Qué papel desempeña cada una de las palabras? ¿Cuáles son las palabras que evolucionan al discurso?, etc.

Algunos estudios bibliográficos han sido realizados anteriormente y el análisis de correspondencias fue el método factorial más utilizado (Kerbaol et al., 2006; Morin, 2004, 2006). Sin embargo, la aplicación de otros métodos tales como: Análisis Factorial Múltiple de Tablas de Contingencias, Clasificación Jerárquica con Restricción de Contigüidad, Clasificación Cronológica, Palabras Cronológicas e Incrementos Específicos, fue una contribución innovadora de esta tesis en el análisis bibliográfico porque proporcionan resultados

relevantes sobre la evolución del vocabulario, los períodos o momentos en los que se producen cambios en la investigación, el uso del vocabulario y los artículos pioneros.

En el caso de textos no estructurados, como el análisis de discursos, la metodología ofrece la posibilidad de descubrir la organización evolutiva del texto y representar su estructura a través de un árbol etiquetado o mediante gráficas de Bertin, con el fin de facilitar el flujo argumentativo del discurso, dado que cada palabra desempeña un papel específico en cada una de las partes en que está dividido.

Todos los métodos presentados en esta tesis se implementaron en R. Los resultados que proporcionan las macro funciones hacen de éstas una herramienta útil y con aplicación en grandes conjuntos de datos. En anexos se encuentra un manual para facilitar la comprensión de cómo utilizar cada una de las funciones e interpretar las salidas, y cada función cuenta con una ayuda en la que se proporciona un ejemplo fácil de interpretar. Posteriormente a la presntación de esta tesis, se pretende incorporar nuevos métodos al análisis de corpus cronológicos, como el Análisis Simultaneo (Zarraga y Goitisolo, 2006; Zárraga y Goitisolo, 2011), crear otras macro funciones y desarrollar dos paquetes en R, BiblioMineR y TextoChrono. Partiendo de la macro función MacroBiblio, la cual combina varios métodos estadísticos para el análisis de datos bibliográficos, se pretende crear BiblioMineR. Este nuevo paquete especializado en análisis bibliográfico, además de incorporar la metodología implementada en MacroBiblio, debe contar con nuevas funciones que ayuden a la importación de bases de datos y a su transformación en un formato que se pueda leer desde R como csv, txt, etc. así como con resultados de salida más accesible a los usuarios y mejores representaciones gráficas.

El otro paquete, TextoChrono, se desarrollará a partir de la función MacroTxChrono. El objetivo es proporcionar una herramienta sencilla y más completa para el análisis de textos cronológicos. Se pretende incorporar las funciones que ya fueron programadas y utilizadas para obtener algunos de los resultados mostrados en la tesis. Estas funciones son las que implementan los métodos propuestos por Huber y Labbé, Índice del Vocabulario y Crecimiento del Vocabulario, las que representan las matrices de reorden de Bertin y las que determinan las funciones de las palabras.

En resumen, esta tesis propone un nuevo procedimiento que combina varios métodos estadísticos para el análisis de corpus cronológicos y un conjunto de funciones en R, en donde se implementa la metodología.

Bibliografía

Aguirre, K. F. (2003). Análisis textual: generación y aplicaciones. *Metodología de Encuestas*, 5(1):55–66.

Bansard, J.-Y., Rebholz-Schuhman, D., Cameron, G., Clark, D., Van Mulligen, E., Beltrame, F., Barbolla, E. D. H., Martin-Sanchez, F., Milanesi, L., Tollis, I., et al. (2007). Medical informatics and bioinformatics: a bibliometric study. *IEEE Transactions on Information Technology in Biomedicine*, 11(3):237.

Bécue-Bertaut, M. (1989). *Un Sistema informático para el análisis estadístico de datos textuales*.

Bécue-Bertaut, M. (1991). *Análisis de datos textuales: Métodos estadísticos y algoritmos*. Cisia.

Bécue-Bertaut, M. (2014). Tracking verbal-based methods beyond conventional descriptive analysis in food science bibliography. a statistical approach. *Food Quality and Preference*, 32:2–15.

Bécue-Bertaut, M., Alvarez-Esteban, R., y Pagès, J. (2008). Rating of products through scores and free-text assertions: Comparing and combining both. *Food Quality and Preference*, 19(1):122–134.

Bécue-Bertaut, M., Kostov, B., Morin, A., y Naro, G. (2014). Rhetorical strategy in forensic speeches: multidimensional statistics-based methodology. *Journal of Classification*, 31(1):85–106.

Bécue-Bertaut, M. y Pagès, J. (2004). A principal axes method for comparing contingency tables: Mfact. *Computational Statistics & Data Analysis*, 45(3):481–503.

Bécue-Bertaut, M. y Pagès, J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis*, 52(6):3255–3268.

Benzécri, J. P. (1973). *Analyse des données. Analyse des correspondances (Vol. 2)*.

Benzécri, J. P. (1977). Histoire et préhistoire de lÂ´analyse des données. partie v: lÂ´analyse des correspondances. *Les Cathiers de lÂ´Analyse des Donées*, 2:9–40.

Benzécri, J. P. (1981). *Pratique de lâ analyse des données. Linguistique & lexicologie (Vol. 3)*.

Bertin, J. (1973). Sémiologie graphique: Les diagrammes-les réseaux-les cartes.

Bertin, J. (1977). *La graphique et le traitement graphique de l'information*, volume 1. Flammarion Paris.

Bertin, J. (1981). *Graphics and graphic information processing*. Walter de Gruyter.

Bouchet-Valat, M. y Bastin, G. (2013). Rcmdrplugin.temis, a graphical integrated text mining solution in r. *The R Journal*, 5(1):188–196. ISSN 2073-4859.

Cazes, P. y Moreau, J. (1991). Analysis of a contingency table in which the rows and the columns have a graph structure. *E. Diday & Y. Lechevallier (éds.) Symbolic-numeric data analysis and learning. New York: Nova Science Publishers*, pages 271–280.

Chauchat, J. y Risson, A. (1995). Amado,a new method and a software integrating jacques bertin's graphics and multidimensional data analysismethods. In *International Conference on Visualization of Categorical Data*.

Chen, C.-H. (2002). Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica*, 12(1):7–30.

Core, T. R. (2014). R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria, 2012.

Escofier, B. (1983). Analyse de la différence entre deux mesures définies sur le produit de deux mêmes ensembles. *Les Cahiers de l'analyse des données*, 8(3):325–329.

Escofier, B. (2003). *Analyse des correspondances*.

Escofier, B. y Pagès, J. (1988). *Analyses factorielles simples et multiples*. Dunod.

Escofier, B. y Pagès, J. (1992). *Análisis Factoriales Simples y Multiples. Objetivos, Métodos e interpretación*. Bilbao.

Falguerolles, A., Friedrich, F., Sawitzki, G., y Heidelberg, S. (1997). A tribute to j. bertin's graphical data analysis. *SoftStat*, 97:11–20.

Feinerer, I. (2008). An introduction to text mining in r. *R News*, 8(2):19–22.

Feinerer, I. y Hornik, K. (2012). tm: Text mining package. *R package version 0.5-7.1*.

for Biotechnology Information, N. C. (2012). Pubmed.

Greenacre, M. y Lewi, P. (2009). Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of classification*, 26(1):29–54.

Hubert, P. y Labbé, D. (1990a). La répartition des mots dans le vocabulaire présidentiel (1981-1988). *Mots*, 22(1):80–92.

Hubert, P. y Labbé, D. (1990b). Note sur líndice de répartition utilisé dans íindex du vocabulaire de f. mitterrand. 7(1):1–135.

Husson, F., Josse, J., Le, S., Mazet, J., y Husson, M. F. (2015). Package, factominer.

Husson, F., Lê, S., y Pagès, J. (2010). *Exploratory multivariate analysis by example using R*. CRC press.

Kerbaol, M., Bansard, J., y Coatrieux, J.-L. (2006). An analysis of ieee publications. *Engineering in Medicine and Biology Magazine, IEEE*, 25(2):6–9.

Kerbaol, M. y Bansard, J.-Y. (2000). Sélection de la bibliographie des maladies rares par la technique du vocabulaire commun minimum. In *Proceedings of JADT2000: 5th Journées Internationales d Analyse Statistique des Données Textuelles Lausanne*.

Kozyrev, S. V., Lewén, S., Reddy, P., Linga, M., Pons-Estel, B., Witte, T., Junker, P., Laustrup, H., Gutiérrez, C., Suárez, A., et al. (2007). Structural insertion/deletion variation in irf5 is associated with a risk haplotype and defines the precise irf5 isoforms expressed in systemic lupus erythematosus. *Arthritis & Rheumatism*, 56(4):1234–1241.

Labbé, D. y Hubert, P. (2016). La repartition du vocabulaire. artículo no publicado (comunicación personal de los autores).

Lê, S., Josse, J., Husson, F., et al. (2008). Factominer: an r package for multivariate analysis. *Journal of statistical software*, 25(1):1–18.

Lebart, L. y Morineau, A. (1984). Spad tome iii: Analyse des données textuelles. *Cisia, Paris*.

Lebart, L., Morineau, A., Bécue-Bertaut, M., y Haeusler, L. (1989). Systéme portable pour l'analyse des données textuelles (spad-t)[portable system for the analysis of textual data]. *Paris: CESIA*.

Lebart, L., Morineau, A., y Piron, M. (1997). *Statistique exploratoire multidimensionnelle*. Dunod.

Lebart, L., Salem, A., y Bécue-Bertaut, M. (2000). *Análisis estadístico de textos*. Editorial Milenio.

Lebart, L., Salem, A., y Berry, L. (1998). *Exploring textual data*, volume 4. Springer Science & Business Media.

Legendre, P. y Legendre, L. (1998). Numerical ecology: second english edition. *Developments in environmental modelling*, 20.

Loiseau, S., Vaudor, L., Decorde, M., Heiden, S., y Decorde, M. M. (2014). Package textometry.

Michalke, M. (2014). korpus: An r package for text analysis. *Version 0.04-40), last verified*, 11:2013.

Morin, A. (2004). Intensive use of correspondence analysis for information retrieval. In *Information Technology Interfaces, 2004. 26th International Conference on*, pages 255–258. IEEE.

Morin, A. (2006). Intensive use of factorial correspondence analysis for text mining: application with statistical education publications. In *ICOTS-7 (International Conference on Teaching Statistics), Salvador, Bahia, Brazil*.

Murtagh, F. (1985). Multidimensional clustering algorithms. *Compstat Lectures, Vienna: Physika Verlag, 1985*.

Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R.* CRC Press.

Murtagh, F., Ganz, A., y McKie, S. (2009). The structure of narrative: the case of film scripts. *Pattern Recognition*, 42(2):302–312.

Murtagh, F., Ganz, A., y Reddington, J. (2011). New methods of analysis of narrative and semantics in support of interactivity. *Entertainment Computing*, 2(2):115–121.

Nakayamada, S., Saito, K., Nakano, K., y Tanaka, Y. (2007). Activation signal transduction by β1 integrin in t cells from patients with systemic lupus erythematosus. *Arthritis & Rheumatism*, 56(5):1559–1568.

Pagès, J. y Bécue-Bertaut, M. (2006). Multiple factor analysis for contingency tables. In Greenacre, M. y Blasius, J., editors, *Multiple correspondence analysis and related methods*, pages 299–326. Chapman & Hall / CRC PRESS.

Perin, C., Dragicevic, P., y Fekete, J.-D. (2014). Revisiting bertin matrices: New interactions for crafting tabular visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2082–2091.

Ratinaud, P. (2009). Iramuteq: Interface de r pour les analyses multidimensionnelles de textes et de questionnaires. *Téléchargeable á l'adresse: http://www. iramuteq. org*.

Risson, A. (1994). *AMADO:Analyse graphique d'une matrice de donnees : guide pratique*. Saint-Mande (1 Av. Herbillon, 94160) : CISIA.

Rouillier, J., Bansard, J.-y., y Kerbaol, M. (2002). Application de lÂ´analyse statistique des données textuelles à une revue bibliographique de la littérature médicale. In *Journées internationales d' Analyse statistique des Données Textuelles*, pages 665–676.

Salem, A. (1987). Le lexicloud. programes pour le traitament lexicométrique des textes. *Ecole Normale Supérieure, Fontenay-Saint Cloud*.

Šilić, A., Morin, A., Chauchat, J.-H., y Bašić, B. D. (2012). Visualization of temporal text collections based on correspondence analysis. *Expert Systems with Applications*, 39(15):12143–12157.

Spinakis, A. y Chatzimakri, A. (2005). Comparative study of text mining tools. *Studies in Fuzziness and Soft Computing*, 185:223.

Tanaka, Y., Yamamoto, K., Takeuchi, T., Nishimoto, N., Miyasaka, N., Sumida, T., Shima, Y., Takada, K., Matsumoto, I., Saito, K., et al. (2007). A multicenter phase i/ii trial of rituximab for refractory systemic lupus erythematosus. *Modern Rheumatology*, 17(3):191–197.

Valencia, X., Yarboro, C., Illei, G., y Lipsky, P. E. (2007). Deficient cd4+ cd25high t regulatory cell function in patients with active systemic lupus erythematosus. *The Journal of Immunology*, 178(4):2579–2588.

van Gemert, J. (2000). Text mining tools on the internet. *ISIS technical report series*, 25.

Zarraga, A. y Goitisolo, B. (2006). Simultaneous analysis: a joint study of several contingency tables with different margins. *Multiple Correspondence Analysis and Related Methods. Boca Raton, IL: Chapman and Hall/CRC*, pages 327–350.

Zárraga, A. y Goitisolo, B. (2011). Simultaneous analysis in s-plus: The simultan package. *Journal of Statistical Software*, 70(11).

Manual de funciones en R

Tabla 1: Índice

CharDocWord	Characteristic Documents and Words (CharDocWord)	3
DocVarTable	Documents by Variables Table (DocVarTable)	7
DocWordTable	Documents by Words Table (DocWordTable)	8
HierarchWords	Hierarchical Words (HierarchWords)	10
MacroBiblio	Analysis of Bibliography (MacroBiblio)	14
MacroCaHcpc	Correspondence Analysis and Hierarchical Clustering (MacroCaHcpc)	20
MacroTxChrono	Chronological Corpus (MacroTxChrono)	22
MDocWordTable	Multiple Document by Words Table (MDocWordTable)	26
META.CA	Metakeys-Metadocs (META.CA)	29
print.DocWordTable	Prints DocWordTable results	32
print.MacroBiblio	Prints MacroBiblio results	34
print.MacroCaHcpc	Prints MacroCaHcpc results	35
print.MacroTxChrono	Prints MacroTxChrono results.	36
print.TxCA	Prints TxCA results	37
print.TxCHCPC	Prints TxCHCPC results	39
print.TxMFACT	Prints TxMFACT results	40
print.VocIndex	Prints VocIndex results	42
summary.DocWordTable	Summary DocWordTable objects	44
summary.MacroBiblio	Summary MacroBiblio objects	45
summary.MacroTxChrono	Summary MacroTxChrono objects	46
summary.TxCA	Summary TxCA object	47
summary.TxMFACT	Summary TxMFACT objects	48
summary.VocIndex	Summary VocIndex objects	49
TxCA	Correspondence Analysis of Lexical Tables (TxCA)	50
TxCharClust	Characteristic Documents and Words of the Clusters (TxCharClust)	57
TxCHCPC	Constrained Hierarchical Clustering (TxCHCPC)	60
TxMFACT	Multiple Factor Analysis Contingency Tables for Textual Data (TxMFACT)	68
uCutDoc	Cut the sentences in homogeneous group (uCutDoc)	71
uHomo.groups	Homogeneous groups (uHomo.groups)	73
uSentences	Arbitrary sentences (uSentences)	76
VocIndex	Index of Vocabulary (VocIndex)	78

CharDocWord *Characteristic Documents and Words (CharDocWord)*

Description

Documents and words characterizing groups or clusters of documents.

Usage

```
CharDocWord(y, x = NULL, num.agg = NULL, proba = 0.05, maxDocs = NULL,
 list.details = FALSE, Fmin = 5, Dmin = 1, stop.word.user = NULL,
stop.word.tm = FALSE,idiom="en")
```

Arguments

y	an object of class DocWordTable
x	an object of class DocVar
num.agg	column index or name of the aggregation column
proba	significance threshold considered to characterize groups or clusters(by default 0.05)
maxDocs	maximum number of characteristic documents by group (by default 10)
list.details	if TRUE, detailed results on the characteristic words
Fmin	minimum threshold on the word frequency (by default 5)
Dmin	minimum threshold on the number of documents using the word (by default 1)
stop.word.user	vector indicating the stopwords chosen by the user
stop.word.tm	boolean, if TRUE the stopword list provided by tm is taken into account
idiom	language of the textual column(s)

Value

CharDoc	list with the characteristic documents of each group
CharWord	list with the characteristic words of each group

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>, Belchin Kostov

References

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.).

Examples

```
## Not run:
data(dataOpen.question)
y<-DocWordTable(dataOpen.question,num.text=c(6,7))
x<-DocVarTable(dataOpen.question,VarSel=c(1,2,3,4))
res.Char<-CharDocWord(y,x,num.agg=3, proba=1)

####CODE  CharDocWord

CharDocWord<-function(y, x = NULL, num.agg = NULL, proba = 0.05, maxDocs = NULL,
list.details = FALSE, Fmin = 5, Dmin = 1, stop.word.user = NULL,
 stop.word.tm = FALSE, idiom="en"){
if (!inherits(y,"DocWordTable")) stop("non convenient DocTerm")
if (!is.null(x)) {
if (!inherits(x,c("DocVarTable","data.frame","HCPC"))) stop("non convenient DocVar")
if (!is.null(num.agg)){
if(is.character(num.agg)) num.agg <- which(colnames(x)
}
}
if(proba<0|proba>1) stop("proba should be between 0 and 1")
if(!is.null(maxDocs)&is.null(num.agg)) warning("As the documents are not grouped,
          characteristic documents extraction has no meaning")
sel.words<-which(y$Nfreqword >= Fmin & y$Ndocword >= Dmin)
     pos.sparse<-which(y$DocTerm$j
y$DocTerm$j<-y$DocTerm$j[pos.sparse]
y$DocTerm$v<-y$DocTerm$v[pos.sparse]
     y$DocTerm$i<-y$DocTerm$i[pos.sparse]
y$DocTerm$dimnames$Terms<-y$DocTerm$dimnames$Terms[sel.words]
recoderFunc<-function (x, from, to){
mapidx <- match(x, from)
     mapidxNA <- is.na(mapidx)
     from_found <- sort(unique(mapidx))
x[!mapidxNA] <- to[mapidx[!mapidxNA]]
     return(x)
}
y$DocTerm$j<-recoderFunc(y$DocTerm$j,sel.words,1:length(sel.words))
y$DocTerm$ncol<-length(sel.words)
DocTermR <- as.matrix(y$DocTerm)
if(sum(apply(DocTermR,1,sum)==0)>0) pos.elim<-which(apply(DocTermR,1,sum)==0)
DocTermR <- DocTermR[-pos.elim,]
if (!is.null(stop.word.user)) DocTermR <- DocTermR[, which(!colnames(DocTermR)
     if (stop.word.tm) {
         stopword <- stopwords(idiom)
DocTermR <- DocTermR[, which(!colnames(DocTermR)
     }
SeparateOverInfra<-function(x){
over <- subset(x,x[,6] > 0)
infra <- subset(x,x[,6] < 0)
res <- list(rownames(over),rownames(infra))
names(res) <- c("Over_represented_word", "Infra_represented_word")
if(sum(unlist(lapply(res,is.null)))==2) return()
else return(res[lapply(res,is.null)==FALSE])
}
if(is.null(num.agg)&inherits(x,c("DocVarTable","data.frame"))){
res.descfreq <- descfreq(DocTermR,proba=proba)
CharWord <- lapply(res.descfreq,SeparateOverInfra)
```

```
}else{
if(inherits(x,c("DocVarTable","data.frame"))) varsel<-x[, num.agg]
else varsel<-x$data.clust$clust
if(length(varsel)!=nrow(DocTermR)) varsel<-varsel[-pos.elim]
if(length(y$corpus)!=nrow(DocTermR)) y$corpus<-y$corpus[-pos.elim]
if(!is.null(maxDocs)&!is.null(num.agg)){
res.descfreq <- descfreq(DocTermR, by.quali = varsel, proba=1)
vlev <- levels(varsel)
nlev <- nlevels(varsel)
motsval <- sapply(res.descfreq,function(x) x[order(row.names(x)),6],simplify = TRUE)
tmotsval <- t(motsval)
colnames(tmotsval) <- colnames(DocTermR)
rownames(tmotsval) <- levels(varsel)
lisresult <- vector(mode="list",length=nlev)
         for (igru in 1:nlev) {
lisresult[[igru]]<-data.frame()
             resp <- which(varsel == vlev[igru])
             ntrep <- min(maxDocs, length(resp))
DocTermcurs <- DocTermR[resp, ]
                 ly <- tmotsval[igru, ,drop=FALSE]
a <- crossprod(t(ly),t(DocTermcurs))
b <- rowSums(DocTermcurs)
repvaltest <- a
                 repvaltest[b > 0] <- a[b > 0]/b[b > 0]
                 ordrep <- order(repvaltest, decreasing = "TRUE")
                 for (i in 1:ntrep) {
                   lisresult[[igru]][i, 1] <- rownames(DocTermR)[resp[ordrep[i]]]
                   lisresult[[igru]][i, 2] <- repvaltest[ordrep[i]]
                   lisresult[[igru]][i, 3] <- y$corpus[resp[ordrep[i]]]
                 }
colnames(lisresult[[igru]]) <- c("DOCUMENT", "CLASS. CRITERION", "CHAR. ANSWER")
             }
         names(lisresult) <- levels(varsel)
if(inherits(x,"HCPC")){
maxDocs<-min(maxDocs,length(x$desc.ind$para[[i]]))
lisresultHCPC<- vector(mode="list",length=2)
lisresultHCPC[[1]] <- vector(mode="list",length=nlev)
lisresultHCPC[[2]] <- vector(mode="list",length=nlev)
for (igru in 1:nlev) {
lisresultHCPC[[1]][[igru]]<-data.frame()
lisresultHCPC[[1]][[igru]][,1] <- names(x$desc.ind$para[[i]])[1:maxDocs]
lisresultHCPC[[1]][[igru]][,2] <- rep(i,maxDocs)
lisresultHCPC[[1]][[igru]][,3] <- y$corpus[which(rownames(DocTermR)
lisresultHCPC[[2]][[igru]]<-data.frame()
lisresultHCPC[[2]][[igru]][,1] <- names(x$desc.ind$dist[[i]])[1:maxDocs]
lisresultHCPC[[2]][[igru]][,2] <- rep(i,maxDocs)
lisresultHCPC[[2]][[igru]][,3] <- y$corpus[which(rownames(DocTermR)
colnames(lisresultHCPC[[1]][[igru]])<- colnames(lisresultHCPC[[2]][[igru]])<- c("DOCUMENT",
 "CLUSTER", "ANSWER")
names(lisresultHCPC)<-c("Close_to_centroid_documents","Far_from_other_clusters_document")
}
}
}
res.descfreq <- descfreq(DocTermR, by.quali = varsel, proba=proba)
CharWord <- lapply(res.descfreq,SeparateOverInfra)
}
if(list.details) resCharWord <- res.descfreq
```

```
else resCharWord <- CharWord
     if(!is.null(maxDocs)&(!is.null(num.agg)|inherits(x,"HCPC"))){
if(!inherits(x,"HCPC")) res <- list(CharDoc = lisresult, CharWord = resCharWord)
else res <- list(CharDoc = lisresult, CharWord = resCharWord, CharHCPC = lisresultHCPC)
cat("CHARACTERISTIC ANSWERS\n(WORDS FREQUENCY CRITERION)\n")
for(i in 1:length(lisresult)){
cat(paste("\n","GROUP ",i,": ",names(lisresult)[i],sep="", "\n"))
cat(paste(rep("-",35)))
cat("\nCLASSIFICATION       DOCUMENT     CHARACTERISTIC\n")
cat("CRITERION                          ANSWER\n")
cat(paste(rep("-",35)))
for(j in 1:nrow(lisresult[[i]])) cat(paste("\n",round(lisresult[[i]][j,2],3),"
    ------     ",lisresult[[i]][j,1],"          ",lisresult[[i]][j,3],sep="","\n"))
}
}else{
res <- list(CharWord = resCharWord)
}
return(res)
}


## End(Not run)
```

`dataBiblio` *dataBiblio (data)*

Description

A total of 386 abstracts on Systemic Lupus Erythematosus published between January 2000 and December 2012.

Usage

```
data("dataBiblio")
```

Format

Data frame with 386 abstracts. The rows represent the document (abstracts), the columns represent the variables which describe the abstracts (Title, Abstract, Year, Author, Journal).

`dataOpen.question` *dataOpen.question (data)*

Description

Data issued from a survey. Three questions are included in the questionnaire: "What do you wish for your family?", "What else?" and "What does culture mean for you?". These questions require free answers.

Usage

```
data("dataOpen.question")
```

Format

Data frame with 100 rows and 8 columns. The rows represent the documents (respondents) and the columns the variables. The first five columns correspond to the individuals characteristics (Gender,Age_interval,Gender_Age,Gender_educ,ageeduc) and the last three columns correspond to the open-ended questions.

---

dataSpeech *dataSpeech (data)*

Description

Speech "I Have a Dream", by Martin Luther King, Jr. delivered on 28 August 1963, at the Lincoln Memorial, Washington D.C.

Usage

```
data("dataSpeech")
```

Format

A data frame with one textual column.

DocVarTable *Documents by Variables Table (DocVarTable)*

Description

Builds a data frame with the contextual variables selected from the base.

Usage

```
DocVarTable(base, VarSel)
```

Arguments

base	data frame with at least one textual column and one contextual variable
VarSel	column index(es) or name(s) of the selected variable(s)

Value

DocVar	data frame with I rows (documents) and J columns (quantitative or categorical variables)

Author(s)

Daria M Hernandez <daria.micaela.hernandez@upc.edu>

Examples

```
## Not run:
data(dataBiblio)
DocVar<-DocVarTable(dataBiblio,VarSel=c(3,4,5))

### CODE DocVarTable
DocVarTable <- function(base,VarSel)
{
if(is.character(VarSel)) VarSel<-which(colnames(base)
if(length(VarSel)==0) stop("Incorrect column name(s)")
if(max(VarSel)>ncol(base)|min(VarSel)<=0)
        stop("Column index(es) should be between 1 and numbre of columns")
DocVar<-as.data.frame(base[,VarSel,drop=FALSE])
class(DocVar) <- c("DocVarTable", "data.frame")
return(DocVar)
}

## End(Not run)
```

`DocWordTable` *Documents by Words Table (DocWordTable)*

Description

Builds a Document by Words Table and a summary about the corpus.

Usage

```
DocWordTable(base, num.text, idiom = "en", lminword = 1, lower = TRUE, Remov.Number = TRUE)
```

Arguments

`base`	data frame with at least one textual column
`num.text`	column index(es) or name(s) of the textual column(s)
`idiom`	language of the textual column(s) (by default English "en")
`lminword`	minimum threshold on the word length (by default 1)
`lower`	boolean, if TRUE the corpus is converted into lowercase
`Remov.Number`	boolean, if TRUE the numbers are removed

Value

`DocTerm`	data frame with I rows (documents) and J columns (words)
`Tfreq`	glossary of words
`Nlength`	corpus size
`Nword`	vocabulary size
`Ndoc`	number of documents
`Nfreqword`	frequencies of words

Ndocword	frequencies of words in documents
corpus	corpus analyzed
idiom	language of the textual column(s)

Author(s)

Daria M. Hernandez<daria.micaela.hernandez@upc.edu>, Belchin Kostov

References

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.).

See Also

print.DocWordTable, summary.DocWordTable, removePunctuation,

regex: shows the class [:punct:] of punctuation characters

Examples

```
## Not run:
data(dataBiblio)
res.DWT<-DocWordTable(dataBiblio,num.text=2, idiom="en",lminword=3,Remov.Number=TRUE,lower=TRUE)
DocTerm<-res.DWT$DocTerm
summary(res.DWT,nword=20)

###CODE DocWordTable

DocWordTable <- function (base, num.text, idiom = "en", lminword = 1, lower = TRUE,
         Remov.Number = TRUE){
if (is.character(num.text)) num.text <- which(colnames(base)
      if (length(num.text) > 1) {
for (i in 1:length(num.text)){
if (i == 1) text1 <- base[, num.text[1]]
                 else text1 <- paste(text1, base[, num.text[i]], sep = ".")
            }
            base[, (ncol(base) + 1)] <- text1
            num.text <- ncol(base)
}
     dtmCorpus <- Corpus(DataframeSource(base[num.text]), readerControl = list(language = idiom))
     filt = "([?]|[[:punct:]]|[[:space:]]|[[:cntrl:]])+"
     dtmCorpus <- tm_map(dtmCorpus, content_transformer(function(x) gsub(filt, " ", x)))
     if (Remov.Number == TRUE) dtmCorpus <- tm_map(dtmCorpus, removeNumbers)
     dtm <- DocumentTermMatrix(dtmCorpus, control = list(tolower = lower,
        wordLengths = c(lminword, Inf)))
    Nfreqword<-tapply(dtm$v,dtm$j,sum)
Ndocword<-tapply(dtm$v>0,dtm$j,sum)
Table <- cbind(Nfreqword,Ndocword)
rownames(Table) <- dtm$dimnames$Terms
    colnames(Table) <- c("Frequency", "N.Documents")
    TFreq <- Table[order(Nfreqword, Ndocword, decreasing = TRUE), ]
    res <- list(DocTerm = dtm, Ndoc = dtm$nrow, Nlength = sum(Nfreqword), Nword = dtm$ncol,
           Tfreq = TFreq, Nfreqword = Nfreqword, Ndocword = Ndocword,
           corpus = base[,num.text], idiom = idiom)
    class(res) <- c("DocWordTable", "list")
return(res)
```

```
}

## End(Not run)
```

`HierarchWords` *Hierarchical Words (HierarchWords)*

Description

Characteristic words of the nodes of the hierarchy.

Usage

```
HierarchWords(res,Table)
```

Arguments

`res`	either the result of a factor analysis, a dataframe, or a vector
`Table`	data frame with I rows (documents) and J columns (words)

Value

`Hierarch.words`	data frame with words characterizing the nodes of the hierarchy

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>, Belchin Kostov

References

Becue-Bertaut, M., Kostov, B., Morin, A., & Naro, G. (2014). Rhetorical Strategy in Forensic Speeches: Multidimensional Statistics-Based Methodology *Journal of Classification*, 31. doi:10.1007/s00357-014-9148-9

See Also

TxCHCPC

Examples

```
## Not run:
data(Biblio)
resT<-DocWordTable(Biblio,num.tex=2,lminword=3)
DocTerm<-resT$DocTerm
DocVar<-DocVarTable(Biblio,VarSel=c(3,4,5))
res.TxCA<-TxCA(DocVar,DocTerm,num.agg=1, Fmin=100,Dmin=20,graph=FALSE,stop.word.tm=TRUE)
Table<-res.TxCA$Table
res<-res.TxCA$res.ca
res.HW<-HierarchWords(res,Table)

###CODE HierarchWords
```

```
HierarchWords <-function (res,Table){
        CoR<-res
        if((is.data.frame(CoR))|(is.matrix(CoR))){
            Rcoor<-CoR
         }else{
if(!is.null(res$row$coord))
Rcoor<-res$row$coord
if(!is.null(res$ind$coord))
Rcoor<-res$ind$coord
       }
   DocTerm<-Table
CHC<-function(X, groups=NULL){

### Similarity Matrix
d<-dist(X)
d0<-as.matrix(d)
maxd<-max(d)
maxd<-maxd+1e-10
Sim<-as.matrix(maxd-d)
Sim0<-Sim
d<-as.matrix(d)

### Constrained Matrix
Cont<-matrix(nrow=nrow(Sim),ncol=ncol(Sim),0)
Cont[1,2]<-1
for (i in 2:(nrow(Cont)-1)){
Cont[i,i+1]<-1
Cont[i,i-1]<-1
}
Cont[nrow(Cont),nrow(Cont)-1]<-1
rownames(Cont)<-rownames(Sim)
colnames(Cont)<-colnames(Sim)

### Similarity matrix used for constrained clustering
SimCont<-Sim*Cont
DistCont<-d*Cont
if (is.null(groups)){
groups<-list()
for (i in 1:nrow(Sim)){
groups[[i]]<-i
}
}
distclust<-numeric()
clust<-list()
i<-1

indice<-nrow(d)-1

while(indice>0){

### Find the position of the maxim similarity
maxsim<-max(SimCont)
posmaxsim<-which(SimCont==maxsim)

if (posmaxsim[1]
fila<-posmaxsim[1]
col<-nrow(SimCont)
```

```
}else{
fila<-posmaxsim[1]
col<-posmaxsim[1]
}

maxfc<-max(fila,col)
minfc<-min(fila,col)
distclust[i]<-DistCont[fila,col]

clust[[i]]<-vector(mode="list",length=2)
clust[[i]][[1]]<-groups[[minfc]]
clust[[i]][[2]]<-groups[[maxfc]]

rownames(Sim)[minfc]<-colnames(Sim)[minfc]<-rownames(d)[minfc]<-colnames(d)[minfc]
                      <-rownames(Cont)[minfc]<-colnames(Cont)[minfc]
                      <-paste(rownames(Sim)[minfc],"-",rownames(Sim)[maxfc])

if (minfc!=1){
Sim[minfc,minfc-1]<-Sim[minfc-1,minfc]
                                <-0.5*Sim[minfc,minfc-1]+0.5*Sim[maxfc,minfc-1]-0.5*abs(Sim[minfc,
                                minfc-1]-Sim[maxfc,minfc-1])
d[minfc,minfc-1]<-d[minfc-1,minfc]
                                <-0.5*d[minfc,minfc-1]+0.5*d[maxfc,minfc-1]+0.5*abs(d[minfc,
                               minfc-1]-d[maxfc,minfc-1])
Cont[minfc-1,minfc]<-Cont[minfc,minfc-1]<-1
}
if (maxfc!=nrow(SimCont)){
Sim[maxfc+1,minfc]<-Sim[minfc,maxfc+1]<-0.5*Sim[minfc,maxfc+1]+0.5*Sim[maxfc
                               ,maxfc+1]-0.5*abs(Sim[minfc,maxfc+1]-Sim[maxfc,maxfc+1])
d[maxfc+1,minfc]<-d[minfc,maxfc+1]<-0.5*d[minfc,maxfc+1]+0.5*d[maxfc,
                               maxfc+1]+0.5*abs(d[minfc,maxfc+1]-d[maxfc,maxfc+1])
Cont[maxfc+1,minfc]<-Cont[minfc,maxfc+1]<-1
}

groups[[minfc]]<-c(groups[[minfc]],groups[[maxfc]])
groups<-groups[-maxfc]
Sim<-Sim[-maxfc,-maxfc]
d<-d[-maxfc,-maxfc]
Cont<-Cont[-maxfc,-maxfc]
i<-i+1

SimCont<-Sim*Cont
DistCont<-d*Cont
indice<-indice-1
             }

clust<-clust

hc<-hclust(dist(X))
hc$height<-distclust
hc$order<-sort(hc$order)
grups_blocs<-list()
grups_blocs[[1]]<-rep(0,nrow(X))
for (i in 1:(length(clust)-1)){
grups_blocs[[i+1]]<-grups_blocs[[i]]
grups_blocs[[i+1]][c(clust[[i]][[1]],clust[[i]][[2]])]<-i
}
```

```
for(i in 1:nrow(hc$merge)){
if (length(clust[[i]][[1]])==1&length(clust[[i]][[2]])==1){
hc$merge[i,1]<-(-clust[[i]][[1]])
hc$merge[i,2]<-(-clust[[i]][[2]])
}else{
if (length(clust[[i]][[1]])==1){
hc$merge[i,1]<-(-clust[[i]][[1]])
}else{
hc$merge[i,1]<-grups_blocs[[i]][clust[[i]][[1]][1]]
}

if (length(clust[[i]][[2]])==1){
hc$merge[i,2]<-(-clust[[i]][[2]])
}else{
hc$merge[i,2]<-grups_blocs[[i]][clust[[i]][[2]][1]]
}
}
}

return (res=list(hc=hc, clust= clust, groups=groups, dist=distclust))
}

objCCC<-CHC(Rcoor)

espcron<-function(objCCC, DocTerm){
    mat<-DocTerm[which(rownames(DocTerm)
    FreqDoc<-apply(mat,2,sum)
    mat<-mat[,which(FreqDoc>0)]

nodos<-descfreq(mat)

juntar<-function(elem){
elem<-unlist(elem)
aux<-as.data.frame(mat)
aux[nrow(aux)+1,]<-apply(aux[elem,],2,sum)
rownames(aux)[nrow(aux)]<-paste(rownames(aux)[min(elem)],"-",
                rownames(aux)[max(elem)],sep="")
aux<-aux[-elem,]}
        numNodo<-vector("list",length(nodos))
  for (i in 1:length(nodos)){
            Nodo<-as.vector(rep(i,nrow(nodos[[i]])))
            nodos[[i]]<- cbind(nodos[[i]],Nodo)
}


pal.car<-lapply(lapply(objCCC$clust,function(el) juntar(el)),function(m) descfreq(m))
  length(pal.car)
length(pal.car[[1]])

     pal.carP<-pal.car
      numNodo<-vector("list",length(pal.carP))
       K <-length(pal.carP)+1
for (i in 1:length(pal.carP)){
          for (j in 1:length(pal.carP[[i]])){
           if(j<length(pal.carP[[i]])){
           Nodo<-as.vector(rep(i,nrow(pal.carP[[i]][[j]])))
```

```
           pal.carP[[i]][[j]]<-cbind(pal.carP[[i]][[j]],Nodo)
           }else{
           if(!is.null(pal.carP[[i]][[j]])){
Nodo<-as.vector(rep(K+1,nrow(pal.carP[[i]][[j]])))
      pal.carP[[i]][[j]]<-cbind(pal.carP[[i]][[j]],Nodo)
           }
     K=K+1
         }

     }

       }

res<-cbind(rep(names(nodos)[1],nrow(nodos[[1]])),nodos[[1]])
for (i in 2:length(nodos)){
res<-rbind(res,cbind(rep(names(nodos)[i],nrow(nodos[[i]])),nodos[[i]]))
}

for (i in 1:(length(pal.car)-1)){
for (j in 1:length(pal.car[[i]])){
if (!is.null(pal.carP[[i]][[j]]))
   res<-rbind(res,cbind(rep(names(pal.carP[[i]])[j],nrow(pal.carP[[i]][[j]])),pal.carP[[i]][[j]]))
}
}

res<-as.data.frame(res)
colnames(res)[1]<-"Group"
res$names<-rownames(res)
res<-res[!duplicated(res[,-8]),]
      for (i in 2:(ncol(res)-1)) res[,i]<-as.numeric(as.character(res[,i]))
res[,1]<-as.character(res[,1])
res<-res[which(res$v.test>0),]
words<-levels(as.factor(res$names))
char.words<-data.frame(matrix(nrow=length(words),ncol=ncol(res)-1))
for (i in 1:length(words))
       char.words[i,]<-res[which(res$p.value==min(res$p.value[which(res$names==words[i])])
       &res$names==words[i]),1:8]
rownames(char.words)<-words
colnames(char.words)<-colnames(res)[1:8]
      char.words <- with(char.words, char.words[order(Nodo, decreasing = TRUE),])
      char.words<-char.words[,c(1,8,2:7)]
      colnames(char.words)[2] <- "Node"
return(char.words)
 }
  Hierarch.words<-espcron(objCCC, DocTerm)
  return(Hierarch.words)

## End(Not run)
```

MacroBiblio *Analysis of Bibliography (MacroBiblio)*

Description

Macro function for the analysis of a bibliographic database.

Usage

```
MacroBiblio(base, num.text = "Abstract", num.agg = "Year", idiom = "en",
lminword = 3, Fmin = 10, Dmin = 5, Fmax = NULL, equivalence = NULL,
stop.word.user = NULL, lmd = 3, lmk = 3, ncp = 10, row.sup = NULL,
col.sup = NULL, graph = TRUE, axes = c(1, 2), proba = 0.01)
```

Arguments

base	data frame with I rows (abstracts/articles) and J columns. The names of the main columns must be: Title, Year, Abstract, Journal; in addition, the database may have other quantitative or categorical variables such as: "Author", "Year_class", etc.
num.text	column index(es) or name(s) of the textual column(s) (by default "Abstract")
num.agg	column index or name of the aggregation column (by default "Year")
idiom	language of the textual column(s) (by default English "en")
lminword	minimum threshold on the word length (by default 3)
Fmin	minimum threshold on the word frequency (by default 10)
Dmin	minimum threshold on the number of documents using the word (by default 5)
Fmax	maximum threshold on the word frequency
equivalence	data frame with n rows and two columns (original word and new word)
stop.word.user	vector indicating the stopwords chosen by the user
lmk	minimum threshold on the contribution for selecting the metakeys (by default 3, which mean contribution 3 times greater than the mean contribution)
lmd	minimum threshold on the contribution for selecting the metadocs (by default 3, which mean contribution 3 times greater than the mean contribution)
ncp	number of dimensions stored in the results (by default 10)
row.sup	vector with the index(es) or name(s) of the supplementary row(s)
col.sup	vector with the index(es) or name(s) of the supplementary frequency column(s)
graph	boolean, if TRUE graphs are displayed
axes	a length 2 vector specifying the dimensions to plot
proba	significance threshold used to select the characteristic words in each category (by default 0.01)

Value

Returns a list including:

Corpus	summary of the information about the corpus
Glossary	glossary of the selected words in frequency order
DocTermR	documents by words (all documents, selected words)
Tagreg	lexical aggregated table
Metakeys.Metadocs	graphical representation of metakeys and metadocs
res.CA	results of direct correspondence analysis
res.CA.Agreg	results of aggregate correspondence analysis by year

CharWord	characteristic words of each category of the aggregation variable
res.CHCPC	results of constrained hierarchical clustering
res.MFACT	result of multiple factor analysis for contingency tables
OrdWord	words order by their coordinates on the first dimension
pioneers	pioneer articles

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>, Belchin Kostov

References

Kerbaol, M. Bansard, JY.;Coatrieux, JL. (2006) An analysis of IEEE publications in biomedical engineering. *IEEE Engineering in Medicine and Biology. Magazine*.

Morin, A.(2006)Intensive Use of Factorial Correspondence Analysis for Text Mining: *Application with Statistical Education Publications*.

Becue-Bertaut, M. (2014). Tracking verbal-based methods beyond conventional descriptive analysis in food science bibliography. A statistical approach. *Food Quality and Preference*,32, 2-15.

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.).

See Also

print.MacroBiblio,summary.MacroBiblio

Examples

```
## Not run:
data(dataBiblio)
res.dataBiblio<-MacroBiblio(dataBiblio, lmd = 6, lmk =6)
print(res.dataBiblio)
summary(res.dataBiblio, nword=20)

###CODE MacroBiblio
MacroBiblio <-function(base, num.text="Abstract", num.agg="Year",
        idiom ="en",lminword=3, Fmin=10, Dmin=5,Fmax=NULL, equivalence=NULL,
        stop.word.user=NULL, lmd=3, lmk=3, ncp=10, row.sup=NULL, col.sup=NULL,
         graph=TRUE, axes=c(1,2), proba=0.01){

if (!is.null(num.text)) {
 if (is.character(num.text))
 num.text<- which(colnames(base)
  if (is.numeric(num.text))
  num.text<- num.text
 if(length(num.text)==1)
  num.text<-num.text
 if(length(num.text)>1){
   for(i in 1:length(num.text)){
    if(i==1)
    text1<-base[,num.text[1]]
    else text1<-paste(text1,base[,num.text[i]],sep=".")
   }
    base[,(ncol(base)+1)]<-text1
    num.text<-ncol(base)
  }
```

```
}

#Contextual variables
 VarSel<-DocVarTable(base, VarSel=c("Year","Journal" ))
 DocVar<-as.data.frame(VarSel[,c("Year","Year")])
 DocVar[,1]<-as.numeric(DocVar[,1])
 DocVar[,2]<-as.factor(DocVar[,2])
   summary(DocVar)

# Corpus and DocWordTable
 DocTerm<-DocWordTable(base, num.text, idiom , lminword,
                       lower = TRUE, Remov.Number = TRUE)

# direct  CA for  Metakeys and Metadocs
 res.TxCA.Dir<-TxCA(DocTerm,DocVar, Fmin, Dmin, Fmax, idiom=idiom,
        equivalence, num.agg = NULL,stop.word.user, lmd,lmk,ncp,
        row.sup=NULL,col.sup ,stop.word.tm=TRUE,graph=FALSE,
         axes,quanti.sup =1,quali.sup=2)

 res.caD<-res.TxCA.Dir$res.ca
       DocTermR<-res.TxCA.Dir$DocTermR
       res.meta<-res.TxCA.Dir$res.meta
       Glossary<-res.TxCA.Dir$Glossary
       DimWords<- res.meta$DimWord[which(res.meta$DimWord[,1]>1),]

# Aggregate CA
 if(num.agg=="Year"){
 res.TxCA.Agreg<-TxCA(DocTerm, DocVar, Fmin, Dmin,Fmax, idiom=idiom,
     equivalence, num.agg=1,stop.word.user, lmd,lmk,ncp,
      row.sup,col.sup ,stop.word.tm=TRUE,graph=FALSE,axes,
     quanti.sup=NULL,quali.sup=NULL)

 }else{
   res.TxCA.Agreg<-TxCA(DocTerm, y=base, Fmin, Dmin,Fmax, idiom=idiom,
     equivalence, num.agg=num.agg,stop.word.user, lmd,lmk,ncp, row.sup,
     col.sup ,stop.word.tm=TRUE,graph=FALSE,axes,quanti.sup=NULL,quali.sup=NULL)
 }

 res.caAg<-res.TxCA.Agreg$res.ca
       Table<-res.TxCA.Agreg$Table

#characteristic words
     CharWord.Compl<-descfreq(Table, by.quali = NULL, proba)
CharWord=vector(mode="list",length=length( CharWord.Compl))
names(CharWord)<-names( CharWord.Compl)
      for(i in 1:length( CharWord.Compl)){
            bWord <- as.data.frame( CharWord.Compl[[i]])
            over <- subset(bWord, bWord$v.test > 0)
            Over <- row.names(over)
            infra <- subset(bWord, bWord$v.test < 0)
            Infra <- row.names(infra)
            OverInfra <- list(Over, Infra)
            names(OverInfra) <- c("Over_represented_word", "Infra_represented_word")
            CharWord[[i]] <- OverInfra
        }
        CharWords<-CharWord
        res.CharWord<-list( CharWord.Compl= CharWord.Compl, CharWords=CharWords)
```

```
# Constrained hierarchical clustering
 res.chcpc<-TxCHCPC(res.caAg,cluster.CA = "rows",nb.clust=-1, graph=FALSE)

#cronological evolution AFMTC
  DocTermR<-res.TxCA.Dir$DocTermR
  DocTermRM<-DocTermR[apply(DocTermR,1,sum)>0,]
  DocVarR<-as.data.frame(DocVar[which(rownames(DocVar)
  MDocWord<-cbind.data.frame(DocTermRM,DocVarR)

  res.mfact<-MFA(MDocWord, group=c(dim(DocTermRM)[2],1,1), type =c("f","s","n"),
                name.group=c("FREQ","YEAR","YEAR_CAT"), num.group.sup = 3,graph=FALSE)

# most contributive word in the DIM1
   WordDim1<-as.data.frame((res.mfact$freq$coord[,1]))
   WordDim1$Words<-as.data.frame(row.names(WordDim1))
   WordDim<-cbind(WordDim1[,2],round(WordDim1[,1],3))
   colnames(WordDim)<-c("Word", "Coord.Dim1")
   WordDim1Comp<-with( WordDim, WordDim[order(Coord.Dim1,Word),])
   sel<-which(res.mfact$freq$contrib[,1]>lmk*mean(res.mfact$freq$contrib[,1]))
   WordMoreContrib<- WordDim1Comp[which( WordDim1Comp[,1]
 res.WordDim1<-list(WordMoreContrib= WordMoreContrib, WordDim1Comp=WordDim1Comp)

#pioneers articles
  SelYear<-as.numeric(rownames(subset(res.mfact$quali.var.sup$coord,
         res.mfact$quali.var.sup$coord[,1]>0)))
  SelYear<-SelYear[1:(length(SelYear)-2)]

  MDocWordA<-MDocWord[which(MDocWord[,ncol(MDocWord)-1]
  MDocWordA<-MDocWordA[,apply(MDocWordA[,1:(ncol(MDocWordA)-2)],2,sum)>0]
mfact.res<-MFA(MDocWordA, group=c((ncol(MDocWordA)-2),1,1), type =c("f","s","n"),
                name.group=c("FREQ","YEAR","YEAR_CAT"), num.group.sup = 3,graph=FALSE)

###  distance between the partial points
G1<-mfact.res$ind$coord.partiel[seq(1,(ncol(MDocWordA)-1),2),]
      G1<-G1[which(apply(G1,1,sum)!=0),]
G2<-mfact.res$ind$coord.partiel[seq(2,(ncol(MDocWordA)),2),]
G2<-G2[which(apply(G2,1,sum)!=0),]
      G<-G1-G2

 if(graph){

#graph of AFMTC
      lim=0.8
         plot(1,1,pch=16,col="white",xlim=c(-0,4),ylim=c(-4,5),xlab=paste("DIM",
      axes[1],"(",round(mfact.res$eig[axes[1],2],2),"
      ylab=paste("DIM",axes[2],"(",round(mfact.res$eig[axes[2],2],2),"
      ,main="Superimposed Representation (partial axes)",cex=0.75)
      points(G2[,1],G2[,2],pch=16,col="white",cex=0.75)
points(G1[,1],G1[,2],pch=16,col="white",cex=0.75)
points(mfact.res$ind$coord[,1],mfact.res$ind$coord[,2],pch=16,col="white",cex=0.75)
segments(G1[,1],G1[,2],mfact.res$ind$coord[,1],mfact.res$ind$coord[,2],lty=2,col="white",cex=0.75)
segments(mfact.res$ind$coord[,1],mfact.res$ind$coord[,2],G2[,1],G2[,2],lty=1,col="white")
abline(h=0,v=0,lty=2)
points(G1[which(G[,1]>lim),1],G1[which(G[,1]>lim),2],pch=16,col="red",cex=0.75)
points(G2[which(G[,1]>lim),1],G2[which(G[,1]>lim),2],pch=16,col="green",cex=0.75)
points(mfact.res$ind$coord[which(G[,1]>lim),1],mfact.res$ind$coord[which(G[,1]>lim),2],
```

```
          pch=16,col="black",cex=0.75)

    text(mfact.res$ind$coord[which(G[,1]>lim),1],mfact.res$ind$coord[which(G[,1]>lim),2],
    rownames(mfact.res$ind$coord[which(G[,1]>lim),]),col="black",cex=0.75,font=2,pos=c(3,4,4,3,1,3,1))

     segments(G1[which(G[,1]>lim),1],G1[which(G[,1]>lim),2],mfact.res$ind$coord[which(G[,1]>lim),1],
      mfact.res$ind$coord[which(G[,1]>lim),2],lty=2,col="red",cex=0.75)
      segments(mfact.res$ind$coord[which(G[,1]>lim),1],mfact.res$ind$coord[which(G[,1]>lim),2],
      G2[which(G[,1]>lim),1],G2[which(G[,1]>lim),2],lty=1,col="green")
      legend("topleft",c("Vocabulary","Year"),lty=c(2,1),text.col=c("red","green"),
      col=c("red","green"))

   plot.MFA(res.mfact, choix = "ind", invisible = "ind",  habillage = "group",
                axes = axes, new.plot = TRUE)
     points(res.mfact$quali.var.sup$coord[,1],res.mfact$quali.var.sup$coord[,2]
                ,type="l",cex=0.75,pch=19,col="blue",lwd=2,lty=2)
      dev.new()
      sel1<-which( (res.mfact$freq$contrib[,1]> lmk*mean(res.mfact$freq$contrib[,1])) |
                    (res.mfact$freq$contrib[,2]>lmk*mean(res.mfact$freq$contrib[,2])) )
       par(cex=0.8)
plot.MFA(res.mfact,choix="freq",invisible="row",select=sel1,axes,
unselect=1,palette=palette(c("black","black","black")),
     col.hab=c("green","blue"),title="Words" )

    plot.MFA(res.mfact, choix = "var", habillage = "group",
                axes = axes, new.plot = TRUE, shadowtext = TRUE)
    plot.MFA(res.mfact, choix = "group", axes = axes, new.plot = TRUE)

     ##Graph of TxCHCPC
     res.chcpc<-TxCHCPC(res.caAg,cluster.CA = "rows",nb.clust=-1)
     #Graph aggregate AC
      res.metaA<- META.CA(res.caAg, naxes = ncp, axe.x = axes[1],  axe.y = axes[2],
              lmd = -Inf, lmk, main = "CA documents/words")
         dev.new()
         plot(res.caAg, invisible = c("col", "col.sup","quali.sup"), axes = axes,
           title = "CA documents")
   points(res.caAg$row$coord[,1],res.caAg$row$coord[,2],type="l",cex=0.75,
           pch=19,col="black",lwd=2,lty=2)
        dev.new()
  #Graph direct AC
         res.meta <- META.CA(res.caD, naxes = ncp, axe.x = axes[1],axe.y = axes[2], lmd, lmk)
         dev.new()
         plot(res.caD, invisible = c("col", "col.sup","quali.sup"), axes = axes,
          title = "CA documents")
         dev.new()
         res.meta1<- META.CA(res.caD, naxes = ncp, axe.x = axes[1],  axe.y = axes[2],
         lmd = Inf, lmk, main = "CA words")

   ###pioneers articles
  pioneers<-base[row.names(base)
  pioneers<- pioneers[,colnames(pioneers)
}
  else pioneers<-NULL
res<-list(Corpus=DocTerm,Glossary=Glossary,DocTermR=DocTermR,Tagreg=Table,
    Metakeys.Metadocs=res.meta, res.CA=res.TxCA.Dir, res.CA.Agreg=res.TxCA.Agreg,
    CharWord=res.CharWord, res.CHCPC=res.chcpc, res.MFACT=res.mfact,
    OrdWord=res.WordDim1,pioneers=pioneers)
```

```
 class(res)<-c("MacroBiblio","list")
return(res)
}


## End(Not run)
```

`MacroCaHcpc` *Correspondence Analysis and Hierarchical Clustering (MacroCaHcpc)*

Description

Macro function for the analysis of data issued from a database with open-ended questions. Automatic chain of the main steps: AC and HCPC.

Usage

```
MacroCaHcpc(base, num.text, idiom = "en", nb.clust = -1, Fmin = 5, Dmin = 5,
 Fmax = NULL, equivalence = NULL, stop.word.user = NULL,
stop.word.tm = FALSE, lmd = 3, lmk = 3, ncp = 5, row.sup = NULL,
col.sup = NULL, quanti.sup = NULL, quali.sup = NULL, axes = c(1, 2))
```

Arguments

`base`	data frame with at least one textual column
`num.text`	column index(es) or name(s) of the textual column(s)
`idiom`	language of the textual column(s) (by default English "en")
`nb.clust`	integer. If 0, the tree is cut at the level the user clicks on. If -1, the tree is automatically cut at the suggested level (by defaul -1)
`Fmin`	minimum threshold on the word frequency (by default 5)
`Dmin`	minimum threshold on the number of documents using the word (by default 5)
`Fmax`	maximum threshold on the word frequency
`equivalence`	data frame with n rows and two columns (original word and new word)
`stop.word.user`	vector indicating the stopwords chosen by the user
`stop.word.tm`	boolean, if TRUE the stopword list provided by tm is taken into account
`lmk`	minimum threshold on the contribution for selecting the metakeys (by default 3, which mean contribution 3 times greater than the mean contribution)
`lmd`	minimum threshold on the contribution for selecting the metadocs (by default 3, which mean contribution 3 times greater than the mean contribution)
`ncp`	number of dimensions stored in the results (by default 5)
`row.sup`	vector with the index(es) or name(s) of the supplementary row(s)
`col.sup`	vector with the index(es) or name(s) of the supplementary frequency column(s)
`quanti.sup`	vector with the index(es) or name(s) of the supplementary quantitative column(s)
`quali.sup`	vector with the index(es) or name(s) of the categorical supplementary column(s)
`axes`	a length 2 vector specifying the dimensions to plot

Value

Returns a list including:

`Corpus`	summary of the information about the corpus
`res.TxCA`	correspondence analysis results
`res.hcpc`	hierarchical clustering results
`res.TxCharClust`	characteristic documents and words of the clusters
`ncp`	number of dimensions preserved (corresponding to the dimensions associated to an over the mean eigenvalues) in the construction of hierarchical clustering

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

References

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.).

See Also

print.MacroCaHcpc

Examples

```
## Not run:
data(dataOpen.question)
res.TxC<-MacroCaHcpc(dataOpen.question, num.text=c(6,7), Fmin=15,Dmin=15)
print(res.TxC)

###CODE MacroCaHcpc

MacroCaHcpc<-function(base,num.text, idiom="en",nb.clust =-1, Fmin=5,
 Dmin=5, Fmax=NULL,equivalence=NULL, stop.word.user=NULL, stop.word.tm=FALSE,
  lmd=3, lmk=3,  ncp=5, row.sup = NULL, col.sup = NULL, quanti.sup=NULL,
   quali.sup = NULL, axes = c(1,2)) {

if (!is.null(num.text)) {
 if (is.character(num.text))
 num.text<- which(colnames(base)
  if (is.numeric(num.text))
  num.text<- num.text
 if(length(num.text)==1)
  num.text<-num.text
 if(length(num.text)>1){
   for(i in 1:length(num.text)){
    if(i==1)
    text1<-base[,num.text[1]]
    else text1<-paste(text1,base[,num.text[i]],sep=".")
   }
    base[,(ncol(base)+1)]<-text1
    num.text<-ncol(base)
  }
}
```

```
##Corpus
DocTerm<-DocWordTable(base,num.text,idiom,lminword=1,Remov.Number=TRUE,lower=TRUE)

# Selection of the words depending from two threshold minimum global
# frequency minimum nr of documents
res.TxCA<-TxCA(DocTerm, y=base, idiom=idiom, Fmin=Fmin,Dmin=Dmin,Fmax=Fmax,
   equivalence=equivalence,stop.word.user=stop.word.user,stop.word.tm=stop.word.tm,
   ncp=ncp,row.sup=row.sup,col.sup= col.sup, quanti.sup,quali.sup=quali.sup,graph=FALSE)
res.ca<-res.TxCA$res.ca

Ch<-TRUE
i=1
while(Ch & i<=nrow(res.ca$eig)){
  if(res.ca$eig[i,1]<mean(res.ca$eig[,1])){
   ncp<-i-1
   Ch<-FALSE
  }
  else i<-i+1
}
 if(ncp==1){
 ncp=2
}

res.TxCA<-TxCA(DocTerm, y=base, idiom=idiom, Fmin=Fmin,Dmin=Dmin,Fmax=Fmax,
   equivalence=equivalence,stop.word.user=stop.word.user,stop.word.tm=stop.word.tm,
   ncp=ncp, row.sup=row.sup,col.sup= col.sup, quanti.sup,quali.sup=quali.sup,graph=FALSE)
#summary(res.TxCA)
DocTermR<-res.TxCA$DocTermR
Table<-res.TxCA$Table
res.ca<-res.TxCA$res.ca

# Classification
res.hcpc<-HCPC(res.ca,cluster.CA="rows", nb.clust =nb.clust, order=TRUE)

res.TxCharClust<-TxCharClust(base,res.hcpc,num.text)
res.TxCharClust
plot(res.ca,invisible="row")
dev.new()
plot(res.ca,invisible="col")

res<-list(Corpus=DocTerm, res.TxCA=res.TxCA, res.hcpc=res.hcpc,
      res.TxCharClust=res.TxCharClust, ncp=ncp)
class(res)<-c("MacroCaHcpc","list")
return(res)
}


## End(Not run)
```

MacroTxChrono *Chronological Corpus (MacroTxChrono)*

Description

Macro function for an analysis of a chronological corpus

Usage

```
MacroTxChrono(base, num.text, divide=TRUE, SentLength=100,  idiom ="en",
 Fmin=5, Dmin=1, Fmax=NULL, equivalence=NULL, stop.word.user=NULL,
stop.word.tm=FALSE, lmk=3,lmd=3, ncp=5,  row.sup = NULL, col.sup = NULL,
quanti.sup=NULL, quali.sup = NULL, graph = TRUE, axes = c(1,2), N=5000,
alfa=0.15, CorChronoDim=0.10, HierWords=TRUE, SegRep=FALSE)
```

Arguments

base	data frame with at least one textual column
num.text	column index(es) or name(s) of the textual column(s)
divide	boolean, if TRUE the whole corpus is divided into arbitrary sentences of size SentLength
SentLength	length of the arbitrary sentences (by default 100 words)
idiom	language of the textual column(s) (by default English "en")
Fmin	minimum threshold on the word frequency (by default 5)
Dmin	minimum threshold on the number of documents using the word (by default 1)
Fmax	maximum threshold on the word frequency
equivalence	data frame with n rows and two columns (original word and new word)
stop.word.user	vector indicating the stopwords chosen by the user
stop.word.tm	boolean, if TRUE the stopword list provided by tm is taken into account
lmk	minimum threshold on the contribution for selecting the metakeys (by default 3, which mean contribution 3 times greater than the mean contribution)
lmd	minimum threshold on the contribution for selecting the metadocs (by default 3, which mean contribution 3 times greater than the mean contribution)
ncp	number of dimensions stored in the results (by default 5)
row.sup	vector with the index(es) or name(s) of the supplementary row(s)
col.sup	vector with the index(es) or name(s) of the supplementary frequency column(s)
quanti.sup	vector with the index(es) or name(s) of the supplementary quantitative column(s)
quali.sup	vector with the index(es) or name(s) of the categorical supplementary column(s)
graph	boolean, if TRUE graphs are displayed
axes	a length 2 vector specifying the dimensions to plot
N	number of permutation tests (by default 5000)
alfa	significance level (by default 0.15)
CorChronoDim	threshold on cor(chronology, dimensions) to select dimensions for constrained clustering (by default 0.10)
HierWords	boolean, if TRUE results characteristic words for each node in the hierarchy are displayed
SegRep	boolean, if TRUE results characteristic segments for each node in the hierarchy are displayed

Value

Returns a list including:

SentenceList	data frame with I rows (sentences) and two columns (group and sentence)
Homo.Groups	homogeneous groups and list of their contents
Corpus	summary of the information about the corpus
Correlation	correlation between chronology and dimensions
res.TxCA	correspondence analysis results
res.chcpc	constrained hierarchical clustering results
HierWord	characteristic words for each node of the constrained hierarchical clustering
HierSegment	characteristic segments for each node of the constrained hierarchical clustering
VocIndex	regular and specialized vocabulary

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

References

Becue-Bertaut, M., Kostov, B., Morin, A., & Naro, G. (2014). Rhetorical Strategy in Forensic Speeches: Multidimensional Statistics-Based Methodology *Journal of Classification*, 31. doi:10.1007/s00357-014-9148-9.

Legendre, P. & Legendre, L. (1998), Numerical Ecology (2nd ed.), Amsterdam: Elsevier Science.

Murtagh, F. (1985), Multidimensional Clustering Algorithms. Vienna-Wurzburg: Physica-Verlag.,

Murtagh, F., Ganz, A., & Mckie, S. (2008). The structure of narrative: The case of film scripts, *Patterns Recognition*, 42, 302-312.

Examples

```
## Not run:
data(dataSpeech)
res.MTxC<-MacroTxChrono(dataSpeech, SentLength=120, num.text=1, HierWords=FALSE)
print(res.MTxC)
summary(res.MTxC)

###CODE MacroTxChrono

MacroTxChrono<-function(base, num.text, divide=TRUE, SentLength=100, idiom ="en",
 Fmin=5, Dmin=1, Fmax=NULL, equivalence=NULL, stop.word.user=NULL,
 stop.word.tm=FALSE, lmk=3,lmd=3, ncp=5, row.sup = NULL, col.sup = NULL,
 quanti.sup=NULL, quali.sup = NULL, graph = TRUE, axes = c(1,2),
 N=5000, alfa=0.15, CorChronoDim=0.10, HierWords = TRUE, SegRep=FALSE){
if(divide){
#Divide the text into sentences homogeneous
 phrase<- uSentences(base, num.text, SentLength)
}else phrase<- as.data.frame(base[,num.text])

#Phrases are grouped into homogeneous parts.
      Homo.Groups<-uCutDoc(phrase,idiom, num.text=1, N, alfa)
      SentGroup<-Homo.Groups$SentGroup
 res.Homo.Groups<-list(composition=Homo.Groups$GrpComposition,
```

```
      nb.groups=Homo.Groups$Num.groups)

#CA Show trajectory of the parties
#Contextual variable
   DocVar<-DocVarTable(SentGroup, VarSel=1)
#DocumentosXword Table
   LexTable<-DocWordTable(SentGroup,num.text=2,idiom)
   DocTerm<-LexTable$DocTerm

#CA lexica-Agreggated table
#Trayectory
res.TxCA<-TxCA(LexTable,DocVar,num.agg=1,idiom=idiom,Fmin=Fmin,
     Dmin=Dmin,Fmax=Fmax, equivalence=equivalence,
     stop.word.user=stop.word.user,stop.word.tm=stop.word.tm, ncp=ncp,
       row.sup=row.sup,col.sup= col.sup,graph =FALSE)
Tagreg<-res.TxCA$Table
#Chrono variable
Chrono<-as.data.frame(1:nrow(Tagreg))
Tagreg<-cbind.data.frame(Tagreg,Chrono)
res.caCh<-CA(Tagreg[apply(Tagreg,1,sum)>0,],quanti.sup=nrow(Tagreg), graph=FALSE)
#define the number of axes
Nc<-res.caCh$quanti.sup$cos2
rownames(Nc)<-"Correlation"
if(Nc[1]< CorChronoDim){
print(Nc)
stop("The correlation between the first dimension and chronology is low. See below the value.")
}else{
Ch<-TRUE
i=1
while(Ch & i<=length(Nc)){
  if(Nc[i]>= CorChronoDim){
   ncp<-i
   Ch<-TRUE
   i<-i+1
  }else Ch<-FALSE
 }
}
 if(ncp==1){
 ncp=2
}
# Hierarchical clustering
Table<-res.TxCA$Table
res.ca<-res.TxCA$res.ca
res.caH<-res.TxCA$res.ca
res.caH$row$coord<-res.caH$row$coord[,1:ncp]
res.chcpc<-TxCHCPC(res.caH, nb.clust = -1)
    if (HierWords) {
        HierWord <- HierarchWords(res.caH, Table)
  res.chcpc$HierWord<-HierWord
    } else HierWord = NULL
#Hegments in the hierarchy
if(SegRep){
res.segment<-SegmentsRep(SentGroup, num.text=2)
Tab.SegR<-res.segment$tab.seg
        num.agg <-1
        agg <- as.factor(SentGroup[, num.agg])
        dis.X <- tab.disjonctif(agg)
```

```
        Tagreg.LexSeg <- t(Tab.SegR)
        TagSeg<- t(Tagreg.LexSeg)
HierSegment<-HierarchWords(res.caH, TagSeg)
}else HierSegment=NULL

dev.new()
res.meta1<-META.CA(res.ca,naxes=ncp,axe.x=axes[1],axe.y=axes[2],lmd=Inf,lmk, main="CA words")
dev.new()
plot(res.ca,invisible="col")
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="black")
dev.new()
sel<-which( (res.ca$col$contrib[,1]> 3*mean(res.ca$col$contrib[,1]))
            | (res.ca$col$contrib[,2]>3*mean(res.ca$col$contrib[,2])) )
plot(res.ca,choix="CA",selectCol=sel,unselect=1)
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="black")

#Vocabulary index
res.VocIndex<-VocIndex(SentGroup, num.text=2, Fmin=Fmin)
#specific.growth
res.specGrowth<-specific.growth(Table, res=res.ca , N,alfa,val=0.3,lim=1.64 )

res<-list(SentenceList=base,Homo.Groups=res.Homo.Groups,Corpus=LexTable,Correlation=Nc,
    ncp=ncp, res.TxCA=res.TxCA,res.chcpc=res.chcpc,HierWord=HierWord,
    HierSegment=HierSegment,VocIndex=res.VocIndex, spec.growth=res.specGrowth)
class(res)<-c("MacroTxChrono","list")
return(res)
}


## End(Not run)
```

MDocWordTable *Multiple Document by Words Table (MDocWordTable)*

Description

Builds a Multiple Document by Words Table.

Usage

```
MDocWordTable(y,x,Fmin=rep(5,length(y)),Dmin=rep(5,length(y)),
idiom=rep("en",length(y)),stop.word.user=vector(mode="list",length(y)),
stop.word.tm=rep(FALSE,length(y)),num.agg=NULL,Fmax=NULL)
```

Arguments

y	list of lexical tables
x	list of contextual tables associated with the lexical tables (one by one)
Fmin	vector of the minimum thresholds on the word frequency (one threshold for each table, by default 5)
Dmin	vector of the minimum documents using the word (one threshold for each table, by default 1)

idiom	vector indicating the language corresponding to each lexical table
stop.word.user	list of the stopword vectors chosen by the user for each lexical table
stop.word.tm	boolean vector, if TRUE the stopword list provided by tm is taken into account in each lexical table
num.agg	column index or name of the aggregation column. Common to all the table (by default NULL, no aggregation)
Fmax	maximum threshold on the word frequency

Value

MDocWord	multiple document by words table
ncolTs	vector with the number of columns in each table after selection
MTagregSep	list with the separate aggregate tables

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

DocWordTable, TxMFACT

Examples

```
## Not run:
data(dataOpen.question)
res.M1<-DocWordTable(dataOpen.question,num.text=c(6,7))
res.M2<-DocWordTable(dataOpen.question,num.text=8)
DocVar<-DocVarTable(dataOpen.question,VarSel=c(1,2,3,4))
MDocTerm=list(res.M1$DocTerm,res.M1$DocTerm)
MDocVar=list(DocVar,DocVar)
MTable<-MDocWordTable(MDocTerm,MDocVar,num.agg=3,idiom=c("en","en"),
 Fmin=c(10,5),Dmin=c(2,1))
names(MTable)

###CODE MDocWordTable

MDocWordTable<-function(y,x,Fmin=rep(5,length(y)),Dmin=rep(5,length(y)),
idiom=rep("en",length(y)),stop.word.user=vector(mode="list",length(y)),
stop.word.tm=rep(FALSE,length(y)),num.agg=NULL,Fmax=NULL){

   MTables<-list()
DocTermVar<-list()
MTagregSep<-list()
      gpo<-vector()
z<-y
for(i in 1:length(z)){
num.agg=num.agg
y<-z[[i]]
DocVar<-x[[i]]
if(!is.null(Fmax))
  sel.words<-which(y$Nfreqword <= Fmax & y$Nfreqword >= Fmin[i] & y$Ndocword >= Dmin[i])
else sel.words<-which(y$Nfreqword >= Fmin[i] & y$Ndocword >= Dmin[i])
     pos.sparse<-which(y$DocTerm$j
```

```
y$DocTerm$j<-y$DocTerm$j[pos.sparse]
y$DocTerm$v<-y$DocTerm$v[pos.sparse]
     y$DocTerm$i<-y$DocTerm$i[pos.sparse]
y$DocTerm$dimnames$Terms<-y$DocTerm$dimnames$Terms[sel.words]
recoderFunc<-function (x, from, to){
mapidx <- match(x, from)
     mapidxNA <- is.na(mapidx)
     from_found <- sort(unique(mapidx))
x[!mapidxNA] <- to[mapidx[!mapidxNA]]
     return(x)
}
y$DocTerm$j<-recoderFunc(y$DocTerm$j,sel.words,1:length(sel.words))
y$DocTerm$ncol<-length(sel.words)
DocTermR <- as.matrix(y$DocTerm)
if (!is.null(stop.word.user[[i]]))
      DocTermR <- DocTermR[, which(!colnames(DocTermR)
     if (stop.word.tm[i]) {
        stopword <- stopwords(y$idiom)
        DocTermR <- DocTermR[, which(!colnames(DocTermR)
     }
DocTermR<-DocTermR[apply(DocTermR,1,sum)>0,]
       DocVar<-DocVar[rownames(DocTermR),]
        Dcol<-ncol(DocTermR)
        gpo[i]<-Dcol
        MTables[[i]]<-DocTermR
        DocVarR<-as.data.frame(DocVar[which(rownames(DocVar)
 if(length(DocVarR)==1)
        colnames(DocVarR)<-colnames(DocVar)
        DocTermV<-cbind.data.frame(DocTermR,DocVarR)
        DocTermVar[[i]]<-DocTermV
        base<-DocVarR
     if(!is.null(num.agg)){
        AggregVar<-num.agg
      if(length(AggregVar)==1)
AggregVar<-AggregVar
    if(length(AggregVar)>1)
AggregVar<-AggregVar[i]
     if(is.character(AggregVar))
           AggregVar<-which(colnames(base)
          if(is.numeric(AggregVar))
             AggregVar<-AggregVar
         agg<-(base[,AggregVar])
         DocTermRA<-DocTermR
         dis.X<-tab.disjonctif(agg)
         Tagreg<-t(DocTermRA)
         Tagreg<-t(Tagreg)
if(i==1) Mtable=Tagreg
             else Mtable<-cbind.data.frame(Mtable,Tagreg)
          MTagregSep[[i]]<-Tagreg

  }
      else{
               if(i==1) Mtable=DocTermR[,apply(DocTermR,2,sum)>0]

             else{
 DocTermR<-DocTermR[which(rownames(DocTermR)
Mtable<-Mtable[which(rownames(Mtable)
```

```
                Mtable<-Mtable[(apply(Mtable,1,sum)>0)&(apply(DocTermR,1,sum)>0),]
                Mtable<-Mtable[,apply(Mtable,2,sum)>0]
                DocTermR<-DocTermR[(apply(Mtable,1,sum)>0)&(apply(DocTermR,1,sum)>0),]
                DocTermR<-DocTermR[,apply(DocTermR,2,sum)>0]
 Mtable<-cbind.data.frame(Mtable,DocTermR)
                gpo[i]<-ncol(DocTermR)

          }
MTagregSep=NULL
       }
  }

    res<-list(MDocWord=Mtable, ncolTs=gpo,MTagregSep=MTagregSep)
return(res)
}


## End(Not run)
```

`META.CA` *Metakeys-Metadocs (META.CA)*

Description

Representation and list of Metakeys and Metadocs: most contributory documents and words

Usage

```
META.CA(x, naxes = 5, axe.x = 1, axe.y = 2, lmd = 3, lmk = 3,
main = "Metakeys & Metadocs", graph = TRUE)
```

Arguments

`x`	either the result of a factor analysis, a dataframe, or a vector (coord, contrib)
`naxes`	number of axes to be considered (by default 5 axes)
`axe.x`	selected horizontal axis (by default "c(1,y)")
`axe.y`	selected vertical axis (by default "c(x,2)")
`lmd`	minimum threshold contribution for the metadocs (by default "3" times greater than the mean contribution)
`lmk`	minimum threshold contribution for the metakeys(by default "3" times greater than the mean contribution)
`main`	title of the graph
`graph`	if TRUE, graph is displayed for the metakeys and metadocs

Value

`Metakeys.Metadocs`	list with metakeys and metadocs on the first "naxes" dimensions
`DimWord`	dimension of the words
`graphMETA`	a graph is displayed visualizing the metakeys and metadocs

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>, Belchin Kostov

References

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.).

Kerbaol, M. Bansard, JY.;Coatrieux, JL. (2006) An analysis of IEEE publications in biomedical engineering. *IEEE Engineering in Medicine and Biology Magazine*.

Morin, A.(2006)Intensive Use of Factorial Correspondence Analysis for Text Mining: *Application with Statistical Education Publications*

Examples

```
## Not run:
data(Biblio)
resT<-DocWordTable(Biblio,num.tex=2,lminword=3)
DocTerm<-resT$DocTerm
DocVar<-DocVarTable(Biblio,VarSel=c(3,4,5))
res.TxCA<-TxCA(DocVar,DocTerm, Fmin=20,Dmin=10,graph=FALSE)
res.ca<-res.TxCA$res.ca
res.META<-META.CA(res.ca, naxes=5,axe.x=1,axe.y=2,lmd=6,lmk=3)

### CODE META.CA

META.CA <-function(x,naxes=5,axe.x=1,axe.y=2,lmd=3,lmk=3,
        main="Metakeys & Metadocs", graph = TRUE){

Ccontr<-x$col$contrib
Ccoor<-x$col$coord
Rcontr<-x$row$contrib
Rcoor<-x$row$coord
eigen<-x$eig

if(naxes>=min(nrow(Rcontr),nrow(Ccontr)))
naxes=min((nrow(Rcontr)-1),(nrow(Ccontr)-1))

# Computing the metakeys: words with contribuions over lmk*average_contribution of words
Metakeys<-vector(mode="list",length=naxes)
for (i in 1:naxes){
Metakeys[[i]]=vector(mode="list",length=2)}
for(i in 1:naxes){
Metakeys[[i]][[1]]<-sort(Ccontr[which(Ccontr[,i]>
         lmk*mean(Ccontr[,i])&Ccoor[,i]>0),i],decreasing=TRUE)
  if (length(Metakeys[[i]][[1]])==1)
      names(Metakeys[[i]][[1]])<-rownames(Ccontr)[which(Ccontr[,i]
      lmk*mean(Ccontr[,i])&Ccoor[,i]>0),i],decreasing=TRUE))]
     Metakeys[[i]][[2]]<-sort(Ccontr[which(Ccontr[,i]>
      lmk*mean(Ccontr[,i])&Ccoor[,i]<0),i],decreasing=TRUE)
  if (length(Metakeys[[i]][[2]])==1)
       names(Metakeys[[i]][[2]])<-rownames(Ccontr)[which(Ccontr[,i]
      lmk*mean(Ccontr[,i])&Ccoor[,i]<0),i],decreasing=TRUE))]
}

# Computing the metadocs :
# documents/answers with contribuions over lmd*average_contribution of documents
```

```
Metadocs<-vector(mode="list",length=naxes)
for (i in 1:naxes){
Metadocs[[i]]=vector(mode="list",length=2)}
for(i in 1:naxes){
Metadocs[[i]][[1]]<-sort(Rcontr[which(Rcontr[,i]>
        lmd*mean(Rcontr[,i])&Rcoor[,i]>0),i],decreasing=TRUE)
if (length(Metadocs[[i]][[1]])==1) names(Metadocs[[i]][[1]])
        <-rownames(Rcontr)[which(Rcontr[,i]
        lmd*mean(Rcontr[,i])&Rcoor[,i]>0),i],decreasing=TRUE))]
Metadocs[[i]][[2]]<-sort(Rcontr[which(Rcontr[,i]>
        lmd*mean(Rcontr[,i])&Rcoor[,i]<0),i],decreasing=TRUE)
if (length(Metadocs[[i]][[2]])==1) names(Metadocs[[i]][[2]])
        <-rownames(Rcontr)[which(Rcontr[,i]
       lmd*mean(Rcontr[,i])&Rcoor[,i]<0),i],decreasing=TRUE))]
}
# metakeys and metadocs
 Metakeys.MetaDocs<-vector(mode="list",naxes)
       names( Metakeys.MetaDocs)<-paste("DIM",1:naxes,sep="")
for (i in 1:naxes){
X <-list(names(Metakeys[[i]][[1]]),names(Metadocs[[i]][[1]]),
                      names(Metakeys[[i]][[2]]),names(Metadocs[[i]][[2]]))
            names(X)<-c("Metakeys+","Metadocs+","Metakeys-","Metadocs-")
           Metakeys.MetaDocs[[i]]<-X
          }

 #Dimension of the words
    mkeys<-numeric()
for (i in 1:naxes){
for (j in 1:2){
mkeys<-c(mkeys,names(Metakeys[[i]][[j]]))
}
}
mkeys<-as.factor(mkeys)
levels(mkeys)
matmkeys<-t(rbind(summary(mkeys,maxsum=10000),rep(naxes,
        length(levels(mkeys))),
        round(summary(mkeys,maxsum=10000)*100/naxes,2)))
colnames(matmkeys)<-c("Dim","Total.Dim","
DimensionWord<-matmkeys[order(matmkeys[,1],decreasing = T),]
if(graph){
# Graph of metakeys and metadocs on the "axes" of the CA
axes<-c(axe.x,axe.y)

# Identify metakeys of the dimensions "axes"
ClusPal1<-which(Ccontr[,axes[1]]
ClusPal2<-which(Ccontr[,axes[1]]
ClusPal3<-which(Ccontr[,axes[2]]
ClusPal4<-which(Ccontr[,axes[2]]
ClusPal<-c(ClusPal1,ClusPal2,ClusPal3,ClusPal4)
ClusPal<-ClusPal[!duplicated(ClusPal)]

# Identify metadocs of the dimensions "axes"
ClusDoc1<-which(Rcontr[,axes[1]]
ClusDoc2<-which(Rcontr[,axes[1]]
ClusDoc3<-which(Rcontr[,axes[2]]
ClusDoc4<-which(Rcontr[,axes[2]]
ClusDoc<-c(ClusDoc1,ClusDoc2,ClusDoc3,ClusDoc4)
```

```
ClusDoc<-ClusDoc[!duplicated(ClusDoc)]


if(length(ClusPal)>0&length(ClusDoc)>0){
xl=c(min(Ccoor[ClusPal,axes[1]],Rcoor[ClusDoc,axes[1]]),
max(Ccoor[ClusPal,axes[1]],Rcoor[ClusDoc,axes[1]]))
yl=c(min(Ccoor[ClusPal,axes[2]],Rcoor[ClusDoc,axes[2]]),
max(Ccoor[ClusPal,axes[2]],Rcoor[ClusDoc,axes[2]]))
}

if(length(ClusPal)>0&length(ClusDoc)==0){
xl=c(min(Ccoor[ClusPal,axes[1]]),max(Ccoor[ClusPal,axes[1]]))
yl=c(min(Ccoor[ClusPal,axes[2]]),max(Ccoor[ClusPal,axes[2]]))
}

if(length(ClusPal)==0&length(ClusDoc)>0){
xl=c(min(Rcoor[ClusDoc,axes[1]]),max(Rcoor[ClusDoc,axes[1]]))
yl=c(min(Rcoor[ClusDoc,axes[2]]),max(Rcoor[ClusDoc,axes[2]]))
}

if(length(ClusPal)>0|length(ClusDoc)>0){
plot(0,0,pch=16,col="white",xlab=paste("DIM",axes[1],
"(",round(eigen[axes[1],2],2),"
"(",round(eigen[axes[2],2],2),"
abline(h=0,v=0,lty=2)
if(length(ClusPal)>0){
points(Ccoor[ClusPal,axes[1]],Ccoor[ClusPal,axes[2]],col="red",pch=17)
text(Ccoor[ClusPal,axes[1]],Ccoor[ClusPal,axes[2]],rownames(Ccoor[ClusPal,]),col="red",
 cex=0.75,font=2,pos=3)
}
if(length(ClusDoc)>0){
points(Rcoor[ClusDoc,axes[1]],Rcoor[ClusDoc,axes[2]],col="blue",pch=16)
text(Rcoor[ClusDoc,axes[1]],Rcoor[ClusDoc,axes[2]],rownames(Rcoor[ClusDoc,]),col="blue"
,cex=0.75,font=2,pos=3)
}

}else{
print("There is no element to be represented")
}}
return (res=list( Metakeys.Metadocs=  Metakeys.MetaDocs, DimWord=DimensionWord))
}


## End(Not run)
```

`print.DocWordTable` *Prints DocWordTable results*

Description

Prints DocWordTable results.

Usage

```
## S3 method for class DocWordTable
print(x, file = NULL, sep = ";", ...)
```

Arguments

x	object of class DocWordTable
file	a connection, or a character string naming the file to print to. If NULL (the default), the results are not printed in a file
sep	character string to insert between the objects to print (if the argument file is not NULL)
...	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

DocWordTable, summary.DocWordTable

Examples

```
## Not run:
data(dataBiblio)
res.DWT<-DocWordTable(dataBiblio,num.tex=2, idiom="en",lminword=3,Remov.Number=TRUE,lower=TRUE)
print(res.DWT)

### CODE print.DocWordTable

print.DocWordTable <-function (x, file = NULL, sep = ";", ...)
{
res.DocWordTable <- x
if (!inherits(res.DocWordTable, "DocWordTable"))
          stop("non convenient data")
     cat("*The results are available in the following objects:\n\n")
        indice <-9
    res <- array("", c(indice, 2), list(1:indice, c("name", "description")))
    res[1, ] <- c("$Ndoc", "Number of documents")
    res[2, ] <- c("$Nlength", "Corpus size")
    res[3, ] <- c("$Nword", "Vocabulary size")
    res[4, ] <- c("$DocTerm", "Documents by Words table")
    res[5, ] <- c("$Tfreq", "Glossary")
    res[6, ] <- c("$Nfreqword", "Frequencies Words")
    res[7, ] <- c("$Ndocword", "Frequencies of word in documents")
    res[8, ] <- c("$corpus", "Corpus analyzed")
    res[9, ] <- c("$idiom", "Language of the textual  column(s)")
    print(res[1:indice, ])
     if (!is.null(file)) {
         write.infile(res.DocWordTable, file = file, sep = sep)
         print(paste("All the results are in the file", file))
     }
}


## End(Not run)
```

print.MacroBiblio *Prints MacroBiblio results*

Description

Prints MacroBiblio results.

Usage

```
## S3 method for class MacroBiblio
print(x, file = NULL, sep = ";", ...)
```

Arguments

x	object of class MacroBiblio
file	a connection, or a character string naming the file to print to. If NULL (the default), the results are not printed in a file
sep	character string to insert between the objects to print (if the argument file is not NULL)
...	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

MacroBiblio, write.infile

Examples

```
## Not run:
data(dataBiblio)
res.dataBiblio<-MacroBiblio(dataBiblio,lmd = 6, lmk =6, graph=FALSE)
print(res.dataBiblio)

### CODE print.MacroBiblio

print.MacroBiblio <-
function (x, file = NULL, sep = ";", ...)
{
res.MacroBiblio <- x
if (!inherits(res.MacroBiblio, "MacroBiblio")) stop("non convenient data")
cat("**Results of Analysis of bibliography  (MacroBiblio)**\n")
    cat("*The results are available in the following objects:\n\n")
indice <-12
    res <- array("", c(indice, 2), list(1:indice, c("name", "description")))
    res[1, ] <- c("$Corpus", "Summary of the information about corpus")
    res[2, ] <- c("$Glossary", "Glossary of words")
    res[3, ] <- c("$DocTermR", "Documents by words table")
    res[4, ] <- c("$Tagreg", "Lexical aggregate table")
    res[5, ] <- c("$Metakeys.Metadocs", "Words/documents with higher contribution ")
```

```
    res[6, ] <- c("$res.CA", "Results of  correspondence analysis direct")
    res[7, ] <- c("$res.CA.Agreg", "Result of Correspondence analysis aggregate")
    res[8, ] <- c("$CharWord", "characteristic words in each group of the aggregation variable")
    res[9, ] <- c("$res.CHCPC", "Result of constrained hierarchical clustering")
    res[10,] <- c("$res.MFACT","Result of multiple factor analysis for contingency tables")
    res[11,] <- c("$OrdWord", "words and their coordinates in the first dimension")
    res[12,] <- c("$pioneers", "pioneers articles")
    print(res[1:indice, ])
     if (!is.null(file)) {
         write.infile(res.MacroBiblio, file = file, sep = sep)
         print(paste("All the results are in the file", file))
     }
}



## End(Not run)
```

`print.MacroCaHcpc` *Prints MacroCaHcpc results*

Description

Prints MacroCaHcpc results.

Usage

```
## S3 method for class MacroCaHcpc
print(x, file = NULL, sep = ";", ...)
```

Arguments

`x`	object of class MacroCaHcpc
`file`	a connection, or a character string naming the file to print to. If NULL (the default), the results are not printed in a file
`sep`	character string to insert between the objects to print (if the argument file is not NULL)
`...`	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

MacroCaHcpc

Examples

```
## Not run:
data(dataOpen.question)
res.TxC<-MacroCaHcpc(dataOpen.question, num.text=c(6,7), Fmin=15,Dmin=15)
print(res.TxC)

### CODE print.MacroCaHcpc

print.MacroCaHcpc <-function (x, file = NULL, sep = ";", ...)
{
    res<- x
    if (!inherits(res, "MacroCaHcpc"))
        stop("non convenient data")
    cat("**Results for the MacroCaHcpc **\n")
    indice <- 5
    res <- array("", c(indice, 2), list(1:indice, c("name", "description")))
    res[1, ] <- c("$Corpus", "Description of corpus")
    res[2, ] <- c("$res.TxCA", "Results of correspondence analysis")
    res[3, ] <- c("$res.TxCharClust", "characteristic documents and words of the clusters")
    res[4, ] <- c("$res.hcpc", "Results of hierarchical clustering")
    res[5, ] <- c("$ncp", "number of dimensions preserved")
    print(res[1:indice, ])
    if (!is.null(file)) {
        write.infile(MacroCaHcpc, file = file, sep = sep)
        print(paste("All the results are in the file", file))
    }
}


## End(Not run)
```

`print.MacroTxChrono` *Prints MacroTxChrono results.*

Description

Prints MacroTxChrono results.

Usage

```
## S3 method for class MacroTxChrono
print(x, file = NULL, sep = ";", ...)
```

Arguments

`x`	object of class MacroTxChrono
`file`	a connection, or a character string naming the file to print to. If NULL (the default), the results are not printed in a file
`sep`	character string to insert between the objects to print (if the argument file is not NULL)
`...`	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

MacroTxChrono, summary.MacroTxChrono,write.infile

Examples

```
## Not run:
data(dataSpeech)
res.MTxC<-MacroTxChrono(dataSpeech, SentLength=120, num.text=1, HierWords=FALSE)
print(res.MTxC)

### CODE MacroTxChrono
print.MacroTxChrono <-function (x, file = NULL, sep = ";", ...)
{
    res<- x
    if (!inherits(res, "MacroTxChrono"))
        stop("non convenient data")
    cat("**Results for the MacroTxChrono**\n")
    indice <- 10
    res <- array("", c(indice, 2), list(1:indice, c("name", "description")))
    res[1, ] <- c("$SentenceList", "dataset with  sentences")
    res[2, ] <- c("$Homo.Groups", "Description homogeneous group")
    res[3, ] <- c("$Corpus", "Description of corpus")
    res[4, ] <- c("$Correlation", "Correlation between chronology and dimensions")
    res[4, ] <- c("$ncp", "axes will be taken into account in the constrained hierarchical clustering")
    res[5, ] <- c("$res.TxCA", "Results of correspondence analysis")
    res[6, ] <- c("$res.chcpc", "Results for the Constrained hierarchical clustering")
    res[7, ] <- c("$HierWord", "Characteristic words for every node of the hierarchy")
    res[8, ] <- c("$HierSegment", "Characteristic Segments for every node of the hierarchy")
    res[9, ] <- c("$VocIndex", "Vocabulary index")
res[10,] <- c("$spec.growth","specific growth of  the vocabulary")
    print(res[1:indice, ])
    if (!is.null(file)) {
        write.infile(MacroTxChrono, file = file, sep = sep)
        print(paste("All the results are in the file", file))
    }
}

## End(Not run)
```

---

print.TxCA *Prints TxCA results*

Description

Prints TxCA results.

Usage

```
## S3 method for class TxCA
print(x, file = NULL, sep = ";", ...)
```

Arguments

x	object of class TxCA
file	a connection, or a character string naming the file to print to. If NULL (the default), the results are not printed in a file
sep	character string to insert between the objects to be printed (if the argument file is not NULL)
...	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

TxCA, summary.TxCA

Examples

```
## Not run:
data(dataBiblio)
DocTerm<-DocWordTable(dataBiblio,num.tex=2,lminword=3)
DocVar<-DocVarTable(dataBiblio,VarSel=c(3,4,5))
res.Dir<-TxCA(DocTerm,DocVar,Fmin=20,Dmin=10,stop.word.tm=TRUE,graph=FALSE)
print(res.Dir)

### CODE print.TxCA

print.TxCA <-function (x, file = NULL, sep = ";", ...)
{
res.TxCA <- x
if (!inherits(res.TxCA, "TxCA")) stop("non convenient data")
cat("**Results for AC and Aggregated Lexical Table (TxCA)**\n")
    cat("*The results are available in the following objects:\n\n")
      indice <-11
   res <- array("", c(indice, 2), list(1:indice, c("name", "description")))
   res[1, ] <- c("$TableSummary", "Summary of the information")
   res[2, ] <- c("$Glossary", "Glossary of words")
   res[3, ] <- c("$Table", "Table was analyzed by CA")
   res[4, ] <- c("$DocTermR", "Table Documents by Words")
   res[5, ] <- c("$res.agg", "Result of aggregation")
   res[6, ] <- c("$Nfreqword", "Frequencies of words")
   res[7, ] <- c("$Ndocword", "Frequencies of words in documents")
   res[8, ] <- c("$res.ca", "Results of correspondence analysis")
   res[9, ] <- c("$VCr","Cramers V")
   res[10,] <- c("$Inertia","Total inertia")
   res[11,] <- c("$res.meta", "Results of Metakeys and Metadocs")
    print(res[1:indice, ])
     if (!is.null(file)) {
         write.infile(res.TxCA, file = file, sep = sep)
         print(paste("All the results are in the file", file))
     }
}


## End(Not run)
```

print.TxCHCPC *Prints TxCHCPC results*

Description

Prints TxCHCPC results.

Usage

```
## S3 method for class TxCHCPC
print(x, file = NULL, sep = ";", ...)
```

Arguments

x	object of class TxCHCPC
file	a connection, or a character string naming the file to print to. If NULL (the default), the results are not printed in a file
sep	character string to insert between the objects to print (if the argument file is not NULL)
...	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

TxCHCPC, write.infile

Examples

```
## Not run:
data(dataBiblio)
DocTerm<-DocWordTable(dataBiblio,num.tex=2,lminword=3)
DocVar<-DocVarTable(dataBiblio,VarSel=c(3,4,5))
res.agg<-TxCA(DocTerm,DocVar,Fmin=50,Dmin=10,num.agg =1,graph=FALSE)
Table<-res.agg$Table
res.CH<-TxCHCPC(res.agg$res.ca,cluster.CA = "rows", graph = FALSE)
print(res.CH)

### CODE print.TxCHCPC

print.TxCHCPC <-function (x, file = NULL, sep = ";", ...)
{
    res.TxCHCPC <- x
    if (!inherits( res.TxCHCPC, "TxCHCPC"))
        stop("non convenient data")
    cat("**Results for the Constrained hierarchical clustering**\n")
    res <- array("", c(24, 2), list(1:24, c("name", "description")))
    res[1, ] <- c("$data.clust", "dataset with the cluster of the individuals")
    res[2, ] <- c("$HierWord", "characteristic words for every node of the hierarchy")
    res[3, ] <- c("$desc.var", "description of the clusters by the variables")
```

```
  indice <- 4
   if (!is.null(res.TxCHCPC$desc.var$quanti.var)) {
       res[indice, ] <- c("$desc.var$quanti.var",
          "description of the cluster var. by the continuous var.")
       res[indice + 1, ] <- c("$desc.var$quanti",
          "description of the clusters by the continuous var.")
       indice <- indice + 2
   }
   if (!is.null(res.TxCHCPC$desc.var$test.chi2)) {
       res[indice, ] <- c("$desc.var$test.chi2",
       "description of the cluster var. by the categorical var.")
       res[indice + 1, ] <- c("$desc.axes$category",
        "description of the clusters by the categories.")
       indice <- indice + 2
   }
   res[indice, ] <- c("$desc.axes", "description of the clusters by the dimensions")
   indice <- indice + 1
   if (!is.null(res.TxCHCPC$desc.axes$quanti.var)) {
       res[indice, ] <- c("$desc.axes$quanti.var",
         "description of the cluster var. by the axes")
       res[indice + 1, ] <- c("$desc.axes$quanti",
        "description of the clusters by the axes")
       indice <- indice + 2
   }
   res[indice, ] <- c("$desc.ind", "description of the clusters by the individuals")
   res[indice + 1, ] <- c("$desc.ind$para", "parangons of each clusters")
   res[indice + 2, ] <- c("$desc.ind$dist", "specific individuals")
   res[indice + 3, ] <- c("$call", "summary statistics")
   res[indice + 4, ] <- c("$call$t", "description of the tree")
   indice <- indice + 4
   print(res[1:indice, ])
   if (!is.null(file)) {
       write.infile(res.TxCHCPC, file = file, sep = sep)
       print(paste("All the results are in the file", file))
   }
}



## End(Not run)
```

---

print.TxMFACT *Prints TxMFACT results*

Description

Prints TxMFACT results.

Usage

```
## S3 method for class TxMFACT
print(x, file = NULL, sep = ";", ...)
```

Arguments

x	an object of class TxMFACT
file	A connection, or a character string naming the file to print to. If NULL (the default), the results are not printed in a file
sep	character string to insert between the objects to print (if the argument file is not NULL
...	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

TxMFACT, write.infile

Examples

```
## Not run:
data(dataOpen.question)
res.M1<-DocWordTable(dataOpen.question,num.text=c(6,7))
res.M2<-DocWordTable(dataOpen.question,num.text=8)
DocVar<-DocVarTable(dataOpen.question,VarSel=c(1,2,3,4))
MDocTerm=list(res.M1$DocTerm,res.M1$DocTerm)
MDocVar=list(DocVar,DocVar)
MTable<-MDocWordTable(MDocTerm,MDocVar,num.agg=3,idiom=c("en","en"),
Fmin=c(10,5),Dmin=c(2,1))
MTable$ncolTs
res.mfact<-TxMFACT(MTable$MDocWord, group=MTable$ncolTs, type =c("f","f"),
name.group=c("Important","Culture"),graph=FALSE)
print(res.mfact)

### CODE print.TxMFACT

print.TxMFACT <- function (x, file = NULL, sep = ";", ...){
    res.mfa <- x
    if (!inherits(res.mfa, "TxMFACT")) stop("non convenient data")
    cat("**Results of the Multiple Factor Analysis (TxMFACT)**\n")
    cat("The analysis was performed on", nrow(res.mfa$call$X),
        "individuals, described by", ncol(res.mfa$call$X), "variables\n")
    cat("*Results are available in the following objects :\n\n")
    res <- array("", c(22, 2), list(1:22, c("name", "description")))
    res[1, ] <- c("$eig", "eigenvalues")
    res[2, ] <- c("$separate.analyses", "separate analyses for each group of variables")
    res[3, ] <- c("$group", "results for all the groups")
    res[4, ] <- c("$partial.axes", "results for the partial axes")
    res[5, ] <- c("$inertia.ratio", "inertia ratio")
    res[6, ] <- c("$ind", "results for the individuals")
    indice <- 7
    if (!is.null(res.mfa["ind.sup"]$ind.sup)){
      res[indice, ] <- c("$ind.sup", "results for the supplementary individuals")
      indice <- indice + 1
    }
    if (!is.null(res.mfa["quanti.var"]$quanti.var)){
      res[indice, ] <- c("$quanti.var", "results for the quantitative variables")
```

```
    indice <- indice + 1
  }
  if (!is.null(res.mfa["quali.var"]$quali.var)){
    res[indice, ] <- c("$quali.var", "results for the categorical variables")
    indice <- indice + 1
  }
  if (!is.null(res.mfa["quanti.var.sup"]$quanti.var.sup)){
    res[indice, ] <- c("$quanti.var.sup", "results for the quantitative supplementary variables")
    indice <- indice + 1
  }
  if (!is.null(res.mfa["quali.var.sup"]$quali.var.sup)){
    res[indice, ] <- c("$quali.var.sup", "results for the categorical supplementary variables")
    indice <- indice + 1
  }

  if (!is.null(res.mfa["freq"]$freq)){
    res[indice, ] <- c("$freq", "results for the frequencies")
    indice <- indice + 1
  }
  if (!is.null(res.mfa["freq.sup"]$freq.sup)){
    res[indice, ] <- c("$freq.sup", "results for the supplementary frequencies")
    indice <- indice + 1
  }


  if (!is.null(res.mfa$quanti.var)){
    res[indice, ] <- c("$summary.quanti", "summary for the quantitative variables")
    indice <- indice + 1
  }
  if (!is.null(res.mfa$quali.var)){
    res[indice, ] <- c("$summary.quali", "summary for the categorical variables")
    indice <- indice + 1
  }
  res[indice, ] <- c("$global.pca", "results for the global PCA")
  print(res[1:indice,])
  if (!is.null(file)) {
    write.infile(res.mfa,file = file, sep=sep)
    print(paste("All the results are in the file",file))
  }
}


## End(Not run)
```

`print.VocIndex` *Prints VocIndex results*

Description

Prints VocIndex results

Usage

```
## S3 method for class VocIndex
print(x, file = NULL, sep = ";", ...)
```

Arguments

x	object of class VocIndex
file	a connection, or a character string naming the file to print to. If NULL (the default), the results are not printed in a file
sep	character string to insert between the objects to print (if the argument file is not NULL)
...	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

VocIndex, summary.VocIndex,write.infile

Examples

```
## Not run:
data(open.question)
res.VcIndex<-VocIndex(open.question,num.tex=c(6,7),Fmin=5)
print.VocIndex(res.VcIndex)
summary.VocIndex(res.VcIndex)

### CODE print.VocIndex

print.VocIndex <-
function (x, file = NULL, sep = ";", ...)
{
res.VocIndex<- x
if (!inherits(res.VocIndex, "VocIndex")) stop("non convenient data")
cat("**Induce of Vocabulary  (VocIndex)**\n")
    indice <-3
    cat("*The results are available in the following objects:\n\n")
   res <- array("", c(indice, 2), list(1:indice, c("name", "description")))
   res[1, ] <- c("$RegVoc", "Regular or stable vocabulary")
   res[2, ] <- c("$LocalVoc","Local or specialized vocabulary")
   res[3, ] <- c("$VocIndex","Vocabulary index table  ")
    print(res[1:indice, ])
     if (!is.null(file)) {
         write.infile(res.VocIndex, file = file, sep = sep)
         print(paste("All the results are in the file", file))
     }
}


## End(Not run)
```

summary.DocWordTable *Summary DocWordTable objects*

Description

Summarizes DocWordTable object.

Usage

```
## S3 method for class DocWordTable
summary(object, nword=50, ordFreq=TRUE, ...)
```

Arguments

object	object of class DocWordTable
nword	number of words of the glossary to be printed
ordFreq	if ordeFreq=FALSE words of the glossary to be printed in alphabetic order
...	further arguments passed to or from other methods, such as cex, cex.main, ...

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

DocWordTable,print.DocWordTable

Examples

```
## Not run:
data(dataBiblio)
res.DWT<-DocWordTable(dataBiblio,num.tex=2, idiom="en")
summary(res.DWT,nword=20)

### CODE summary.DocWordTable

summary.DocWordTable<-function (object, nword=50, ordFreq=TRUE, ...)
{
  res <- object
     if (!inherits(res, "DocWordTable"))
        stop("non convenient object")
      cat("\nDocWordTable summary\n")
      cat("\nNumber of documents\n")
cat(" ",res$Ndoc,"\n")
cat("\nCorpus size\n")
cat("",res$Nlength,"\n")
cat("\nVocabulary size\n")
cat("",res$Nword,"\n")
 if(ordFreq){
      nword<-min(nword,nrow(res$Tfreq))
      cat("\nGlossary of the ",nword," most frequent  words\n")
      print(res$Tfreq[c(1:nword),])
   }else{
```

```
    Tabfq <- cbind(as.data.frame(res$Nfreqword), as.data.frame(res$Ndocword))
    colnames(Tabfq) <- c("Frequency", "N.Documents")
    nword<-min(nword,nrow(Tabfq))
    cat("\nGlossary of the ",nword," most frequent  words\n")
    print(Tabfq[c(1:nword),])
  }

}

## End(Not run)
```

---

summary.MacroBiblio *Summary MacroBiblio objects*

Description

Summarizes MacroBiblio objects.

Usage

```
## S3 method for class MacroBiblio
summary(object, nword=50, nEig=5, ...)
```

Arguments

object	object of class MacroBiblio
nword	number of words to be printed
nEig	number of eigenvalues to be printed (by default nEig=5)
...	further arguments passed to or from other methods, such as cex, cex.main, ...

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

MacroBiblio

Examples

```
## Not run:
data(dataBiblio)
res.dataBiblio<-MacroBiblio(dataBiblio, lmd = 6, lmk =6, graph=FALSE )
summary(res.dataBiblio, nword=20)

### CODE summary.MacroBiblio

summary.MacroBiblio<-function (object, nword=50, nEig=5, ...)
{
  res1 <- object
     if (!inherits(res1, "MacroBiblio"))
     cat("\nBiblioMineR summary\n")
        cat("\nCORPUS\n")
```

```
 summary(res1$Corpus, nword)
   cat("\nCORRESPONDENCE ANALYSIS CA\n")
       summary(res1$res.CA,nEig)
      cat("\nMETAKEYS-METADOCS\n")
       print(res1$Metakeys.Metadocs$Metakeys.Metadocs)
    cat("\nAGGREGATE CA\n")
 summary(res1$res.CA.Agreg, nEig)
     cat("\nCHARACTERISTIC WORDS OF AGGREGATED VARIABLE\n")
 print(res1$CharWord$CharWords)
     cat("\nMOST CONTRIBUTIVE",nword,"WORD IN THE DIM1\n")
         print(res1$OrdWord$WordMoreContrib[1:nword,])
    cat("\nPIONEER ARTICLES\n")
        print(res1$pioneers)
}


## End(Not run)
```

---

summary.MacroTxChrono *Summary MacroTxChrono objects*

Description

Summarizes MacroTxChrono objects.

Usage

```
## S3 method for class MacroTxChrono
summary(object, nword=20, nEig=5, ...)
```

Arguments

object	object of class MacroTxChrono
nword	number of words to be printed
nEig	number of eigenvalues to be printed (by default nEig=5)
...	further arguments passed to or from other methods, such as cex, cex.main, ...

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

MacroTxChrono

Examples

```
## Not run:
data(dataSpeech)
res.MTxC<-MacroTxChrono(dataSpeech, SentLength=120, num.text=1, HierWords=FALSE)
summary.MacroTxChrono(res.MTxC)

### CODE summary.MacroTxChrono
```

```
summary.MacroTxChrono<-function (object, nword=20, nEig=5, ...)
{
  res <- object
     if (!inherits(res, "MacroTxChrono"))
   stop("non convenient object")
    cat("\nMacroTxChrono summary\n")
    summary(res$Corpus,nword)
    summary(res$res.TxCA,nEig)
    cat("\nCorrelation between chronology and dimensions\n")
    print(res$Correlation)
   cat("\n ", res$ncp, " axes will be taken into account in the constrained hierarchical clustering\n")
    cat("\nHierarchical words\n")
    print(res$HierWord)
  if (!is.null(res$HierSegment)) {
        cat("\nHierarchical segments\n")
        print(res$HierSegment)
 }
    summary(res$VocIndex, nword)
}



## End(Not run)
```

---

summary.TxCA *Summary TxCA object*

Description

Summarizes TxCA objects.

Usage

```
## S3 method for class TxCA
summary(object, nb.dec = 3, nEig =5, ordDim = 1, order = FALSE, nbelements = 10,
nbind = nbelements, ncp = 3, align.names = TRUE, nword = 10, file = "", ...)
```

Arguments

object	object of class TxCA
nb.dec	number of decimals to be printed
nEig	number of eigenvalues to be printed (by default nEig=5)
ordDim	axis number which is used to rank the documents depending on their coordinate on this axis
order	if TRUE the documents are ranked depending on their contribution on ordDim axis
nbelements	number of elements whose coordinates are listed (variables, categories, frequencies); use nbelements = Inf to have all the elements
nbind	number of documents whose coordinates are listed; use nbind = Inf to have the results for all the documents and nbind = 0 if you do not want the results for any document

ncp	number of dimensions for which the results are printed
align.names	boolean, if TRUE the names of the objects are written using the same number of characters
nword	number of words in the glossary of frequency to be printed
file	a connection, or a character string naming the file to print to
...	further arguments passed to or from other methods

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

TxCA,print.TxCA

Examples

```
## Not run:
data(dataBiblio)
DocTerm<-DocWordTable(dataBiblio,num.tex=2,lminword=3)
DocVar<-DocVarTable(dataBiblio,VarSel=c(3,4,5))
res.Dir<-TxCA(DocTerm,DocVar,Fmin=20,Dmin=10,stop.word.tm=TRUE,graph=FALSE)
summary(res.Dir)

## End(Not run)
```

---

summary.TxMFACT *Summary TxMFACT objects*

Description

Summarizes TxMFACT objects.

Usage

```
## S3 method for class TxMFACT
summary(object, nb.dec = 3, nbelements = 10, nbind = nbelements,
 ncp = 3, align.names = TRUE, file = "", ...)
```

Arguments

object	object of class TxMFACT
nb.dec	number of decimals to be printed
nbelements	number of elements written (variables, categories, frequencies); use nbelements = Inf if you want to have all the elements
nbind	number of documents written; use nbind = Inf to have the results for all the documents and nbind = 0 if you do not want the results for documents
ncp	number of dimensions to be printed
align.names	boolean, if TRUE the names of the objects are written using the same number of characters
file	a connection, or a character string naming the file to print to
...	further arguments passed to or from other methods, such as cex, cex.main, ...

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

TxMFACT,print.TxMFACT

Examples

```
## Not run:
data(dataOpen.question)
res.M1<-DocWordTable(dataOpen.question,num.text=c(6,7))
res.M2<-DocWordTable(dataOpen.question,num.text=8)
DocVar<-DocVarTable(dataOpen.question,VarSel=c(1,2,3,4))
MDocTerm=list(res.M1$DocTerm,res.M1$DocTerm)
MDocVar=list(DocVar,DocVar)
MTable<-MDocWordTable(MDocTerm,MDocVar,num.agg=3,idiom=c("en","en"),
Fmin=c(10,5),Dmin=c(2,1))
MTable$ncolTs
res.mfact<-TxMFACT(MTable$MDocWord, group=MTable$ncolTs, type =c("f","f"),
name.group=c("Important","Culture"),graph=FALSE)
summary(res.mfact)

## End(Not run)
```

---

summary.VocIndex *Summary VocIndex objects*

Description

Summarizes VocIndex objects

Usage

```
## S3 method for class VocIndex
summary(object, nword = 20, ...)
```

Arguments

object	object of class IndVocab
nword	number of words to be printed
...	further arguments passed to or from other methods, such as cex, cex.main, ...

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

VocIndex

Examples

```
## Not run:
data(open.question)
res.VcIndex<-VocIndex(open.question,num.tex=c(6,7),Fmin=5)
print.VocIndex(res.VcIndex)
summary.VocIndex(res.VcIndex,nword=5)

### CODE summary.VocIndex

summary.VocIndex <-
function (object, nword=20, ...)
{
   res <- object
     if (!inherits(res, "VocIndex"))
        stop("non convenient object")
     cat("\nVocIndex summary\n")
     cat("\nUse of vacabulary\n")
     Nword1<-min(nword,nrow(res$RegVoc))
cat("\n",Nword1,"Regular Words\n")
       print(res$RegVoc[c(1:Nword1),])
      Nword2<-min(nword,nrow(res$LocalVoc))
cat("\n",Nword2,"specialized Words\n")
       print(res$LocalVoc[c(1:Nword2),])
}



## End(Not run)
```

TxCA *Correspondence Analysis of Lexical Tables (TxCA)*

Description

Correspondence analysis (CA) of lexical tables.

Usage

```
TxCA(y,x=NULL,num.agg=NULL, idiom="en", Fmin=5, Dmin=5, Fmax=NULL,
 equivalence=NULL, stop.word.user=NULL, stop.word.tm=FALSE,
  lmd=3, lmk=3,  ncp=5, row.sup = NULL, col.sup = NULL, quanti.sup=NULL,
   quali.sup = NULL, graph = TRUE, axes = c(1,2), row.w = NULL)
```

Arguments

y	an object of class DocWordTable
x	an object of class DocVar
num.agg	column index or name of the aggregation column, By default there is no aggregation
idiom	language of the textual column(s) (by default English "en")
Fmin	minimum threshold on the word frequency (by default 5)

Dmin	minimum threshold on the number of documents using the word (by default 5)
Fmax	maximum threshold on the word frequency
equivalence	data frame with n rows and two columns (original word and new word)
stop.word.user	vector indicating the stopwords chosen by the user
stop.word.tm	boolean, if TRUE the stopword list provided by tm is taken into account
lmk	minimum threshold on the contribution for selecting the metakeys (by default 3, which mean contribution 3 times greater than the mean contribution)
lmd	minimum threshold on the contribution for selecting the metadocs (by default 3, which mean contribution 3 times greater than the mean contribution)
ncp	number of dimensions stored in the results (by default 5)
row.sup	vector with the index(es) or name(s) of the supplementary row(s)
col.sup	vector with the index(es) or name(s) of the supplementary frequency column(s)
quanti.sup	vector with the index(es) or name(s) of the supplementary quantitative column(s)
quali.sup	vector with the index(es) or name(s) of the categorical supplementary column(s)
graph	boolean, if TRUE graphs are displayed
axes	a length 2 vector specifying the dimensions to plot
row.w	an optional row weights (by default, a vector of 1 and each row has a weight equal to its margin)

Value

Returns a list including:

TableSummary	summary of the results
Glossary	glossary of the selected words in frequency order
DocTermR	documents by words (all documents, selected words)
Table	table summitted to CA (non-empty documents, aggregated or not, by selected words)
Nfreqword	vector with frequencies of words
Ndocword	vector with frequencies of words in documents
res.Vcr	Cramer's V coefficient
Inertia	total inertia
res.ca	correspondence analysis results
res.agg	aggregation results
res.meta	Metakeys and Metadocs results

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

References

Benzecri, J, P. (1981). Pratique de l'analyse des donnees. Linguistique & lexicologie (Vol.3). (P. Dunod., Ed).

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.).

Murtagh F. (2005). Correspondence Analysis and Data Coding with R and Java. Chapman & Hall/CRC.

Husson F., Le S,, Pages J. (2011). Exploratory Multivariate Analysis by Example Using R. Chapman & Hall/CRC.

See Also

print.TxCA, summary.TxCA

Examples

```
## Not run:
data(dataBiblio)
DocTerm<-DocWordTable(dataBiblio,num.tex=2,lminword=3)
DocVar<-DocVarTable(dataBiblio,VarSel=c(3,4,5))
##Direct analysis
res.Dir<-TxCA(DocTerm,DocVar,Fmin=20,Dmin=10,stop.word.tm=TRUE)
print(res.Dir)
summary(res.Dir,nEig=5)
###Aggregate analysis by year
res.agg<-TxCA(DocTerm, DocVar,Fmin=20,Dmin=10,stop.word.tm=TRUE, num.agg="Year")
print(res.agg)
summary(res.agg,nEig=5, order=TRUE)


### CODE TxCA

TxCA <-function(y,x=NULL,num.agg=NULL, idiom="en", Fmin=5, Dmin=5, Fmax=NULL,
 equivalence=NULL, stop.word.user=NULL, stop.word.tm=FALSE,
  lmd=3, lmk=3,  ncp=5, row.sup = NULL, col.sup = NULL, quanti.sup=NULL,
   quali.sup = NULL, graph = TRUE, axes = c(1,2), row.w = NULL){

  DocTerm<-y$DocTerm
  DocVar<-x
  #controls
  if(is.null(DocVar)) DocVar<-as.data.frame(1:nrow(DocTerm))
  if(nrow(DocTerm)<3 | ncol(DocTerm)<3)
  stop("DocTerm must have at least three rows and three columns")
  if(nrow(DocTerm)!= nrow(DocVar))
  stop("Different number of rows in DocTerm and DocVar dataframes")
  if (!is.null(num.agg)&!is.null(quanti.sup))
    stop("When an aggregated CA is performed, quanti.sup  should be NULL")

  # Treatment of the equivalences
    if (!is.null(equivalence)){
DocTermB<-DocTerm
      equivalence<-equivalence[which(equivalence[,1]
      if(nrow(equivalence)<1 | ncol(equivalence)<2 )
      stop("equivalence must have at least one rows and two columns")
 for(i in 1:nrow(equivalence)){
      if (equivalence[i,1]
```

```
        DocTermB[,which(colnames(DocTermB)==equivalence[i,2])]<-(DocTermB[,
            which(colnames(DocTerm)==equivalence[i,1])]
              + DocTermB[,which(colnames(DocTerm)==equivalence[i,2])])
          }else{
            pos<-which(colnames(DocTermB)==equivalence[i,1])
            colnames(DocTermB)[pos]<-as.character(equivalence[i,2])
        }
       }
        DocTerm<-DocTermB[,which(!colnames(DocTermB)
     }
      #minimum threshold on the word frequency and
 #minimum threshold on the number of documents using the word
   FreqWord<-y$Nfreqword
   NumDoc<-y$Ndocword
   DocTermR<-DocTerm[,which(FreqWord>=Fmin&NumDoc>=Dmin)]
  if (nrow(DocTermR) < 3 | ncol(DocTermR) < 3)
        stop(cat("\n After selection the dimension of data frame is ",
         dim(DocTermR), "\n  It must have at least three rows and three columns.\n"))

  #maximum threshold on the word frequency
     if(!is.null(Fmax)){
  FrecMax<-apply(DocTermR,2,sum)
        PalSel<-which(FrecMax>=Fmax)
        DocTermR<-DocTermR[,-PalSel]
   }
  # stopwords chosen by the user
  if (!is.null(stop.word.user)){
DocTermR<-DocTermR[,which(!colnames(DocTermR)
   }
  #stopword list provided by tm
  if(stop.word.tm){
   stopword<-stopwords(idiom)
DocTermR<-DocTermR[,which(!colnames(DocTermR)
   }

  #control after selection
 if (nrow(DocTermR) < 3 | ncol(DocTermR) < 3)
    stop(cat("\n After selection the dimension of data frame is ",
     dim(DocTermR), "\n  It must have at least three rows and three columns.\n"))

 # Printint a first summay
      TableSummary<-matrix(c(sum(DocTerm),sum(DocTermR),nrow(DocTerm),nrow(DocTermR),
       ncol(DocTerm),ncol(DocTermR),round(sum(DocTerm)/nrow(DocTerm),2),
       round(sum(DocTermR)/nrow(DocTermR),2)), nrow =2, ncol = 4, byrow = FALSE,
       dimnames = list(c("Before-selection", "After-selection"),
                    c("Occurrences", "Documents", "Words", "Mean-length")))
  # frequencies of words,
    dtmMR<-as.matrix(DocTermR)
    Nfreqword <- apply(dtmMR, MARGIN = 2, FUN = sum)

  #frequencies of words in documents
    dtmAR <- dtmMR
    dtmAR[dtmMR[, ] > 0] <- 1
    Ndocword <- apply(dtmAR, MARGIN = 2, FUN = sum)

  #glossary of words
    Nfreqwords <- as.data.frame(Nfreqword)
```

```
    Ndocwords <- as.data.frame(Ndocword)
    Table <- cbind(Nfreqwords, Ndocwords)
    colnames(Table)<-c("Frequency", "N.Documents")
    Glossary<-with(Table,Table[order(Nfreqwords,Ndocwords,decreasing=TRUE),])

#if there are supplementary frequency columns
      if (!is.null(col.sup)){
         if(is.character(col.sup))
           col.sup<-which(colnames(DocTermR)
          if(is.numeric(col.sup))
     col.sup<-col.sup
       }
   ##there is aggregatio##
       base<-DocVar
   #num.agg is a column index or name of the aggregation column
   if(!is.null(num.agg)){
         if(is.character(num.agg))
           num.agg<-which(colnames(base)
          if(is.numeric(num.agg))
     num.agg<-num.agg
      base[,num.agg]<-as.factor(base[,num.agg])

   #if there is "NA" in the aggregation column, "NA" is converted into categorical
      if (any(is.na(base[,num.agg]))) {
             levels(base[,num.agg]) <- c(levels(base[,num.agg]),"NA")
             base[is.na(base[,num.agg]),num.agg] <- "NA"
       }
#if there is "" in the aggregation column, "" is converted into "NA"
       if (levels(base[,num.agg])[1] == "") levels(base[,num.agg])[1] = "NA"

    # building of aggregate table
agg<-as.factor(base[,num.agg])
       DocTermRA<-DocTermR
       dis.X<-tab.disjonctif(agg)
 Tagreg<-t(DocTermRA)
 Tagreg<-t(Tagreg)
 Torig<-t(DocTerm)
 Torig<-t(Torig)

     #If there are supplementary rows
     if (!is.null(row.sup)){
        if(is.character(row.sup))
         row.sup<-which(rownames(Tagreg)
         if(is.numeric(row.sup))
     row.sup<-row.sup
    }

    #if there are categorical supplementary columns
    Tagreg1<-Tagreg
    if (!is.null(quali.sup)){
  if(is.character(quali.sup))
           quali.sup<-which(colnames(base)
         if(is.numeric(quali.sup))
    quali.sup<-quali.sup

# building of aggregate tables of
# categorical supplementary columns
```

```
       for(i in 1:length(quali.sup)){
       nquali<-quali.sup[i]
       qag<-as.factor(base[,nquali])
       DocTermRA<-DocTermR
       dis.X<-tab.disjonctif(qag)
 Tquali<-t(DocTermRA)
 Tquali<-t(Tquali)
       if(i==1){
        Tagquali<-Tquali
        }else{
        Tagquali<-rbind(Tagquali,Tquali)
        }
       }
       TagregEx<-rbind(Tagreg,Tagquali)
     #index of supplementary rows when there is quali.sup
       if (!is.null(row.sup)&!is.null(quali.sup)){
       row.sup=c(row.sup,(nrow(Tagreg)+1):nrow(TagregEx))
       }else{
         row.sup=c((nrow(Tagreg)+1):nrow(TagregEx))
        }
        Tagreg1<-TagregEx
    }

 #correspondence analysis of aggregate table
ncp <- min(ncp, (nrow(Tagreg1) - 1), (ncol(Tagreg1) - 1))
       res.ca<-CA(Tagreg1[apply(Tagreg1,1,sum)>0,],ncp, row.sup, col.sup,
           quanti.sup=NULL, quali.sup=NULL,graph=FALSE)

    #tables (DocTerm+Categ, DocTermR+Categ)
    # used for the summary of the categories
     Categ<- as.data.frame(base[,num.agg])
     colnames(Categ)<-"GrupoCat"
     DocTermRA<- cbind(DocTermRA,Categ)
     DocTermRA$GrupoCat<-as.factor(DocTermRA$GrupoCat)
     DocTermOrg<- cbind(DocTerm,Categ)
     DocTermOrg$GrupoCat<-as.factor(DocTermOrg$GrupoCat)
     res.agg<-list(Tagreg=Tagreg,Torig=Torig, DocTermRA=DocTermRA,DocTermOrg=DocTermOrg)
}else{

    ##there is not aggregation, but there are
 #quali.sup or  quanti.sup or or both
       if (!is.null(quali.sup)|!is.null(quanti.sup)){
       if (!is.null(quali.sup)){
  if(is.character(quali.sup))
           quali.sup<-which(colnames(base)
         if(is.numeric(quali.sup))
    quali.sup<-quali.sup
          }
       if (!is.null(quanti.sup)){
  if(is.character(quanti.sup))
           quanti.sup<-which(colnames(base)
         if(is.numeric(quanti.sup))
    quanti.sup<-quanti.sup
       }
     # building of DocTerm incorporating quali.sup and quali.sup
DocTermRAC<-DocTermR[apply(DocTermR,1,sum)>0,]
            DocTermRAC<-DocTermRAC[,apply(DocTermRAC,2,sum)>0]
```

```
  base<-as.data.frame(base[rownames(DocTermRAC),])
         if (!is.null(quali.sup)&!is.null(quanti.sup))
VarAgreg<-base[,c(quali.sup,quanti.sup)]
   if (!is.null(quali.sup)& is.null(quanti.sup))
            VarAgreg<-base[,quali.sup]
         if (is.null(quali.sup)& !is.null(quanti.sup))
    VarAgreg<-base[,quanti.sup]
 VarAgreg<-as.data.frame(VarAgreg)
           DocTermRA<-cbind.data.frame(DocTermRAC, VarAgreg)
       if (!is.null(quali.sup)&!is.null(quanti.sup)){
  q<-length(quali.sup)
           quali.sup<-(ncol(DocTermR)+1):(ncol(DocTermRA)-length(quanti.sup))
           quanti.sup<-(ncol(DocTermR)+1+q):ncol(DocTermRA)
  }
if (!is.null(quali.sup)& is.null(quanti.sup))
         quali.sup<-(ncol(DocTermR)+1):ncol(DocTermRA)
       if (is.null(quali.sup)& !is.null(quanti.sup))
     quanti.sup<-(ncol(DocTermR)+1):ncol(DocTermRA)

   #If there are supplementary rows
          if (!is.null(row.sup)){
             if(is.character(row.sup))
             row.sup<-which(rownames(DocTermRA)
             if(is.numeric(row.sup))
        row.sup<-row.sup
             }
 #correspondence analysis  direct of  table DocTermR
 #when there are quanti.sup or quali.sup or both
  ncp <- min(ncp, (nrow(DocTermRA) - 1), (ncol(DocTermRA) - 1))
           res.ca<-CA(DocTermRA, ncp, row.sup, col.sup,
               quanti.sup, quali.sup,graph)
   res.agg=NULL
              Tagreg=NULL
         }else{
      #there is not aggregation##
  #If there are supplementary rows
  if (!is.null(row.sup)){
         if(is.character(row.sup))
          row.sup<-which(rownames(DocTermR)
          if(is.numeric(row.sup))
     row.sup<-row.sup
       }
#correspondence analysis  direct of  table DocTermR
 ncp <- min(ncp, (nrow(DocTermR) - 1), (ncol(DocTermR) - 1))
 res.ca<-CA(DocTermR[apply(DocTermR,1,sum)>0,],ncp, row.sup, col.sup,
    quanti.sup, quali.sup,graph)
        res.agg=NULL
        Tagreg=NULL
      }
 }
      if(!is.null(num.agg))
         Table<-Tagreg[apply(Tagreg,1,sum)>0,]
       else Table<-DocTermR[apply(DocTermR,1,sum)>0,]

  #Cramers V coefficient and total inertia
      Inertia<-round(sum(res.ca$eig[,1]),4)
  VCr<-round(sqrt(sum(res.ca$eig[,1])/min((nrow(Table)-1),(ncol(Table)-1))),4)
```

```
       InVc<-cbind(Inertia,VCr)
       colnames(InVc)<-c("Inertia", "Cramer V")
       rownames(InVc)<-"Total"
if(graph){  #graphs are displayed

barplot(res.ca$eig[,1],main="Eigenvalues",
                  names.arg=paste("dim",1:nrow(res.ca$eig)))

            if((!is.null(col.sup))&(is.null(num.agg))){
dev.new()
              plot(res.ca,invisible=c("col","row","row.sup"),selectCol = "cos2 30",
              unselect=1,axes=axes, title="Supplementary" )
            }

            dev.new()
       res.meta<-META.CA(res.ca,naxes=ncp,axe.x=axes[1],axe.y=axes[2],
                         lmd=-Inf,lmk, main="CA documents/words")

dev.new()
plot(res.ca,axes=axes, title="CA documents/words" )

dev.new()
res.meta1<-META.CA(res.ca,naxes=ncp,axe.x=axes[1],axe.y=axes[2],
                        lmd=Inf,lmk, main="CA words")

dev.new()
plot(res.ca,invisible=c("col","col.sup"),axes=axes, title="CA documents" )

  }else{
#Metakeys and Metadocs
res.meta<-META.CA(res.ca,naxes=ncp,axe.x=axes[1],axe.y=axes[2],
                    lmd,lmk, main="CA documents/words",graph = FALSE)

}
 #Returns a list including the values
    res<-list(TableSummary=TableSummary,DocTermR=DocTermR,Tagreg=Tagreg,
                Table=Table,Nfreqword=Nfreqword,Ndocword=Ndocword,
                 Glossary=Glossary, res.ca=res.ca, VCr=VCr, res.meta=res.meta,
                  Inertia=Inertia, res.agg=res.agg, Inertia.VCr=InVc)
    #function is an object of class TxCA and list
 class(res)<-c("TxCA","list")

return(res)

## End(Not run)
```

TxCharClust *Characteristic Documents and Words of the Clusters (TxCharClust)*

Description

Characteristic documents and words of the clusters.

Usage

```
TxCharClust(base, res, num.text)
```

Arguments

base	data frame with at least one textual column
res	HCPC or TxCHCPC results
num.text	column index(es) or name(s) of the textual column(s)

Value

Char_word_doc	list with characteristic documents and words of each cluster
CharWord_details	detail list with the characteristic words of each cluster

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

References

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.).

See Also

MacroCaHcpc

Examples

```
## Not run:
data(dataOpen.question)
DocTerm<-DocWordTable(dataOpen.question,num.text=c(6,7))
DocVar<-DocVarTable(dataOpen.question,VarSel=c(1,2,3,4))
res.TxCA<-TxCA(DocTerm,DocVar,Fmin=10,Dmin=5,stop.word.tm=TRUE,graph=FALSE)
res.hcpc<-HCPC(res.TxCA$res.ca,cluster.CA = "rows", graph=FALSE)
res.CharClust<-TxCharClust(dataOpen.question, res.hcpc,num.text=c(6,7))
res.CharClust$Char_word_doc

### CODE TxCharClust

TxCharClust <-function(base, res, num.text){
if (!is.null(num.text)) {
 if (is.character(num.text))
 num.text<- which(colnames(base)
  if (is.numeric(num.text))
  num.text<- num.text
 if(length(num.text)==1)
  num.text<-num.text
 if(length(num.text)>1){
   for(i in 1:length(num.text)){
    if(i==1)
    text1<-base[,num.text[1]]
    else text1<-paste(text1,base[,num.text[i]],sep=".")
   }
    base[,(ncol(base)+1)]<-text1
    num.text<-ncol(base)
  }
}
res.hcpc<-res
```

```
n<-length(res.hcpc$desc.ind$para)
m<-length(res.hcpc$desc.ind$dist)
resPara<-vector(mode="list",n)
resdist<-vector(mode="list",m)
DocumentPara<-vector(mode="list",n)
DocumentDist<-vector(mode="list",m)
for(i in 1:n){
IndPara<-as.data.frame(res.hcpc$desc.ind$para[i])
       resp<-base[which(rownames(base)
       resPara[[i]]<-resp
       IndPara$Resp<-resp
       DocumentPara[[i]]<-rownames(IndPara)
      IndPara$Document<-as.vector(rownames(IndPara))
IndPara$cluster<-rep(i,length(res.hcpc$desc.ind$para[[i]]))
colnames(IndPara)[1]<-"X"
if(i==1) tP<-IndPara
else tP<-rbind(tP,IndPara)
}

for(i in 1:m){
       IndDist<-as.data.frame(res.hcpc$desc.ind$dist[i])
respD<-base[which(rownames(base)
       resdist[[i]]<-respD
       IndDist$RespD<-respD
       DocumentDist[[i]]<-rownames(IndDist)
      IndDist$Document<-as.vector(rownames(IndDist))
IndDist$cluster<-rep(i,length(res.hcpc$desc.ind$dist[[i]]))
colnames(IndDist)[1]<-"X"
      IndDist<-IndDist[,c(2,3,4)]
if(i==1) tD<-IndDist
else tD<-rbind(tD,IndDist)
}
    data.clust<-res.hcpc$data.clust
    desc.var <- descfreq(data.clust[, -which(sapply(data.clust,
        is.factor))], data.clust[, ncol(data.clust)], proba = 0.05)
     CharWord1 = desc.var
     nClust<-length(CharWord1)
     CharClust<-vector(mode="list",nClust)
     names(CharClust)<-paste("clust", 1:nClust, sep = "")
     OverCharWord = vector(mode = "list", length = length(CharWord1))
     names(OverCharWord) <- names(CharWord1)
     InfraCharWord = vector(mode = "list", length = length(CharWord1))
     names(InfraCharWord) <- names(CharWord1)
 for (i in 1:length(CharWord1)){
        bWord<-as.data.frame(CharWord1[[i]])
        over<-subset(bWord, bWord$v.test>0)
        infra<-subset(bWord, bWord$v.test<0)
        OverCharWord[[i]] <- row.names(over)
        InfraCharWord[[i]] <- row.names(infra)
           Para<-resPara[[i]]
           DocP<-DocumentPara[[i]]
           Dist<-resdist[[i]]
           DocD<-DocumentDist[[i]]
       X <-list(OverCharWord[[i]],InfraCharWord[[i]],DocP,Para,DocD,Dist)
          names(X)<-c("Over_represented_word","Infra_represented_word",
          "Close_to_centroid_documents_(label)","Close_to_centroid_documents_(content)",
           "Far_from_other_clusters_document_(label)" ,
```

```
           "Far_from_other_clusters_document_(content)")
          CharClust[[i]]<-X
         }
    res=list(Char_word_doc=CharClust,CharWord_details=CharWord1)
 return(res)
}

## End(Not run)
```

TxCHCPC *Constrained Hierarchical Clustering (TxCHCPC)*

Description

Agglomerative constrained hierarchical clustering of documents starting from the results CA or MFACT. The chronology corresponds to the order of the documents in the analysed data frame.

Usage

```
TxCHCPC(res, Table = NULL, nb.clust = 0, min = 3, max = NULL, metric = "euclidean",
 nb.par = 5, graph = TRUE, proba = 0.05, ...)
```

Arguments

res	CA or MFACT result
Table	data frame with I rows (documents) and J columns (words)
nb.clust	integer. If nb.clust=0, the tree is cut at the level the user clicks on. If nb.clust=-1, the tree is automatically cut at the suggested level (see details). If a (positive) integer, the tree is cut with nb.cluters clusters.
min	integer. Minimum number of suggested clusters
max	integer. Maximum number of suggested clusters; by default the maximum is between 10 and the number of documents divided by 2.
metric	used metric. The currently available options are "euclidean" and "manhattan"
nb.par	integer. Number of edited paragons
graph	if TRUE, graphis are displayed
proba	probability used to select axes and variables in cluster description (see catdes for details)
...	arguments passed to or from other methods

Value

Returns a list including:

data.clust	the original data with a supplementary row called cluster containing the partition
desc.axes	description of the clusters by the factors (axes). (See catdes for detail)
desc.var	description of the clusters by the variables. (See catdes for details)
call	list or parameters and internal objects; call$t gives the results for the hierarchical tree

ind.desc	paragons (para) and the more typical documents of each cluster. (See details)
HierWord	characteristic words for each node of the hierarchy

Returns the tree and a barplot of dissimilarity levels, the document factor map with the tree (3D), the factor map with documents colored by cluster (2D).

Author(s)

Daria M Hernandez <daria.micaela.hernandez@upc.edu>

References

Legendre, P. & Legendre, L. (1998), Numerical Ecology (2nd ed.), Amsterdam: Elsevier Science.

Becue-Bertaut, M., Kostov, B., Morin, A., & Naro, G. (2014). Rhetorical Strategy in Forensic Speeches: Multidimensional Statistics-Based Methodology *Journal of Classification*, 31. doi:10.1007/s00357-014-9148-9.

Murtagh F. (1985). Multidimensional Clustering Algorithms. Vienna: Physica-Verlag, COMPSTAT Lectures.

Murtagh F. (2005). Correspondence Analysis and Data Coding with R and Java. Chapman & Hall/CRC.

See Also

TxCharClust, print.TxCHCPC

Examples

```
## Not run:
data(dataBiblio)
DocTerm<-DocWordTable(dataBiblio,num.tex=2,lminword=3)
DocVar<-DocVarTable(dataBiblio,VarSel=c(3,4,5))
res.agg<-TxCA(DocTerm,DocVar,Fmin=100,Dmin=20,num.agg =1,graph=FALSE)
Table<-res.agg$Table
res.CH<-TxCHCPC(res.agg$res.ca, nb.clust = -1)
#With characteristic words for every node of the hierarchy
res.CH<-TxCHCPC(res.agg$res.ca, nb.clust = -1, Table)
res.CH$HierWord
print(res.CH)

### CODE TxCHCPC

TxCHCPC <-function (res,Table=NULL, nb.clust = 0,  min =3, max = NULL,
 metric = "euclidean",nb.par = 5, graph = TRUE, proba = 0.05, ...)
{
 method = "complete"
 graph.scale= "inertia"
 kk = Inf
 cluster.CA = "rows"
 Chc.out.tree= function(res, min, max, metric, method, cla = NULL, ...)
         {

      X = as.data.frame(res$ind$coord)
d<-dist(X)
d0<-as.matrix(d)
maxd<-max(d)
```

```
maxd<-maxd+1e-10
Sim<-as.matrix(maxd-d)
Sim0<-Sim
d<-as.matrix(d)

### Constrained Matrix
Cont<-matrix(nrow=nrow(Sim),ncol=ncol(Sim),0)
Cont[1,2]<-1
for (i in 2:(nrow(Cont)-1)){
Cont[i,i+1]<-1
Cont[i,i-1]<-1
}
Cont[nrow(Cont),nrow(Cont)-1]<-1
rownames(Cont)<-rownames(Sim)
colnames(Cont)<-colnames(Sim)

### Similarity matrix used for constrained clustering
SimCont<-Sim*Cont
DistCont<-d*Cont
groups<-list()
for (i in 1:nrow(Sim)){
groups[[i]]<-i
}

distclust<-numeric()
clust<-list()
i<-1

indice<-nrow(d)-1

while(indice>0){

### Find the position of the maxim similarity
maxsim<-max(SimCont)
posmaxsim<-which(SimCont==maxsim)

if (posmaxsim[1]
fila<-posmaxsim[1]
col<-nrow(SimCont)
}else{
fila<-posmaxsim[1]
col<-posmaxsim[1]
}

maxfc<-max(fila,col)
minfc<-min(fila,col)
distclust[i]<-DistCont[fila,col]

clust[[i]]<-vector(mode="list",length=2)
clust[[i]][[1]]<-groups[[minfc]]
clust[[i]][[2]]<-groups[[maxfc]]

rownames(Sim)[minfc]<-colnames(Sim)[minfc]<-rownames(d)[minfc]
                <-colnames(d)[minfc]<-rownames(Cont)[minfc]<-colnames(Cont)[minfc]
                <-paste(rownames(Sim)[minfc],"-",rownames(Sim)[maxfc])

if (minfc!=1){
```

```
Sim[minfc,minfc-1]<-Sim[minfc-1,minfc]<-0.5*Sim[minfc,minfc-1]
                +0.5*Sim[maxfc,minfc-1]-0.5*abs(Sim[minfc,minfc-1]-Sim[maxfc,minfc-1])
d[minfc,minfc-1]<-d[minfc-1,minfc]<-0.5*d[minfc,minfc-1]
               +0.5*d[maxfc,minfc-1]+0.5*abs(d[minfc,minfc-1]-d[maxfc,minfc-1])
Cont[minfc-1,minfc]<-Cont[minfc,minfc-1]<-1
}
if (maxfc!=nrow(SimCont)){
Sim[maxfc+1,minfc]<-Sim[minfc,maxfc+1]<-0.5*Sim[minfc,maxfc+1]+
               0.5*Sim[maxfc,maxfc+1]-0.5*abs(Sim[minfc,maxfc+1]-Sim[maxfc,maxfc+1])
d[maxfc+1,minfc]<-d[minfc,maxfc+1]<-0.5*d[minfc,maxfc+1]+
               0.5*d[maxfc,maxfc+1]+0.5*abs(d[minfc,maxfc+1]-d[maxfc,maxfc+1])
Cont[maxfc+1,minfc]<-Cont[minfc,maxfc+1]<-1
}

groups[[minfc]]<-c(groups[[minfc]],groups[[maxfc]])
groups<-groups[-maxfc]
Sim<-Sim[-maxfc,-maxfc]
d<-d[-maxfc,-maxfc]
Cont<-Cont[-maxfc,-maxfc]
i<-i+1

SimCont<-Sim*Cont
DistCont<-d*Cont
indice<-indice-1
             }

clust<-clust

hc<-hclust(dist(X))
hc$height<-distclust
hc$order<-sort(hc$order)
grups_blocs<-list()
grups_blocs[[1]]<-rep(0,nrow(X))
for (i in 1:(length(clust)-1)){
grups_blocs[[i+1]]<-grups_blocs[[i]]
grups_blocs[[i+1]][c(clust[[i]][[1]],clust[[i]][[2]])]<-i
}

for(i in 1:nrow(hc$merge)){
if (length(clust[[i]][[1]])==1&length(clust[[i]][[2]])==1){
hc$merge[i,1]<-(-clust[[i]][[1]])
hc$merge[i,2]<-(-clust[[i]][[2]])
}else{
if (length(clust[[i]][[1]])==1){
hc$merge[i,1]<-(-clust[[i]][[1]])
}else{
hc$merge[i,1]<-grups_blocs[[i]][clust[[i]][[1]][1]]
}

if (length(clust[[i]][[2]])==1){
hc$merge[i,2]<-(-clust[[i]][[2]])
}else{
hc$merge[i,2]<-grups_blocs[[i]][clust[[i]][[2]][1]]
}
   }

}
```

```
            coord = as.data.frame(res$ind$coord)
coord2<-coord
marge.row<-res$call$row.w
marge.row2<-marge.row
moy.p <- function(V, poids) {
          res <- sum(V * poids)/sum(poids)
     }
 Total.inertia <- sum(apply(sweep(coord^2, 1, marge.row, "*"), 2, sum))
        Between <- vector()
        Within<-vector()
        for (i in 1:(length(clust))) {
            coord2[clust[[i]][[1]][1], ] <- apply(coord2[c(clust[[i]][[1]],
                clust[[i]][[2]]), ], 2, moy.p, marge.row2[c(clust[[i]][[1]],
                clust[[i]][[2]])])
            coord2[c(clust[[i]][[2]]), ] <- 0
            marge.row2[clust[[i]][[1]][1]] <- sum(marge.row2[clust[[i]][[1]]]) +
                sum(marge.row2[clust[[i]][[2]]])
            marge.row2[c(clust[[i]][[2]])] <- 0
            coord3 <- coord2[apply(coord2, 1, sum) != 0, ]
            marge.row3 <- marge.row2[which(marge.row2 != 0)]
            Inter <- sum(apply(sweep(coord3^2, 1, marge.row3, "*"),
                2, sum))
            Intra<-Total.inertia-Inter
            Between[i] <-round(Inter, 10)
            Within[i]<-round(Intra,10)
        }
 Within<-rev( Within)
 Between<-rev(Between)
        quot = Within[min:(max)]/Within[(min - 1):(max - 1)]
        nb.clust = which.min(quot) + min - 1
        inert.gain <-rev(hc$height)
        return(list(res = res, tree = hc, merge=hc$merge, nb.clust = nb.clust,
        Within.inertia = Within,
    Between.inertia=Between, dissimilarity.levels = inert.gain, quot = quot))
    }
  res.origen<-res
  coord.construction = function(coord.centers, coord.ind, clust) {
        coord.centers = as.data.frame(coord.centers)
        for (i in 1:nrow(coord.centers)) rownames(coord.centers)[i] = paste("center",
            i)
        coord.ind = cbind(coord.ind, clust)
        return(list(coord.ind = coord.ind, coord.centers = coord.centers))
    }
    select = function(Y, default.size, method, coord.centers) {
        clust = Y[1, ncol(Y)]
        Y = Y[, -ncol(Y)]
        Z = rbind(Y, coord.centers)
        if (nrow(Y) == 1) {
            distance = data.frame(0, row.names = "")
            colnames(distance) = rownames(Z[1, ])
        }
        else {
            distance = as.matrix(dist(Z, method = method))
            distance = distance[(nrow(Y) + 1):nrow(distance),
                -((nrow(Y) + 1):ncol(distance))]
            distance = sort(distance[clust, ], decreasing = FALSE)
```

```
    }
    if (length(distance) > default.size)
        distance = distance[1:default.size]
    else distance = distance
}
distinctivness = function(Y, default.size, method, coord.centers) {
    clust = as.numeric(Y[1, ncol(Y)])
    Y = Y[, -ncol(Y)]
    Z = rbind(Y, coord.centers)
    if (nrow(Y) == 1) {
        distance = as.matrix(dist(Z, method = method))
        ind.car = vector(length = 1, mode = "numeric")
        ind.car = min(distance[-c(1, (clust + 1)), 1])
        names(ind.car) = rownames(Z[1, ])
    }
    else {
        distance = as.matrix(dist(Z, method = method))
        distance = distance[(nrow(Y) + 1):nrow(distance),
            -((nrow(Y) + 1):ncol(distance))]
        if (nrow(distance) == 2)
            center.min = distance[-clust, ]
        else center.min = apply(distance[-clust, ], 2, min)
        ind.car = sort(center.min, decreasing = TRUE)
    }
    if (length(ind.car) > default.size)
        ind.car = ind.car[1:default.size]
    else ind.car = ind.car
}
if (is.vector(res)) {
    res = cbind.data.frame(res, res)
    res = PCA(res, scale.unit = FALSE, ncp = Inf, graph = FALSE)
    vec = TRUE
}
else vec = FALSE
if (is.matrix(res))
    res <- as.data.frame(res)
cla <- NULL
if (is.data.frame(res)) {
    res <- res[, unlist(lapply(res, is.numeric))]
    if (kk < nrow(res)) {
        cla <- kmeans(res, centers = kk, iter.max = 100,
            nstart = 4)
        res <- PCA(cla$centers, row.w = cla$size, scale.unit = FALSE,
            ncp = Inf, graph = FALSE)
    }
    else res <- PCA(res, scale.unit = FALSE, ncp = Inf, graph = FALSE)
}
if (inherits(res, "CA")) {
    if (cluster.CA == "rows")
        res = PCA(res$row$coord, scale.unit = FALSE, ncp = Inf,
            graph = FALSE, row.w = res$call$marge.row * sum(res$call$X))
    if (cluster.CA == "columns")
        res = PCA(res$col$coord, scale.unit = FALSE, ncp = Inf,
            graph = FALSE, row.w = res$call$marge.col * sum(res$call$X))
}
if (is.null(max))
    max = min(10, round(nrow(res$ind$coord)/2))
```

```
        max = min(max, nrow(res$ind$coord) - 1)
    if (inherits(res, "PCA") | inherits(res, "MCA") | inherits(res,
        "MFA") | inherits(res, "HMFA") | inherits(res, "FAMD")) {
        if (!is.null(res$call$ind.sup))
            res$call$X = res$call$X[-res$call$ind.sup, ]
        t =  Chc.out.tree(res, min = min, max = max, metric = metric,
            method = method, cla = cla)

    }
    else stop("res should be from PCA, MCA, FAMD, MFA, or HMFA class")

    if(!is.null(Table))
                HierWord=HierarchWords(res,Table)
          else HierWord=NULL

if (inherits(t$tree, "agnes"))
        t$tree <- as.hclust(t$tree)
    if (inherits(t$tree, "hclust")) {
        if (graph.scale == "inertia") {
            nb.ind = nrow(t$res$ind$coord)
            inertia.height = rep(0, nb.ind - 1)
            for (i in 1:(nb.ind - 1)) inertia.height[i] = t$dissimilarity.levels[(nb.ind -
                i)]
            inertia.height = sort(inertia.height, decreasing = FALSE)
            t$tree$height = inertia.height
        }
        auto.haut = ((t$tree$height[length(t$tree$height) - t$nb.clust +
            2]) + (t$tree$height[length(t$tree$height) - t$nb.clust +
            1]))/2
        if (graph) {
            if (!nzchar(Sys.getenv("RSTUDIO_USER_IDENTITY")))
                dev.new()
            par(mar = c(0.5, 2, 0.75, 0))
            lay = matrix(ncol = 5, nrow = 5, c(2, 4, 4, 4, 4,
                2, 4, 4, 4, 4, 2, 4, 4, 4, 4, 2, 4, 4, 4, 4,
                1, 3, 3, 3, 3))
            layout(lay, respect = TRUE)
            layout.show(n = 4)
            barplot(t$dissimilarity.levels[1:max(15, max)], col = c(rep("black",
                t$nb.clust - 1), rep("grey", max(max, 15) - t$nb.clust +
                1)), rep(0.1, max(max, 15)), space = 0.9)
            plot(x = 1, xlab = "", ylab = "", main = "", col = "white",
                axes = FALSE)
            text(1, 1, "Constrained Hierarchical Clustering", cex = 2)
            plot(x = 1, xlab = "", ylab = "", main = "", col = "white",
                axes = FALSE)
            legend("top", "Dissimilarity levels", box.lty = NULL, cex = 1)
        }
        else {
            if (nb.clust == 0 | nb.clust == 1)
                nb.clust = -1
        }
        if ((nb.clust == 0) | (nb.clust == 1)) {
            if (!nzchar(Sys.getenv("RSTUDIO_USER_IDENTITY"))) {
                plot(t$tree, hang = -1, main = "Click to cut the tree",
                  xlab = "", sub = "")
                abline(h = auto.haut, col = "black", lwd = 3)
```

```
            coupe = locator(n = 1)
            while (coupe$y < min(t$tree$height)) {
              cat("No class \n")
              coupe = locator(n = 1)
            }
            y = coupe$y
        }
        else {
            plot(t$tree, hang = -1, main = "Tree and suggested number of clusters",
              xlab = "", sub = "")
            abline(h = auto.haut, col = "black", lwd = 3)
            y <- auto.haut
        }
    }
    else {
        if (graph)
            plot(t$tree, hang = -1, main = "Constrained Hierarchical Clustering",
              xlab = "", sub = "")
        if (nb.clust < 0)
            y = auto.haut
        else y = (t$tree$height[length(t$tree$height) - nb.clust +
            2] + t$tree$height[length(t$tree$height) - nb.clust +
            1])/2
    }
}
else stop("The tree should be from hclust or agnes class.")
clust = cutree(as.hclust(t$tree), h = y)
nb.clust = max(clust)
X = as.data.frame(t$res$ind$coord)
if ((graph) & !nzchar(Sys.getenv("RSTUDIO_USER_IDENTITY"))) {
    rect = rect.hclust(t$tree, h = y, border = seq(1, nb.clust,
        1))
    clust = NULL
    for (j in 1:nb.clust) clust = c(clust, rep(j, length(rect[[j]])))
    clust = as.factor(clust)
    belong = cbind.data.frame(t$tree$order, clust)
    belong = belong[do.call("order", belong), ]
    clust = belong$clust
    clust = as.factor(clust)
}
    list.centers = by(X, clust, colMeans)
    centers = matrix(unlist(list.centers), ncol = ncol(X),
        byrow = TRUE)
    colnames(centers) = colnames(X)
    coordon = coord.construction(centers, X, clust)

cluster = coordon$coord.ind$clust
para = by(coordon$coord.ind, cluster, simplify = FALSE, select,
    default.size = nb.par, method = metric, coord.centers = coordon$coord.centers)
dist = by(coordon$coord.ind, cluster, simplify = FALSE, distinctivness,
    default.size = nb.par, method = metric, coord.centers = coordon$coord.centers)
desc.ind = list(para = para, dist = dist)
clust = as.factor(clust)
X = cbind.data.frame(X, clust)
data.clust = cbind.data.frame(t$res$call$X, clust)
if (vec)
    data.clust = as.data.frame(data.clust[, -2])
```

```
    desc.var = catdes(data.clust, ncol(data.clust), proba = proba)
    desc.axe = catdes(X, ncol(X), proba = proba)
    call = list(t = t, min = min, max = max, X = X, vec = vec,
        call = sys.calls()[[1]])
    data.clust  = cbind.data.frame(res.origen$call$X, clust)
        res.TxCHCPC = list(data.clust = data.clust, desc.var = desc.var,
        desc.axes = desc.axe, call = call, desc.ind = desc.ind,HierWord=HierWord)
    if ((graph) & !nzchar(Sys.getenv("RSTUDIO_USER_IDENTITY"))) {
        if (vec)
            plot.HCPC(res.TxCHCPC, choice = "3D.map", t.level = "all",
                angle = 0, ind.names = FALSE, new.plot = TRUE)
        else {
            plot.HCPC(res.TxCHCPC, choice = "3D.map", t.level = "all",
                ind.names = TRUE, new.plot = TRUE)
            plot.HCPC(res.TxCHCPC, choice = "map", draw.tree = FALSE,
                label = "ind", new.plot = TRUE)
        }
    }
    class(res.TxCHCPC) = c("TxCHCPC","HCPC")
    return(res.TxCHCPC)

## End(Not run)
```

TxMFACT *Multiple Factor Analysis Contingency Tables for Textual Data (TxMFACT)*

Description

Multiple Factor Analysis in the sense of Escofier and Pages with supplementary documents and supplementary groups of variables. Groups of variables can be quantitative, categorical or contingency tables. Missing values in numeric variables are replaced by the column mean. Missing values in categorical variables are treated as an additional level.

Usage

```
TxMFACT(MDocWord, group, type = rep("s", length(group)), col.sup = NULL, ind.sup = NULL,
 ncp = 5, name.group = NULL, num.group.sup = NULL, graph = TRUE, weight.col.mfa = NULL,
 row.w = NULL, axes = c(1, 2), tab.comp = NULL)
```

Arguments

MDocWord	data frame with I rows (documents) and J columns (variables)
group	vector with the number of columns in each group
type	type of group: "c" or "s" for quantitative variables ("s" variables are scaled to unit variance), "n" for categorical variables and "f" for frequency (columns tables); by default, all variables are quantitative and scaled to unit variance
ind.sup	vector indicating the indexes of supplementary documents
col.sup	vector indicating the indexes or names of supplementary columns
ncp	number of dimensions stored in the results (by default 5)

name.group	vector with the names of the groups (by default, NULL and the group are named group.1, group.2 and so on)
num.group.sup	the indexes of the illustrative groups (by default, NULL and no group are illustrative)
graph	boolean, if TRUE graphs are displayed
weight.col.mfa	vector of weights, useful for HMFA method (by default, NULL and a MFA is performed)
row.w	an optional row weights (by default, a vector of 1 for uniform row weights)
axes	a length 2 vector specifying the dimensions to plot
tab.comp	object obtained from the imputeMFA function of the missMDA package that allows to handle missing values

Value

Returns a list including:

summary.quali	summary of the results for the categorical variables
summary.quanti	summary of the results for the quantitative variables
separate.analyses	results for the separate analyses
eig	matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance
group	list of matrices containing all the results for the groups
rapport.inertie	inertia ratio
ind	list of matrices containing all the results for the active documents
ind.sup	list of matrices containing all the results for the supplementary documents
quanti.var	list of matrices containing all the results for the quantitative variables
freq	list of matrices containing all the results for the frequencies
quanti.var.sup	list of matrices containing all the results for the supplementary quantitative variables
freq.sup	list of matrices containing all the results for the supplementary frequencies
partial.axes	list of matrices containing all the results for the partial axes
global.pca	result of the analysis when it is considered as a unique weighted PCA

Author(s)

Belchin Kostov <badriyan@clinic.ub.es>, Daria M. Hernandez

References

Escofier, B., Pages, J. (1990). Analyses factorielles simples et multiples: objectifs, methodes, interpretation. Dunod, Paris

Becue-Bertaut, M.,Pages, J. (2004). A principal axes method for comparing multiple contingency tables: MFCT *Computational Statistics and Data Analysis*, 45, 481-503.

Becue-Bertaut, M., Pages, J. (2008). Multiple factor analisis and clustering of a mixture of quantitative, categorical and frequency data. *Computational statistics and Data Analysis*, 52, 3255-3268

See Also

summary.TxMFACT

Examples

```
## Not run:
data(dataOpen.question)
res.M1<-DocWordTable(dataOpen.question,num.text=c(6,7))
res.M2<-DocWordTable(dataOpen.question,num.text=8)
DocVar<-DocVarTable(dataOpen.question,VarSel=c(1,2,3,4))
MDocTerm=list(res.M1$DocTerm,res.M1$DocTerm)
MDocVar=list(DocVar,DocVar)
MTable<-MDocWordTable(MDocTerm,MDocVar,num.agg=3,idiom=c("en","en"),
Fmin=c(10,5),Dmin=c(2,1))
res.mfact<-TxMFACT(MTable$MDocWord, group=MTable$ncolTs, type =c("f","f"),
name.group=c("Important","Culture"))
summary(res.mfact)

### CODE TxMFACT

TxMFACT <-function(MDocWord, group, type = rep("s",length(group)),
        col.sup=NULL, ind.sup = NULL,  ncp = 5, name.group = NULL,
        num.group.sup = NULL, graph = TRUE, weight.col.mfa = NULL,
        row.w = NULL,axes = c(1,2),tab.comp=NULL)
 {
    if (!is.null(col.sup)){
         if(is.character(col.sup))
           col.sup<-which(colnames(MDocWord)
          if(is.numeric(col.sup))
     col.sup<-col.sup
       MSup<-MDocWord[,col.sup]
       MDocWordR<-MDocWord[,-col.sup]
       MDocWord<-cbind.data.frame(MDocWordR,MSup)
       }


    if (!is.null(ind.sup)){
         if(is.character(ind.sup))
           ind.sup<-which(rownames(MDocWord)
          if(is.numeric(ind.sup))
     ind.sup<-ind.sup
       }

  mfact<- MFA (MDocWord, group, type, ind.sup, ncp, name.group, num.group.sup,
              graph, weight.col.mfa, row.w,axes,tab.comp)
 if(graph){
      dev.new()
sel1<-which( (mfact$freq$contrib[,1]> 2*mean(mfact$freq$contrib[,1])) |
                   (mfact$freq$contrib[,2]>2*mean(mfact$freq$contrib[,2])) )

      par(cex=0.7)
plot.MFA(mfact,choix="freq",invisible="row",select=sel1,axes,
    unselect=1,palette=palette(c("black","black","black")),
            col.hab=c("green","blue"),title="Words" )

if (!is.null(col.sup)){
```

```
            selSup<-colnames(MSup)
            sel2<-which(rownames(mfact$freq.sup$cos2)
     sel<-c(sel1,sel2)
       dev.new()
par(cex=0.7)
plot.MFA(mfact,choix="freq",invisible="row",select=sel,axes,
unselect=1,palette=palette(c("black","black","black")),
          col.hab=c("green","blue"),title="Words")
}
dev.new()
      plot.MFA(mfact,choix="ind",axes, title="Documents")
}
  res<-mfact
 class(res)<-c("TxMFACT","MFA","list")
  return(res)
}


## End(Not run)
```

---

uCutDoc *Cut the sentences in homogeneous group (uCutDoc)*

Description

Group sentences or documents into lexically homogeneous groups through chronological clustering.

Usage

```
uCutDoc(base, num.text, idiom = "en", N = 1000, alfa = 0.05)
```

Arguments

base	data frame, or a vector with a least one textual variable
num.text	column index or name of the textual variable
idiom	language of the textual variable
N	number of permutation tests (by default 1000)
alfa	significance level (by default 0.05)

Value

SentGroup	data frame with I rows and 2 columns (groups and sentence)
GroupWord	aggregate table documents by words
Num.groups	number of groups or parts were formed
GrpComposition	composition of groups were formed

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

References

Becue-Bertaut, M., Kostov, B., Morin, A., & Naro, G. (2014). Rhetorical Strategy in Forensic Speeches: Multidimensional Statistics-Based Methodology. *Journal of Classification*, 31. doi:10.1007/s00357-014-9148-9.

See Also

uSentences

Examples

```
## Not run:
##uCutDoc
data(dataSpeech)
res<-DocWordTable(dataSpeech,num.text=1)
DocTerm<-res$DocTerm
Grps<-uCutDoc(speech,DocTerm, num.text=1, N=1000,alfa=0.05)

### CODE uCutDoc

uCutDoc <-function(base, num.text, idiom="en", N=1000,alfa=0.05){

if (!is.null(num.text)) {
 if (is.character(num.text))
 num.text<- which(colnames(base)
  if (is.numeric(num.text))
  num.text<- num.text
 if(length(num.text)==1)
  num.text<-num.text
 if(length(num.text)>1){
   for(i in 1:length(num.text)){
    if(i==1)
    text1<-base[,num.text[1]]
    else text1<-paste(text1,base[,num.text[i]],sep=".")
   }
    base[,(ncol(base)+1)]<-text1
    num.text<-ncol(base)
  }
}

Doc<-DocWordTable(base, num.text, idiom)
DocTermR<-Doc$DocTerm
res.ca<-CA(DocTermR[apply(DocTermR,1,sum)>0,], graph=FALSE)

Homo.groups<-uHomo.groups(res.ca$row$coord,N,alfa)
    GroupComposition<- Homo.groups$groups
  nb_groups<-length(Homo.groups$groups)
bloque<-vector("list",nb_groups)
for(i in 1:nb_groups){
bloque[[i]]<- rep(i,length(Homo.groups$groups[[i]]))

}
Group<-as.vector(unlist(bloque))
Group<-as.factor(Group)
Relim<-as.numeric(rownames(DocTermR[apply(DocTermR,1,sum)<=0,]))
```

```
if(length(Relim)==0)
SentGroup<-cbind.data.frame(Group,base[,num.text])
else SentGroup<-cbind.data.frame(Group,base[-Relim,num.text])
   colnames(SentGroup)<-c("Group", "Sentence")
  PW<-matrix(nrow=length(Homo.groups$groups),ncol=ncol(DocTermR))
for(i in 1:nb_groups){
       elem<-vector()
      elem<-(Homo.groups$groups[[i]])
if (length(elem)==1) {
                 PW[i,]<-DocTermR[Homo.groups$groups[[i]],]
                   } else {
       PW[i,]<-apply(DocTermR[Homo.groups$groups[[i]],],2,sum)
}
        }

rownames(PW)<-paste("P",1:nb_groups,sep="")
colnames(PW)<-colnames(DocTermR)
res = list(GroupWords=PW,SentGroup=SentGroup, GrpComposition=GroupComposition,
        Num.groups=nb_groups)
      return(res)
  }


## End(Not run)
```

---

uHomo.groups *Homogeneous groups (uHomo.groups)*

Description

Grouping parts into lexically homogeneous groups through chronological clustering.

Usage

```
uHomo.groups(res, N =1000,alfa=0.05)
```

Arguments

res	factor analysis results or coordinates of the factor analysis (by default, all dimensions are considered)
N	number of permutation tests (by default 1000)
alfa	significance level (by default 0.05)

Value

groups	list with content of groups or parts
NumHomogGroups	number of groups or parts were formed

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>, Belchin Kostov

References

Becue-Bertaut, M., Kostov, B., Morin, A., & Naro, G. (2014). Rhetorical Strategy in Forensic Speeches: Multidimensional Statistics-Based Methodology. *Journal of Classification*, 31. doi:10.1007/s00357-014-9148-9.

Legendre, P., Legendre, L. (1998). Numerical Ecology. (A. E. Science., Ed.) (2nd ed.).

Examples

```
## Not run:
data(dataBiblio)
DocTerm<-DocWordTable(dataBiblio,num.tex=2,lminword=3)
DocVar<-DocVarTable(dataBiblio,VarSel=c(3,4,5))
##Direct analysis
res.Dir<-TxCA(DocTerm, DocVar,Fmin=20,Dmin=10,stop.word.tm=TRUE,graph=FALSE)
res.ca<-res.TxCA$res.ca
res.Gps<-uHomo.groups(res.ca, N =1000,alfa=0.05)
res.Gps

### CODE uHomo.groups
uHomo.groups <-function(res, N =1000,alfa=0.05){
   CoR<-res
        if((is.data.frame(CoR))|(is.matrix(CoR))){
            Rcoor<-CoR
         }else{
if(!is.null(res$row$coord))
Rcoor<-res$row$coord
if(!is.null(res$ind$coord))
Rcoor<-res$ind$coord
       }

     X<- Rcoor

### Similarity Matrix
d<-dist(X)
d0<-as.matrix(d)
maxd<-max(d)
maxd<-maxd+1e-10
Sim<-as.matrix(maxd-d)
Sim0<-Sim
d<-as.matrix(d)

### Constrained Matrix
Cont<-matrix(nrow=nrow(Sim),ncol=ncol(Sim),0)
Cont[1,2]<-1
for (i in 2:(nrow(Cont)-1)){
Cont[i,i+1]<-1
Cont[i,i-1]<-1
}
Cont[nrow(Cont),nrow(Cont)-1]<-1
rownames(Cont)<-rownames(Sim)
colnames(Cont)<-colnames(Sim)

### Similarity matrix used for constrained clustering
SimCont<-Sim*Cont
DistCont<-d*Cont
```

```
groups<-list()
for (i in 1:nrow(Sim)){
groups[[i]]<-i
}

distclust<-numeric()
clust<-list()
i<-1

### Clustering
while(sum(SimCont!=0)>0){

### Find the position of the maxim similarity
maxsim<-max(SimCont)
posmaxsim<-which(SimCont==maxsim)

if (posmaxsim[1]
fila<-posmaxsim[1]
col<-nrow(SimCont)
}else{
fila<-posmaxsim[1]
col<-posmaxsim[1]
}

maxfc<-max(fila,col)
minfc<-min(fila,col)

distclust[i]<-DistCont[fila,col]

### Permutation test
perm<-numeric()
if (length(groups[[minfc]])==1&&length(groups[[maxfc]])==1){
pvalid<-TRUE
}else{
##library(gdata)
maux<-d0[c(groups[[minfc]],groups[[maxfc]]),c(groups[[minfc]],groups[[maxfc]])]
dmed<-median(upperTriangle(maux))
mdist<-d0>dmed
uns<-sum(mdist[groups[[minfc]],groups[[maxfc]]])
ncomb<-factorial(length(groups[[minfc]])+length(groups[[maxfc]]))/
                        (factorial(length(groups[[minfc]]))*factorial(length(groups[[maxfc]])))
if (ncomb>N|ncomb==Inf|is.na(ncomb)){
for (j in 1:N){
elemperm<-sample(c(groups[[minfc]],groups[[maxfc]]),
                                        length(c(groups[[minfc]],groups[[maxfc]])))
perm[j]<-sum(mdist[elemperm[1:length(groups[[minfc]])],
                                        elemperm[(length(groups[[minfc]])+1):length(elemperm)]])
}
if (sum(perm>=uns)<(N*alfa)) pvalid<-FALSE
else pvalid<-TRUE
}else{
     combin1<-combn(c(groups[[minfc]],groups[[maxfc]]),
                                  length(groups[[minfc]]),simplify=FALSE)
combin2<-combn(c(groups[[minfc]],groups[[maxfc]]),
                                  length(groups[[maxfc]]),simplify=FALSE)

for (j in 1:length(combin1)){
```

```
perm[j]<-sum(mdist[combin1[[j]],combin2[[length(combin2)-j+1]]])
}

if (sum(perm>=uns)<(length(combin1)*alfa)) pvalid<-FALSE
else pvalid<-TRUE
}
}

### Join two clusters if permutation test is ok

if (pvalid){
clust[[i]]<-vector(mode="list",length=2)
clust[[i]][[1]]<-rownames(Sim)[minfc]
clust[[i]][[2]]<-rownames(Sim)[maxfc]
rownames(Sim)[minfc]<-colnames(Sim)[minfc]<-rownames(d)[minfc]<-colnames(d)[minfc]
                        <-rownames(Cont)[minfc]<-colnames(Cont)[minfc]<-paste(rownames(Sim)[minfc],"-",
                        rownames(Sim)[maxfc])

if (minfc!=1){
Sim[minfc,minfc-1]<-Sim[minfc-1,minfc]<-0.5*Sim[minfc,minfc-1]+
                                0.5*Sim[maxfc,minfc-1]-0.5*abs(Sim[minfc,minfc-1]-Sim[maxfc,minfc-1])
d[minfc,minfc-1]<-d[minfc-1,minfc]<-0.5*d[minfc,minfc-1]+
                                     0.5*d[maxfc,minfc-1]+0.5*abs(d[minfc,minfc-1]-d[maxfc,minfc-1])
}
if (maxfc!=nrow(SimCont)){
Sim[maxfc+1,minfc]<-Sim[minfc,maxfc+1]<-0.5*Sim[minfc,maxfc+1]+
                                0.5*Sim[maxfc,maxfc+1]-0.5*abs(Sim[minfc,maxfc+1]-Sim[maxfc,maxfc+1])
d[maxfc+1,minfc]<-d[minfc,maxfc+1]<-0.5*d[minfc,maxfc+1]+
                                0.5*d[maxfc,maxfc+1]+0.5*abs(d[minfc,maxfc+1]-d[maxfc,maxfc+1])
Cont[maxfc+1,minfc]<-Cont[minfc,maxfc+1]<-Cont[maxfc,maxfc+1]
}
groups[[minfc]]<-c(groups[[minfc]],groups[[maxfc]])
groups<-groups[-maxfc]
Sim<-Sim[-maxfc,-maxfc]
d<-d[-maxfc,-maxfc]
Cont<-Cont[-maxfc,-maxfc]
i<-i+1

### If permutation test is not ok, change Constrained matrix

}else{
Cont[maxfc,minfc]<-Cont[minfc,maxfc]<-0
}

SimCont<-Sim*Cont
DistCont<-d*Cont
}
return (res=list(NumHomogGroups=length(groups),groups=groups))

## End(Not run)
```

---

uSentences *Arbitrary sentences (uSentences)*

Description

Divides the corpus in arbitrary sentences

Usage

```
uSentences(base, num.text, SentLength)
```

Arguments

base	data frame with a least one textual variable
num.text	column index or name of the textual variable
SentLength	length of the arbitrary sentence

Value

Sentences	data frame with I rows (number of sentence and phrases) and one columns

Author(s)

Daria M. Hernandez <daria.micaela.hernandez@upc.edu>

See Also

uCutDoc

Examples

```
## Not run:
data(dataSpeech)
sentences<-uSentences(dataSpeech, num.text=1,SentLength=50)

### CODE uSentences

uSentences <-
function(base, num.text, SentLength){
  if (!is.null(num.text)) {
 if (is.character(num.text))
 num.text<- which(colnames(base)
  if (is.numeric(num.text))
  num.text<- num.text
 if(length(num.text)==1)
  num.text<-num.text
 if(length(num.text)>1){
   for(i in 1:length(num.text)){
    if(i==1)
    text1<-base[,num.text[1]]
    else text1<-paste(text1,base[,num.text[i]],sep=".")
   }
    base[,(ncol(base)+1)]<-text1
    num.text<-ncol(base)
  }
}

     r<-length(rownames(base))
p1 <- unlist(strsplit(as.character(base[1:r,num.text]),split=" "))
```

```
sel <- which(p1=="")
if (length(sel)==0){
          p1<-tolower(p1)
}
 if (length(sel)!=0){
     p1 <- p1[-sel]
p1 <- tolower(p1)
}
       Tfrase<-SentLength
listagrupos<-list()
ngrupos<-length(p1)
        for(i in 1:ngrupos){
if (i!=ngrupos){
listagrupos[[i]]<-p1[((i-1)*Tfrase+1):(i*Tfrase)]
       }else{
listagrupos[[i]]<-p1[((i-1)*Tfrase+1):length(p1)]
listagrupos[[i]]

}
}

listita<-listagrupos
  for (i in 1:length(listita)){
   names(listita[[i]])<-c(1:length(listita[[i]]))
  }
numeracio<-paste(c(1:length(p1[((i-1)*Tfrase+1):length(p1)])))
listita2 <- lapply(listita, FUN=function(x) x[numeracio])
Frase<- do.call("cbind", listita2)
row.names(Frase) <-numeracio
Fr<-t(Frase)
Frse<-vector("list",nrow(Fr))
for(i in 1:nrow(Fr)){
for(j in 1:length(numeracio)){
 if (!is.na(Fr[i,j]))Frse[[i]]<-paste(Frse[[i]],Fr[i,j])
}
}

for (i in 1:length(Frse)){
names(Frse[[i]])<-c(1:length(Frse[[i]]))
}
numeracio<-paste(c(1:1))
listita<- lapply(Frse, FUN=function(x) x[numeracio])
Frse<- do.call("cbind", listita)
row.names(Frse) <-"Sentence"
Frse<-as.data.frame(t(Frse))
return (Sentences=Frse)
}


## End(Not run)
```

VocIndex *Index of Vocabulary (VocIndex)*

Description

Computing the regular and specialized vocabulary

Usage

```
VocIndex(base, num.text, Fmin=5, sep.punctuation=TRUE, separ=NULL)
```

Arguments

base	data frame with a least one textual variable
num.text	column index or name of the textual variable
Fmin	minimum threshold on the word frequency (by default 5)
sep.punctuation	boolean, if TRUE delete the punctuation
separ	removes vector of characters used as word separation (punctuation marks, symbols, etc.)

Value

VocIndex	data frame with n rows and 3 columns (words, frequencies and index)
RegVoc	data frame with the regular vocabulary
LocalVoc	data frame with the local vocabulary

Author(s)

Daria M Hernandez <daria.micaela.hernandez@upc.edu>, Nuria Planel

References

Hubert, P. (1988). A model of vocabulary partition *Literary and Linguistic Computing*, 3(4), 223-225

See Also

summary.VocIndex, print.VocIndex

Examples

```
## Not run:
data(open.question)
res.VcIndex<-VocIndex(open.question,num.tex=c(6,7),Fmin=5)
print(res.VcIndex)
summary(res.VcIndex)

### CODE VocIndex

VocIndex<-function(base, num.text, Fmin=5, sep.punctuation=TRUE, separ=NULL){

if (!is.null(num.text)) {
 if (is.character(num.text))
 num.text<- which(colnames(base)
  if (is.numeric(num.text))
  num.text<- num.text
```

```
 if(length(num.text)==1)
  num.text<-num.text
 if(length(num.text)>1){
   for(i in 1:length(num.text)){
    if(i==1)
    text1<-base[,num.text[1]]
    else text1<-paste(text1,base[,num.text[i]],sep=".")
   }
    base[,(ncol(base)+1)]<-text1
    num.text<-ncol(base)
  }
}

    r<-length(rownames(base))
     base[,num.text]<-str_replace_all( base[,num.text],"[]", " ")
     p1 <- unlist(strsplit(as.character(base[1:r,num.text]),split=" "))
    if(sep.punctuation){
                filt="([?\n<U+202F><U+2009>]|[[:punct:]]|[[:space:]]|[[:cntrl:]])+"
                if (!is.null(separ)) filt=paste(separ,filt,sep="|")
      p2<-str_replace_all(p1,filt, "")
                    p3<-str_replace_all(p2,"[,;.:0123456789*]", "")
                   p1<-p3
}
sel <- which(p1=="")
if (length(sel)==0){
          p1<-tolower(p1)
}
 if (length(sel)!=0){
     p1 <- p1[-sel]
pOK <- tolower(p1)
}
Lpal <- length(pOK)
uLpal <- unique(pOK)

indices <- vector()
freq<-vector()
Posit<-vector("list",length(uLpal))
Distancias<-vector("list",length(uLpal))
      Exponencial<-vector()

   for(i in 1:length(uLpal)){
  pos <- which(pOK==uLpal[i])
F <- length(pos)
  freq[i]<-F
T <- Lpal/F
posIni <- pos
posEnd<-c(pos[2:length(pos)],Lpal+pos[1])
D <- posEnd-posIni
k <- which(D>T)
Distancias[[i]]<-D

if(length(k)==0){
  Nprim <- Lpal
  }

  if(length(k)!=0){
  Di <- D[k]
```

```
   Nprim <- Lpal - sum(Di-T)
  }

  R <- (Nprim-F)/(Lpal-F)
  indices[i] <- round(R,2)
Posit[[i]]<-pos

}
IndVocab<-cbind.data.frame(uLpal, indices, freq)
IndVocab<-subset(IndVocab, IndVocab$freq>=Fmin)
colnames(IndVocab)<-c("Word","Index","Freq")
IndComp<-cbind.data.frame(uLpal, indices, freq)
IndVocab1<-subset(IndComp, IndComp$freq>=Fmin)
colnames(IndVocab1)<-c("Word","Index","Freq")
IndVocab<-with(IndVocab,IndVocab[order(Word),])
Voc.Regular<-with(IndVocab1, subset(IndVocab1, Index >=0.60))
Voc.Regular<-with(Voc.Regular,Voc.Regular[order(Index, decreasing=TRUE),])
Voc.Local<-with(IndVocab1, subset(IndVocab1, Index <= 0.40))
Voc.Local<-with(Voc.Local,Voc.Local[order(Index,Freq),])

res<-vector(mode=list)
 res$VocIndex<-IndVocab
    res$RegVoc<-Voc.Regular
 res$LocalVoc<-Voc.Local
         class(res)<-c("VocIndex","list")
return(res)

}



## End(Not run)
```