UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Máster en Análisis Avanzado de Datos Multivariantes y Big Data

Trabajo Fin de Máster

Minería de Texto con software libre: DtmVic & IRaMuTeQ

Estudio comparativo aplicado al cambio climático

Autores

Manuel Souto Juan

Ricardo Augusto del Cid Mancio

Tutora

María Purificación Galindo Villardón

2020

Resumen

Este trabajo de fin de máster presenta una comparación entre dos programas informáticos para el análisis o minería de texto: DtmVic e IRaMuTeQ. Surge la necesidad del mismo dado que las ciencias sociales se apoyan cada vez más en el uso de ordenadores para investigación cualitativa. Existen en el mercado muchas opciones de software para tal efecto: de pago y gratuitas, de código abierto y cerrado, con diferentes capacidades de análisis léxico y gráfico, etc.; se han escogido los programas mencionados al ser gratuitos y de código abierto.

Este trabajo hace un análisis de las herramientas seleccionadas y sus capacidades, así como un estudio comparativo de las mismas por medio de su aplicación a una misma base de datos construida a partir de noticias publicadas en medios digitales sobre un tema de actualidad: el cambio climático. Ambas herramientas emplean técnicas estadísticas multivariantes descriptivas e incluyen métodos de tratamiento de datos léxicos mencionados en la literatura. Finalmente, se destacan las limitaciones de ambos programas y se muestran alternativas posibles a los mismos que permitan en el futuro un análisis estadístico mas potente.

PALABRAS CLAVE: *text mining*, grupos léxicos, *corpus* de texto, DtmVic, IRaMuTeQ.

Abstract

This Master's Thesis makes a comparison between two computer programs for text analysis or mining: DtmVic and IRaMuTeQ. The need for it arises as the social sciences rely more and more on the use of computers for qualitative research. There are many software options on the market for this purpose: paid and free, open and closed source, and with different lexical and graphic analysis capabilities, etc.; both softwares have been chosen since they are free and open source.

This work analyzes the selected tools and their capabilities, as well as a comparative study of them, through their application to the same database built from news published in digital media on a current issue: climate change. Both tools use various descriptive multivariate statistical techniques and include lexical data treatment methods mentioned in the literature. Finally, the limitations of both software are highlighted and possible alternative techniques are suggested that allow for a more powerful statistical analysis in the future.

KEYWORDS: text mining, lexical groups, text corpus, DtmVic, IRaMuTeQ.

Índice general

Índice de figuras

Introducción

El análisis o minería de textos (*text mining*) es campo novedoso, que intenta extraer información significativa de textos en lenguaje natural. Las bases de datos tradicionales almacenan datos que son estructurados, a partir de los cuales es fácil extraer información significativa por medio de lenguajes de programación diseñados para tal efecto. Sin embargo, en la civilización moderna, el modo más común y formal de intercambiar información es por medio de la palabra escrita (texto), de modo que una gran parte de la información disponible se encuentra almacenada en ordenadores en forma de texto no estructurado. Surge entonces la necesidad de extraer información de forma automatizada, aunque sea de forma parcial, de textos cuya extensión sea demasiado larga para ser procesada en forma rápida y eficiente por un investigador humano.

El término 'minería de texto' generalmente se refiere cualquier sistema informático que es capaz de analizar grandes cuerpos de texto en lenguaje natural y detectar grupos o patrones léxicos, intentando extraer alguna información útil del mismo. Es importante en este punto señalar las diferencias entre 'minería de texto' y otros términos ampliamente usados en el campo del Big Data, como 'minería de datos' (*data mining*) y 'procesamiento de lenguaje natural' (*natural language processing*).

La minería de datos puede ser definida como la extracción de información implícita (no obvia a simple vista por lo menos), desconocida y probablemente útil de los datos mediante el uso de técnicas estadísticas. En el caso de la minería de texto, en cambio, la información no está implícita ni oculta (un autor suele esforzarse en utilizar un lenguaje fácilmente entendible por sus lectores), el problema en este caso es que la información esta organizada de una manera que no es procesable por medio de Software. Cuando se emplean técnicas de minería de datos se espera obtener un resultado 'útil'; por ejemplo, hacer predicciones estadísticas sobre nuevos datos provenientes de la misma fuente (este es el campo del *Machine Learning* o Aprendizaje de Máquina). Sin embargo, en el caso de la minería de texto la utilidad puede tener diversas interpretaciones, como se verá más adelante.

El procesamiento del lenguaje natural es un campo de la inteligencia artificial que busca mecanismos eficientes para la interacción entre computadoras y humanos por medio del lenguaje natural, lo cual puede implicar profundizar en aspectos cognitivos del lenguaje humano. La minería de textos, en cambio, trata de aplicar técnicas estadísticas a cuerpos de textos, para extraer alguna información útil o interpretable por el investigador.

Otra diferencia importante entre la minería de datos y la minería de textos es que el resultado de la primera suele ser de naturaleza cuantitativa, mientras que los resultados de la segunda suelen ser cualitativos, si bien estos resultados cualitativos se obtienen solo después de haber aplicado técnicas estadísticas.

Muchos son los tipos de datos que es posible analizar por medio de las técnicas de minería de texto: textos literarios, cuestionarios, noticias, transcripciones de entrevistas, etc. En el trabajo desarrollado se presenta un análisis de noticias publicadas en diarios digitales, dado que las mismas desempeñan un papel importante en la construcción del imaginario colectivo, creando símbolos que sirven como representación de la realidad en la mente de los lectores.

En la actualidad existen muchos programas dedicados al análisis de textos, entre los cuales se ha decidido estudiar dos en particular: DtmVic (*Data and Text Mining: Visualization, Inference, Classification*) e IRaMuTeQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*); en parte porque son gratuitos y se basan en los principios del software libre: libre uso, distribución, copia y modificación, y también porque ambos ofrecen análisis estadísticos y gráficos de textos.

El presente trabajo emplea los conocimientos adquiridos en el Máster de Análisis Avanzado de Datos Multivariantes y Big Data de la Universidad de Salamanca, y su estructura es la siguiente:

- El primer capítulo estará centrado en la descripción del análisis léxico de textos, mostrando al lector la terminología empleada y el uso de los distintos métodos que permiten trasladar estos datos a un formato susceptible de ser analizado por técnicas estadísticas más conocidas. Su propósito es poner en contexto al lector en la terminología del análisis textual.

- El segundo capítulo expondrá el funcionamiento y las diversas técnicas estadísticas presentes en ambos programas. Este capítulo trata de mostrar una base teórica de estas técnicas que permita comprender los análisis realizados por los programas e interpretar sus resultados.

- En el tercer capítulo se presenta una comparación de los resultados obtenidos por ambos programas aplicados a la base de datos de noticias sobre el cambio climático recopilada, mostrando las ventajas y limitaciones de ambos programas en este aspecto.

Finalmente se concluye con una comparación de ambos programas y se indican ciertas recomendaciones sobre ambos que permitan extender el marco de análisis que contienen estos programas.

Un apunte importante sobre el desarrollo y la estructura de este trabajo debe comentarse en este punto. Una parte principal en este trabajo ha sido el conocimiento del uso de los programas y el desarrollo de un manual traducido que emplea la terminología de los autores, cuyo fin era proporcionar al lector un guión a seguir para el desarrollo de futuros trabajos empleando estos programas. Por motivos de límites de hojas no puede incluirse directamente en este trabajo,

pero se indica a continuación un código QR donde los interesados pueden acceder a la versión completa de este trabajo que incluye estos manuales. También se incluyen en este trabajo las referencias bibliográficas empleadas en el desarrollo de estos capítulos adicionales.

Objetivos

El objetivo general de este trabajo es la presentación de los programas libres DtmVic e IRaMuTeQ como una herramienta en el contexto del Análisis de Datos Textuales. Se hará uso de una base de datos para ejemplificar el uso de los programas y destacar las diversas posibilidades que permiten.

Entre los objetivos específicos perseguidos, se destacan:

- Explorar el estado actual de las tecnologías de análisis textual.
- Comparar las capacidades de ambos en el Análisis de Datos Textuales.
- Sugerir variaciones de las técnicas presentes que, en desarrollos futuros, permitan mejorar las conclusiones obtenidas en este ámbito.

Metodología

Para hacer la comparación de capacidades de los programas escogidos, se ha escogido analizar una base de datos creada recopilando noticias, porque se piensa que esta es la aplicación que puede interesar a mayor número de usuarios. Entre los muchos temas de actualidad se ha escogido el cambio climático porque este tendrá el mayor impacto a largo plazo en el desarrollo de la historia humana, limitando el alcance a las publicadas en España a finales de 2019 y la primera mitad del año 2020, dado que es en este momento en que la pandemia del COVID-19 ha recibido la mayor parte de la atención pública, dejando otros problemas urgentes relegados. Se han recopilado las noticias publicadas por los diarios digitales con mayor número de lectores en España, que de acuerdo con la Asociación para Los Medios de Comunicación (`http://reporting.aimc.es/index.html#/main/diarios`) y como se muestra en la figura 1 son El País, El Mundo y La Vanguardia; se han escogido dichos periódicos dado que los periódicos Marca y El As son diarios de carácter deportivo y no centrados en noticias de índole general. La creación de dicha base de datos se mostrará en el Capítulo 3 aunque no de forma extensa dado que el objetivo es comparar las capacidades de análisis de los dos programas y no alcanzar conclusiones definitivas sobre el fenómeno del cambio climático.

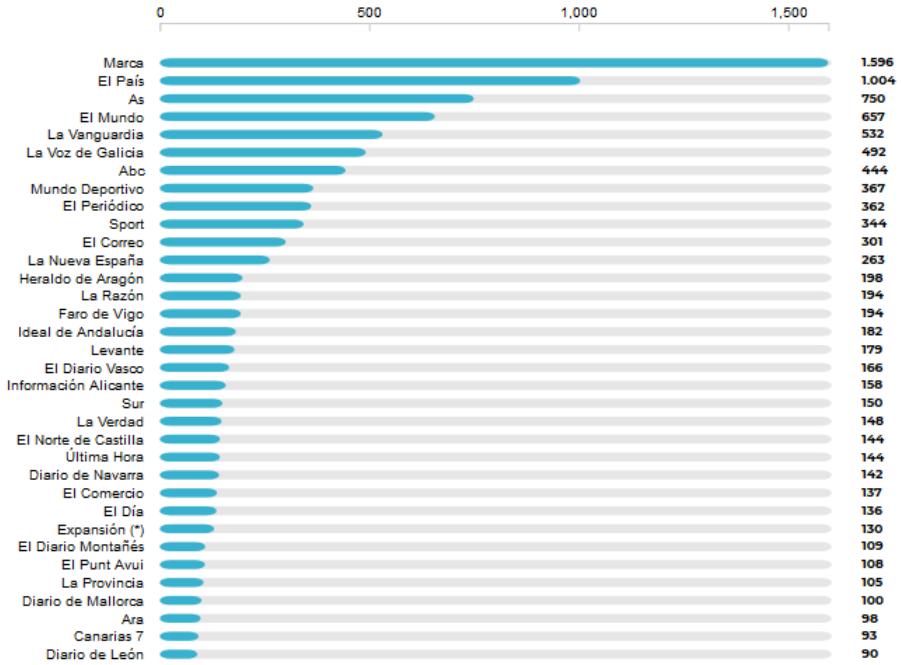


Figura 1: Medios digitales ordenados por número de lectores diarios, de acuerdo a la AIMC.

Glosario

ACM	Análisis de Correspondencias Múltiple. Véase Capítulo 2.
ACP	Análisis de Componentes Principales.
AFC	Análisis Factorial de Correspondencias. Véase Capítulo 2.
Análisis automático	El software hace la mayor parte de los análisis con mínima intervención humana.
Análisis manual	El investigador tiene control total sobre las tareas.
Análisis semiautomático	Requiere más intervención humana que el automático.
Corpus	Conjunto textos que sirven de base a la investigación. Archivo de texto que sirve de entrada al programa de análisis léxico.
Estadística Textual	Véase Textometría.
Forma	Unidad léxica. Palabra con la que trabaja el programa en caso de optar por lematizar el *corpus*.
Hápax	Formas cuya frecuencia de aparición en el texto es igual a uno.
Lematizar	Reducir las palabras principales a sus raíces, derivando así 'morfemas léxicos'.
Lenguaje natural	Lenguaje humano o hablado.
Lexicometría	Véase Textometría.
Logometría	Véase Textometría.
Minería de texto	Rama específica de la minería de datos dedicada al análisis y la derivación de nueva información de los textos.
Morfema léxico	Raíz de la palabra que remite a su familia de procedencia.
SOM	Mapa auto-organizado de Kohonen. Véase Capítulo 2.
Textometría	Forma actual de la lexicometría. Área que propone procedimientos de ordenamiento y de cálculos estadísticos para el estudio de un *corpus* de texto digitalizado.
Unidad léxica	Todo elemento situado entre dos espacios de la cadena escrita, o entre un espacio y un signo de puntuación. Palabra o forma.
UCE	Unidad de Contexto Elemental. Véase Capítulo 1.

Capítulo 1

Análisis Estadístico de Datos Textuales

En este capítulo se presentarán la terminología y metodología empleada en el análisis de datos de tipo textual en la literatura [1]. El objetivo de este apartado es poner en contexto al lector que no hayan profundizado en el tema y servir como recordatorio para aquellos más especializados, ofreciendo un marco común de expresión.

Dado que en este trabajo se presentan dos programas desarrollados por autores diferentes, a pesar de que en análisis realizado por ambos sea similar, los términos y metodología empleados en ambos difieren y pueden confundir al lector. Por ello, en este apartado se distinguirá por una parte la teoría de la Estadística Textual de mano de Lebart y, por otro lado, los mundos léxicos de Reinert y la Teoría de las Representaciones Sociales de Moscovici.

1.1. Descripción y metodología del análisis del *corpus*. Unidades estadísticas textuales de Lebart

En la actualidad, un gran número de estudios están fundamentados en datos textuales. Por ello, es importante para los investigadores encontrar métodos que les permitan analizar y clasificar la información contenida en estos textos y relacionarla con sus autores, como encuestados o personas entrevistadas, por ejemplo. Por tanto, la unidad fundamental en este campo es el estudio del texto obtenido en el estudio; en este campo recibe el nombre de *corpus*. La índole del *corpus* puede ser variada en función del campo estudiado: textos científicos, literarios, periodísticos, etc; un ejemplo podrían ser las noticias presentes en los periódicos digitales españoles en relación a la crisis del COVID-19.

Las principales técnicas de análisis en el campo del Análisis de Datos Textuales se deben a la escuela francesa de Estadística, principalmente de la mano de Benzécri, quien fue padre de diversas técnicas en el campo de la Estadística Multivariante; destaca entre ellas el Análisis de Correspondencias [2]. Lebart & Salem (1994) continúan el desarrollo en este campo en busca de una forma de automatización de los procesos dado que, en aquel momento, aún se abogaba por una post-codificación manual de las respuestas a preguntas abiertas que se realizaba. En este ámbito, Lebart desarrolla el programa DtmVic (2012) como una interfaz gráfica de su metodología con fines académicos.

Posteriormente, Reinert propuso el método informatizado Alceste (Análisis Lexical de Co-ocurrencias en Enunciados Simples de un Texto - empleado en IRaMuTeQ). En el análisis automático, el software hace la mayor parte de los análisis con un mínimo de intervención humana (es necesario dedicar un tiempo a la preparación de los archivos de entrada y a la interpretación de los resultados). Estos hacen análisis estadísticos textuales (buscan patrones y correlaciones entre las formas o palabras de un *corpus* de texto), y fueron desarrollados como respuesta a las problemáticas de las ciencias sociales y la investigación documental: analizar grandes cuerpos de texto procedentes de discursos, respuestas a entrevistas, encuestas y preguntas abiertas.

En este apartado del trabajo también cabe mencionar el análisis semiautomático, que obviamente implica menor intervención por parte del investigador que en el análisis manual, pero más que en el automático. Una forma de análisis semiautomático es donde el usuario hace una categorización de las palabras, tomando como referencia por ejemplo un diccionario, creando relaciones entre temas de interés, antes de emplear alguna herramienta de software para su análisis. Este trabajo se centrará en los dos programas mencionados para análisis automático.

Dada la complejidad léxica del *corpus* al completo, un análisis estadístico directo no es posible. Por ello, se definen una serie de conceptos y unos métodos de lematización que se introducirán a continuación, y que permitirán simplificar la información relevante en el texto y realizar un análisis estadístico sobre el *corpus* simplificado. En primer lugar, el *corpus* se segmenta por palabras o combinaciones de palabras que se consideran indivisibles a efectos del análisis; estas combinaciones deben ser elegidas *a priori* por el investigador en función de su criterio. Estas unidades textuales reciben el nombre de *formas gráficas*. La definición de un conjunto de palabras como una única forma gráfica permite definir como unidad indivisible a un conjunto de palabras; por ejemplo, sería el caso de palabras compuestas como "Banco Central Europeo".

Las formas gráficas se entienden como una serie de caracteres entre dos delimitadores en sus extremos, compuestas por caracteres en el alfabeto del idioma que corresponde al *corpus* concreto. En la Estadística Textual propuesta por Lebart & Salem (1994), esta forma gráfica es la unidad estadística básica en el análisis y recibe el nombre de *palabra*. El lector debe tener en cuenta que, siguiendo el ejemplo anterior, 'Banco Central Europeo' sería una única palabra en este contexto. En este contexto se habla de *ocurrencias* como el número de presencias de palabras en el *corpus*, *tamaño del corpus* como el número de ocurrencias y *vocabulario* como el número de palabras distintas presentes en el *corpus*. Estos términos proporcionan una idea sobre el tamaño del texto y de la riqueza léxica del mismo.

Otra unidad básica del lenguaje en este contexto se denomina *lema*, y se entiende como la raíz de la palabra en su entrada al diccionario. Esta definición permite clasificar las palabras del *corpus* en función de su tipo: adjetivos, verbos, sustantivos, etc. Se comentará en mayor profundidad esta definición en el apartado sobre la lematización de los textos.

Por último, podemos estudiar la estructura del *corpus* en relación a las respuestas a preguntas abiertas. Este tipo de preguntas se refieren a las preguntas con una respuesta libre, que buscan la espontaneidad en la respuesta del entrevistado. La clasificación de estos *corpus* se realiza en ese caso mediante las respuestas a preguntas cerradas, o preguntas con respuestas fijadas de antemano. De esta forma, la estructura del *corpus* dependería del caso concreto de cada estudio, donde cada unidad considerada correspondería a un texto diferente. En el ejemplo anterior, se podría considerar el *corpus* como un conjunto de textos, cada uno correspondiente a una noticia en un periódico.

1.1.1. Tablas léxicas y pre-procesado de textos. Lematización

El análisis estadístico a realizar en este caso debe estar basado en las ocurrencias de las palabras en el *corpus* analizado. Por tanto, dado que se trata de una serie de frecuencias, parece lógico que la metodología pase por emplear unas tablas de contingencia específicas. En este ámbito, dichas tablas reciben el nombre de *tablas léxicas*. Estas tablas se construyen a partir de las formas gráficas del texto, definiendo una *variable léxica* cuyos valores posibles serán las formas gráficas del *corpus* correspondiente. Por tanto, las tablas léxicas muestran las frecuencias de uso de cada palabra por cada individuo en el *corpus*, y su objetivo es comparar los perfiles léxicos de las respuestas. En el caso de disponer de una segmentación del *corpus* en varios textos, para cada uno de ellos se puede construir una tabla con las frecuencias de sus palabras de forma similar; en este caso reciben el nombre de *tablas léxicas agregadas*. Su objetivo es comparar los perfiles léxicos de los textos formados en la segmentación por la variable categórica correspondiente.

El análisis estadístico estará basado, por tanto, en un análisis sobre estas tablas de contingencia. Cabe destacar que se tratará de tablas *sparse* o dispersas, en que sólo habrá unos pocos datos en las celdas de la tabla. Este hecho se tiene en cuenta en los diversos programas disponibles que exploran esta metodología, incluyendo el caso de DtmVic. Al tratarse de una metodología basada en tablas de contingencia y desarrollado por un autor de la escuela francesa de Estadística, es lógico imaginar que el procedimiento comparativo propuesto será basado en la comparación de perfiles léxicos de palabras e individuos; por tanto, la literatura procede mediante una aplicación del método de Análisis de Correspondencias. Esta información estará complementada por la inclusión de segmentaciones por variables categóricas, que

modelarán las respuestas a preguntas cerradas en los textos y permitirán una riqueza mayor en las conclusiones obtenidas.

En relación al contenido propiamente dicho de los textos, se cuenta con información dispersa y sujeta a muchas condiciones: palabras homónimas, contexto de la respuesta, etc.; problemas que no son fácilmente solucionables de forma automática y requieren la intervención humana. Por otro lado, una lematización es un proceso estándar en el área de lingüística computacional, que permite reducir el número de palabras poco frecuentes y con poca información y agrupar las palabras por clases similares, permitiendo un análisis más sencillo de los textos. Por tanto, debe tenerse en cuenta que este tipo de programas debe siempre contar con una supervisión humana en la definición y posterior interpretación de los resultados.

Existen diversos procedimientos algorítmicos para realizar un adecuado pre-procesamiento del *corpus* estudiado. Uno de ellos es el estudio de concordancias, donde se listan todos los contextos de una misma palabra. De esta forma, el análisis tendría en cuenta la existencia de palabras homógrafas, o palabras con la misma grafía pero distinto significado; por tanto, el programa evitaría problemas de clasificación de las palabras en cuanto al contexto ofrecido.

Otro estudio previo necesario es ls reducción de palabras poco frecuentes, ya que si recordamos estamos tratando con un Análisis de Correspondecias y éste es susceptible de variaciones elevadas por la aparición de frecuencias pequeñas en la tabla de contingencia, como consecuencia de los pesos obtenidos en la distancia χ^2. Por ello, estos métodos definen un umbral para las frecuencias de las palabras y buscan aumentar las frecuencias del resto de palabras, por ejemplo, sumando todas las palabras sinónimas en el *corpus* en una sola. También se eliminan las denominadas *palabras herramienta*: palabras como artículos, preposiciones, etc. cuya frecuencia es elevada y no aportan información.

Por último, se realiza una lematización del texto. Este procedimiento pasa por sustituir las palabras en el *corpus* por su lema correspondiente. Los lemas de cada tipo gramatical se definen como su raíz, por lo que serán:

- Sustantivos en singular.
- Adjetivos en singular masculino.
- Verbos en infinitivo.

Esta lematización permite reducir el vocabulario del *corpus* de una manera apreciable, agrupando las palabras correspondientes a lemas idénticos. Es un proceso estándar y se encuentra programado en los software dedicados a este tema.

1.2. El Método Alceste y los mundos léxicos

Los programas de análisis textual que nos ocupan fueron concebidos con fines pragmáticos, pero tienen fundamentos teóricos que los sustentan. Reinert [7] propone que todo discurso se compone de lo que llama 'mundos léxicos', que dan coherencia y racionalidad al mismo. Un mundo léxico está formado por las palabras que se repiten en el discurso, independientemente de su construcción sintáctica. Por ejemplo, en el siguiente fragmento de Don Quijote de la Mancha:

> *La libertad, Sancho, es uno de los más preciosos dones que a los hombres dieron los cielos; con ella no pueden igualarse los tesoros que encierra la tierra ni el mar encubre; por la libertad, así como por la honra se puede y debe aventurar la vida, y, por el contrario, el cautiverio es el mayor mal que puede venir a los hombres.*

Si se toman las palabras *libertad*, *Sancho*, *dones*, *cielos*, *tesoro*, *tierra*, *mar*, *vida*, etc. (nótese que se han eliminado los adverbios, artículos y preposiciones) estas evocan una concepción muy romántica de la vida por parte del locutor (Don Quijote), y el simple hecho de conocerlas nos ayuda a formar un juicio sobre el personaje y su carácter, o sobre el carácter y contenido de cualquier discurso en general. El objetivo del método de clasificación propuesto por Reinert (Alceste) es el de poner en evidencia los mundos léxicos ocultos en el discurso bajo estudio. Esto se hace encontrando que formas o palabras en el *corpus* son redundantes o co-ocurrentes, sin tomar en cuenta la sintaxis del discurso. Es importante señalar que Reinert hace la distinción (aceptada en lingüística), entre 'palabras principales' y 'palabras relacionales', siendo las primeras sustantivos, verbos y adjetivos que indican el contenido de un discurso, y las segundas conjunciones, artículos y preposiciones que crean las relaciones entre las palabras principales.

El método Alceste parte de la suposición de que de que el análisis de palabras principales en un discurso permitirá diferenciar los 'lugares de enunciación' o mundos léxicos del mismo. Observando la frecuencia con que aparecen las palabras principales puede encontrarse qué mundos léxicos son mas citados que otros, o bien qué mundos léxicos existen en 'oposición' a otros; es decir, un mundo léxico no se define de forma aislada, sino que existe en relación a otros presentes en el discurso.

Con el fin de descubrir la estructura de co-ocurrencias en las palabras de un discurso, se parte del supuesto de que el mismo esta formado por enunciados simples o elementales. Las palabras son trasladadas a una tabla binaria, que organiza en filas los enunciados simples y en columnas el vocabulario o palabras principales para poner en evidencia al discurso como un conjunto de enunciados y a los enunciados como un conjunto de palabras principales.

1.2.1. Unidades de Contexto Elementales y morfemas léxicos

Dado que es difícil definir o delimitar el concepto de 'enunciado simple', Reinert en su momento definió las UCE's o *unidades de contexto elementales* como los segmentos de texto compuestos de palabras principales, cuyo tamaño puede ser definido por el investigador (número de palabras) o bien por signos de puntuación. Un enunciado es entonces reemplazado por una UCE para poder tratarlo estadísticamente a nivel de software, y las palabras del texto son reemplazadas por 'formas simples', que son las antes mencionadas palabras principales y relacionales.

El análisis se hace entonces a partir de las palabras principales, que son a continuación lematizadas (reducidas a sus raíces), creando así los llamados 'morfemas léxicos'. La lematización elimina la variabilidad en las formas de una palabra, conservando únicamente el significado fundamental. Esta reducción en necesaria para luego hacer la clasificación jerárquica descendente, similar al análisis de *cluster*.

1.3. La Teoría de las Representaciones Sociales

Como la teoría de las representaciones sociales pertenece al campo de la sociología y no es el propósito de este trabajo poner a prueba teorías de esa área, se hará una breve descripción de la misma dado que el Software IRaMuTeQ fue concebido basándose en dicha teoría, y se harán algunos comentarios sobre las representaciones sociales al final del trabajo.

Sergei Moscovici propuso la teoría en Francia en la década de 1960, en el contexto de un estudio hecho que llevó a la publicación de su libro 'El psicoanálisis, su imagen y su público', en el cual contiene sus hallazgos sobre lo que pensaba el publico francés sobre la teoría del psicoanálisis.

Según Moscovici [10], la representación social corresponde a un acto del pensamiento en el cual el individuo se relaciona con un objeto el cual luego es sustituido por un símbolo en la mente del mismo. El objeto entonces existe de manera simbólica en su mente, pero no simplemente como una copia del objeto real porque, en opinión de Jodelet [9]; la representación social implica transformación o construcción de los símbolos, ya que en el proceso de representación, la interpretación de la realidad hecha por la mente del sujeto esta afectada por sus valores, religión, necesidades, posición en la jerarquía social, y otros aspectos económicos y culturales. Es decir, cuando un individuo interpreta la realidad, no copia, sino que transforma y construye, agregando contenido original de su mente; este comportamiento está asociado a su lenguaje y cultura. Esto nos lleva a otro punto importante, al hecho de que las representaciones sociales no existen únicamente en las mentes de los individuos, sino que pasan de estas a formar parte de la cultura. En consecuencia, las ideas que se intercambian en las comunicaciones

personales y en los medios de comunicación (noticias) afectan los modos de pensar de la población. Las representaciones sociales no son simples construcciones simbólicas estáticas, sino que evolucionan influenciadas por las personas y el continuo intercambio de ideas entre las mismas, y su función principal es la de dar sentido a la realidad, de transformar lo desconocido en conocido.

León [11], da a las representaciones sociales un enfoque estructuralista, afirmando que algunas de sus funciones son:

- Convertir objetos, personas y eventos desconocidos en cosas convencionales y conocidas, es decir, convertir realidades extrañas en realidades familiares.
- Facilitar la comunicación entre los individuos.
- Tejer el pensamiento colectivo, que lleva a la creación de la identidad social (conocimiento del grupo al que uno pertenece).
- Justificar las conductas sociales, sean estas positivas o negativas.

Es por esto, por su carácter compartido, que la teoría se ha llamado de representaciones *sociales* y no 'individuales'. Retomando el tema del impacto de los medios de comunicación en la creación de representaciones sociales, estos fomentan la renovación continua de conocimientos de la población, por lo cual, a la vez que contribuyen a crear las representaciones sociales, evitan que las mismas cristalicen y pasen a formar parte permanente de la cultura, como es el caso de los mitos. En el ámbito del análisis de noticias, (como un objetivo secundario de este trabajo) se intentará buscar, por medio del programa, cuáles son las representaciones sociales existentes respecto al 'cambio climático'.

Capítulo 2

Métodos estadísticos presentes

En este apartado del trabajo se presentarán los distintos métodos estadísticos presentes en los programas DtmVic e IRaMuTeQ de forma teórica. Su objetivo es mostrar al lector un marco de referencia de las técnicas empleadas por estos programas, tanto para la descripción de técnicas no conocidas como por su traducción en el lenguaje de la escuela francesa a la forma de describir las técnicas usualmente empleado en la escuela de la Universidad de Salamanca. Se procederá a la descripción de los métodos presentes, entre los que destacamos: el Análisis de Correspondencias y el Análisis de Correspondencias Múltiple; el Análisis de *Clusters*, incluyendo *Kohonen maps*, *minimum spanning tree*, el algoritmo *chain search* de Lebart y la Clasificación Jerárquica Descendente de Reinert; y el proceso de seriación de una tabla de datos. Se ha decidido no incluir un apartado sobre el Análisis de Componentes Principales ya que, a pesar de ser la base de muchos de los métodos aquí mencionados, no se emplea explícitamente como técnica en los programas de análisis de datos textuales aquí estudiados.

2.1. Análisis Factorial de Correspondencias

([5], [12]) Se trata de la técnica principal empleada en el ámbito del Análisis de Datos Textuales. Su definición se debe a Benzécri (1973), quien la desarrolla como una técnica de análisis de tablas de contingencia, o tablas de clasificación de individuos mediante dos variables categóricas, mediante la representación simultánea de sus filas y columnas en un espacio de dimensión reducida; de esta última parte recibe el nombre de factorial. Por tanto, podría describirse como una técnica multivariante que estudia relaciones de dependencia entre variables categóricas a través del análisis de tablas de contingencia (Albert GIFI).

En este sentido, la técnica se emplea cuando existe una asociación significativa entre las variables categóricas estudiadas. Cuando esto ocurre, su objetivo es representar las filas y columnas de la tabla de contingencia en dos espacios vectoriales reducidos, para posteriormente superponerlos y obtener la representación conjunta de ambos. La figura 2.1 muestra un esquema de esta metodología.

En el caso presente en este trabajo, un ejemplo de punto de partida correspondería a una tabla léxica de dimensión $\mathbf{nxp}$ con filas y columnas correspondientes a palabras y textos del *corpus*, respectivamente. En ese caso, la asociación existente entre diversas palabras y textos se entiende en términos de **distancias** en el espacio vectorial correspondiente. Para definir estas distancias de forma que los puntos fila y columna tengan distancias comparables, es preciso

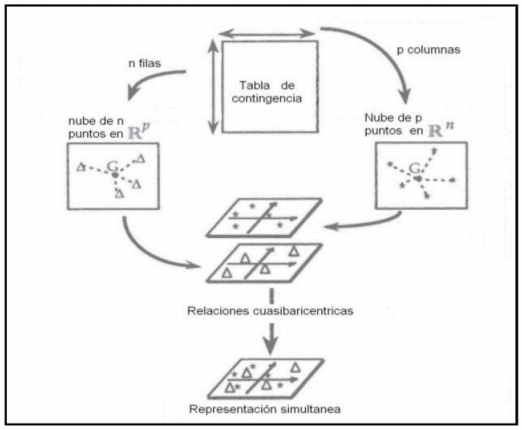


Figura 2.1: [13] Esquema simplificado del Análisis de Correspondencias.

introducir unos conceptos que dan sentido a esta comparación; todos ellos estarán basados en los totales marginales de filas y columnas y en el tamaño de muestra. Estos conceptos son:

- Perfil: distribución de frecuencias relativas de la fila o columna en relación a su total marginal. Por tanto, se trata de una relativización de cada fila o columna en función de su total marginal correspondiente. De esta forma, si f_{ij} representa el valor cruzado de la fila i y la columna j, el perfil fila correspondiente a la fila i-ésima sería la fila

$$\frac{f_{ij}}{f_{i\cdot}},\ \forall\ j = 1 \dots p$$

- Masa: término de ponderación de cada perfil fila o columna respecto al total muestral. Por tanto, la masa del perfil fila i-ésimo sería el valor

$$\frac{f_{i\cdot}}{f_{\cdot\cdot}}$$

- Distancia χ^2: distancia entre perfiles fila o columna en cada espacio vectorial. Se define en cada caso como

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2$$

  Esta distancia se trata de una distancia euclídea ponderada, que permite la comparación de perfiles fila y columna posterior que se pretende realizar.

- Inercia: medida de dispersión de los perfiles en el espacio multidimensional; se calcula como la suma de la masa de cada punto por su distancia al centroide de la nube elevada al cuadrado.

El cálculo de los ejes principales y las coordenadas de los perfiles correspondientes implica una diagonalización y obtención de valores propios en su determinación. Estos valores propios indican la inercia de cada dimensión, definida como se indica anteriormente. Esto implica que es de utilidad una representación de dichos valores propios para cada componente principal en un gráfico denominado *scree plot*. Este gráfico permite, de una forma visual, comprender la información recogida por cada eje principal; por tanto, es un elemento esencial en la determinación del número de ejes retenidos en el análisis. De esta forma, diferencias apreciables entre valores propios contiguos indican ejes separados y, por tanto, susceptibilidad a ser interpretados correctamente.

A raíz de este análisis, surgen dos términos de gran interés que permiten profundizar en la interpretación de los resultados obtenidos. El primero de ellos se conoce como la *contribución de la categoría al factor* (CREF), tratándose de la coordenada correspondiente a la proyección de un perfil sobre un eje principal concreto. Este término indica qué parte de la varianza del eje viene explicada por dicho perfil, por lo que permite detectar las categorías que intervienen de una manera superior en la definición de cada eje factorial.

El otro término es similar y se denomina la *contribución del eje al factor* (CAEF) y se obtiene dividiendo la inercia correspondiente al punto por el total de la fila o columna correspondiente. Por tanto, este término indica la contribución del eje en la definición de cada categoría, indicando sobre qué ejes está mejor definida.

Otro factor a tener en cuenta en este análisis es la *calidad de representación*, entendida como el grado de asociación del perfil con los ejes del plano. Se calcula como el coseno al cuadrado del ángulo que forma dicho punto en el plano con el eje correspondiente. Una calidad de representación mayor indica que la información que contiene el punto está mejor representada en los ejes, por lo que los puntos con calidades bajas pueden contener su información en otros ejes no considerados.

Un hecho importante a tener en cuenta, más acusado en el caso de tablas léxicas que se estudia en este trabajo, es la representación de los puntos mediante la distancia χ^2. Esta distancia está ponderada por los inversos de los marginales fila o columna correspondientes, por lo que perfiles bajos implican un mayor peso; este hecho es problemático ya que, por la definición de inercia, implicará que dichos puntos se representarán en torno al origen de coordenadas. Estos puntos harán que el resto de puntos, mejor representados por los ejes, se desplacen para acomodar esta inercia y puede llevar a resultados o asociaciones erróneas. Por ello, en el caso del Análisis de Datos Textuales, los diversos autores aconsejan definir un

umbral mínimo de ocurrencias, de forma que palabras o lemas poco repetidos en los textos no se tengan en cuenta para el análisis. Estos elementos pueden introducirse como categorías suplementarias, siendo representadas en el plano factorial sin haber participado activamente en su definición.

2.2. Análisis de Correspondencias Múltiple

([14], [15]) Esta técnica surge como una extensión evidente del Análisis de Correspondencias en el caso de tablas de contingencia multidimensionales. Por tanto, este método permite el análisis de varias variables categóricas con una o más categorías.

En el marco del Análisis de Datos Textuales, esta técnica comprende datos en que las categorías de cada variable son mutuamente excluyentes; en ese sentido, los valores de las variables pueden codificarse de forma binaria en presencia o ausencia. Un ejemplo de este caso sería una encuesta y preguntas con varias respuestas posibles; en este caso se codificaría cada pregunta como una variable categórica y cada respuesta posible como una categoría y, suponiendo que sólo se admite una respuesta, podría codificarse como se ha indicado anteriormente.

Este método parte de una matriz $\mathbf{Z}$ de dimensiones $\mathbf{nxp}$ donde n indica el número de individuos y p el número de categorías totales de todas las variables categóricas; en el caso del ejemplo anterior, p sería el número de respuestas posibles del cuestionario al completo. Esta matriz se construye como la yuxtaposición de las submatrices $\mathbf{nxp_i}$ con p_i el número de categorías correspondientes a la variable categórica i-ésima. Cabe destacar que, por su propia construcción, la matriz $\mathbf{Z}$ será una matriz *sparse*, hecho a tener en cuenta en la elaboración de algoritmos de resolución del método. A partir de esta matriz se define la *matriz de contingencia de Burt* como $B = Z'Z$, que por su construcción estará agrupada en bloques en función de la tabulación cruzada resultante:

- Los bloques diagonales, correspondientes al producto de las submatrices $Z_i'Z_i$, toman la forma de matrices diagonales y representan la frecuencia de respuesta a cada categoría de la variable.
- Los bloques no diagonales, correspondientes al producto de las submatrices $Z_i'Z_j$, representan la relación entre las preguntas i-ésima y j-ésima.

En la figura 2.2 se muestra esta estructura en un ejemplo concreto. La tabla de *Burt* puede interpretarse de forma análoga a la matriz de covarianzas en el caso de variables continuas. El desarrollo del método comienza con el estudio para dos preguntas, donde se demuestra la equivalencia entre un AFC de las matrices $\mathbf{Z}$, $\mathbf{B}$ y la matriz cruzada $\mathbf{Z_i'Z_j}$, válido también en el caso de más de dos preguntas. Este resultado permite, por tanto, generalizar al caso de más

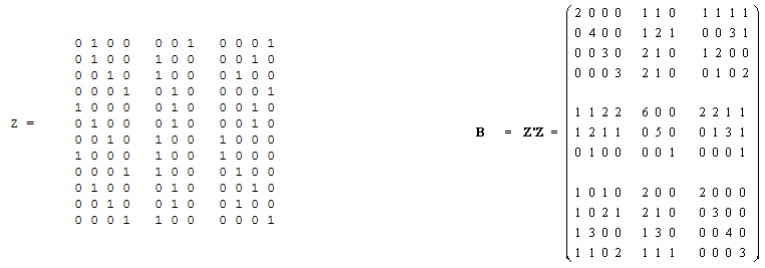


Figura 2.2: [15] Tabla de codificación (izquierda) y tabla de *Burt* asociada (derecha).

preguntas mediante el estudio de las posiciones relativas de los dos subespacios vectoriales generados por dos variables realizado para todas las combinaciones posibles de dos variables. Una presentación extensa de los resultados y el procedimiento del método puede consultarse en [15].

Esta técnica tiene varias propiedades heredadas del AFC simple que muestran la utilidad del método. Una de estas propiedades es que los puntos de las categorías correspondientes a una variable tienen el mismo centro de gravedad que el resto de categorías, por lo que todas las variables comparten el centro de gravedad.

Otra propiedad interesante surge de la mano de las categorías suplementarias: al igual que en el caso del AFC, es posible proyectar elementos que no participan activamente en la definición de los ejes factoriales sobre ellos, de forma que estos elementos se describan sin alterar la posición de los ejes. En este caso, los autores pretenden contrastar la hipótesis de que las observaciones correspondientes a la categoría suplementaria son aleatorias (sin reemplazamiento). La motivación de definir la hipótesis nula de esta forma es que, al tratarse de variables con codificación binaria, la posición de las categorías suplementarias en un eje concreto es su media aritmética propiamente escalada. A partir de dicha hipótesis, los autores presentan un test que denominan *test-value* o *t-value*. El estadístico se comporta como una variable t de *Student* estandarizada; un resultado significativo indicaría que la categoría suplementaria tiene una ordenación significativa respecto a un eje, por lo que indicaría una representación elevada en él. En el caso del programa DtmVic los valores de corte para la significación son mayores de 2 y menores de -2, lo que indica confianzas de aproximadamente un 95 %.

2.3. Análisis de *Clusters*

([16], [17]) El Análisis de *Clusters* o Análisis de Conglomerados es una de las técnicas principales en el área de la Estadística Multivariante. Su objetivo es clasificar las variables correspondientes en unos conjuntos, tratando de maximizar la homogeneidad presente en el interior de los grupos y maximizando la heterogeneidad entre los grupos; de esta forma, se

dispondrá de grupos homogéneos y separados entre sí a la mayor distancia posible. En general, esta técnica no se emplea en solitario, sino que se utiliza como un método de clasificación sobre los resultados obtenidos mediante un análisis previo; de esta forma, los resultados obtenidos proporcionan más información relevante al investigador. La ventaja del Análisis de *Clusters* es que los grupos o *cluster* definidos por el método sólo dependen de la estructura interna de los datos y no requieren especificar explícitamente una organización previa de los mismos.

El método de creación de los *clusters* está basado en el concepto de *disimilaridad* o distancia entre individuos. Esta medida indicará la relación que existe entre dos individuos en función de los valores que toman en las variables consideradas, y permitirán su clasificación en los *cluster* definidos. También pueden emplearse ciertos coeficientes de *similaridad*, donde en este caso valores elevados indicarán relaciones fuertes entre los individuos, aunque este caso es menos empleado. Las principales distancias empleadas en este ámbito son:

- **Distancia de Minkowski**:

$$D_{ij} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^n \right)^{1/n}$$

  En el caso $k = 1$ esta distancia recibe el nombre de distancia Manhattan o distancia ciudad, mientras que en el caso $k = 2$ se reduce a la distancia euclídea usual.

- **Distancia euclídea estandarizada**:

$$D_{ij}^2 = \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{S_k}$$

  donde S_k indica la desviación típica de la variable k-ésima.

- **Distancia de Mahalanobis**:

$$D_{ij} = (\mathbf{x}_i - \mathbf{x}_j)' S^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

  donde S representa la matriz de covarianzas dentro del grupo. La ventaja de estas dos últimas distancias es que son invariables frente a cambios de escala.

Los algoritmos de *clustering* se pueden clasificar en dos grandes tipos: jerárquicos y no jerárquicos. A continuación se expondrán las características principales de cada uno de ellos y diversos métodos de clasificación de cada tipo, aunque en el trabajo se centrará la explicación en describir los métodos planteados en el programa DtmVic. Existen diversas consideraciones en el uso de este tipo de técnicas, que se indican a continuación:

- Es conveniente realizar una estandarización de las variables, dado que estos algoritmos son sensibles a diferencias en las escalas de medición.

- Los algoritmos son sensibles a la presencia de *outliers* o valores atípicos. Este hecho puede corregirse empleando modificaciones a los métodos *cluster* estándar.

- Es recomendable comprobar los resultados por diferentes métodos, de forma que se estime si la estructura interna de los datos es pobre o no. En última instancia la validez de los *cluster* está sujeta a una interpretación del investigador, por lo que puede inducir en subjetividad.

2.3.1. Algoritmos de clasificación no jerárquicos. Método *k-means*.

Los métodos no jerárquicos requieren la imposición previa de un número de *clusters* a definir. Esto se debe a que este tipo de técnicas definen previamente una serie de centroides iniciales, de forma aleatoria, y asignan los valores al *cluster* definido por su centroide más cercano. Tras clasificarlos, se produce una reasignación de los individuos respecto a una regla de parada específica para cada algoritmo. Por tanto, las diferencias entre este tipo de algoritmos se basan en la asignación de los centroides iniciales y la definición de esta regla o criterio.

El método *k-means* (Forgy, 1965) es el principal método de clasificación no jerárquico empleado, del cual surgen la mayoría de los algoritmos refinados actuales. Está basado en el mismo principio enunciado anteriormente, tomando como criterio de reasignación de los individuos un cálculo de los centroides tras la clasificación. De esta forma, tras terminar el paso de clasificación de los individuos por su distancia al centroide, éste se recalcula como el centro de gravedad asociado a los puntos del *cluster*, y se repite el paso de clasificación. Este algoritmo termina cuando dos iteraciones devuelven la misma partición de individuos o se cumple un criterio de parada especificado. Una variación de este método es la propuesta por MacQueen (1967), que propone tomar los centroides iniciales aleatoriamente entre los individuos, en lugar de ser puntos totalmente aleatorios. La desventaja de esta técnica es que presupone que los grupos pueden ser separados y con formas esféricas, aunque existen métodos alternativos que exploran este problema. Debido a su rápida ejecución y que el resultado obtenido depende en gran parte de los centroides iniciales, este método se suele realizar varias veces para observar el cambio en los resultados.

2.3.2. Algoritmos de clasificación jerárquicos. Método de Ward.

Los métodos de clasificación jerárquicos son algoritmos de clasificación en los que no es necesario imponer un número determinado de *clusters* inicial. A su vez, estos métodos pueden ser de dos tipos en función del tipo de agrupación realizada: aglomerativos, que crean los *clusters* introduciendo sucesivos datos, o divisivos, que crean un *cluster* de inicio con todos los datos y particionan en sucesivos menores. Este tipo de métodos se emplean con carácter

exploratorio de los datos, ya que no es necesario conocer su estructura interna a priori para su clasificación. Las clasificaciones de estos métodos suelen presentarse en forma de un diagrama de clasificación denominado **dendrograma**; este gráfico en forma de árbol presenta los diversos objetos de estudio como hojas y su tronco representa las clasificaciones de los individuos. A medida que se sube por el tronco, el número de clases disminuye al agrupar los individuos; por tanto, el investigador define un punto de corte que determina el número de *clusters* tomados finalmente.

Los principales métodos de clasificación jerárquicos son aglomerativos, sus diferencias se basan en la definición de la distancia entre un individuo y un grupo o entre grupos de individuos. Entre este tipo de algoritmos destacan, principalmente: *single linkage* o vecino más próximo, *complete linkage* o vecino más lejano, *group average* o grupo promedio, el método del centroide y el método de Ward. Dado que este trabajo se centra en el estudio del programa DtmVic y éste emplea el algoritmo jerárquico de Ward y *single linkage*, se describirán con detalle estos método a continuación.

El método de los vecinos más próximos (Florek et al., 1951) es uno de los primeros métodos de clasificación que aparece en la literatura y se trata de uno de los más sencillos. Está basado en la definición de distancia entre *clusters* como la distancia entre los dos puntos más próximos entre ambos. Se trata de un método sencillo de interpretar y programar, pero que no está preparado para el trabajo con grandes bases de datos y presenta inconvenientes similares al resto de métodos. El motivo de introducir este sencillo algoritmo es debido a que, como se mostrará posteriormente, es equivalente al método conocido como *minimum spanning tree*. Una desventaja de este método es que tiende a producir *clusters* demasiado grandes.

El método de Ward (1963) es un método jerárquico aglomerativo cuya forma de proceder es distinta a la empleada por los mencionados anteriormente. Si denotamos los n puntos a clasificar como $\mathbf{x}$, vectores de p componentes, con una masa asociada m_i, su varianza total se define como:

$$I = \sum_{i=1}^{n} m_i ||x_i - g||^2$$

donde g indica el centro de gravedad de los puntos. En ese caso, al descomponer los puntos en q *clusters*, se obtiene una separación de dicha varianza total en dos partes, un término entre grupos y otro intra-grupo, respectivamente:

$$I = \sum_{q} m_q ||g_q - g||^2 + \sum_{q} \sum_{i \in q} m_i ||x_i - g_q||^2$$

Se demuestra que el criterio de mínima separación entre dos elementos equivale a una minimización de la varianza relativa entre ambos; en términos generales, eso implica que el

método de Ward busca la mejor configuración minimizando la varianza intra-grupo y maximizando la varianza entre grupos. Una desventaja de este método es que, por su definición, favorece la creación de *clusters* esféricos aunque estos tengan una forma distinta.

El uso de ambos tipos de técnicas de clasificación debe ser complementario y no excluyente, ya que un empleo previo de un método jerárquico como el algoritmo de Ward permite obtener una estructura de clases previa, permitiendo fijar ese número de clases en un algoritmo no jerárquico, como *k-means*, y obtener una clasificación mejor que si se realizase únicamente un análisis.

2.3.3. Clasificación Jerárquica Descendente

La Clasificación Jerárquica Descendente (en francés *Classification Hiérarchique Descendante*, CHD en adelante), se basa en el método propuesto por Max Reinert (1983) [6], y es un algoritmo que trabaja de forma descendente: comienza con un grupo global en el que se incluyen todos los elementos y luego va haciendo divisiones sucesivas hasta terminar con tantos grupos como elementos (palabras o formas) haya en el *corpus* (también llamado algoritmo divisivo por ello, trabaja al revés que los algoritmos ascendentes, que forman grupos o *clusters* cada vez más grandes a partir de elementos individuales). De este modo se obtienen clases que ponen en evidencia mundos léxicos, los cuales están relacionados entre sí de acuerdo al sentido y a la proximidad del vocabulario incluido en ellos (Moreno y Ratinaud [4]).

Las clases están máximamente relacionadas en el interior, a la vez que mínimamente asociadas entre sí. En el caso que nos ocupa, cuando los elementos a clasificar son palabras o Unidades de Contexto Elementales (UCE's), estas son primero lematizadas (reducidas a su forma raíz, tal como se describe en el Capítulo 1), luego estas palabras lematizadas o formas son utilizadas para construir una matriz o tabla binaria de tipo *sparse* (muchos elementos de valor nulo), para proceder luego con la CHD que divide sucesivamente el *corpus* hasta formar clases.

La tabla binaria cruza todas las UCE's en columnas y todas las palabras en filas, de modo que un cero en una de sus celdas indica la presencia de una palabra en particular en el texto, y un uno el caso contrario (Figura 2.3). Luego se divide el contenido de esta tabla en dos clases, que maximizan la similitud de las UCE's dentro de la misma clase, a la vez que maximizan la diferencia entre las mismas (la CHD se basa en la suposición inicial de que el *corpus* de palabras a clasificar tiene una distancia que separa los objetos entre sí, y luego establece unas reglas para calcular dicha distancia).

El método busca dividir el *corpus* minimizando las palabras traslapadas. El traslape es medido mediante el valor χ^2 de una tabla con dos filas, comparando las distribuciones observadas de las palabras con las distribuciones esperadas. Si las palabras son diferentes entre dos clases,

Palabras		UCE's						
		1	2	3	4	5	...	j
	1	0	1	1	0	0	...	0
	2	1	1	0	1	0	...	1
	3	0	0	0	0	0	...	1
	4	0	0	0	1	0	...	0
	...	...	...	...	...	...	...	...
	i	0	1	1	0	0	...	1
	Totales	2	51	54	451	35	...	20

Figura 2.3: Ejemplo de tabla binaria de cruce entre palabras y UCE's.

la distribución observada diferirá de la distribución esperada. Luego la CHD maximiza el valor de χ^2 mediante divisiones sucesivas, proceso que finaliza cuando un número determinado de iteraciones ya no produce valores significativos estadísticamente o se alcanza un umbral previamente determinado. El procedimiento de la CHD se ilustra a continuación mediante un ejemplo sencillo.

Si definimos como I la filas de la matriz (UCE's), y como J las columnas (formas o palabras lematizadas), sea K_{ij} el valor de la intersección de la fila i-ésima y la columna j-ésima, y sea entonces la dimensión de la matriz IxJ. Se busca encontrar una partición (I_1, I_2) de I para dar lugar a dos clases diferenciadas, maximizando el valores χ^2 de la tabla de contingencia de las filas de la matriz IxJ (Figura 2.4).

Clases		j			
	1	I1	L1j	....	L1
	2	I2	L2j	....	L2
			Kj		

Figura 2.4: Tabla de contingencia tras realizar una partición en dos clases.

Aquí el valor L_{1J} corresponde a la intersección de I_1 con la forma j, es el número de UCE's de la clase I_1 en que está presente la forma j. Cada valor de la tabla se calcula como:

$$L_{1j} = \sum_{i \in I_1} K_{ij} \quad ; \quad L_1 = \sum_{j=1}^{J} L_{1j} \quad ; \quad K = L_1 + L_2$$

$$L_{2j} = \sum_{i \in I_2} K_{ij} \quad ; \quad L_2 = \sum_{j=1}^{J} L_{2j} \quad ; \quad K_j = L_{1j} + L_{2j}$$

El valor de χ^2 asociado se obtiene como:

$$\chi^2 = L_1 L_2 \sum_{j=1}^{J} \frac{\left(\frac{L_{1j}}{L_1} - \frac{L_{2j}}{L_2}\right)^2}{K_j}$$

y permite contrastar los perfiles de ambas clases. Encontrar la partición (I_1, I_2), que maximiza el valor de χ^2, es suficiente. La técnica empleada da una buena aproximación de este. Se procede del siguiente modo:

- Por medio del AFC, calcular el primer eje factorial de N(I), en el espacio $\mathbb{R}^J$, con la métrica χ^2.
- Se busca el hiperplano que separa la nube de puntos N(I) en dos subnubes (N(I_1) y N(I_2)), maximizando la inercia interclases. Este valor máximo resulta ser casi igual al valor χ^2 de la tabla de contingencia asociada.
- Dado que los centros de gravedad de las dos nubes de puntos no están exactamente situados sobre el primer eje factorial, la inercia interclases puede ser aumentada intercambiando puntos entre ambas nubes de forma iterativa, para aproximarse un poco más al χ^2 deseado. Las iteraciones consisten en comprobar si el cambio de clase de cada punto i aumenta o disminuye la inercia interclases, cambiándolo en caso positivo.

Una vez encontrada la partición se repite el mismo procedimiento, dividiendo por dos la más grande de las clases de UCE's resultantes. Se repite el mismo proceso por un número de iteraciones fijado por el programa.

Para decidir qué clases se considerarán en el resultado final, se hace una comparación entre dos clases, haciendo la comparación con unidades de contexto de diferente tamaño, es decir, construyendo dos matrices de datos para unidades de contexto de diferente tamaño. Puede fijarse el número de clases terminales, para tener un criterio de parada para el algoritmo.

La comparación de clases para determinar cuales tienen mayor estabilidad se hace en base a cuantas UCE's han sido incluidas en cada una. Para ello, se comparan las clases obtenidas en la primera clasificación con las obtenidas en la segunda, utilizando el valor χ^2 para relacionarlas. De este modo se obtienen las parejas de clases (I_L, I_H) en la que el valor χ^2 es máximo, lo cual indica que su grado de asociación es también máximo con respecto a cualquier otra pareja de clases encontrada en la jerarquía. La comparación entre dos clases, I_L e I_H, obtenidas en la primera y segunda clasificación se hace por medio de una tabla de contingencia (figura 2.5) donde los valores no indicados pueden calcularse por diferencias.

En dicha tabla, n_1 son las UCE incluidas en la clase I_L tras la primera clasificación, n_2 el número de UCE's incluidas en I_H tras la segunda clasificación, n_{12} las UCE's presentes al mismo tiempo en la clase I_L de la primera y la clase I_H de la segunda, y n el total de UCE's.

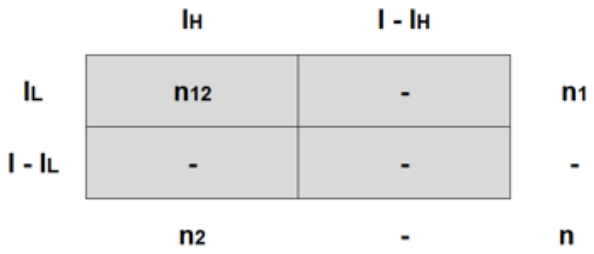


Figura 2.5: Calculo del valor χ^2 para la asociación entre clases.

El algoritmo compara las parejas de clases y selecciona aquellas para las cuales el valor χ^2 sea máximo. Se considera una clase como estable cuando esta contiene UCE's que están así mismo presentes simultáneamente en dos clases que forman una pareja de máxima asociación. Las clases estables se determinan simplemente eligiendo a las que hayan sido creadas en una misma partición, lo cual asegura que las unidades seleccionadas pertenezcan a una sola clase estable.

Una clase es definida por las formas reducidas más comunes presentes en la misma. Para encontrar las formas más comunes se calcula un coeficiente de asociación de forma a clase: χ^2 a partir de una tabla de contingencia construida para cada forma, cruzando la forma analizada en la UCE con la presencia de dicha UCE a la clase en cuestión. Por ejemplo, con la tabla de contingencia de la figura 2.6, se busca la asociación de la forma Fa con la clase I_L.

En esta tabla tenemos que:

$n_1 =$ número de UCE's contenidas en la clase.

$n_2 =$ número de UCE's en las que está presente la palabra o forma en cuestión

$n_{12} =$ número de UCE's de las clases donde se presenta la forma.

$n =$ número de UCE's clasificadas.

El valor n_{12} es comparado al valor teórico $n_1 n_2/n$ al calcular el valor de χ^2, y al resultado se suma el signo de la diferencia $n_{12} - (n_1 n_2/n)$ para determinar presencia o ausencia de la forma. Luego de esto, las formas que resultan específicas a una determinada clase, que son seleccionadas teniendo en cuenta los valores de χ^2 y si resultan con signo positivo en la diferencia anterior, pasan a formar el árbol de clasificación o dendrograma. La interpretación de este dendrograma se reduce, según Reinert, a las palabras con mayor proximidad semántica, dado que estas dan una idea global del contenido de cada una de las clases.

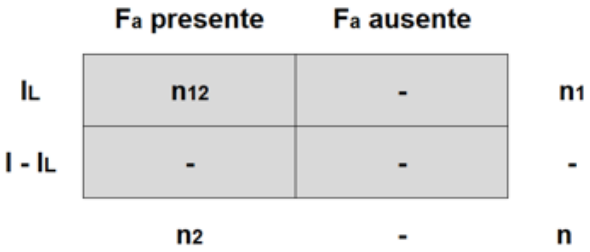


Figura 2.6: Cálculo del valor χ^2 de asociación entre una forma y una clase.

2.3.4. Algoritmo *chain search*

Este algoritmo se introduce en DtmVic como una alternativa a los métodos de clasificación jerárquicos usuales introducidos anteriormente. Dado que los nodos se construyen de uno en uno, el requerimiento computacional cuando se trabaja con bases de datos grandes es muy costoso, por lo que este algoritmo trata de reducir el número de cálculos necesarios para optimizar este problema.

Está basado en el concepto de *reciprocal neighbours* o vecinos recíprocos (McQuitty, 1966), en que dos puntos o grupos son *reciprocal neighbours* si uno es el vecino más próximo del otro y viceversa. Por tanto, el algoritmo crea, en cada paso, tantos nodos como vecinos recíprocos existen. Los autores indican que este algoritmo reduce el número de cálculos necesarios del orden de n^3 en los algoritmos anteriores al orden de n^2 en este caso, donde n indica el número de puntos a clasificar. La creación del algoritmo empleada en DtmVic, conocido como "*chain search*", se debe a Benzécri (1982). Los pasos que sigue el algoritmo se indican a continuación:

- Paso 1: se comienza en un elemento aleatorio x_1 de la muestra, y se forma la cadena

$$x_1 x_2 \dots$$

  de forma que x_i es el vecino más próximo de x_{i-1} $\forall i$. Esta cadena es necesariamente finita y termina cuando se cumple que x_{j-1} es el vecino más próximo de x_j; por tanto, ambos serían vecinos recíprocos y se agregan a un *cluster*.

- Paso 2: Si $j = 2$, y por tanto x_1 y x_2 son los vecinos recíprocos, se toma un nuevo elemento de inicio y se repite el paso 1, agregando los nodos en un nuevo *cluster*.

- Paso 3: Si $j > 2$, se extiende la cadena del paso 1 comenzando en x_{j-2}. El algoritmo termina tras $n - 1$ *clusters*.

Una consideración a tener en cuenta es que este algoritmo no haga desaparecer la relación de vecinos más próximos entre los pares anteriores de la cadena que no han sido agrupados;

siguiendo la notación empleada, entre los pares x_i $(i = 1 \dots j-2)$. Para asegurar este hecho, es necesario que no se produzca una inversión en la cadena, es decir, que el nodo creado al agregar dos elementos no esté más cerca de un tercer elemento de lo que estaban ambos individualmente de éste. Esta condición se presenta como:

$$d(a,b) < \text{ínf}\{d(a,c),\ d(b,c)\} \implies \text{ínf}\{d(a,c),\ d(b,c)\} < d([a;b],c)$$

donde a, b, c indican puntos a clasificar y $[a;b]$ representa el nodo formado por los vecinos recíprocos a, b. Esta condición se cumple para los principales métodos de clasificación jerárquicos, como *single linkage*, *complete linkage*, *group average* o el método de Ward. En el caso de este último, la distancia definida que asegura este hecho es:

$$d(a,b) = \frac{m_a m_b}{m_a + m_b} d(g_a, g_b)$$

con m_a, m_b las masas de los puntos correspondientes y g_a, g_b sus centros de gravedad.

2.3.5. *Kohonen maps*

([21], [22]) Los mapas auto-organizados o *self-organizing maps* (SOM), también conocidos como *Kohonen maps* en honor a su creador Teuvo Kohonen, componen una técnica en auge con un gran desarrollo en una gran variedad de campos a finales del siglo XX; este hecho puede observarse en [23]. Esto es debido a que se trata de una red neuronal no supervisada cuyo objetivo es la visualización de datos multidimensionales, mostrando los resultados en un mallado de dos dimensiones. El método preserva la estructura topológica de los datos multidimensionales en su representación bidimensional y también es capaz de encontrar relaciones textuales en frases. Estas posibilidades hacen que esta técnica haya tenido un desarrollo tan importante y que hayan sido implementadas en el programa DtmVic.

En la terminología del método, se conoce como *modelo* a cada celda del mallado representado, correspondiente a cada nodo de la red neuronal subyacente. La forma del mapa, en casos usuales de tipo rectangular o hexagonal, sólo es indicativo del número de conexiones en la red y no tiene relevancia en la aplicación del método sino en su precisión; un modelo hexagonal permite una mayor distinción en las diferencias entre celdas que un modelo rectangular. Estos modelos reciben una serie de *inputs* correspondientes a los datos y obtienen la salida presentada en la celda correspondiente del mapa. El propio algoritmo tiende a ordenar las celdas en función de que los modelos sean similares, proporcionando una estructura de similaridades o *clusters* que los autores defienden como un proceso adaptativo similar al encontrado en el cerebro de las especies inteligentes. Un ejemplo de este tipo de organización puede observarse en [21], donde se muestra que modelos similares se organizan de forma próxima en la red. Esta

técnica puede emplearse en su forma básica usando un único mapa, o puede refinarse mediante un sistema que emplee varios mapas que permite obtener resultados más potentes.

El modelo comienza con la definición de una serie de vectores modelo $\mathbf{m}_i \in \mathbb{R}^n$ en el espacio de las variables observables con valores $\mathbf{x} \in \mathbb{R}^n$. La actualización de los vectores modelo que mejor representa a $\mathbf{x}$ se obtiene por el método del descenso del gradiente:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t)[\mathbf{x}(t) - \mathbf{m}_i(t)]$$

donde t indica el paso de la regresión y la función $h_{c(x),i}(t)$, conocida como *neighborhood function*, representa el *kernel* que genera el modelo de regresión. Su primer subíndice, $c = c(x)$ representa el modelo $\mathbf{m}_c(t)$ que mejor se ajusta a $x(t)$ en términos de distancia, llamado *ganador*. Esta función suele tomarse como gaussiana, de forma que:

$$h_{c(x),i} = \alpha(t)\ exp\Big(-\frac{||r_i - r_c||^2}{2\sigma^2(t)}\Big)$$

donde α es el factor de aprendizaje y se compone de una serie de escalares monótonos decrecientes cumpliendo $0 < \alpha(t) < 1\ \forall t$, r_i y r_c son vectores bidimensionales que representan las posiciones espaciales de los modelos en el mapa y $\sigma(t)$ indica la anchura del *kernel* gaussiano empleado, que se reduce en cada iteración.

Respecto a la inicialización de los vectores modelo $\mathbf{m}_i$, en primera instancia es posible asignarlos de forma aleatoria y el modelo convergerá, pero se remarca en [21] que una asignación basada en valores contenidos en el subespacio vectorial correspondiente al plano principal de los datos pueden hacer más rápida la convergencia del método en varios órdenes de magnitud; este hecho puede observarse en el ejemplo de aplicación de esta técnica en [16], donde se observa que la estructura léxica obtenida de un SOM cuadrado de dimensión 3 representa los *clusters* de datos obtenidos mediante la proyección de las palabras en el primer plano factorial correspondiente a un Análisis de Componentes Principales. Sin embargo, en el caso de DtmVic no se incluye una opción de permitir estas inicializaciones, dado que está basado en las búsquedas iniciales con este método.

2.3.6. *Minimum spanning tree*

Se trata de una técnica estadística visual para la clasificación de puntos en el espacio, de forma similar a las técnicas anteriores. Está basado en el conocimiento de una distancia o disimilaridad entre individuos, de forma que se representa una línea continua que une pares de individuos cuya longitud se corresponde a dicha medida de distancia o disimilitud.

En el caso de un gráfico completo, todos los pares estarían unidos por líneas continuas y la interpretación sería complicada. En ese caso, se trata de extraer un grafo parcial, que

contenga los nodos pero un número reducido de líneas que muestre la información relevante de una forma más sencilla. Entre los grafos parciales extraíbles, son relevantes los grafos de tipo *árbol*, dado que pueden representarse en dos dimensiones. Según la teoría de grafos, este tipo de árboles cumplen las propiedades:

- Es un grafo conexo, cada vértice está unido por al menos un camino al resto. También puede interpretarse como que es un grafo con una única componente conexa.
- No tiene ciclos, es decir, no existe un camino cuyo origen y final sea el mismo nodo que no pase dos veces por un nodo. Un ejemplo de un ciclo sería un grafo en forma de triángulo con tres nodos conectados entre sí por pares.

El método trata de buscar un árbol concreto, denominado *minimum spanning tree*, que en teoría de grafos se interpreta como el árbol de coste mínimo. En este caso, el coste de las aristas del grafo es el índice de disimilaridad o la distancia entre vértices, y por coste mínimo se entiende que la suma de los costes de las aristas es mínimo. Por tanto, en su obtención se emplean los algoritmos clásicos de obtención de este árbol: los algoritmos de *Kruskal*, *Prim* y *Florek*.

En [16] se demuestra la relación existente entre el método de los vecinos más próximos o *single linkage* y la creación del *minimum spanning tree*, empleando la conexión y no existencia de ciclos en este último. Esto permite, a partir de la construcción del árbol, obtener la jerarquía de clasificación del método *single linkage*, separando los objetos comenzando por los correspondientes a la arista más larga y descendiendo en orden decreciente de longitud. De esta forma, el par de objetos separados en la última iteración representa los objetos que se clasifican en el primer *cluster* en el método *single linkage* aglomerativo.

Este método es útil cuando no se introduce un umbral mínimo de representación en la representación en el plano principal tras realizar un análisis de reducción de dimensionalidad, como el Análisis de Componentes Principales. En ese caso, es posible que puntos próximos en el plano tengan, en realidad, una separación mostrada en un eje no considerado; en ese caso, el algoritmo mostraría ambos puntos no unidos por una arista e indicaría la presencia de dicha separación.

2.4. *Bootstrap* o técnicas de remuestreo

([27], [28]) En los campos de investigación actuales, especialmente en el campo del Análisis de Datos Textuales, es común que las bases de datos estudiadas no tengan una estructura conocida a priori. En este sentido, las técnicas de *clustering* y de componentes principales proporcionan una herramienta que permite encontrar patrones y clasificar objetos, permitiendo

obtener conclusiones más precisas. Sin embargo, este tipo de técnicas requieren una validación adicional en el campo de los datos textuales, donde los *corpus* empleados están compuestos de una gran variedad de unidades léxicas que dificultan su interpretación. Esta validación puede ser enfocada en dos formas distintas:

- Externa: validación mediante datos externos o empleando un subconjunto de los datos como *aprendizaje* del método; esto último es usual en el caso de los modelos de aprendizaje supervisados o no supervisados. Permiten la complementación de la información conocida para obtener conclusiones más fiables.
- Interna: se refiere a un remuestreo o *bootstrap* de los datos disponibles, ofreciendo regiones de confianza. La ventaja de este tipo de técnicas es conocida, ya que los modelos teóricos a menudo son irreales e imprecisos y, de esta forma, también se elimina la complejidad de estos. Por otro lado, son sencillos de implementar dado que no es necesario conocer la distribución subyacente a los datos y, al ser iterativos, obtener las mejores aproximaciones a los valores reales.

En el contexto de este apartado del trabajo, nos centraremos en introducir el tipo de técnicas *bootstrap* incluidas en el programa DtmVic en el contexto del Análisis de Datos Textuales. Aunque se supone conocido por el lector, los métodos *bootstrap* no paramétricos tienen la siguiente metodología:

1. Obtener K muestras con reposición de tamaño n a partir de la muestra original de datos de tamaño n.
2. Por la obtención de las muestras, se supone como la nueva población a la muestra de datos original.
3. Se estima el valor del parámetro deseado (media, varianza, ...) en cada muestra K, obteniendo K estimaciones del parámetro.
4. Se obtiene el resultado estadístico deseado (estimación, intervalo de confianza) a partir de los K valores del parámetro.

La eficacia de este método está basada en la aproximación a la distribución real por medio de la distribución empírica de la muestra. Los costes computacionales actuales son muy reducidos en comparación al momento en que se desarrolló la técnica, por lo que su utilidad se ha visto incrementada con el paso de los años. Existen variantes de esta técnica para el caso del ACP tanto en categorías activas como suplementarias basados en *test-values* similares a los definidos en el apartado sobre ACM anterior. De esta forma, se puede obtener un intervalo de confianza

para cada punto del subespacio en su proyección sobre los ejes principales considerados que se representa en el plano como elipses o envolventes convexas (*convex hull*) de confianza.

A continuación se introducirán los dos métodos disponibles en el programa para la realización de *bootstrap* sobre los datos textuales.

2.4.1. *Bootstrap* parcial

Este tipo de *bootstrap* emplea la proyección de los datos replicados sobre los ejes principales obtenidos de una descomposición en valores singulares de la matriz de covarianzas. El desarrollo de este método en ACP se debe a Chateau & Lebart [29] mientras que en el caso del AFC y ACM se realizaron importantes avances de la mano de Greenacre [30].

Las técnicas de reducción de la dimensionalidad permiten eliminar la mayor parte del *ruido* presente en los datos, mostrando la información relevante para la investigación. Por tanto, la proyección de estos puntos replicados sobre el eje permitirá una mayor confianza en las estimaciones obtenidas. Para cada par de ejes principales, este método de remuestreo presenta una zona de confianza de cada observación; estas se fijan de forma que contengan aproximadamente el 90 % de las réplicas.

La presentación mediante elipses o envolventes convexas es complementaria, en el primer caso se muestra la densidad de los puntos réplica obtenidos y en el segundo permite observar posibles puntos *outliers* en cada región. Es importante destacar que, al igual que en el caso del *bootstrap* tradicional, la superposición de dos regiones de confianza indica que no existe una diferencia apreciable estadística entre ambos objetos.

2.4.2. *Bootstrap* total

Esta técnica varía respecto al caso parcial en el hecho de que se repite el método estadístico empleado para cada repetición; por ejemplo, en el caso de un AFC, se repetiría este análisis sobre cada muestra *bootstrap*. En este caso es obvio que no se dispone de un espacio común de representación, lo que complica la interpretación de los resultados; la diagonalización de las matrices de covarianzas sucesivas puede llevar a problemas como cambios de signo en los ejes o rotaciones en los mismos.

Por ello, este método requiere una serie de transformaciones que identifican los ejes comunes en la realización del método para cada muestra. En función del tipo de transformación aplicada, se habla de tres tipos de *bootstrap* total:

- *Bootstrap* total tipo 1: se realizan cambios de signo en los ejes considerados homólogos con los ejes principales de la muestra poblacional. Este método es el más conservador, y

supone que no existen problemas de rotación y reordenación de los ejes principales, por lo que sólo debería ser empleado en datos con estructuras internas robustas.

- *Bootstrap* total tipo 2: incluye un tratamiento para posibles reordenaciones de los ejes. El criterio se basa en estudiar la correlación de cada eje replicado con el eje de la muestra poblacional y casar ambos ejes cuando el valor absoluto de la correlación es máximo. Tras eso se incluye una transformación de tipo 1 para corrección en los signos de los ejes. Este tipo de técnica es útil en el caso de ejes considerados como variables latentes, donde la ordenación de los autovalores puede cambiar pero no se espera una rotación, que indicaría una composición distinta con las variables consideradas.

- *Bootstrap* total tipo 3: emplea una rotación procrustes [31], basadas en rotaciones o escalados de la configuración hasta minimizar la suma de distancias al cuadrado de los puntos a un centroide obtenido a partir de ellos. Este ajuste no sólo puede realizarse en el plano, sino que la minimización puede realizarse en dimensiones mayores al mismo tiempo. Este método es el más adaptable de los tres, y permite comparar el subespacio en su conjunto con el perteneciente a la muestra poblacional haciendo coincidir varios ejes al mismo tiempo. En este sentido, este método es el más similar al *bootstrap* parcial anterior.

Una comparación entre los tipos de *bootstrap* total y parcial introducidos puede consultarse en el texto de Lebart [28]. En él, se observa que existe similitud entre el *bootstrap* parcial y el total tipo 3, como se había indicado, y que las elipses correspondientes a los otros métodos son más grandes y, por tanto, producen resultados más conservadores.

2.4.3. *Bootstrap* jerárquico

Este método de remuestreo es interesante en el campo de los datos textuales, ya que su motivación es el análisis de datos con unidades estadísticas diferentes. En el caso del Análisis de Datos Textuales, existen al menos dos unidades estadísticas distintas: los individuos u observaciones (encuestados, entrevistas, usuarios Web...) y las ocurrencias estudiadas (palabras, lemas...). Esta dualidad implica que el método *bootstrap* usual, empleado en este caso sobre las palabras del *corpus*, podría ser usado también sobre los individuos; esto permitiría obtener patrones y conclusiones que por estudios del otro nivel de datos no serían interpretables. Este tipo de técnica puede ser observada de nuevo en el estudio de Lebart [28].

2.5. Seriación de una tabla de datos

Los métodos de seriación de una tabla de datos se emplean con el objetivo de representar estructuras subyacentes en matrices de datos. Están basados en permutaciones de las filas y columnas de la tabla estudiada y muestra bloques homogéneos de valores elevados o pequeños de las variables consideradas. También pueden mostrar una progresión en los perfiles estudiados observando la evolución de los bloques.

El método de seriación empleado en el programa DtmVic se basa en el trabajo de Hill [32], que basa la ordenación de los puntos fila y los puntos columna mediante su ordenación en el primer eje principal en el caso de un AFC. De esta forma, la seriación de la tabla léxica correspondiente ofrece una metodología visual para la agrupación de los datos estudiados.

Capítulo 3

Análisis léxico aplicado a noticias sobre el cambio climático

En este último capítulo del trabajo se procede a hacer la comparación entre ambos programas, en la mayor parte de los casos se ha decidido emplear tablas para facilitar la interpretación. Se discuten primero las características generales de ambos antes de pasar al ejemplo específico de las noticias en la prensa digital española sobre el cambio climático, para luego hacer una discusión final.

Se han analizado 15 noticias de los tres medios digitales más leídos (según se discute en la metodología), correspondientes al final del año 2019 y la primera mitad del 2020. Noticias con temas muy divergentes o específicos han sido excluidas con el fin de que se incluyesen informaciones de temática general, dejándonos con las mostradas en la figura 3.1; se han codificado de acuerdo a un número asignado cronológicamente y al medio digital de publicación.

Número	Título	Periódico	Codificación
1	Refugiados invisibles y "migraciones traumáticas", el cambio climático que ya sufren millones en todo el mundo	ElMundo	**** *ElMundo_1
2	El ministro astronauta: "Ese 'tapón de hielo' de 1.582 km2 en la Antártida no es por el cambio climático"	ElMundo	**** *ElMundo_2
3	La ONU advierte de que hay que cambiar la dieta para parar el cambio climático	ElMundo	**** *ElMundo_3
4	Islandia dice adiós a Ok, su primer glaciar desaparecido por el calentamiento global	ElMundo	**** *ElMundo_4
5	Los incendios forestales en Brasil aumentan un 83% a causa de la deforestación y la sequía en el Amazonas	ElMundo	**** *ElMundo_5
6	El Mediterráneo sufrirá el cambio climático un 20% más que el resto del mundo	ElMundo	**** *ElMundo_6
7	Cambio climático, desertificación y la COP25	ElPais	**** *ElPais_7
8	Cambio climático y responsabilidades	ElPais	**** *ElPais_8
9	La salud humana y la del planeta van de la mano	ElMundo	**** *ElMundo_9
10	La vegetación se apodera del Everest a causa del cambio climático	LaVang	**** *LaVang_10
11	¿Qué supondrá la declaración de emergencia climática en España?	ElMundo	**** *ElMundo_11
12	El hambre vuelve a crecer por culpa del clima	ElMundo	**** *ElMundo_12
13	El Polo Sur se calienta tres veces más rápido que el conjunto del planeta	LaVang	**** *LaVang_13
14	El Gobierno presenta su hoja de ruta para que España sobreviva a la crisis climática	ElMundo	**** *ElMundo_14
15	La crisis climática está transformando ya los bosques, con árboles más jóvenes y de menor altura	ElMundo	**** *ElMundo_15

Figura 3.1: Características generales de la base de datos empleada.

3.1. Comparación de características generales

En esta sección se compararán las características generales de ambos programas, de forma que el lector obtenga una idea clara de los puntos en común entre ambos programas y qué aspectos influyen en diferencias entre ambos. Para la elaboración de la tabla de características generales mostrada en la figura 3.2 se han tenido en cuenta los aspectos que pueden ser de mayor interés a la comunidad académica. Estos también pueden consultarse en los sitios Web de DtmVic e IRaMuTeQ. En general ambos programas presentan características similares, siendo DtmVic el basado en un lenguaje de programación más antiguo y obsoleto, aunque también recibe una mayor actualización por parte de los autores.

En la figura 3.3 se muestran los diversos métodos de importación de datos en ambos programas y los tipos de resultados que se obtienen en su salida. En este caso es importante hacer un comentario sobre la compatibilidad de IRaMuTeQ con otros programas de tratamiento de datos usados por estudiantes e investigadores, tales como MS Excel, MS Word, Libre Office y SPSS. Si bien cuando se trabaja en SPSS es posible exportar los datos a MS Excel y viceversa, si se desea exportar a este programa debe pasarse antes por un editor de texto (se presenta un ejemplo con IRaMuTeQ en los anexos). Este problema no existe con DtmVic dado que existen importaciones directas de archivos en formato *.csv* correspondientes a Excel, R o SPSS e incluso archivos tipo XML comúnmente usados en bases de datos SQL.

	IRaMuTeQ	DTMVIC
Última versión	0.7 alpha 2 / 2014	6.2 / 2020
Autores	Pierre Ratinaud & Sébastien Déjean Université de Touluse	L. Lebart, A. Morineau
Plataformas	Windows, Macintosh & Linux	Windows (Macintosh & Linux mediante programa adicional)
Manuales en castellano	http://www.iramute q.org/documentatio n/fichiers/guiairamuteq	http://www.dtmvic.com/doc/DtmVic_E nglish_Manual_2016.pdf
Enlace de descarga	http://www.iramuteq.org	http://www.dtmvic.com/05_SoftwareE.html
Año de lanzamiento	2008	SPAD (1987)
Licencia	GNU/ Software Libre	GNU/ Software Libre
Idiomas	Francés, Portugués, Castellano, Inglés.	Inglés, Francés, Castellano, Italiano.
Código fuente	Python, R	Fortran 77

Figura 3.2: Características generales de los programas comparados.

	IRaMuTeQ	DTMVIC
Ficheros de entrada	Textos literarios Entrevistas Cuestionarios Noticias periodísticas	Textos literarios Entrevistas Cuestionarios Noticias periodísticas
Formato de entrada	Archivo en formato Unicode (UTF-8) creado por un editor de texto	Archivo en formato ANSI creado por un editor de texto Archivo csv con tabla de variables frente a individuos Archivo XML de bases de datos en línea (SQL)
Ficheros de salida	- Conteo de palabras o lemas del corpus - Gráfico de Análisis de Similitud - Gráfico de recuento de palabras - Análisis multivariados: Análisis de correspondencias múltiple y cluster.	- Archivo imp con resultados del análisis - Ficheros de los pasos intermedios del análisis

Figura 3.3: Comparación de ficheros de entrada y salida de ambos programas.

3.2. Análisis léxico con DtmVic

En este apartado se describirá el análisis del *corpus* de texto correspondiente a las noticias sobre el cambio climático mediante el programa DtmVic. En primer lugar se realizará una breve descripción del proceso de importación y pre-procesamiento del texto, incluyendo la lematización empleada, y a continuación se describirá el análisis realizado y se mostrarán los gráficos y conclusiones obtenidas.

En caso de requerir una importación de encuestas con diversas preguntas abiertas y disponer de preguntas cerradas adicionales, el proceso de importación a seguir es similar al descrito en el manual anexo de este trabajo. Sin embargo, en el caso de trabajar con simples noticias de texto, el proceso de importación es más sencillo. En este caso ha bastado con copiar y pegar los textos correspondientes en un editor de texto y separarlos mediante el formato requerido por el programa (internamente conocido como tipo 1). Por tanto, basta con asegurar los puntos mostrados a continuación:

- Comprobar que el archivo comienza en la primera línea del fichero de texto.
- El archivo comienza con cuatro asteriscos delimitando el inicio de cada texto (****) seguidos de **cuatro** espacio en blanco.
- Tras los espacios en blanco se introduce el ID de cada texto sin emplear caracteres especiales.
- Entre dos textos existe una separación de una línea en blanco.
- Al final del último texto, se deja una línea en blanco y se insertan cuatro símbolos de igualdad (====).

- Por último, se guarda el archivo de texto con codificación ANSI, de esta forma las tildes y demás caracteres especiales serán reconocidos por DtmVic.

Este archivo de texto se encuentra en el formato adecuado para ser analizado por DtmVic. En esta etapa se realizan dos pasos importantes para el análisis mediante el programa, con herramientas que se encuentran en las opciones de pre-procesado de textos. El primero de ellos es el cambio de la longitud de las líneas de texto, dado que DtmVic sólo permite una longitud máxima de 200 caracteres por línea de texto. Este paso no involucra ningún cambio en el *corpus* ya que es puramente técnico para el funcionamiento del programa, por lo que un número estándar como 150 es suficiente. El segundo paso es la transformación de los caracteres a minúsculas, para evitar posibles problemas en la posterior lematización del programa; este paso también se encuentra en la misma pestaña y se realiza de forma sencilla.

Una vez realizados estos pasos, el último paso antes de comenzar al análisis del texto es la lematización. En este caso podrían emplearse unos análisis previos de las concordancias (**CORDA**) y los segmentos repetidos (**SEGME**) en el texto, aunque se ha observado que no aportan información especialmente relevante en los textos considerados; en caso de contar con un número mayor de textos serían de gran utilidad en este apartado. Para llevar a cabo la lematización se siguen los pasos indicados en la sección correspondiente con el programa **WinTreeTagger**, donde el único apunte a realizar es la retirada de ciertas palabras o símbolos del texto que ocurre en la lematización del texto, como se muestra en la parte izquierda de la figura **??**. En este caso se indicarán los apartados que, en carácter general, tanto en este trabajo como en posteriores se deberían marcar siempre en primer lugar: BACKSLASH, CARD, CM, COLON, DASH, ITJN, LP, PERCT, QT, RP, SEMICOLON, SLASH, SYM; estas selecciones retiran todos los símbolos y marcadores de puntuación como comas y barras y también eliminan los números (cardinales) dado que no aportan información en el caso del Análisis Textual. Otras opciones que se marcarán en este caso son: ADJ, ADV, ART, CC, CCAD, CCNEG, CQUE, DM, INT, PPC, PPO, PPX, PREP, QU, REL, SE, UMMX y todas las formas verbales de los verbos ser, estar y haber; estas opciones retiran la mayor parte de las formas léxicas que aportan poca información en las conclusiones obtenidas, como preposiciones, adverbios, artículos, etc.

Una vez en este punto ya es posible realizar un análisis concreto mediante el programa y se puede proceder a obtener clasificaciones de los textos considerados. En el caso presente se mostrarán las diversas opciones posibles:

- Realizar un AFC simple sobre la tabla léxica correspondiente (método **VISUTEX**).
- Tomando como variable categórica el periódico al que pertenece cada noticia, realizar un AFC sobre las tablas léxicas correspondientes a cada categoría (método **ANALEX**).

- De forma similar al anterior pero realizando el procedimiento **VISURECA** e introducir la variable de forma suplementaria. Este caso permite seleccionar el número de *clusters* a realizar.

Se han estudiado los tres posibles análisis por separado y, basado en las conclusiones tomadas en cada caso, se ha estimado que el método que más información aporta en este ejemplo es el primero de todos. La razón de no emplear los dos últimos, que en principio parece que aportar mejoras frente al método simple **VISUTEX** no es debido a que sea mejor o peor, sino simplemente más adecuado en este caso. En este trabajo se dispone de un número reducido de noticias, y como se ha mostrado en la figura 3.1 la mayor parte se corresponden con el diario El Mundo; esto produce un sesgo a la hora de tratar de clasificar las noticias y compararlas en función de su periódico de procedencia, que sería el objetivo principal de los métodos **ANALEX** y **VISURECA**. En general, estos últimos métodos son más indicados para el estudio de problemas más complejos, con un mayor número y variedad de textos y disponiendo de más información externa (que se interpreta como variables suplementarias); y dado que en el caso planteado sólo se trata de mostrar las capacidades de DtmVic en este tipo de análisis, se considera más indicado realizar el análisis simple **VISUTEX**.

En la ejecución del análisis se ha tomado un umbral mínimo de **cuatro ocurrencias** por palabra; en caso de que el umbral sea mucho menor se corre el riesgo de que las masas involucradas en el AFC deformen el espacio y coloquen los puntos más cerca del origen. Tras la ejecución del análisis se obtienen diversos resultados de importancia que se muestran a continuación, y que como se ha indicado en parte se encuentran en el archivo *.imp* de salida del programa. En primer lugar deben estudiarse los pasos previos del análisis que proporcionan resúmenes importantes del *corpus* de texto (Figura 3.4), donde se puede observar el peso de cada texto y la presencia de hápaxes y las palabras retenidas, proporcionando una idea básica del comportamiento de cada texto.

Por otro lado, la información de los pasos **Aplum** y **Mocar** (o **Recar** si se ha incluido manualmente) es el resultado principal buscado con este análisis, donde se obtendrá la representación conjunta en el plano de componentes principales mediante el AFC y su descripción mediante palabras y respuestas características. En la figura 3.5 se muestran las representaciones por separado de los textos y las palabras (variables e individuos, respectivamente) donde se puede observar su clasificación y las relaciones de parecido entre ellas. Una representación conjunta es claramente posible pero también es poco informativa dado la gran cantidad de palabras representadas; en caso de disponer de más noticias este problema es más acusado aún. Por ello, es importante notar que las palabras y respuestas características son los principales resultados de este análisis que, combinados con la representación factorial del plano de los

repartition of terms in texts/ ----------

number of text	identifier	number of words	/1000 of total	mean per response	number of words (distinct)	/1000 words of text	number of words kept
1 =	elmundo_1	469	91.5	10.4	132	281.4	246
2 =	elmundo_2	407	79.4	10.4	112	275.2	255
3 =	elmundo_3	195	38.0	10.8	69	353.8	115
4 =	elmundo_4	259	50.5	10.8	83	320.5	143
5 =	elmundo_5	149	29.1	10.6	43	288.6	67
6 =	elmundo_6	242	47.2	10.5	80	330.6	150
7 =	elpais_7	331	64.6	9.7	104	314.2	179
8 =	elpais_8	477	93.0	10.6	137	287.2	266
9 =	elmundo_9	292	57.0	10.1	85	291.1	143
10 =	lavang_10	270	52.7	11.3	80	296.3	154
11 =	elmundo_11	551	107.5	10.6	131	237.7	296
12 =	elmundo_12	626	122.1	11.0	150	239.6	352
13 =	lavang_13	246	48.0	10.7	75	304.9	153
14 =	elmundo_14	349	68.1	10.6	117	335.2	202
15 =	elmundo_15	264	51.5	10.6	88	333.3	162
	g l o b a l	5127	1000.0	10.6			2883

Figura 3.4: Descripción básica de la composición de los textos sobre el cambio climático considerados.

textos, proporcionan las conclusiones adecuadas en el estudio de las similitudes y composición de los textos.

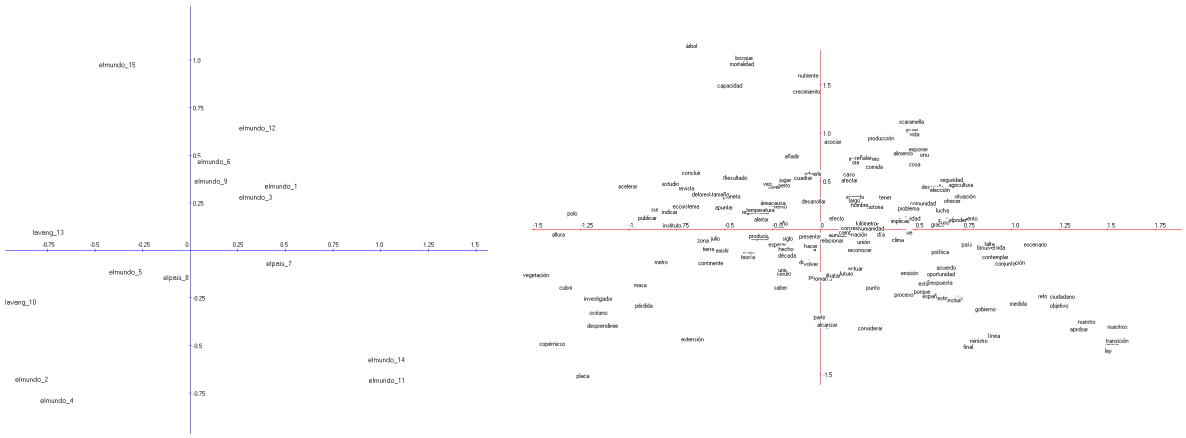
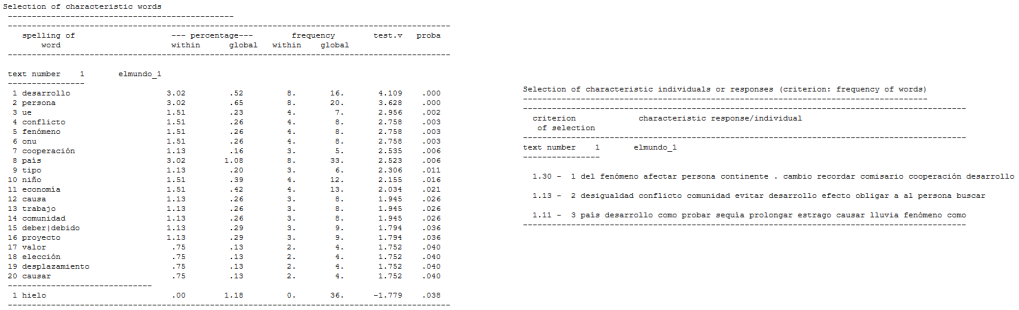


Figura 3.5: Plano principal del Análisis Textual del *corpus* de texto. Representación del plano de noticias (izquierda) y plano de formas léxicas (derecha).

En la figura 3.6 se muestran las palabras y respuestas características que definen el primer texto, tal como se ha descrito en los manuales anexos. Esta información combinada con la representación de los textos mostrada en la figura 3.5 proporciona los estándares de clasificación de los textos y permite interpretar los parecidos y diferencias que existen entre ellos. Las palabras características muestran las principales formas léxicas presentes en cada texto y son

útiles en la determinación de los temas principales de cada texto y en asociar parecidos entre ellos, mientras que las respuestas características muestran fragmentos del texto que permiten obtener una mejor idea del tema tratado en cada texto. Cabe destacar que la información conjunta de estas palabras y respuestas y la representación gráfica puede consultarse en la representación **ClusterView**, que muestra estos resultados de forma más directa al tratar con el programa de forma interactiva.



Selection of characteristic words

spelling of word	--- percentage within	--- global	frequency within	frequency global	test.v	proba
text number 1 elmundo_1						
1 desarrollo	3.02	.52	8.	16.	4.109	.000
2 persona	3.02	.65	8.	20.	3.628	.000
3 ue	1.51	.23	4.	7.	2.956	.002
4 conflicto	1.51	.26	4.	8.	2.758	.003
5 fenómeno	1.51	.26	4.	8.	2.758	.003
6 onu	1.51	.26	4.	8.	2.758	.003
7 cooperación	1.13	.16	3.	5.	2.535	.006
8 país	3.02	1.08	8.	33.	2.523	.006
9 tipo	1.13	.20	3.	6.	2.306	.011
10 niño	1.51	.39	4.	12.	2.155	.016
11 economía	1.51	.42	4.	13.	2.034	.021
12 causa	1.13	.26	3.	8.	1.945	.026
13 trabajo	1.13	.26	3.	8.	1.945	.026
14 comunidad	1.13	.26	3.	8.	1.945	.026
15 deber|debido	1.13	.29	3.	9.	1.794	.036
16 proyecto	1.13	.29	3.	9.	1.794	.036
17 valor	.75	.13	2.	4.	1.752	.040
18 elección	.75	.13	2.	4.	1.752	.040
19 desplazamiento	.75	.13	2.	4.	1.752	.040
20 causar	.75	.13	2.	4.	1.752	.040
1 hielo	.00	1.18	0.	36.	-1.779	.038

Selection of characteristic individuals or responses (criterion: frequency of words)

criterion of selection	characteristic response/individual
text number 1	elmundo_1
1.30 - 1	del fenómeno afectar persona continente . cambio recordar comisario cooperación desarrollo
1.13 - 2	desigualdad conflicto comunidad evitar desarrollo efecto obligar a al persona buscar
1.11 - 3	país desarrollo como probar sequía prolongar estrago causar lluvia fenómeno como

Figura 3.6: Distribución de palabras (izquierda) y respuestas (derecha) características correspondientes al primer texto del periódico El Mundo según la codificación empleada.

Aunque no se muestran todas las palabras y respuestas características obtenidas de cada texto por falta de espacio, en la tabla 3.7 se muestran las tres principales palabras características de los textos estudiados, que proporcionan una idea del tema representado en cada una y aportan sentido a la clasificación mostrada en la figura 3.5. Acompañado de la información presentada por las respuestas características, se encuentra una agrupación de textos en bloques:

- **Bloque de deshielo**: está formado por las noticias 2, 4, 10 y 13 y está marcado por una temática de preocupación frente a problemas de cambio climático relacionados con el deshielo y el cambio del nivel del mar.

- **Bloque de evidencias**: este bloque se sitúa de forma central en el plano y se constituye de noticias de carácter general dedicadas a mostrar evidencias de la existencia del cambio climático; este es el caso, por ejemplo, de la noticia 9 sobre los efectos en la salud, la noticia 12 sobre la desnutrición o la noticia 6 sobre el nivel del agua del Mediterráneo, entre otras.

- **Bloque económico**: este bloque, formado por las noticias 11 y 14, se centra en noticias de carácter político y administrativo relacionadas con el cambio climático, como se evidencia por su posición en el plano alejada del resto.

Textos	Palabras características
ElMundo_1	desarrollo, persona, UE
ElMundo_2	hielo, plataforma, antártida
ElMundo_3	suelo, contribuir, documento
ElMundo_4	glaciar, islandia, perder
ElMundo_5	incendio, cifra, deforestación
ElMundo_6	grado, agua, mar
ElPais_7	consecuencia, territorio, atmósfera
ElPais_8	larsen, teoría, juzgar
ElMundo_9	salud, contaminación, asociar
LaVang_10	nieve, vegetación, altura
ElMundo_11	declaración, emergencia, medida
ElMundo_12	desnutrición, familia, hambre
LaVang_13	polo, calentamiento, indicar
ElMundo_14	borrador, recoger, transición
ElMundo_15	árbol, bosque, mortalidad

Figura 3.7: Tres principales palabras características asociadas a cada texto estudiado según el criterio de frecuencia mostrado.

Una vez mostrada esta clasificación, podemos interpretar fácilmente el carácter de cada componente principal, de forma que se conozca la interpretación latente de cada una de ellas. Basado en la clasificación mostrada, podemos interpretar los ejes como:

- Eje horizontal: este eje representa el **carácter económico y político** de las noticias, separando el lado izquierdo del plano, más liberal, del lado derecho más concentrado en consecuencias económicas y proposiciones políticas al respecto.
- Eje vertical: este eje muestra el **carácter humanista** de las noticias, donde las noticias presentadas en la parte superior del gráfico se centran en la situación humanitaria y las consecuencias directas sobre las personas de estos cambios, y la parte inferior muestra noticias más deshumanizadas y centradas en interpretaciones económicas del cambio climático y sus consecuencias.

Por último, se muestra en la figura 3.8 el mapa auto-organizado o *SOM* de Kohonen correspondiente a las noticias estudiadas en este trabajo. En él se puede observar una clasificación en casillas de un mallado que representan de una forma fiel las similitudes entre bloques contiguos mediante un proceso no supervisado, como se ha comentado en los manuales anexos. En este caso se observan resultados que concuerdan con las conclusiones obtenidas anteriormente, ya que se observan claramente el bloque de deshielo (parte inferior izquierda de la figura), el bloque de evidencias (se concentra en la diagonal del mallado de celdas) y el bloque económico

(parte superior izquierda). Esta representación refuerza la clasificación obtenida anteriormente y, en caso de disponer de más noticias, sería de gran utilidad en su clasificación.

Figura 3.8: Mapa auto-organizado de las noticias y las palabras involucradas en el análisis de dimensión $(5, 5)$.

3.3. Análisis léxico con IRaMuTeQ

En primer lugar, se comenzará mostrando el proceso de importación de la base de datos en un formato legible por IRaMuTeQ. Para ello, se hará uso de un archivo Excel de la siguiente forma: se comienza pegando en una hoja de cálculo las noticias recolectadas en el Internet sobre el cambio climático, creando una columna (en amarillo) para concatenar, siguiendo el estilo '**** *var_mod', el nombre del periódico de publicación ('variable'), el número de noticia ('modalidad', ordenadas cronológicamente), y por último, un columna (en verde), donde

concatena 'var_mod' con el cuerpo de texto. Un extracto de este documento se muestra en la figura 3.9. La última columna creada es lo que se guarda en un editor de texto para ser analizado por IRaMuTeQ.

#	Periodico	Año	Mes	Título	Subtítulo / resu	var_mod	Cuerpo	Concatenar
1	ElMundo	2019	8	Refugiados invisibles y "migi	Aunque en los p	**** *ElMundo_1	La mayor cris	**** *ElMundo_1 La mayor crisi
2	ElMundo	2019	8	El ministro astronauta: "Ese	El ministro de C	**** *ElMundo_2	El pasado 25	**** *ElMundo_2 El pasado 25 d
3	ElMundo	2019	8	La ONU advierte de que hay	El modo en que	**** *ElMundo_3	La meta del A	**** *ElMundo_3 La meta del A
4	ElMundo	2019	8	Islandia dice adiós a Ok, su primer glaciar de		**** *ElMundo_4	Okjokull, prin	**** *ElMundo_4 Okjokull, prim
5	ElMundo	2019	8	Los incendios forestales en	Las cifras surgen	**** *ElMundo_5	Brasil ha regi	**** *ElMundo_5 Brasil ha regis
6	ElMundo	2019	10	El Mediterráneo sufrirá el ca	La Unión por el	**** *ElMundo_6	La región me	**** *ElMundo_6 La región mec
7	ElPais	2019	11	Cambio climático, desertific	Estamos ante u	**** *ElPais_7	Hace unos 12	**** *ElPais_7 Hace unos 12.00
8	ElPais	2019	12	Cambio climático y responsi	Si no se actúa, e	**** *ElPais_8	Una posible c	**** *ElPais_8 Una posible defir
9	ElMundo	2019	12	La salud humana y la del pla	El calentamient	**** *ElMundo_9	No se trata d	**** *ElMundo_9 No se trata de
10	LaVang	2020	1	La vegetación se apodera de	Un nuevo estud	**** *LaVang_10	La cumbre m	**** *LaVang_10 La cumbre más

Figura 3.9: Creación de base de datos en Excel, previo a su exportación.

Lo siguiente es el resumen global de todo el texto que arroja IRaMuTeQ tras la importación de la base de datos compuesta por 15 artículos. Puede hacerse una distinción entre las formas activas, que son las que se considerarán en el análisis, y las formas suplementarias (a no tenerse cuenta por ser palabras como artículos y preposiciones). Esto se determina con la opción de 'limpieza de texto'. Estos resúmenes se muestran en la gráfica 3.10.

Resumen	
Número de Textos	15
Número de ocurrencias	11981
Número de formas	2220
Número de hápax	1193
Média de ocurrencias por texto	798.73

Forma	Frecuencia	Tipo
cambio	88	sustantivo
climático	83	adjetivo
más	78	advervio
grande	49	adjetivo
hielo	38	sustantivo
año	36	sustantivo
efecto	36	sustantivo
ya	34	advervio
país	32	sustantivo
hacer	31	verbo

Figura 3.10: Resumen de las ocurrencias (izquierda) y de las 10 formas activas mas comunes (derecha) en el *corpus* de texto.

El gráfico de la ley de Zipf mostrado en la figura 3.11 resume la información de la figura anterior mostrando en el eje de las abscisas los logaritmos de rangos y en el eje de las ordenadas los logaritmos de frecuencias.

3.3.1. Análisis de especificidades

El análisis de especificidades se hace en función de las variables que se hayan definido inicialmente (los nombres de los periódicos), y asocia los textos del *corpus* (definidos como modalidades, por número), y muestra como se usan las formas específicas. Puede interesar saber qué formas específicas se usan en cada discurso; como se dispone de un gran número, se muestra el caso de la forma *climático* y de múltiples formas en la figura 3.12. De esta forma, por ejemplo, se observan usos altos de la palabra *climático* en los textos 1, 3 y 7 y poco uso en las noticias 5, 10 y 15. Estos resultados pueden arrojar más información sobre la estructura

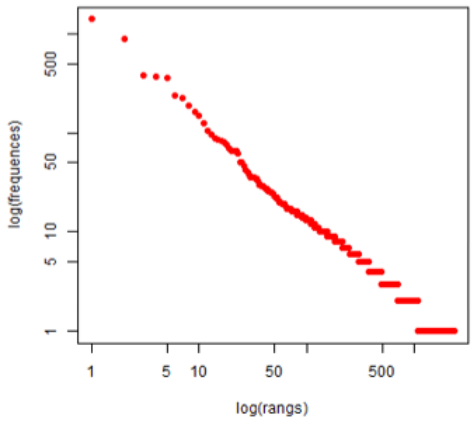


Figura 3.11: Análisis de frecuencias del *corpus* de texto mediante la ley de Zipf.

de las noticias estudiadas al ponerse en conjunto al resto de resultados que se muestran en este apartado.

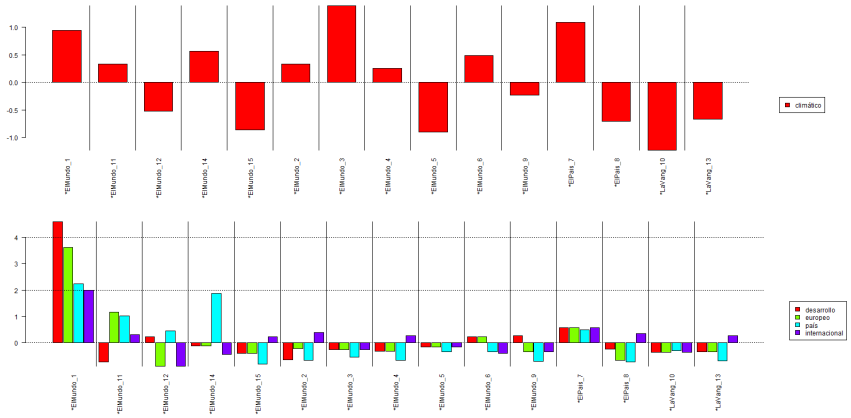
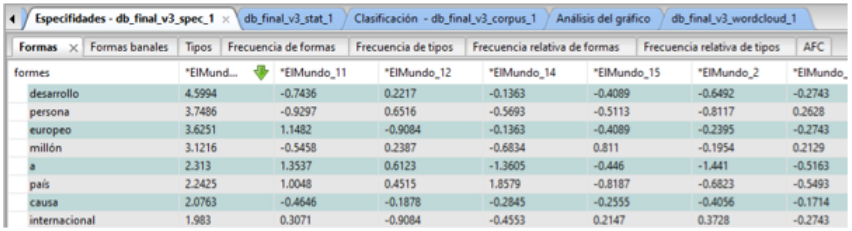


Figura 3.12: Gráfico de similitud para la forma *climático* (superior) y gráfico de similitud para múltiples formas (inferior).

Estos resultados pueden complementarse con las salidas obtenidas de este análisis, como se muestra en la figura 3.13.

3.3.2. Análisis de similitudes

El análisis de similitudes muestra en un diagrama de árbol las relaciones de los núcleos semánticos por co-ocurrencias de palabras. Las formas son representadas en los nodos de la gráfica, y los enlaces entre los nodos representan co-ocurrencia de las mismas. El tamaño de las palabras es proporcional a su frecuencia, y el grosor de los enlaces es proporcional a la co-ocurrencia. IRaMuTeQ permite hacer un análisis de todas las formas encontradas en el *corpus*,



Especifidades - db_final_v3_spec_1 | db_final_v3_stat_1 | Clasificación - db_final_v3_corpus_1 | Análisis del gráfico | db_final_v3_wordcloud_1

Formas | Formas banales | Tipos | Frecuencia de formas | Frecuencia de tipos | Frecuencia relativa de formas | Frecuencia relativa de tipos | AFC

formes	*ElMund...	*ElMundo_11	*ElMundo_12	*ElMundo_14	*ElMundo_15	*ElMundo_2	*ElMundo_
desarrollo	4.5994	-0.7436	0.2217	-0.1363	-0.4089	-0.6492	-0.2743
persona	3.7486	-0.9297	0.6516	-0.5693	-0.5113	-0.8117	0.2628
europeo	3.6251	1.1482	-0.9084	-0.1363	-0.4089	-0.2395	-0.2743
millón	3.1216	-0.5458	0.2387	-0.6834	0.811	-0.1954	0.2129
a	2.313	1.3537	0.6123	-1.3605	-0.446	-1.441	-0.5163
país	2.2425	1.0048	0.4515	1.8579	-0.8187	-0.6823	-0.5493
causa	2.0763	-0.4646	-0.1878	-0.2845	-0.2555	-0.4056	-0.1714
internacional	1.983	0.3071	-0.9084	-0.4553	0.2147	0.3728	-0.2743

Figura 3.13: Resultados del análisis de especificidades.

pero al ser demasiadas, se obtiene un gráfico saturado e ilegible. Para el gráfico mostrado en la figura 3.14, se ha limitado el análisis a todas las formas con una frecuencia de aparición igual o superior a diez, revelándose como nodos principales las formas *más*, *cambio* y *climático*; enlazados textualmente entre sí por enlaces mas gruesos, siendo, como era esperarse en estos discursos, más fuerte el nexo entre *cambio* y *climático*, y el nexo entre estas dos y la forma *más*, implicando que las noticias analizadas esperan o pronostican un aumento de las formas relacionadas directamente con la palabra más: sequía, carencia, futuro, niño (fenómeno de...), etc., revelándose así la importancia de las consecuencias esperadas del cambio climático.

También se permite destacar cualquiera de las ramas periféricas del gráfico. Por ejemplo, en el extremo superior, se muestran formas léxicas relacionadas con la emisión de gases de efecto invernadero a la atmósfera (cambio ->climático ->efecto ->invernadero ->gas ->emisión ->bosque).

3.3.3. Clasificación Jerárquica Descendente

Como se ha discutido ya en el anexo, la CHD se basa en la tesis de Reinert de que todo discurso se expresa a partir de un conjunto de palabras o formas que constituyen unidades de significado, independientemente de su construcción sintáctica (los 'mundos léxicos', o lugares de enunciación, los cuales se definen por oposición entre sí); por lo cual ya no se elaborará más sobre el tema, pasando a discutir los resultados del análisis con IRaMuTeQ. Para la CHD se usa todo el *corpus* de texto sin distinción de variables, ya que interesa conocer qué grupos se generan de todo el discurso.

La clasificación obtenida se muestra en la figura 3.15, que encuentra cinco clases o *clusters* en el *corpus*. Puede verse que las formas que predominan en cada una de ellas se refieren a temas similares, lo cual permite adivinar cuales son las representaciones sociales presentes en cada uno de ellos, de modo que puede asignarse un nombre que las defina, por ejemplo:

- **Clase 1: Impacto humano**. El menor grupo con solo el 12.1 % de formas, este agrupa palabras que podemos considerar representan el impacto humano del cambio climático: 'niño', 'persona', 'escasez', 'contaminación', 'sequía', etc.

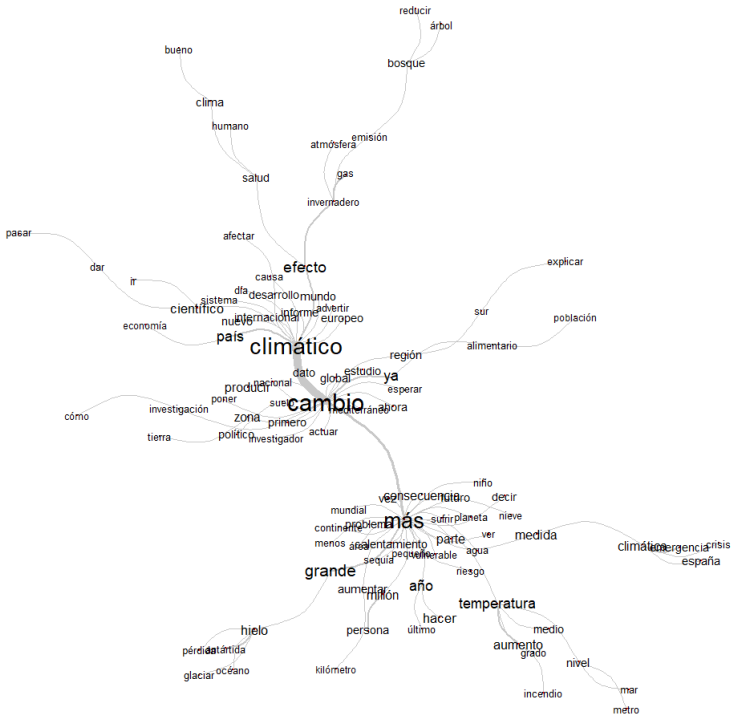


Figura 3.14: Análisis de similitudes de las formas con frecuencia mayor o igual a 10.

- **Clase 2: Catástrofes**. Muy cerca de la Clase 1, esta clase agrupa formas como 'advertir', 'moralidad', 'catástrofe', relacionadas a consecuencias negativas.

- **Clase 3: Economía**. Esta clase agrupa términos relacionados con el impacto económico y la sociedad.

- **Clase 4: Política**. Aquí vemos una agrupación de términos relacionados con acciones políticas en respuesta al problema, tales como 'solución', 'nación', 'estrategia', 'cooperación'.

- **Clase 5: Polos**. se agrupan acá palabras relacionadas con el impacto del fenómeno en los polos, tales como 'hielo', 'Antártida', 'océano', 'desprendimiento', 'iceberg'.

- **Clase 6: Atmósfera**. Agrupa términos relacionados a la contaminación atmosférica, tales como 'gas', 'invernadero', 'temperatura', 'carbono'.

- **Clase 7: Relaciones Públicas**. Términos relacionados a como los medios de comunicación representan el problema: 'emergencia', 'medida', 'gobierno', 'declaración', 'público', etc.

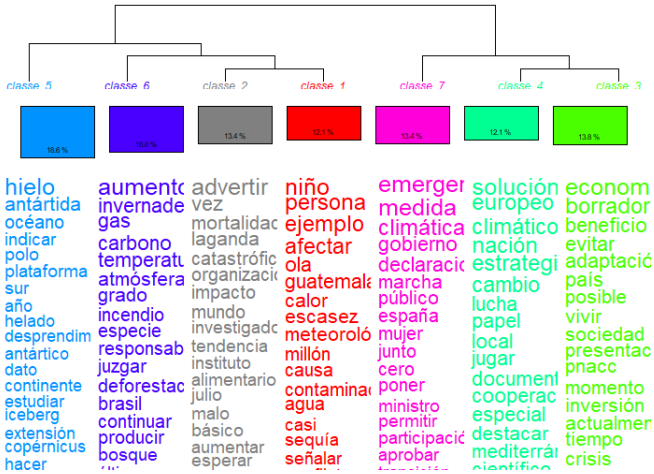
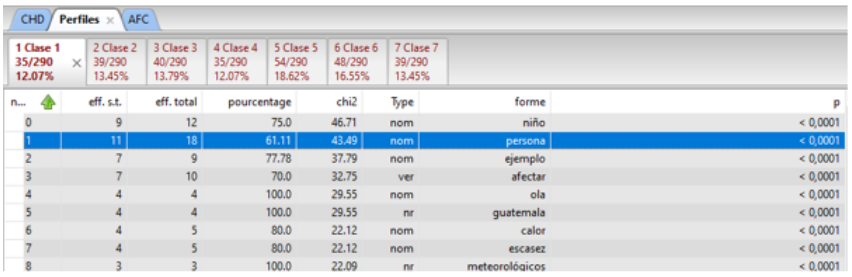


Figura 3.15: Clasificación jerárquica descendente de la base de datos estudiada.

Es de notar el modo en que la clasificación descendente ha separado los grupos poniendo en evidencia los mundos léxicos del discurso, primero fusionando 'Catástrofe' e 'Impacto Humano' con las dos clases relacionadas con la ecología, por un lado; y por el otro uniendo las clases relacionadas con la política y economía. Como una observación final sobre las clases, puede notarse que la distribución de formas está balanceada entre ellas (aproximadamente 15 % por clase).

Un último resultado obtenido de este análisis son los perfiles léxicos correspondientes a las clases creadas por el CHD. Este resultado es muy similar al obtenido por DtmVic mediante las **palabras características** (Figura 3.7) de cada clase, mostrando de nuevo un acercamiento entre ambos.


CHD | Perfiles | AFC

1 Clase 1 35/290 12.07%	2 Clase 2 39/290 13.45%	3 Clase 3 40/290 13.79%	4 Clase 4 35/290 12.07%	5 Clase 5 54/290 18.62%	6 Clase 6 48/290 16.55%	7 Clase 7 39/290 13.45%

n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	9	12	75.0	46.71	nom	niño	< 0,0001
1	11	18	61.11	43.49	nom	persona	< 0,0001
2	7	9	77.78	37.79	nom	ejemplo	< 0,0001
3	7	10	70.0	32.75	ver	afectar	< 0,0001
4	4	4	100.0	29.55	nom	ola	< 0,0001
5	4	4	100.0	29.55	nr	guatemala	< 0,0001
6	4	5	80.0	22.12	nom	calor	< 0,0001
7	4	5	80.0	22.12	nom	escasez	< 0,0001
8	3	3	100.0	22.09	nr	meteorológicos	< 0,0001

Figura 3.16: Perfiles léxicos asociados a las noticias.

3.3.4. Análisis Factorial de Correspondencias

El AFC se basa en la probabilidad de aparición simultánea de formas o palabras, cuya cercanía en el gráfico indica cercanía de significado, mismas que confirman lo encontrado con la clasificación jerárquica descendente. El AFC reúne las formas co-ocurrentes en grupos,

acercándolos o alejándolos de otros grupos de formas co-ocurrentes, a modo de identificar las asociaciones semánticas entre los mismos. Las distancias de las proyecciones de las formas sobre los ejes es proporcional su semejanza en cuanto a frecuencia de aparición en cada texto o noticia.

Se muestran a continuación los resultados del AFC con IRaMuTeQ, en tres gráficos. Los dos primeros se muestran en la figura 3.17, donde la figura izquierda confirma claramente la estructura mostrada en el dendrograma de la figura 3.15. La figura derecha muestra la posición de las noticias individuales en relación al contenido a las clases, misma que el programa hace evidentes al asignarles distintos colores en base al grupo en donde predominan las formas léxicas que contienen.

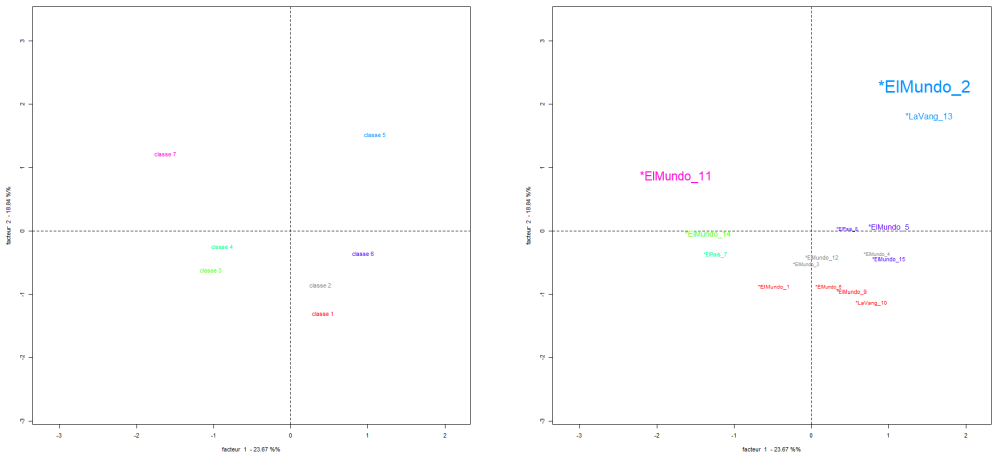


Figura 3.17: Análisis Factorial de Correspondencias sobre las clases (izquierda) y sobre las noticias individualmente (derecha).

Finalmente, en el gráfico 3.18, se muestran las palabras más comunes mapeadas de acuerdo a las clases a las que pertenecen. Esta figura permite una interpretación conjunta de los gráficos de la figura 3.17 y el dendrograma mostrado en la figura 3.15; su intepretación es muy similar a la obtenida en DtmVic mediante el uso de las palabras características asociadas a cada clase.

3.3.5. Nube de Palabras

La nube de palabras (Figura 3.19) no es en sí mismo un análisis léxico, pero se incluye dado que es interesante desde el punto de vista estético. Representa las formas en proporción directa a su frecuencia de aparición. La interpretación obtenida de este gráfico es similar a la mostrada en el análisis de similitudes anterior en la figura 3.14, donde las formas con mayor frecuencia y relación aparecen más juntas y en mayor tamaño de fuente.

Figura 3.18: Análisis Factorial de Correspondencias sobre las formas léxicas.

3.4. Comparación de aspectos específicos

Tras la discusión de los análisis léxicos de ambos programas, podemos proceder a discutir aspectos específicos. Para empezar, la figura 3.20 muestra una comparación de las características específicas de ambos programas. De esta forma, se muestra que ambos programas realizan análisis similares, mostrando una mayor variedad de técnicas estadísticas en DtmVic y unas capacidades gráficas de superiores en IRaMuTeQ.

También son de gran importancia los gráficos obtenidos mediante los análisis léxicos presentes; para comparar los mismos se han escogido los resultados de la representación factorial y la clasificación jerárquica (*clusters*), y se muestran en la figura 3.21. En este último caso, se observa que la estructura factorial obtenida por ambos métodos es muy similar, clasificando las noticias en base a sus temáticas como ha sido comentado en los apartados anteriores, y formando *clusters* en posiciones similares en ambos planos.

Sin embargo, se observan ciertas diferencias en las posiciones de los puntos en ambos planos, estas se deben principalmente a las elecciones de los umbrales mínimos de ocurrencias de las palabras en el *corpus*. Se recuerda que, en este caso, las noticias con DtmVic habían empleado un margen razonable de **cuatro** repeticiones; en el caso de IRaMuTeQ, la elección viene por defecto en **dos** repeticiones y no permite ser cambiado. Este hecho puede dar lugar a deformaciones en el plano y sus puntos, ya que la ponderación asociada a la distancia en el AFC está basada en el concepto de *masa*, que es inversamente proporcional a la frecuencia de aparición de una palabra; una explicación más detallada se encuentra en la sección pertinente del capítulo 2.

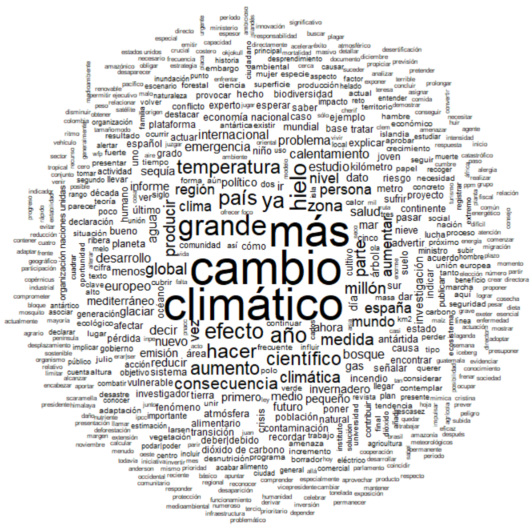


Figura 3.19: Nube de palabras sobre las formas léxicas presentes en las noticias estudiadas.

Por último, se muestra en la figura 3.22 los principales puntos fuertes y débiles de ambos programas, de forma que sean conocidos tanto sus posibilidades como sus deficiencias en ciertos aspectos. En este caso se destacan puntos fuertes similares, donde se aprecia que IRaMuTeQ es un programa con un punto fuerte basado en las representaciones gráficas mientras que DtmVic está basado en una importación de datos robusta y la aplicación de una mayor variedad de técnicas.

	IRaMuTeQ	DTMVIC
Herramientas	-Clusters -Análisis de frecuencias -Análisis de especificidades -Análisis Factorial de Correspondencias -Clasificación jerárquica descendente (método Reinert) -Análisis de similitud -Nubes de palabras	-Lematizador WinTreeTagger -Análisis de frecuencias -Análisis Factorial de Correspondencias - Análisis de Correspondencias Múltiple -Clasificación jerárquica (Ward y algoritmo chain-search) -Kohonen maps -Bootstrap -Minimum spanning tree -Nearest neighbours
Gráficos	-Gráficos 2D -Dendrogramas -Matriz de similitud -Frecuencias -Nubes de palabras	-Gráficos 2D -Dendrogramas (en archivo de resultados y mediante herramienta SplitsTree) -Tabla léxica (palabras x textos) -Kohonen maps -Bootstrap
Corpus de texto	-Textos literarios (novelas, cuentos) -Entrevistas -Cuestionarios -Noticias -Leyes -Preguntas abiertas	-Textos literarios (novelas, cuentos) -Entrevistas -Cuestionarios -Noticias -Leyes -Preguntas abiertas

Figura 3.20: Características específicas de los programas comparados.

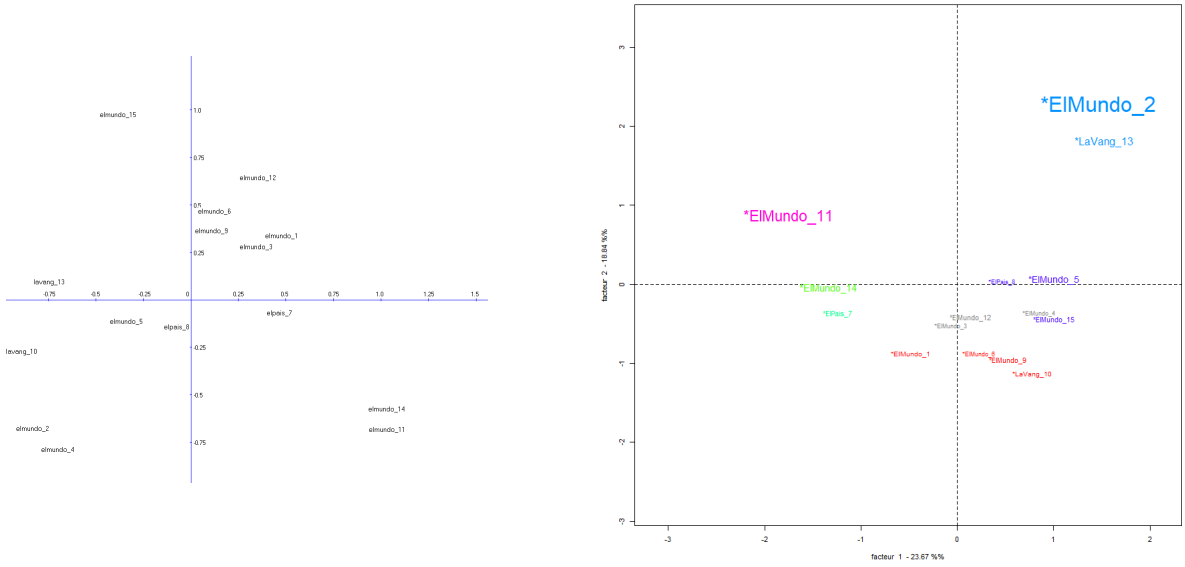


Figura 3.21: Comparación de gráficos de Representación Factorial en DtmVic (izquierda) e IRaMuTeQ (derecha).

	IRaMuTeQ	DTMVIC
Puntos fuertes	-Interfaz amigable con el usuario -Análisis bastante completo y robusto -Adecuado para investigación en ciencias sociales -Uso recurrente por parte de investigadores -Múltiples análisis léxicos -Gratuito -De código abierto -Compatible con múltiples sistemas operativos -Gráficos coloridos y fáciles de interpretar -Múltiples idiomas (inglés incluido) -Manual en castellano	-Incluye herramientas externas para suplir pasos importantes en el análisis (WinTreeTagger, SplitsTree) -Incluye técnicas novedosas respecto a otros programas de su índole (Kohonen maps o bootstrap) -Empleado por investigadores -Análisis léxicos basados en AFC y ACM -Gratuito -De código abierto -Compatible con Windows (plataforma de uso en Macintosh & Linux disponible) -Interfaz en inglés -Manual en castellano
Puntos susceptibles de mejora	-No es compatible con las últimas versiones de R -Preparación del archivo de entrada es complicada -No hay nuevas actualizaciones desde 2014 -No hay soporte en caso de dudas o mal funcionamiento -No hay soporte para más que cuatro idiomas -Algunos menús siguen estando en francés	-Basado en un lenguaje de programación en desuso -Proceso inicial de comprensión del programa complejo -Las técnicas centrales del análisis son las usuales en el campo - Gráficos poco elaborados y visualmente faltantes -Soporte del lematizador para diversos idiomas pero el programa sólo permite lematizar en cuatro.

Figura 3.22: Comparación de los principales puntos fuertes y débiles de ambos programas.

Conclusiones

El presente trabajo tenía como objetivo explorar las capacidades de los programas DtmVic e IRaMuTeQ, crear manuales de instalación y uso que sirvan como complemento a los existentes en castellano y hacer un análisis comparativo, con la esperanza de contribuir a la difusión de los mismos entre la comunidad académica y los investigadores en el campo de las ciencias sociales que estén interesados en obtener datos cualitativos de cuerpos de texto. Si bien ya existen trabajos académicos en donde se menciona el uso de ambos, ninguno hasta la fecha explora todas las posibilidades de los mismos. Los dos programas para análisis léxico estudiados deben ser utilizados como exploratorias de un *corpus* de datos, no se puede extraer de ellos conclusiones definitivas sobre el tema de investigación.

Al tratarse en ambos casos de software de licencia libre (GNU), ambos tienen la ventaja de ser gratuitos pero el grave inconveniente de no tener soporte ni actualización constante, en caso de problemas los usuarios son dejados a sus propios medios. Ambos requieren una preparación cuidadosa del *corpus* de texto a analizar (fichero o archivo de entrada) y dependen de otras plataformas para su funcionamiento: IRaMuTeQ de instalar la versión correcta de R, mientras que DtmVic está basado en el lenguaje Fortran.

Análisis de correspondencias: tanto un programa como el otro realizan análisis de correspondencias, comparando perfiles-fila y perfiles-columna para sintetizar las características distribucionales de las formas léxicas (palabras). Una comparación gráfico contra gráfico para el mismo *corpus* de texto se da en el Capítulo 3. Las distancias en ambos casos son calculadas con la Chi-Cuadrado.

Clasificaciones jerárquicas: ambos programas realizan clasificación jerárquica descendente, IRaMuTeQ por medio del algoritmo descrito por Reinert, y DtmVic por medio del algoritmo *chain search* de búsqueda de vecinos más próximos. Nuevamente, una comparación gráfica contra gráfico para el mismo *corpus* de texto se da en el Capítulo 3.

En ambos casos, ambos programas puede ayudar a un investigador en ciencias sociales a identificar las representaciones sociales presentes en los *corpus* de texto (monografías, artículos, entrevistas, etc.) simplemente identificando las palabras más usadas en los mismos. Los medios de comunicación hacen uso de mundos léxicos muy diferentes para abordar el mismo problema global: el cambio climático (queda en manos de los sociólogos juzgar si las representaciones encontradas corresponden a objetos en la realidad objetiva). En una comparación entre ambos, las conclusiones obtenidas son:

- Ambos programas no emplean técnicas novedosas: DtmVic tiene su base en estudios realizados en el año 1987 aunque sigue estando bajo supervisión (ya que su última actualización es del año 2020), e IRaMuTeQ está desarrollado en años posteriores y

es más reciente, pero se encuentra bajo una menor supervisión. En ambos casos, los programas emplean técnicas que se utilizan en la mayoría de los software de Análisis Textual, por lo que con el paso de los años no han introducido otros conjuntos de técnicas que se pueden emplear en este campo para obtener mejores análisis y resultados.

- Respecto a los gráficos obtenidos, es claro que IRaMuTeQ posee una interfaz visual más reciente y con un mayor número de posibilidades visuales al estar basada en R, mientras que DtmVic se encuentra basado en un código fuente antiguo y en un lenguaje actualmente obsoleto, por lo que sus gráficos han quedado desfasados respecto a los de IRaMuTeQ.

- Respecto a los análisis realizados, en el caso de IRaMuTeQ están principalmente basados en las herramientas gráficas disponibles (Análisis de Similitud, Nube de Palabras) mientras que en el estudio puramente estadístico se encuentra el AFC y el método de clasificación jerárquica desarrollada por Reinert. En el caso de DtmVic, a pesar de contar con herramientas gráficas que son necesarias para las interpretaciones finales, su fuerte se basa en la disponibilidad de diversas técnicas que amplían las posibles conclusiones obtenidas, como los SOM, el bootstrap sobre las palabras de los textos o el *minimum spanning tree*; en este sentido, las conclusiones obtenidas de este programa son más completas.

- Respecto a la importación de datos, en el caso de IRaMuTeQ sólo es posible pasando en primer lugar por un editor de textos y transformando manualmente los textos requeridos a su formato interno. Sin embargo, DtmVic permite importar directamente mediante herramientas internas archivos en formato *.csv* provenientes de programas como R, SPSS o Excel, y permite la importación de archivos XML comúnmente empleados por programas de bases de datos de tipo SQL.

- Por último, respecto a la lematización, en el caso de IRaMuTeQ se realiza mediante la herramienta de Estadísticos Clásicos tras seleccionar el idioma empleado y las categorías de formas gramaticales que serán usadas activamente (o de forma suplementaria) o serán retiradas del análisis. En el programa DtmVic esta lematización ocurre mediante un programa externo denominado **WinTreeTagger** con soporte para diversos idiomas, aunque el programa finalmente sólo tiene soporte para los indicados en la figura 3.2.

Ambos programas son similares y requerirían una ampliación de sus técnicas a campos más novedosos, ya que tras su desarrollo no parece haber habido ninguna intención de ampliar o actualizar su repertorio de técnicas. Entre otras, los principales puntos que deberían ser considerados por parte de ambos programas son:

- **CLUSTERS**: el campo de estudio de las técnicas de clasificación es muy amplio y existen infinidad de métodos alternativos, muchos de ellos derivados de la necesidad de nuevos métodos en el contexto de *Machine Learning* o clasificaciones no supervisadas. Estos métodos solucionan problemas posibles (como es el tratamiento de *outliers*) y están optimizados para el trabajo con grandes bases de datos. Ejemplos de este tipo de algoritmos más recientes pueden ser: *two-step cluster* [18], CLARANS [19] o DBSCAN [20].

- **Tablas de tres vías**: en el caso de DtmVic, el análisis de tablas de contingencia de tres vías se realiza mediante el método **TALEX**. Sin embargo, dado que la simple yuxtaposición de tablas puede conducir a resultados erróneos debido a no considerar la estructura de covariación de las variables, es preciso indicar que existen una serie de métodos en la literatura conocidos como métodos STATIS precisamente desarrollados para este tipo de análisis, aunque deberían ser adaptados al caso del Análisis de Datos Textuales. Este hecho puede extenderse al caso de IRaMuTeQ, donde no se encuentra una alternativa similar para este caso.

- **Kohonen maps**: este tipo de técnicas de clasificación basadas en redes neuronales cobran mucha importancia en los temas de investigación actuales, por lo que se desarrollo es prácticamente necesario. El programa DtmVic está basado en los algoritmos empleados en los inicios de la técnica, mientras que en la literatura se han seguido desarrollando alternativas basadas en ellos que mejoran los tiempos de programación y las clasificaciones obtenidas. Estos métodos y sus variantes se han empleado en diversos campos de forma reciente, como pueden ser el Análisis de Datos Textuales basado en *corpus* [24], ecología en estudios de calidad del agua [25] o sobre la calidad de vida [26].

Por tanto, en este trabajo se concluye que los programas DtmVic e IRaMuTeQ constituyen potentes alternativas de código libre en el contexto del Análisis de Datos Textuales, un tema en constante desarrollo donde cada vez surgen más investigaciones al respecto. A pesar de las posibles limitaciones de los programas planteados, constituyen alternativas fiables frente a los programas competidores de pago para llevar a cabo investigaciones basadas en el análisis de textos, obteniendo resultados fiables que permitan extraer conclusiones acertadas sobre la investigación llevada a cabo. Ambos programas analizados tienen aplicación en múltiples campos, se sugiere algunos de ellos:

- Periodismo: análisis de artículos de prensa, discursos, debates políticos, leyes.

- Administración: análisis de reuniones, artículos de prensa relacionados a la economía, obtención de información estratégica de artículos publicados en Internet.

- Psicología y Sociología: análisis de entrevistas, estudios socio-económicos, *Currículums Vitae* de posibles empleados, estudio de representaciones sociales.

- Mercado y publicidad: procesado respuestas a preguntas abiertas, minutas de reuniones, estudios de posicionamiento de marca, encuestas.

- Medicina: análisis de entrevistas a pacientes y familiares o a trabajadores de la salud, perfiles psicológicos de pacientes.

Referencias y bibliografía consultada

[1] Pardo, C. E., Ortíz, J., & Cruz, D. (2012, Julio). Análisis de datos textuales con DtmVic. En XXII Simposio Internacional de Estadística (pp. 1-42).

[2] Benzécri, J.-P. (1973). *L'Analyse des Données. Volume II. L'Analyse des Correspondances.* París: Dunod.

[3] Lebart, L., & Piron, M. (2012). Exploring Numerical and Textual Data in Practice with Dtm-Vic.

[4] Moreno, M., & Ratinaud, P. (2015). Manual uso de Iramuteq. Versión 0.7 alpha 2. Recuperado a partir de `http://iramuteq.org/documentation/fichiers/guiairamuteq/at_download/file`

[5] Morineau, A., Lebart, L., & Warwick, K. (1984). Multivariate descriptive statistical analysis and related techniques for large matrices. Chapter 4: Correspondence Analysis.

[6] Reinert, M. (1983). Une méthode de classification descendante hérarchique: application à l'analyse lexicale par contexte. Les cahiers de l'analyse des donnes 8, (2), 187-198.

[7] Reinert, M. (1998). Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste. En JADT (pp. 557-569). Recuperado a partir de `http://w3dev.ualg.pt/~lfaisca/SMAD03/JADT_Reinert_1998.pdf`

[8] Reinert, M. (2003). Le rôle de la répétition dans la représentation du sens et son approche statistique par la méthode ALCESTE. Semiotica-La Haye Then Berlin-, 147(1/4), 389-420

[9] Jodelet, D. (1985). La representación social: fenómenos, conceptos y teoría. En J.-C. Abric & S. Moscovici (Eds.), Psicología social (pp. 469-494). Barcelona [etc.]: Paidós.

[10] Moscovici, S. (1961). El psicoanálisis, su imagen y su público. Buenos Aires: Huemul.

[11] León, M. (2002). Representaciones sociales: actitudes, creencias, comunicación y creencia social. En: Psicología Social: Buenos Aires: Prentice Hall.

[12] Marin, Z. C. O. (2006). Contribuciones al análisis de datos textuales (Doctoral dissertation, Universidad de Salamanca). (pp. 33-48).

[13] Galindo, M. P. (2016). Análisis Factorial de Correspondencias. Máster de Análisis Avanzado de Datos Multivariantes y Big Data.

[14] Greenacre, M., & Blasius, J. (Eds.). (2006). Multiple correspondence analysis and related methods. CRC press.

[15] Morineau, A., Lebart, L., & Warwick, K. (1984). Multivariate descriptive statistical analysis and related techniques for large matrices. Chapter 5: Multiple Correspondence Analysis.

[16] Morineau, A., Lebart, L., & Warwick, K. (1984). Multivariate descriptive statistical analysis and related techniques for large matrices. Chapter 6: Clustering Techniques.

[17] Mellado, I. B. (2020). Análisis de Cluster. Máster de Análisis Avanzado de Datos Multivariantes y Big Data.

[18] Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. International conference on Knowledge discovery and data mining (pp. 263-268). San Francisco, USA.

[19] Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. IEEE transactions on knowledge and data engineering, 14(5), 1003-1016.

[20] Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery, 2(2), 169-194.

[21] Kohonen, T. (1998). The self-organizing map. Neurocomputing, 21(1-3), 1-6. https://doi.org/10.1016/S0925-2312(98)00030-7

[22] Kohonen, T. (1990). The self-organizing map. (neural network architecture) (technical). Proceedings of the IEEE, 78(9), 1464-1480. https://doi.org/10.1109/5.58325

[23] Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981-1997. Neural computing surveys, 1(3&4), 1-176.

[24] Zhao, X., Li, P., & Kohonen, T. (2011). Contextual self-organizing map: software for constructing semantic representations. Behavior Research Methods, 43(1), 77-88. https://doi.org/10.3758/s13428-010-0042-z

[25] Li, T., Sun, G., Yang, C., Liang, K., Ma, S., & Huang, L. (2018). Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes. Science of the Total Environment, 628-629, 1446-1459. https://doi.org/10.1016/j.scitotenv.2018.02.163

[26] Carboni, O., & Russu, P. (2015). Assessing Regional Wellbeing in Italy: An Application of Malmquist-DEA and Self-organizing Map Neural Clustering. Social Indicators Research, 122(3), 677-700. https://doi.org/10.1007/s11205-014-0722-7

[27] Lebart L. (2003) Validation Techniques in Text Mining. Text Mining and its Applications, Spiros Sirmakessis, Springer. 169-178.

[28] Lebart L. (2007) Which bootstrap for principal axes methods? Selected Contributions in Data Analysis and Classification, P. Brito et al. Springer, 581-588.

[29] Chateau F. , & Lebart L. Assessing sample variability and stability in the visualization techniques related to principal component analysis; bootstrap and alternative simulation methods. COMPSTAT 1996, Prat A. (ed), Physica Verlag, Heidelberg (1996), 205-210.

[30] Greenacre, M. Theory and Applications of Correspondence Analysis. Academic Press, London (1984).

[31] Gower J. C., & Dijksterhuis G. B. (2004). Procrustes Problems, Oxford Univ. Press, Oxford.

[32] Hill M. O. (1974). Correspondence analysis: a neglected multivariate method. Applied Statistics, 23, 340-354.

[33] Pukelsheim, F., & Simeone, B. (2009). On the iterative proportional fitting procedure: Structure of accumulation points and L1-error analysis.

[34] Huson D. H., Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution, 23 (2): 254 - 267. Software available from `www.splitstree.org`.

[35] Reinert, M. (1985). Un logiciel d'analyse lexicale [ALCESTE]. Cahiers de l'Analyse des Données, p. 471-484.

[36] Schmid, H. TreeTagger, a part-of-speech tagger. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[37] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, Intl. In Conference on New Methods in Language Processing. Manchester, UK.

[38] O'Duibhin, C. Windows interface for TreeTagger. `http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/`