Aplicación de técnicas de *Análisis Estadístico de Textos* a una encuesta con preguntas abiertas: comparación de los resultados obtenidos con post-codificación

Montevideo, Uruguay
Abril 2008

Facultad de Ciencias Económicas y de Administración.
Licenciatura en Estadística.

UNIVERSIDAD DE LA REPUBLICA
URUGUAY

Silvina Rodríguez
Daniel Alessandrini
Tutor: Ramón Álvarez

Índice general

Índice general i

I Introducción 1

1. Planteamiento del Problema 2
1.1. Introducción . . . 2
1.2. Objetivos . . . 3
1.2.1. Objetivo general . . . 3
1.2.2. Objetivos específicos . . . 3
1.3. Antecedentes . . . 4

II Aspectos Metodológicos 6

2. Datos 7
2.1. Datos . . . 7
2.1.1. Descripción de la base . . . 7
2.1.2. Tratamiento de variables . . . 8
2.2. Breve reseña del trabajo realizado por la CPP sobre la Salud Mental de los niños uruguayos . . . 9
2.2.1. Muestra . . . 10
2.2.2. Cuestionario . . . 10

3. Marco Teórico 12
3.1. Técnicas de Análisis Multivariante . . . 12
3.1.1. Un Poco de Historia . . . 12
3.1.2. Análisis Factorial . . . 13
3.1.3. Análisis Factorial de Correspondencias . . . 16
3.1.4. Análisis de Clusters o Clasificación de Grupos . . . 19
3.2. Lingüística y Estadística: ¿Complementarias u Opuestas? . . . 19
3.2.1. Nociones de Lingüística . . . 19

3.2.2. Relaciones entre Lingüística y Estadística . . . 21
3.3. Análisis Estadístico de Textos . . . 22
3.3.1. Reseña Histórica . . . 22
3.3.2. Metainformación . . . 24
3.3.3. Preguntas abiertas "versus" preguntas cerradas . . . 25
3.3.4. Unidades Léxicas y segmentación del texto . . . 27
3.3.5. Documentos Lexicométricos . . . 29
3.3.6. Tablas Léxicas y de Segmentos Repetidos . . . 36
3.3.7. Análisis de Correspondencias sobre Tablas Léxicas . . . 36
3.3.8. Análisis de Grupos sobre Tablas Léxicas . . . 38
3.3.9. Visualización de Datos Textuales . . . 40
3.3.10. Elementos Característicos y Respuestas Modales . . . 44
3.3.11. Otras Técnicas Situadas en el ADT . . . 49

III Resultados **60**

4. ACM Preguntas Abiertas Post-Codificadas **61**
4.1. Análisis de Correspondencia Múltiple . . . 61
4.1.1. Descripción del Análisis . . . 62
4.1.2. Análisis de Cluster . . . 66
4.2. Análisis de Correspondencia Simple . . . 68

5. Análisis Textual de Preguntas Abiertas **71**
5.1. Visualización de tablas léxicas . . . 71
5.2. Correspondencia Múltiple . . . 75
5.2.1. Comparación de las respuestas postcodificadas y textuales . . . 77
5.3. Concordancias . . . 80

6. Conclusiones **83**
6.1. Recomendaciones . . . 84
6.2. Temas pendientes . . . 85
Bibliografía . . . 86

IV Anexos **89**

A. Glosario **90**

B. Definición de las variables **97**

C. Análisis descriptivo de las variables **106**

D. Salidas **109**
D.1. Análisis Correspondencias . . . 109
D.1.1. Primer análisis de corresponcencia múltiple . . . 109
D.1.2. Segundo análisis de corresponcencia múltiple . . . 113
D.1.3. Correspondencias Simples . . . 117
D.1.4. Clusters . . . 118
D.2. Análisis estadístico de textos . . . 119
D.2.1. Tablas léxicas . . . 119
D.2.2. Análisis de Correspondencia Múltiple . . . 121
D.2.3. Ejemplos de concordancias . . . 124

Índice alfabético **127**

Índice de figuras **128**

Índice de cuadros **129**

Agradecimientos

Difícil es creer cuando el momento llega... pero finalmente llegó. Ha sido sin dudas un período aleccionador para nosotros, desde diferentes puntos de vista: profesional, por plasmar en un trabajo tantas cosas aprendidas y aprehendidas a lo largo de toda una carrera; personal, por nutrirnos de fuerza en los momentos difíciles, sabiduría al darnos el lujo de equivocarnos y aprender de esos errores y alegría cuando las cosas salen como uno las planea.

Queremos, en primer lugar, agradecer a la Clínica de Psiquiatría Pediátrica (CPP) del Hospital Pereira Rossell –dependiente de la Facultad de Medicina de la UDELAR– por permitirnos acceder a los formularios y cedernos la base de datos, insumo indispensable para el presente trabajo.

Luego, a nuestro tutor Ramón Álvarez, por el apoyo brindado durante este largo y arduo camino, tanto de forma profesional como espiritual.

A nuestras familias y amigos, que supieron sobrellevar nuestra ausencia en todo este tiempo, aguantarnos cuando parecía que nunca íbamos a llegar, y porque siempre, pero siempre estuvieron allí.

Y finalmente, a *Pilar*, que nos acompañó en todo momento.

¡Gracias!

Resumen

El estudio de grandes bases de datos textuales es materia de sumo interés para investigadores de diversos campos. Los textos insertos en éstas pueden haber sido relevados mediante encuestas socioeconómicas o entrevistas, estudios literarios o políticos o ser parte de archivos históricos o de bases documentales. Consecuentemente, el objetivo de partida de este trabajo de pasantía es hacer una *"puesta a punto"* de alguna de las herramientas que faciliten la gestión y la descripción de corpus de gran tamaño y que permitan a su vez derivar información de ellos desde un punto de vista estadístico. Estas herramientas pertenecen a la mencionada técnica de *Análisis Estadístico de Textos*.

De este modo, se presenta un caso práctico particular: el análisis estadístico de un cuestionario utilizado para medir trastornos de conducta en niños y adolescentes donde se aplican, además de una batería de preguntas cerradas –compuestas por los ítems que conforman el test más preguntas sociodemográficas habituales–, dos preguntas de respuesta libre que, mediante la técnica de postcodificación serán analizadas en una primera instancia, y comparadas luego con los resultados obtenidos a partir de la instrumentación del análisis estadístico de textos.

Palabras clave: *Análisis Textual*, *Análisis Estadístico de Textos*, *Análisis de preguntas abiertas*, *Postcodificación*, *Técnicas multivariantes*.

Parte I

Introducción

Capítulo 1

Planteamiento del Problema

1.1. Introducción

Los seres humanos se distinguen de otras especies vivientes en cuantiosos aspectos como por ejemplo la capacidad de razonar. Pero sin duda, una de las diferencias sobresalientes está estrechamente relacionada con el lenguaje como forma de expresión, tanto personal como interpersonalmente. Por ejemplo, cuando se quiere hacer notar sentimientos sobre algún tema en especial o simplemente cuando se desea que alguien realice una tarea, se trasmite a través de palabras, frases, en forma de prosas o versos, o quizás en una simple esquela en una servilleta de papel.

A algo que parece tan simple o tan elemental, por muchos años no se le dio la real trascendencia que tiene: la Lingüística, ciencia que estudia el lenguaje y todo lo relacionado a éste, comenzó a tener relevancia recién a comienzos del siglo pasado, cuando pasó de ser simplemente una ciencia de carácter histórico –al comparar distintas lenguas en cuanto a su evolución u origen– a una disciplina formal en el sentido estricto de la palabra, enfocándose en cuestiones propias de la lengua *per sé*, considerada ésta como un "sistema de signos", dejando de lado al texto como su único objeto de estudio. Consecuentemente, el lenguaje comienza a tomar importancia como fuente de información ya no sólo desde el punto de vista histórico. Si bien hacia la antigüedad el interés por descifrar significados entre líneas de texto ya estaba presente, la subjetividad intrínseca en el punto de partida del análisis juega muchas veces en contra de éstos, en el sentido que el sesgo se introducía desde el momento que el o los investigadores decidían cuáles palabras contar y cuáles no, o bien qué partes de un corpus textual eran retenidas.

Por otro lado, la Estadística se ha desarrollado fuertemente en el siglo que pasó, hasta tal punto de constituir una ciencia imprescindible para numerosos aspectos de la vida práctica: difícil es hoy en día que alguien no pregunte cómo se comportará el clima al día siguiente o que muchos actores de la sociedad no saquen conclusiones a raíz del nivel de inflación estimado por organismos oficiales. Todo ello surge a través de la aplicación de modelos matemáticos y

estadísticos, que tienen como fin la obtención de información sobre cierta temática.

Curiosamente, dos ciencias con objetivos en común no cruzaron sus lazos hasta la segunda mitad del siglo XX, cuando un grupo de analistas franceses comenzaron a estudiar grandes bases de datos, cuya fuente eran numerosos textos literarios. De allí en más distintas técnicas han surgido en diferentes campos de aplicación: la búsqueda de información (relacionada con la informática), el análisis del discurso (vinculado a las ciencias sociales o la medicina), estimación del volumen de vocabulario (relacionado con la lingüística), por citar algunos ejemplos.

En el caso del trabajo de pasantía que se presenta a continuación, el objetivo central es, por un lado, introducir en modo general la Textometría o Análisis Estadístico de Datos Textuales destacando las principales ramas de investigación que funcionan en la actualidad, y por otro mostrar una aplicación particular, comparando la técnica de post-codificación para el análisis de respuestas a preguntas abiertas, ampliamente utilizada en cuestionarios de diversa índole, con la información obtenida a través de técnicas estadísticas descriptivas tales como el análisis de correspondencias sobre una tabla léxica o la creación y estudio pormenorizado de segmentos o locuciones artificiales. En este contexto, el insumo a utilizar es una base de datos proveniente de una encuesta sobre salud mental infantil.

Finalmente, este trabajo se estructura de la siguiente manera: a continuación, se exponen los objetivos del mismo (Sección 1.2), junto con los antecedentes en la materia (Sección 1.3). En la segunda parte, se describen los datos utilizados (Capítulo 2) y el marco teórico y conceptual (Capítulo 3). Por último, en la tercera parte se exponen los resultados (Capítulos 4 y 5) y las conclusiones obtenidas a partir de éstos (Capítulo 6).

1.2. Objetivos

La pregunta de investigación del presente trabajo se centra en saber si los resultados del análisis de preguntas abiertas postcodificadas coincide en alguna medida con los resultados del análisis de las mismas, mediante técnicas de análisis de datos textuales.

1.2.1. Objetivo general

Introducir algunas de las diferentes técnicas y aplicaciones del análisis estadístico de textos, AET.

1.2.2. Objetivos específicos

- Aplicación de las técnicas antes mencionadas a un caso particular: respuestas a preguntas abiertas en un cuestionario autoadministrado.

- Comparar los resultados obtenidos mediante estas técnicas con los obtenidos luego de poscodificar las respuestas abiertas.

1.3. Antecedentes

A nivel mundial, los primeros esbozos en estudios de textos se remontan a la antigüedad. En Alejandría los grámaticos llegaron a elaborar listados de los *hápax*[1] de Homero, además de inventariar todas las palabras de La Biblia. Más cerca en el tiempo, los lingüistas anglosajones se han dedicado, durante el primer tercio del Siglo XX, al análisis de las concordancias (contexto discursivo que figura alrededor de una palabra) de ciertos vocablos en los grandes autores literarios. Recientemente, el enfoque se centró en descubrir leyes empíricas acerca de la distribución de las palabras en determinado corpus. Los trabajos realizados por Zipf (*The psychobiology of language, an introduction to dynamic philology* en 1935), Yule (*A statistical study of vocabulary* en 1944), Guiraud (*Les caractères statistiques du vocabulaire* en 1954) y Müller (*Initiation aux méthodes de la statistique linguistique* en 1973) establecieron un incremento sustantivo en la formalización de los estudios sobre el *léxico* empleado en los textos. Se establecieron leyes empíricas respecto a las características del vocabulario, como la *ley de Zipf* que se explicará más adelante.

En los últimos años, la rápida expansión del acceso a la informática así como también el desarrollo de nuevas técnicas de investigación –en particular las que nacieron de la interacción entre diferentes disciplinas– han permitido un crecimiento sostenido del análisis de textos en sus más diversos formatos. Podemos citar por ejemplo la detección de unidades físicas dentro de los textos; tarea que era, hasta hace poco, tediosa por la gran cantidad de tiempo empleado en el recuento de las mismas. Otro ejemplo es el *tiempo léxico*, línea de estudio emparentada con el análisis de series temporales en donde se mide la evolución del vocabulario empleado en los textos producidos a lo largo del tiempo. También la inteligencia artificial a través de las aplicaciones para estudiar textos en tiempo real (para mejorar el "diálogo hombre-máquina") ha crecido notoriamente.

En América Latina, la técnica de Análisis Estadístico de Textos ha ganado adeptos. En algunas universidades sudamericanas[2] existen cursos ya sea de grado o postgrado en donde se imparten los conocimientos básicos sobre la materia. Desde el punto de vista de las investigaciones realizadas, podemos decir que varias son las áreas alcanzadas por los diferentes cientistas: investigaciones de mercado, medicina, análisis del discurso, etc.

No obstante, en Uruguay las tendencias de la región no han sido tomadas. Si bien existen trabajos referidos al análisis textual, éstos no hacen referencia a la técnica estadística citada líneas arriba, sino a una definición más general de una rama de aplicación de la lingüística[3], donde

[1] Vocablo que aparece sólo una vez en un texto.

[2] En Argentina y Colombia, entre otras.

[3] "*Análisis Textual estratégico en el cuento 'Continuidad de los Parques', de J. Cortázar*", de Claudia Rodríguez Reyes es

básicamente se realiza una lectura exhaustiva de un texto, se cuenta la cantidad de palabras, se buscan relaciones sintagmáticas o de otra índole, pero en ningún momento aparecen mencionadas, por ejemplo, las nociones de distancia entre elementos (palabras dentro de un corpus) o la asociación entre elementos de un texto y variables que agrupen individuos según determinadas características. En este contexto, es vital tener presente la diferencia entre "análisis textual" –como un concepto más general– y "análisis estadístico de textos" o simplemente "análisis de datos textuales" –aplicación de la ciencia estadística a la lingüística–.

un buen ejemplo. Disponible en http://www.ucm.es/info/especulo/numero34/cparques.html (versión hipertexto) o en http://es.geocities.com/cuentohispano_zip2/articulos/art49.pdf (versión PDF).

Parte II

Aspectos Metodológicos

Capítulo 2

Datos

2.1. Datos

El insumo utilizado para la realización del presente trabajo monográfico, es un estudio que se centra en obtener información epidemiológica respecto a la salud mental infantil en nuestro país. El mismo fue realizado por un comité interdisciplinario de investigación, formado por miembros de la Clínica de Psiquiatría Pediátrica, el Departamento de Métodos Cuantitativos, ambos de la Facultad de Medicina; el Instituto de Estadística de la Facultad de Ciencias Económicas, todos ellos de la Universidad de la República (UDELAR) y una extensa colaboración de equipos de investigación extranjeros.

Dicha base comporta información extraida del "Cuestionario sobre el comportamiento de niños de 6 a 18 años"[1].

Para el procesamiento de toda la base fue utilizado el *Data and Text Mining (DTM)*, software desarrollado por Ludovic Lebart y colaboradores. Este es considerado la "versión libre" del renombrado software *Système Pour l'Analyse des Données (SPAD)*.

2.1.1. Descripción de la base

La base original a la cual se tuvo acceso contenía 1308 individuos y 227 variables, que se dividían en tres bloques: el primero contenía variables tales como edad, grado, sexo, etc. del niño; el segundo hacía referencia a las preguntas específicas del CBCL[2] y la última provenía de un cuestionario complementario en donde se relevaba información sobre vivienda, composición familiar e instrucción de los padres.

Esta base fue sometida a diferentes cambios, hasta llegar a la definitiva consistente en la mis-

[1] Ver Anexo B. Se recuerda que, aunque dicho cuestionario está enfocado a niños de 6 a 18 años, en este caso particular la muestra apuntó a niños en edad escolar, con lo que el rango de edades es sensiblemente menor.

[2] Siglas de Child Behavior Checklist.

ma cantidad de individuos, pero un número sensiblemente menor de variables, 19. Las variables descartadas tenían un enfoque que escapaba a los objetivos del presente estudio, con lo cual su remoción no incidió de manera alguna en los resultados.

Para aplicar la técnica de análisis de correspondencias se procedió al pre-tratamiento usual de las variables: convertir a todas aquellas variables cuantitativas en cualitativas. Por ejemplo, recodificar la edad en tramos, cruzar variables –como es el caso de *Edad-Sexo*–, codificar las respuestas a preguntas abiertas, como ocupación de los padres o las preguntas A y B[3].

Cabe destacar que no se optó por excluir individuos de la base, afín de preservar la información contenida sobre todo en las dos preguntas abiertas, analizadas en primera instancia post-codificadas y luego en su estado natural.

2.1.2. Tratamiento de variables

Postcodificación de preguntas abiertas

En la base de datos original, las respuestas a las dos preguntas abiertas A y B del cuestionario fueron digitadas resumiendo en pocas palabras la información contenida[4]. A raíz de esto se decidió volver a transcribir textual y totalmente estas dos respuestas, teniendo en cuenta correcciones ortográficas pertinentes, para luego someterlas a post-codificación.

Construcción de otras variables

Edad

En esta variable se han detectado numerosos errores de digitación. Para corregirlos, se tomaron en cuenta otras variables tales como *fecha de realización de la encuesta*, *fecha de nacimiento*, *grado que cursa* y *cantidad de años repetidos*, si corresponde. Asimismo, se detectó una cantidad nada despreciable de datos faltantes en la variable *fecha de nacimiento*. En estos casos se verificó la edad del niño y, en caso de detectar errores, se utilizó como apoyo el mismo procedimiento antes mencionado.

Por último, se convirtió esta variable en categórica, recodificándola en 3 tramos: de 5 a 7 años, 8 y 9 años y de 10 a 13 años. Estos intervalos fueron seleccionados en función de la cantidad de observaciones al interior de cada uno: aproximadamente un tercio de niñas y niños caen en cada intervalo[5].

[3] Ver anexo de descripción de variables.

[4] Por ejemplo, si se respondió "Que lleva todas las situaciones hasta el limite. Los celos." a la pregunta *¿Qué es lo que más le preocupa de su hijo/a?*, los digitadores optaron por dejar solamente "Celos".

[5] Ver Anexo C.

Sexo-Edad

Con la variable *Edad* recodificada, procedióse a crear una nueva, cruzando a ésta con la variable dicotómica *Sexo*, formando así *Sexo-Edad*. La misma consta de 6 categorías.

Trabajo de los padres

En el formulario, la pregunta *Trabajo usual de los padres* aparece como abierta; es decir, no hay categorías definidas *a priori* a ser seleccionadas por el encuestado/a. Se procedió a su post-codificación, tomando como referencia el Manual Guía para la Codificación de Ocupaciones de Actividad (CIUO-88), adaptada a Uruguay (CNUO-95), a un dígito[18]. En una segunda instancia, se colapsaron alguna de las categorías registradas debido a su baja frecuencia.

2.2. Breve reseña del trabajo realizado por la CPP sobre la Salud Mental de los niños uruguayos

El instrumento de *screening* utilizado fue el *Child Behavior Checklist (CBCL)*, creado por los profesores Thomas Achenbach y Leslie Rescorla del Child Psychiatry Department de la Universidad de Vermont, EEUU[1]. Este cuestionario fue escogido por el equipo de investigación debido a su uso extendido a nivel mundial a la hora de hacer comparaciones en estudios epidemiológicos. Consta de un formulario autoadministrado por los padres, donde éstos evaluan aspectos relacionados al comportamiento y las relaciones sociales de sus hijos, valorando los últimos seis meses previos a la entrevista.

En la Introducción de la versión escrita de la primera presentación pública de este trabajo de investigación, se sintetiza que "la Salud Mental es un componente básico de la salud integral del individuo y la sociedad"[33]. En particular, en la niñez y la adolescencia se cimentan las bases para su desarrollo. Es claro que la realidad social y contextual de determinado país influye en esos cimientos, induciendo al individuo a cargar con distintas problemáticas a lo largo de su vida. Así, el enfoque del trabajo citado se dirige a establecer políticas en salud mental infantil para así posibilitarle a la población acceso a una mejor calidad de vida.

Paradójicamente, a nivel internacional los estudios epidemiológicos, fundamentales para poder establecer un punto de partida del problema, son tanto escasos como de díficil comparación entre ellos. En el caso del Uruguay, la inexistencia de estudios epidemiológicos resulta un gran obstáculo para planificar y organizar, en forma eficiente, políticas públicas en relación a este tema.

Se trazaron varios objetivos, siendo el primordial medir la utilidad y aplicabilidad en nuestro país de un instrumento de *screening*, técnica usada en Medicina para explorar cierta patología o enfermedad y aplicar, basada en las conclusiones extraídas del estudio, políticas para la preven-

ción y minimización de estas patologías o enfermedades. En función de éste, obtener información epidemiológica en salud mental y dejar las puertas abiertas para la formulación de nuevas hipótesis de trabajo en el tema.

Es importante resaltar que en este momento el trabajo se encuentra en fase preliminar, pues antes de la versión definitiva éste será sometido a una etapa de validación externa.

2.2.1. Muestra

Todos los aspectos concernientes a la muestra estuvieron determinados por el departamento educativo de ANEP. Se desarrolló un muestreo *PPS* (probabilidad proporcional al tamaño) en dos etapas. El tamaño de muestra a manejar fue de 1400 estudiantes de 1° a 6° año asistentes a escuelas de educación común de todo el país.

En la primera etapa, se sortearon 70 escuelas de educación primaria, con probabilidad proporcional al tamaño de cada una de ellas. El tamaño de la escuela venía dado por la matrícula total de 1° a 6° grado. En cada una de las escuelas seleccionadas, el equipo de investigación solicitó un listado de los alumnos asistentes. Con este marco de lista, originalmente se debía proceder a través de un muestreo sistemático seleccionando 20 niños por escuela, teniendo de esa manera un *Intervalo de Muestreo* (IM) variable. En la práctica, el equipo de investigación en el trabajo de campo relevó información de 72 escuelas, pero no se respetó la fracción costante de 20 alumnos por escuela, sino que se tomó un número variable, con lo cual para el posterior análisis se debió realizar una reponderación, al tener los tipos de escuelas manejados en la primera etapa de muestreo desbalanceados.

Esos nuevos pesos no fueron considerados en el análisis que se hace en este informe de pasantía, ya que este aspecto de ponderación de las respuestas fue desarrollado cuando ya se había avanzado notoriamente en dicho informe, y por otro lado el software usado no permite hacer esta ponderación. Este aspecto hace que los resultados encontrados deban ser tomados con precaución y necesariamente el análisis deba ser hecho nuevamente, teniendo en cuenta los pesos muestrales.

2.2.2. Cuestionario

Lo primero a tener en cuenta es que el instrumento usado (CBCL) no hace diagnósticos psiquiátricos, sino que recoge la percepción de los padres sobre la presencia o ausencia de síntomas comportamentales y emocionales de su hijo en los últimos 6 meses.

Su utilidad se centra en:

- establecer la media de la población y así poder comparar esta con los promedios de otras

poblaciones;

- comparar grupos dentro de una misma población;
- determinar líneas de corte para discriminar en áreas, según si existe o no patología, o si esos niños están en una "zona de riesgo".

En líneas generales, el estudio permite mostrar una alta correlación entre las dificultades en el aprendizaje y la presencia de problemas emocionales y de comportamiento en los escolares de todo el país, y dar cuenta al mismo tiempo de la escasa derivación de los casos patológicos en centros especializados, tal cual es esperable en estas circunstancias.

En cuanto a la validación del instrumento, existen ciertas diferencias con los percentiles para los puntos de corte entre la población uruguaya y los obtenidos en el estudio original.

Capítulo 3

Marco Teórico

3.1. Técnicas de Análisis Multivariante

Para establecer una definición en términos generales se puede decir, citando a Peña[30], que el análisis de datos multivariantes "comprende el estudio estadístico de determinado número de variables medidas en elementos de una población", siendo sus objetivos:

- Resumir los datos mediante un reducido conjunto de nuevas variables, construídas estas como transformación de las originales, con la mínima pérdida de información;
- Encontrar grupos en los datos, si existen;
- Clasificar nuevas observaciones en grupos definidos;
- Relacionar dos conjuntos de variables (por ejemplo, si se quiere buscar la conexión entre variables de comportamiento con otras sociodemográficas y en función de ello conocer cuántas dimensiones tiene esa relación)

3.1.1. Un Poco de Historia

El análisis multivariante es una herramienta de gran utilidad para muchas disciplinas. Su punto de partida es un tanto difuso, aunque se cita generalmente a Harold Hotelling (1895-1973) como el iniciador de los métodos de reducción de la dimensión. Atraído por la Estadística, Hotelling viaja en 1929 a la estación de investigación agrícola de Rothamsed, Reino Unido, para trabajar con el estadístico R. A. Fisher en investigaciones sobre tratamientos en suelos. Basado en éstas, descubrió la relación que estos estudios tenían con el problema al cual se enfrento Karl Pearson (1857-1936) unos años antes, cuando buscaba determinar si dos grupos de personas de las cuales se conocían ciertas medidas físicas, pertenecían al mismo grupo étnico. Siguiendo esta línea, Hotelling se contactó, luego de volver a los EEUU, con el profesor Truman Kelley el cual le planteó el dilema de encontrar los factores capaces de explicar los resultados obtenidos en un test de inteligencia. A raíz de esto Hotelling desarrolla el método de los *Componentes Principales*,

indicadores que resumen de manera óptima un conjunto amplio de variables y éstos a su vez generan el nacimiento del análisis factorial. Además, Hotelling generaliza esta idea al introducir el análisis de **correlaciones canónicas**, que permiten resumir simultáneamente dos conjuntos de variables o más.

Por otra parte, la primera solución al problema de la clasificación fue otorgada por Fisher en 1933, cuando buscaba resolver un problema de discriminación en antropología, partiendo de las medidas de un cráneo encontrado en una excavación, del cual se quería saber si pertenecía a un homónido o no. Entonces, basándose en el análisis de varianza, Fisher encontró una variable indicadora, combinación lineal de las originales que conseguía máxima separación entre dos poblaciones.

Todas las ideas anteriores tenían su entorno de aplicación en variables numéricas, hasta que en la década de 1940 se comienza a imputar también en variables de tipo cualitativas. Autores como Fisher (en 1940), Guttman (en 1941) son los precursores en este sentido. Mientras el primero utilizó sus ideas de *análisis discriminante en tablas de contingencia*, el segundo presentó un procedimiento para construir escalas numéricas en variables cualitativas, relacionado ésto de alguna forma con los trabajos previos de Fisher. Años mas tarde, Jean Paul Benzécri, estadístico y matemático francés, introduce un enfoque *geométrico* del análisis de correspondencia, prescindiendo ademas de distribuciones de probabilidad y métodos inferencistas muy usados hasta ese momento a la hora del análisis, logrando así generalizar varios conceptos vertidos previamente por otros autores.

Es claro además que la revolución informática que se desató desde la pasada década de 1970 ha contribuido a una radical transformación en el mundo del análisis multivariante, pasando de una visión simplista del análisis de datos a otra más cercana a la realidad. Escuetamente podemos decir que hoy en día la transformación de la disciplina apunta en dos direcciones: por un lado, al disponer de grandes masas de datos, algunas aplicaciones conducen al desarrollo de métodos de aproximación local, que no requieren hipótesis generales sobre el conjunto de observaciones; y por otro, se prescinde de las hipótesis sobre la distribución de los datos y se cuantifica la incertidumbre mediante métodos de cálculo informático intenso.

3.1.2. Análisis Factorial

Desde un punto de vista general, la Estadística tiene como objetivos resumir, simplificar y -eventualmente- explicar. Desde los albores de esta disciplina, se han desarrollado modelos matemáticos que trataban de representar fielmente el fenómeno en estudio, pero en muchos de los casos los investigadores centraban sus ideas en cómo los datos se ajustaban al modelo en cuestion. Para evitar este problema y encaminarse hacia lo que Jean Paul Benzecri llamó el *segundo principio del análisis de datos*, el cual establece que "el modelo debe ajustarse a los datos, y no al re-

vés"[1], se recomendaba realizar un resumen descriptivo, previo al análisis riguroso y al desarrollo de modelos que expliquen el comportamiento de la base que es sometida a estudio. Dentro de estas "técnicas de resumen" resaltaban los gráficos, instrumentos que ayudan sin lugar a dudas a una mayor y mejor interpretación de la realidad a ser descrita, en comparación a los modelos numéricos. Gracias a ellos, el estadístico puede cumplir con los objetivos establecidos líneas arriba: resumir una gran cantidad de datos, simplificar la naturaleza de los mismos basándose en la habilidad innata del hombre en absorber imágenes rápidamente (*"una imagen vale mas que mil números"*) y generar una visión global del fenómeno en estudio, dejando la puerta abierta para posibles explicaciones. No obstante, un conjunto de histogramas (u otro tipo de figuras) para resumir adecuadamente la información resulta insuficiente, ya que se trabaja en espacios de dimensiones muy grandes. De todas formas, está claramente demostrado que la "relación costo-beneficio" de estas técnicas son favorables, ya que lo que se gana en interpretación excede con creces la pérdida de información resultante de aplicar las mismas.

Si bien las técnicas factoriales son muchas (Componentes Principales, Análisis de Correspondencia Simple o Múltiple, por citar algunos), los principios son únicos: partiendo de tablas rectangulares de individuos por variables, se generan dos nubes de puntos (fila y columna) y éstas a su vez son proyectadas sobre nuevos espacios, en los cuales la representación de filas y columnas están estrechamente relacionadas entre sí.

En tanto los insumos de los diferentes métodos factoriales son: el *espacio*, los *puntos*, las *masas* asignadas a los pesos y la *métrica*; los productos son: *ejes de inercia*, las *coordenadas* de los puntos sobre los ejes, a los que se denominarán factores, y diferentes indicadores que permitirán interpretar los resultados.

Ahora bien, los principales objetivos del análisis factorial se resumen en:

- obtener información sobre de la nube de puntos, eliminando la información redundante,
- reducir la cantidad de dimensiones tratadas, perdiendo la menor información posible,
- diferenciar lo mejor posible a los elementos entre sí

Procedimiento general del Análisis Factorial

Se parte de una matriz de datos rectangular X_{IJ}, donde I representa a los individuos, y J a las variables. A partir de está matriz se puede diferenciar dos nubes de puntos, nube de las filas y nube de las variables representadas en diferentes espacios, R^J y R^I respectivamente.

[1]El primer principio del análisis de datos -según Benzecri- establece que *"Estadística no es Probabilidad"*. Citado en [15].

NUBE DE LAS FILAS

Dada una nube de I puntos, que se denominará N_I, se busca un conjunto de ejes tales que la inercia de dicha nube proyectada en esos ejes sea máxima. Las coordenadas de los I puntos definen una función numérica sobre I, a la que se la llamará *factor*. De forma análoga, se repite el mismo procedimiento para la nube de columnas, N_J.

Las coordenadas x_{ij} de los I puntos, forman la matriz $\mathbf{X}$. Al espacio R^J se le asigna la métrica euclídea, que deriva de un producto escalar en donde la matriz es llamada $\mathbf{M}$.

Si $\mathbf{M}$ es diagonal, la distancia entre dos puntos i y k se escribe de la siguiente manera:

$$d^2(i,k) = \sum_j (x_{ij} - x_{kj})^2 m_j \tag{3.1}$$

donde m_j son los elementos de la diagonal de $\mathbf{M}$. A estos elementos se los denomina *pesos de las columnas*, ya que son los coeficientes que ponderan la influencia de cada columna en el cálculo de las distancias entre elementos.

Se define el siguiente producto escalar entre vectores asociado a $\mathbf{M}$:

$$\langle u, v \rangle_M = u^{'} M v = v^{'} M u \tag{3.2}$$

Hasta ahora se definieron las matrices $\mathbf{X}$ y $\mathbf{M}$, y los vectores u y v, aunque existe un tercer elemento que interviene en el cálculo de los ejes: los pesos de los puntos N_I, los que se denominarán p_i y pertenecerán a una matriz diagonal, D, matriz de *pesos de las filas*.

Proyección de una nube sobre un eje u

Se llamará F_s al vector de dimensión I, formado por las coordenadas de las proyecciones de los puntos de la nube N_I sobre u_s, siendo u_s un vector director de un eje de R^J.

$$F_s(i) = x_i^{'} M u_s \tag{3.3}$$

donde $x_i^{'}$ es la fila i de la matriz $\mathbf{X}$. Matricialmente, esto se puede escribir:

$$F_s = XMu_s \tag{3.4}$$

Inercia de la nube proyectada

Se puede decir entonces, que la inercia de la nube proyectada es:

$$\sum_i p_i [F_s(i)]^2 \tag{3.5}$$

que escrito de forma matricial queda de la siguiente manera es $F_s^{'}DF_s$. Cuando se sustituye F_s por su equivalente indicado en la fórmula (3.4), se obtiene:

$$Inercia = u_s^{'}MX^{'}XMu_s \tag{3.6}$$

Ahora bien, maximizar la inercia es equivalente a buscar un vector **u** unitario, según la métrica **M** de forma que haga máxima (3.6).

NUBE DE LAS COLUMNAS

El desarrollo en la nube de las columnas, es análogo al desarrollo en la nube de las filas: basta sustituir a X por su traspuesta, X' y la métrica y los pesos de la nube de columnas serán, respectivamente, los pesos y la métrica de la nube de filas.

Si se define G_s como el equivalente a F_s en la nube de columnas, se obtiene la siguiente relación:

$$u_s = \frac{1}{\sqrt{\lambda_s}} G_s \tag{3.7}$$

Ahora, se puede decir que la relación entre los factores F_s y G_s surge de la relación entre ejes y factores y puede escribirse de la siguiente manera:

$$F_s = \frac{1}{\sqrt{\lambda_s}} XMG_s \tag{3.8}$$

En función de lo anterior, el cuadro 3.1 es un resumen de las relaciones entre las dos nubes, sus ejes de inercia y sus factores:

3.1.3. Análisis Factorial de Correspondencias

El análisis factorial de correspondencia o simplemente análisis de correspondencia es un método descriptivo gráfico para representar tablas de contingencia, es decir aquellas matrices de datos donde se recoge la frecuencia absoluta de dos o más variables cualitativas en un conjunto de elementos.

Análisis de Correspondencias Simple (ACS)

Esta es una técnica que aplica un método factorial a datos de variables cualitativas organizados en tablas de contingencia.

Como su nombre lo indica, se intenta medir la correspondencia entre modalidades de dos variables cualitativas, es decir la proximidad o la lejanía entre ellas. Para ello, se verifica cuán cerca o lejos se está de la hipótesis de independencia,

	Nube N_I	Nube N_J
Espacio	R^J	R^I
Métrica	$M_{J\times J}$	D
Coordenadas	$X_{I\times J}$	$X^{'}$
Pesos	$D_{I\times J}$	M
Ejes de Inercia	u_s	v_s
Ecuación	$X^{'}DXMu_s = \lambda_s u_s$	$XMX^{'}Dv_s = \lambda_s v_s$
Norma	$\|u_s\|_M = 1$	$\|v_s\|_D = 1$
Factores	$F_s = XMu_s$	$G_s = X^{'}Dv_s$
Ecuaciones	$XMX^{'}DF_s = \lambda_s F_s$	$X^{'}DXMG_s = \lambda_s G_s$
Norma	$\|F_s\|_D = \sqrt{\lambda_s}$	$\|F_s\|_D = \sqrt{\lambda_s}$
Ortogonalidad	$\sum_i F_t(i)F_s(i)p_i = 0 \;\; si \;\; s \neq t$	$\sum_j G_s(j)G_t(j)m_j = 0 \;\; si \;\; s \neq t$
Inercia sobre el eje s	λ_s	λ_s
Inercia total	$\sum_s \lambda_s = \sum_i \sum_j x_{ij}^2 p_i p_j$	$\sum_s \lambda_s = \sum_i \sum_j x_{ij}^2 m_i m_j$
Fórmulas de transición	$u_s = \frac{1}{\sqrt{\lambda_s}} G_s$	$v_s = \frac{1}{\sqrt{\lambda_s}} F_s$

Cuadro 3.1: Relaciones entre la nube fila y la nube columna

$$f_{ij} = f_{i.} f_{.j} \tag{3.9}$$

siendo $f_{ij} = \frac{k_{ij}}{n}$ la frecuencia relativa –donde k_{ij} es el elemento correspondiente a la fila i y columna j de la tabla de contingencia–, $f_{i.}$ y $f_{.j}$ –los denominados *perfiles medios*– las proporciones de la población total que poseen la modalidad i y j, respectivamente.

La transformación de los valores absolutos en relativos no es suficiente para aplicar la técnica descrita. Para completar la cadena de cálculo, es necesario el siguiente eslabón: la construcción de los llamados *perfiles fila* y *perfiles columna*. En cada fila i se dividen las frecuencias relativas f_{ij} entre la marginal $f_{i.}$, y en las columnas entre $f_{.j}$. Así, son construidas las nubes de *perfiles fila* y *perfiles columna*, respectivamente, sobre las que se aplicará el análisis factorial.

En este caso, la distancia usada para medir la similitud entre dos perfiles será la distancia χ^2. Se presenta dicha distancia, en forma respectiva, para las filas y las columnas:

$$d^2(i,l) = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2 \tag{3.10}$$

$$d^2(j,k) = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ik}}{f_{.k}} \right)^2 \tag{3.11}$$

¿Por qué la distancia χ^2?

Respecto a la distacia euclídea, la χ^2 presenta una propiedad fundamental: la de *equivalencia distribucional*. Esta dictamina lo siguiente: dadas dos filas cualesquiera i y h, si se cumple:

$$\frac{f_{ij}}{f_{i.}} = \frac{f_{hj}}{f_{h.}} \quad \forall j, j = 1, 2, ..., J \tag{3.12}$$

es decir, que las filas i y h tienen el mismo perfil, entonces se dice que existe *equivalencia distribucional*. En otros términos, si dos filas son proporcionales, la distancia entre dos columnas cualesquiera no se modifica agrupando las dos filas en una sola, con peso igual a la suma de los pesos. Lo mismo vale para las columnas.

Análisis de Correspondencia Múltiple (ACM)

El análisis de correspondencia múltiple es otra técnica factorial. Las respuestas son almacenadas en grandes tablas de datos, que pueden ser de dos tipos:

1. **Tabla Disyuntiva Completa:** en sus filas se hallan los individuos y en sus columnas las modalidades de las variables en estudio. Al ser las categorías exhaustivas y mutuamente excluyentes, por cada variable el individuo k tomará sólamente un valor, con lo cual cada entrada se define del siguiente modo:

$$z_{ik} = \begin{cases} 1 & \text{si el individuo } i \text{ posee la modalidad } k \\ 0 & \text{en otro caso} \end{cases} \tag{3.13}$$

2. **Tabla de Burt:** esta matriz entrelaza las modalidades del total de variables del estudio. Es una yuxtaposición de todas las posibles tablas de contingencia que se pueden formar cruzando las diferentes variables 2 a 2. Siendo **Z** la TDC, se cumple la siguiente relación:

$$\mathbf{B = Z'Z} \tag{3.14}$$

Cabe destacar que los análisis realizados a partir de cada una de ellas son equivalentes.

Para la puesta en práctica del ACM intervienen tres objetos: los individuos, las variables y las modalidades.

- *Individuos*: uno de los objetivos del ACM es lograr la caracterización de los individuos a través de la siguiente noción de similitud: dos individuos serán próximos entre sí cuanto más modalidades compartan.
- *Variables*: existen dos posibles puntos de vista al respecto: estudiar de forma directa la relacion entre variables, o sintetizar todas las variables en un pequeño grupo, combinacion lineal de las anteriores.

- *Modalidades*: el estudio de las modalidades equivale a sintetizar sus similitudes. Las modalidades pueden ser vistas de dos formas: por un lado como variable indicadora definida sobre el conjunto de los individuos, aquí dos modalidades se parecen más cuando ellas están presentes -o ausentes- al mismo tiempo en una gran cantidad de individuos; y por otro lado, las modalidades pueden ser vistas como clase de individuos de los cuales se conoce su distribución en el conjunto de categorías, en este caso dos modalidades se asemejan tanto más cuando se asocian en mayor o menor medidad a las mismas modalidades.

3.1.4. Análisis de Clusters o Clasificación de Grupos

El análisis de clusters es una técnica exploratoria multivariante cuya finalidad es agrupar elementos en función de las semejanzas entre ellos. Cabe destacar que la clasificación obtenida dependerá de las variables consideradas.

Este análisis presenta tres problemas:

- **partición de los datos:** se dispone de datos heterogéneos y se busca dividir a éstos en un número prefijado de grupos, de manera que cada elemento pertenezca a uno y sólo uno de los grupos, todo elemento quede clasificado y cada grupo sea homogéneo al interior y heterogéneo al exterior de sí mismo.
- **construcción de jerarquías:** implica un orden de los datos en niveles, es decir que los niveles superiores contienen a los inferiores. Si bien –estrictamente hablando– este tipo de métodos no definen grupos directamente, la jerarquía construída permite obtener particiones de los datos en grupos.
- **clasificación de variables:** este estudio exploratorio inicial puede ser útil para reducir dimensiones. Las variables pueden clasificarse en grupos o estructuras jerárquicas.

3.2. Lingüística y Estadística: ¿Complementarias u Opuestas?

3.2.1. Nociones de Lingüística

Según la definición de la Real Academia Española, Lingüística es simplemente *la ciencia del lenguaje*. Esta disciplina se forma a su vez de cinco principales ramas, a saber:

1. **Gramática** : estudio de la forma (*morfología*) y de funciones (*sintaxis*) de elementos lingüísticos.
2. **Semántica**: estudio de significado y significación de elementos.
3. **Fonemática**: estudio de sonidos de una lengua (*fonología*) o de hablas concretas (*fonética*)
4. **Pragmática**: estudio de relaciones de implicación entre el acto de comunicación y uso concreto de cierta lengua.

5. **Normativa**: conjunto de normas que rigen la corrección del uso de la lengua, tanto oral (*ortología*) como escrito (*ortografía*).

Cada una de las disciplinas descritas son de un entramado sumamente complejo, derivando a su vez en ramas, sub–ramas, etc.. Sin pretender llegar al núcleo de éstas, se busca aquí una idea general de la Lingüística como ciencia para entender mejor hacia donde se apunta en el presente trabajo.

Cabe hacer una pequeña reseña histórica respecto a la evolución de la Lingüística como objeto de estudio.

Evolución de la Lingüística

Fase I - Lengua como objeto de especulación. Hacia tiempos remotos, las civilizaciones más avanzadas se centraron en un enfoque filosófico de las lenguas. Como ejemplo, los griegos establecían que el razonamiento sobre la condición original de la lengua prevalecía sobre el estudio de su funcionamiento. Esta tendencia dominó durante siglos el centro de la disciplina, hasta bien entrado el denominado "siglo de las luces".

Fase II - "Ciencia histórica". La lingüística se centra en el estudio de la evolución de las formas lingüísticas a lo largo y ancho de la historia. El origen de esta concepción se remonta hacia finales del Siglo XVIII, donde el descubrimiento del sánscrito como lengua -por parte de Occidente- genera toda una revolución de las concepciones originarias de las lenguas; a partir de este momento se formula la "hipótesis indoeuropeísta" la cual establece que gran parte de las lenguas europeas -y muchas de las asiáticas- tiene un origen común. Podría decirse que, en esta fase, la lingüística es equivalente a la gramática comparativa.[2]

Fase III - Constitución de la Lingüística en ciencia formal. A comienzos del siglo XX, el suizo Ferdinand de Saussure con su obra *Cours de Linguistique Générale* publicada en 1916, dotó a la disciplina de un objeto intrínseco: analizar a la lengua en sus elementos formales propios, sin hacer alusión a presupuestos históricos. En el citado trabajo, Saussure dictamina que *la lengua forma un sistema, compuesto de elementos formales articulados en combinaciones variables según ciertos enunciados*. Aparece con ello el concepto de "estructura" en el lenguaje, definiendo a éste como "sistema de signos". Además, la distinción entre "lengua" en sentido estricto -definido como la parte social del lenguaje- y "habla" -lo individual y accidental dentro del lenguaje- son avances muy importantes respecto de la visión meramente diacrónica[3] que imperaba hasta ese momento en el mundo. De esta forma, la Lingüística pasó a ocuparse de un postulado más general: la posibilidad de estudiar al sistema de la lengua en cierto momento de su evolución, y no solamente

[2]Ver anexo A.

[3]Que se desarrolla a través del tiempo.

compararlo con otras lenguas sin detenerse al interior de cada una. Todo esto devino en el surgimiento del Estructuralismo, escuela con gran peso en la lingüística desde los años 20 hasta fines de la década de 1950, donde si bien los postulados entre las distintas sub–corrientes partían de distintos orígenes (Círculo de Praga, Estructuralismo Norteamericano), tenían en común los principios de inmanencia[4], la utilización del método taxonómico o de clasificación y la concepción de la lengua como un conjunto de elementos definibles por relaciones y oposiciones que sostienen entre sí. Con la idea de solventar las limitaciones explicativas del enfoque estructuralista, Noam Chomsky –un destacado lingüista estadounidense– publica en 1957 *Syntactic Structures*, dando nacimiento a una nueva teoría, la Gramática Generativo-Transformacional (GGT), la cual define a la Gramática como un sistema de reglas formalizado con precision matemática. El foco de atención pasa a ser la lengua como producto de la mente del hablante, la capacidad innata (genética) para aprender y usar una lengua, la *competencia*. Según lo anterior, toda propuesta de modelo lingüístico debe adecuarse al problema global del estudio de la mente humana, por esta razón al generativismo se lo describe como una escuela mentalista o racionalista.

Al plantear a la lengua como un sistema autónomo, las corrientes saussureana y chomskiana –que en conjunto forman la denominada *escuela formalista*– chocan con la *escuela funcionalista*, corriente lingüística que toma fuerza a finales del Siglo XX. Los autores funcionalistas –algunos de los cuales proceden de la antropología o la sociología– consideran que el lenguaje no puede ser estudiado sin tener en cuenta su principal función: la comunicación humana. La figura más relevante dentro de esta corriente es el lingüísta holandés Simon Dik, autor del libro *Functional Grammar*. Esta posición funcionalista acerca la lingüística al ámbito de lo social, dando importancia a la pragmática, al cambio y a la variación lingüística.

La escuela generativista y la funcionalista han configurado el panorama de la lingüística actual: de ellas y de sus mezclas arrancan prácticamente todas las corrientes de la lingüística contemporánea. Tanto el generativismo como el funcionalismo persiguen explicar la naturaleza del lenguaje, y no sólo la descripción de las estructuras lingüísticas.

3.2.2. Relaciones entre Lingüística y Estadística

Si bien parecen ciencias completamente diferentes, existe una interesante relación entre las mismas. Más aún, la Lingüística interactúa con muchas otras ramas de investigación, muy familiares para los estadísticos.

Según Marcial Terrádez Gurrea[31], destacado Doctor en Filología Hispánica y profesor asociado del Departamento de Filología Española de la Facultad de Filología de la Universidad de Valencia, una división certera de la lingüística en ramas conexas es la siguiente:

[4]Expuesto de F. de Saussure, donde se destacan nociones estrictamente gramaticales en el estudio de cierta lengua, dejando de lado argumentos basados en otras disciplinas, como la lógica, la filosofía, etc..

- **Lingüística Matemática:** rama de la lingüística que busca la fundamentación metódica de esta ciencia y la formación teórica estricta, aplicando modelos matemáticos como las series de Fourier, entre otras.

- **Lingüística Estadística o Estadística Lingüística:** enfoque creado por los estilistas[5], dedicado al estudio cuantitativo de obras literarias de grandes autores; en ellas se compara el léxico entre distintos escritores, se mide la evolución y extensión del vocabulario empleado, por citar algunos ejemplos[6].

- **Lingüística Computacional:** actualmente se habla, según los partidarios de la concepción estricta de la disciplina, del procesamiento informático del lenguaje natural como equivalente de esta rama.

- **Ingeniería Lingüística:** es toda aquella aplicación potencialmente comercial que implique el uso de nuevas tecnologías y lenguas. Aquí se incluyen la edición electrónica (diccionarios, diarios, etc.), los productos multimedia, etc..

- **Lingüística Algebraica:** el objetivo de esta rama es investigar y definir gramáticas formales, basándose en métodos matemáticos para la fijación de teorías abstractas (teoría de conjuntos, de grafos, de funciones recursivas y de autómatas).

3.3. Análisis Estadístico de Textos

Brevemente se puede decir que el Análisis de Datos Textuales es un conjunto de técnicas estadísticas que permite la exploración y análisis de textos.

3.3.1. Reseña Histórica

A lo largo de la historia el hombre ha centrado su atención en el estudio minucioso de textos por distintas razones, ya sea por motivos lingüísticos o puramente retóricos, pero en todos estos casos sólo se prestó atención al tipo y volumen de vocabulario utilizado, no así a la descripción, interpretación y crítica del contenido de los mismos. Gracias al desarrollo de la informática y a la introducción de un marco teórico consistente, se amplían las posibilidades de análisis a otros campos más diversos, como por ejemplo al estudio de *preguntas abiertas* incluidas en encuestas, *discusiones de grupo*, *entrevistas* o *documentos diversos*.

El origen de esta técnica de análisis –desde el punto de vista formal– se remonta a la década de 1930, con los aportes teóricos de Zipf, Yule y Guiraud entre otros, en cuanto al descubrimiento

[5]Los estilistas son aquellos especialistas en lingüística que buscan caracterizar el estilo con que las distintas obras investigadas han sido escritas.

[6]Como se verá más adelante, Benzécri creó uno de los principales nexos entre ambas disciplinas, al introducir el análisis de correspondencias.

de *leyes empíricas* acerca de la distribución de las palabras. Con estos principios se logró resolver problemas planteados por los llamados "estilistas" que se centraban en el estudio de obras literarias de grandes autores franco parlantes.

Hacia la década de 1960, Jéan-Paul Benzécri plantea –en un curso de lingüística matemática impartido en la Facultad de Ciencias de la Universidad de Rennes– la técnica de Análisis Factorial de Correspondencias, método inductivo y algebraico útil para el tratamiento de grandes bases de datos (en principio lingüísticas), en base a las posibilidades de cálculo de los ordenadores de aquella época. El profesor Benzécri, influenciado por las corrientes lingüísticas de siglo XX –en particular por la escuela formalista–, toma los postulados de Chomsky y quien fuera maestro de éste, Zellig Harris –el cual afirmaba que todo o casi todo de una lengua puede obtenerse analizando los hechos distribucionales de verbos, sustantivos, etc., pero sin necesariamente recurrir al sentido de ésta–, aborda este nuevo método para el tratamiento de estas tablas de frecuencias. Las mismas tenían las siguientes características: estaban formadas por I nombres (de personas) en las filas, y J verbos o adjetivos en las columnas. Los respectivos perfiles fila y columna eran calculados y, si por ejemplo dos nombres tenían un mismo perfil, estos serían *sinónimos* (desde el punto de vista de asociación con los verbos). Tanto la propiedad de equivalencia distribucional[7], como la representación espacial simplificada y un elevado nivel de abstracción cuantitativa, hacían del análisis de correspondencias una herramienta poderosa a la hora de estudiar y extraer información de grandes corpus textuales.

A mediados de la década de 1970, el análisis de correspondencias comienza a tomar fuerza a nivel del análisis textual, cuando en los coloquios celebrados en Grenoble en 1973 y Montpellier en 1976 se dan a conocer distintos trabajos basados en esta técnica, teniendo como base datos textuales. Además, en este último año comienza la publicación de la revista *Les cahiers de l'analyse des données* fundada y dirigida por Benzécri hasta 1997, en la cual son recogidos diferentes artículos de investigación en lingüística y lexicología, utilizando técnicas del análisis de correspondencias.

A partir de la década de los 80, mientras el análisis de correspondencias trascendía fronteras tanto geográficas como de aplicación –sobre todo a partir de las publicaciones en inglés *Multivariate Descriptive Statistical Analysis*, de Lebart, Morineau y Warwick y el libro de Greenacre advocado al análisis de correspondencias (ver [15])–, otras técnicas comienzan a desarrollarse, como por ejemplo el análisis de clusters y la proyección de elementos suplementarios, técnicas complementarias a las antes mencionadas. Esto ayudó, desde un punto de vista práctico, a los interesantes debates que se centraban, por un lado, en establecer o no un umbral mínimo de frecuencias para las palabras, situadas en las filas de la matriz léxica, y por otro fijar reglas de segmentación que fijan estos mismos elementos (es decir, si quedarse sólo con las denominadas

[7] En este contexto, la *equivalencia distribucional* establace que, en caso de identificar dos verbos (j, k) que son sinónimos en cuanto a su distribución (tienen el mismo perfil), la distancia $d(i, l)$ *no* se altera si se reemplazan las columnas (j, k) por una nueva, j', suma de las anteriores y proporcional a ambas.

"palabras plenas" (verbos, sustantivos, adjetivos, etc.) y descartar o no las "palabras herramienta" (pronombres, preposiciones o artículos)). Además, los inventarios de segmentos repetidos comenzaban a sonar fuerte en el mundo lingüístico, ya que colaboraban en obtener mejores descripciones de los ejes factoriales.

Hacia 1990, el desarrollo de programas informáticos de diversa índole, junto con los métodos y sus aplicaciones a grandes bases de datos sufren un continuo desarrollo. Es en este año además que comienzan, con la primer conferencia en Barcelona, las *Journées Internationales d'Analyse Statistique des Données Textuelles*, *JADT*, ámbito de interacción permanente entre las ciencias relacionadas con el lenguaje y tantas otras, como la informática, la estadística, la matemática, etc.. Se realizan de forma bianual y la última fue realizada en Lyon, en el mes de marzo de 2008. Es en estos seminarios entre otros, donde se destacan las últimas tendencias en análisis de textos: particiones longitudinales de un corpus, series de tiempo textuales, análisis discriminante textual, semiometría, minería de datos, etc..

3.3.2. Metainformación

Para los estadísticos el texto se debe analizar a partir de recuentos. En la mayoría de las aplicaciones estadísticas se encadenan cuatro grandes etapas:

problema $\longrightarrow$ **datos** $\longrightarrow$ **tratamiento** $\longrightarrow$ **interpretación**

Cuadro 3.2: Cadena del tratamiento estadístico

donde cada etapa plantea diferentes problemas dependiendo del contexto, la problemática y el campo de aplicación[8].

- El **problema** en estudio puede dar lugar a la formalización "a priori" de un modelo estadístico, o ya venir formulado en términos generales
- Los **datos** pueden ser experimentales o pueden derivarse de la observación
- El **tratamiento** puede consistir en:
  - poner a prueba hipótesis o modelos, desde un punto de vista inferencial
  - una reorganización de los datos destinada a poner en evidencia sus rasgos estructurales más relevantes, si se trata de un análisis descriptivo o exploratorio
- La **interpretación** se reduce a tener en cuenta las conclusiones de una prueba de hipótesis. Para el caso de un análisis exploratorio, esta fase incluirá una reflexión sobre la validez de las estructuras observadas y, de manera eventual, un replanteamiento de las hipótesis formuladas.

[8]Basado en [28].

La necesidad de automatizar este proceso, con vista a implementar un tratamiento estadístico, ha llevado a introducir la noción de *metadata* o *metainformación*, que se puede decir, se trata de toda la información conocida sobre la matriz de datos y que no se encuentra en la propia tabla. Cuando se trata de datos de encuestas, la metainformación se utiliza para comprobar la coherencia de los datos. La metainformación es particularmente abundante en el caso de los datos textuales. A cada palabra utilizada, corresponden varias líneas, a veces varias páginas, en un diccionario enciclopédico.

3.3.3. Preguntas abiertas "versus" preguntas cerradas

La utilización de preguntas abiertas, supone un mayor esfuerzo de transcripción que cuando se trabaja con preguntas cerradas.

Se sabe que las palabras empleadas en el enunciado de las preguntas juegan un papel fundamental en un cuestionario relativo a actitudes u opiniones. Así, los trabajos de Rugg[9] muestran que la respuesta *"yes"* a la pregunta *"Do you think the United States should forbid public speeches against democracy?"* obtiene 21 puntos (sobre 100) menos que la respuesta *"no"* a la pregunta *"Do you think the United States should allow public speeches against democracy?"*.

Las listas de ítems juegan un papel positivo cuando se tiene que recurrir a la memoria. Por otro lado, es decir dejar la pregunta abierta, puede también jugar un papel positivo, ya que permite crear un clima de confianza y comunicación en ciertos temas.

¿Cuándo utilizar preguntas abiertas?

Uno de los usos más comunes de la utilización de preguntas abiertas es en la fase preparatoria de un estudio, en donde la finalidad consiste en poner a punto una batería de ítems de respuestas para una pregunta cerrada, pero es muy poco usada debido a su alto costo.

Existen al menos tres situaciones tipo para las cuales es bueno utilizar preguntas abiertas:

- *Para reducir el tiempo de la entrevista*: las preguntas abiertas son más económicas en cuanto a tiempo de entrevista se refiere, y hacen el clima de la entrevista más ameno.
- *Para recabar información espontánea*: Por ejemplo en los cuestionarios aplicados a estudios de mercado, las preguntas del tipo ¿que piensa Ud. de...?
- *Para explicitar y entender la respuesta a una pregunta cerrada*: se trata de la clásica pregunta complementaria "¿Por qué?".

En resumen, las respuestas a preguntas abiertas constituyen una prolongación indispensable de los cuestionarios cuando lo que se quiere obtener no son sólo conteos, sino que se va más allá y se trata de explorar y profundizar sobre un tema complejo o mal conocido.

[9] *Experiments in wording questions*, Public Opinion Quarterly, 5, 1941. Citado en [28].

ANÁLISIS DE PREGUNTAS ABIERTAS

Poscodificación manual de las respuestas abiertas

La técnica más habitual para el tratamiento de las respuestas abiertas, consiste en armar un conjunto de ítems a partir de una muestra de las respuestas, para luego codificar el conjunto de las respuestas abiertas de tal forma que se sustituye la respuesta abierta por una o más respuestas cerradas. De esta forma los datos son fácilmente explotables, no obstante, este tipo de tratamiento posee algunos defectos:

- *Mediación del codificador*. A la mediación del encuestador se le suma la del codificador que debe decidir qué interpretar;
- *Destrucción de la forma*. Se pierde la calidad de la expresión, el registro del vocabulario y la tonalidad general de la entrevista;
- *Empobrecimiento del contenido*. Cuando la pregunta permite respuestas de gran diversidad, la información queda sesgada por la poscodificación;
- *Las respuestas poco frecuentes se eliminan "a priori"*. Las respuestas raras, originales y/o poco frecuentes, son poscodificadas con códigos residuales, que por ser muy heterogéneos pierden valor.

Reagrupación de las respuestas libres o abiertas

El tratamiento de las respuestas libres constituye un reto tanto para los estadísticos como para otros especialistas de análisis textual. Limitarse sólo a recuentos es insuficiente; en cambio, reagrupar las respuestas por categorías permitirá contrastar los diferentes *"discursos artificales"* cuyo significado es más claro cuanto más homogéneas sean las categorías. Por lo tanto el problema ahora es cómo reagrupar las respuestas de manera de facilitar la interpretación de los reagrupamientos realizados.

¿Cómo reagrupar las respuestas?

Existen diferentes estrategias posibles para encontrar la agrupación pertinente. Se pueden utilizar los criterios considerados discriminantes en función del tema estudiado, o bien se puede buscar una partición que sea lo más universal posible teniendo en cuenta el tamaño de la muestra, por ejemplo por edad, por sexo, por región o nivel de instrucción, etc.. La operación de agregación de las respuestas facilita la lectura del texto original, aunque es útil disponer de ayuda para la comparación de textos obtenidos por reagrupamientos.

El análisis de respuestas reagrupadas

Este tipo de análisis es muy similar al análisis de textos literarios, políticos e históricos. Esto se debe a la gran variedad de reagrupamientos y, por tanto, al gran número de lecturas posibles.

Cabe destacar que la operación elemental de agrupamiento de respuestas, facilita mucho a la lectura del texto original. Es útil entonces disponer de una ayuda para la comparación de textos obtenidos por reagrupamiento. Para el analista será muy útil poder determinar palabras características de diferentes categorías, así como también diferenciar los grupos que se expresan de similar manera. Para esto, el material textual se debe preparar de manera que se puedan definir nuevas unidades que puedan éstas ser fácilmente reconocidas.

3.3.4. Unidades Léxicas y segmentación del texto

El método estadístico se basa en medidas y recuentos sobre objetos a ser comparados. Para poder aplicar este principio, es necesario definir ciertos lineamientos en pos de identificar las unidades a contar, además de implementar la fragmentación del texto en unidades mínimas, operación denominada *segmentación del texto*[10]. En una fase posterior, denominada *identificación*, se reagrupan las unidades idénticas.

La primera etapa del análisis refiere al recuento de las unidades léxicas, lo que entraña una dificultad ligada a la complejidad del léxico: éste se compone de un conjunto de unidades que se considera abierto. Además, el léxico varía entre locutores y, a su vez, en un mismo locutor puede variar de una situación a otra.

Incluso, los lingüistas se enfrentan a dificultades a la hora de definir el elemento de base del léxico. La unidad léxica sólo encuentra una definición rigurosa en su manifestación tipográfica.

A continuación se presentan los diferentes tipos de unidades léxicas.

Unidades Léxicas Simples

- *Forma Gráfica*: es, según Lebart y Salem[27], "una secuencia de caracteres no delimitadores (letras, en general) comprendida entre dos caracteres delimitadores (espacios o signos de puntuación)."
- *Lema*: son aquellos vocablos que cuentan con una misma raíz y un significado equivalente. Con este tipo de unidad se eliminan ambigüedades en las formas gráficas homónimas (ej. como –partícula comparativa– y como –del verbo comer–) y el hecho que no siempre la forma gráfica como unidad tiene correspondencia biunívoca con la palabra, unidad básica de lenguaje. El uso de plurales o singulares, los tiempos verbales y otros sufijos afectan a una palabra, y ésta puede aparecer con varias formas (ej. niño, niñas, niña).

  De esta manera, *lematizar* equivale a reagrupar las distintas formas que corresponden a un mismo lema, diferenciando las formas homógrafas y separando las palabras formadas mediante procedimientos aglomerativos. Para lematizar, se procede asi:

[10] Ver AnexoA.

- pasar los verbos a infinitivo
- los sustantivos a singular
- adjetivos al masculino singular

Unidades Léxicas Complejas

La necesidad de tener en cuenta las palabras compuestas, modismos y expresiones estereotipadas ha conducido a la búsqueda de estrategias incorporables a los análisis automáticos de textos. Los criterios más relevantes para poner a prueba la unidad de palabras compuestas son las siguientes:

1. **Inseparabilidad**: denota la imposibilidad de insertar otra unidad léxica en el interior. Ejemplo: no se puede decir *mesa de* ***mucha*** *luz*
2. **Conmutación**: permite sustituir un elemento por otro. Ejemplo: *la mesa de luz se vende en mueblerías*; en este caso *mesa de luz* se puede sustituir por *cama*, *placard*, etc.

Según Muller[11], estos criterios implican la definición de normas aceptadas tanto por el lingüista como por el estadístico, aunque la simplificación buscada por la estadística va, en general, contra la continuidad del lenguaje (en particular a nivel léxico). Por otra parte, en el caso informatico se requiere, por más software de alta calidad disponible, de la presencia del investigador para corregir posibles ambigüedades fruto de la complejidad del lenguaje.

Las siguientes definiciones buscan un procedimiento automatizable, totalmente independiente de la lengua y del investigador, ya que dependen unicamente de criterios graficos.

- *Segmento Repetido*: es una secuencia de dos o más palabras, no separadas por un delimitador de secuencia, que aparecen más de una vez en un corpus de datos textuales. Estas unidades sirven para recomponer secuencias clave de un discurso y reducir la ambigüedad asociada a la polisemia[12]. Ejemplo: *crema de enjuague*
- *Cuasisegmento*: según Bécue, estas unidades se constituyen de palabras que aparecen en una determinada secuencia, existiendo entre ellas una distancia máxima de separación fijada, medida en número de palabras. Con esta definición se evitan alteraciones dadas por la inclusión de ciertas palabras en segmentos repetidos. Ejemplo: *formación de profesores* y *formación de los profesores* son diferentes, pero ambos constituyen el mismo cuasisegmento.

Con esto se tiene en cuenta las expresiones estereotipadas en las cuales no es aplicable el criterio de inseparabilidad. Ahora bien, las unidades complejas antes mencionadas producen, a la hora de fragmentar el texto, formas léxicas que se solapan entre sí.

[11] Citado en [28], pág. 45.
[12] Ver anexo A.

Diccionarios de Palabras Compuestas y Locuciones

Para facilitar la segmentación de cualquier texto en *unidades léxicas*, algunos autores tales como Bolasco y Morrone, proponen el uso de "*corpus* de referencia" para complementar los criterios estadísticos, a la hora de evitar la creación de unidades léxicas en demasía. A modo de ejemplo, la Real Academia Española[13] dispone de dos grandes corpus textuales, tales como el Corpus de Referencia del Español Actual (CREA, escrito y oral) y el Corpus Diacrónico del Español (CORDE). Ambos conjuntos son complementarios ya que mientras que el CREA contiene textos pertenecientes a los últimos veinticinco años de historia del español, el CORDE incluye textos de todos los períodos anteriores.

3.3.5. Documentos Lexicométricos

EL ESTUDIO CUANTITATIVO DEL VOCABULARIO

El corpus

El corpus *P*, presentado a continuación, sirve para mostrar las diferentes definiciones referidas a este tema[14]. Dicho corpus está formado por un único texto, donde las diferentes palabras vienen representadas por letras mayúsculas, ya que para lo que sigue no tiene importancia conocer las palabras exactas.

El "CORPUS" P

A	B	C	D	A	E,	B	C	D.	B	A	E	D	A	C	F.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Cuadro 3.3: Ejemplo de un Corpus

Como se puede observar, este corpus tiene dieciseis *ocurrencias* numeradas de 1 a 16; algunas son ocurrencias de la misma palabra. La letra *T* designa el tamaño del corpus, en este caso entonces *T=16*. Además, el vocabulario *V* del corpus se compone por seis palabras distintas.

Frecuencias

En el corpus *P* la palabra E tiene dos ocurrencias, por lo que se puede decir entonces, que su frecuencia es igual a dos y las *direcciones* en el corpus de dicha palabra son: 6 y 12.

En este corpus, la palabra *A* tiene 4 ocurrencias y es la palabra más frecuente, por lo que se puede decir que la frecuencia máxima es 4:

[13]Por más información referirse a: http://www.rae.es/rae/Noticias.nsf/Portada3?ReadForm&menu=3

[14]Este ejemplo se basa en el texto de L. Lebart, A. Salem y M. Bécue, ver [28]

$$F_{max} = 4$$

Las palabras cuya frecuencia sea igual a 1 se denominarán *hápax*. Así con V_i se denota el efectivo de palabras de frecuencia i. En este ejemplo:

$$V_1 = 1 \qquad V_2 = 1 \qquad V_3 = 3 \qquad V_4 = 1$$

Las frecuencias y los correspondientes efectivos mantienen relaciones generales. La suma de los efectivos correspondientes a cada una de las frecuencias es igual al tamaño del vocabulario del corpus, y se escribe de la siguiente manera:

$$\sum_{i=1}^{Fmax} V_i = V \tag{3.15}$$

Si se quiere conocer el tamaño del corpus, basta con sumar los productos frecuencia x efectivos para todas las frecuencias desde 1 a F_{max}:

$$\sum_{i=1}^{Fmax} V_i \times i = T \tag{3.16}$$

Índice de un corpus o Glosario

Un índice es una reorganización de las palabras y ocurrencias de un texto. Puede ser útil para una rápida localización de las ocurrencias de cada palabra. Para esto, se utiliza un sistema de coordenadas numéricas relacionadas con la edición del texto: volumen, página, línea y posición en la línea. En un índice, las palabras pueden ser ordenadas de diferentes maneras

- *índice alfabético*, donde las palabras aparecen ordenadas alfabéticamente como su propio nombre lo indica
- *índice jerárquico*, donde las palabras aparecen ordenadas por frecuencia decreciente, si dos palabras tienen la misma frecuencia se recurre al orden alfabético.

Contextos, concordancias

Una vez localizadas las ocurrencias de cada unidad léxica en el texto, puede resultar interesante estudiar sistemáticamente los contextos inmediatos en los que aparecen las ocurrencias, aunque para una palabra con frecuencia alta puede resultar una tarea difícil.

Según Bécue[11], es posible reorganizar las unidades léxicas y las ocurrencias del texto de tal forma que las ocurrencias de una misma palabra se reagrupen acompañadas de un fragmento de

su contexto más inmediato, cuya longitud varía según las necesidades del análisis. Este proceso, produce lo que se llama *concordancia*.

Se denomina *palabra-pivote* o *forma-polo*, a la palabra cuyos contextos se reagrupan. Se buscan aquí expresiones o ideas que surgen desde los textos, buscando algún sentido en el empleo de las mismas. En la elaboración de las concordancias, se necesita partir de una o varias *formas-polo* seleccionadas a partir de un interés temático o basadas en criterios de frecuencia. Las líneas de concordancia están organizadas según el orden alfabético de la palabra-pivote. La siguiente tabla es un ejemplo de concordancia de la palabra-pivote *A*.

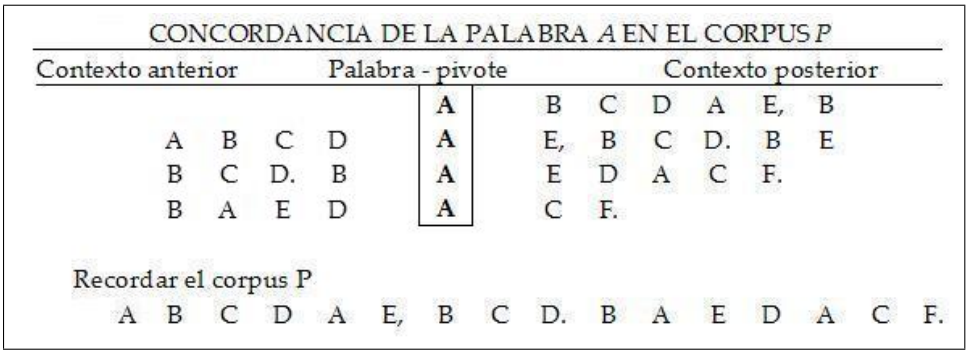

Figura 3.1: Concordancias de la palabra *A*

Inventarios de Segmentos Repetidos

Son listados de todos los segmentos repetidos en donde se indica su frecuencia o el número de formas simples de que constan. Hay varios criterios para ordenar estos ficheros: alfabéticamente, en forma jerárquica, por frecuencias, cronológicamente, o según ciertas *formas-polo*.

El incremento del vocabulario

Uno de los ámbitos tradicionales de la estadística léxica, es el estudio del incremento del vocabulario –número de palabras distintas–. Se puede observar que a medida que un corpus aumenta, su vocabulario tiende a incrementarse.

El número de palabras en un corpus no es proporcional a su tamaño, dado que cuando se tiene un corpus de gran tamaño, el incremento del vocabulario está sujeto a una doble influencia:

- añadir nuevas ocurrencias tiende a aumentar el número total de palabras
- a medida que el corpus aumenta, el número de palabras nuevas aportadas por cada nuevo fragmento de la misma longitud tiende a disminuir.

Dos de las variables lexicométricas para caracterizar un corpus son:

- Las variables lexicométricas del tipo T, que crecen en forma aproximadamente proporcional al tamaño de un texto.
- Las variables lexicométricas del tipo V, que tienden a crecer cada vez menos a medida que aumenta el tamaño del texto.

En la figura 3.2 que se muestra a continuación, se puede ver que las dos curvas de incremento presentan una forma similar, aunque la curva referente al vocabulario en lemas, se encuentra por debajo de la otra –vocabulario medido en palabras–. Esto implica que, en lo que refiere a las principales características lexicométricas, se pueden comparar distintos corpus, siempre y cuando se hayan utilizado idénticas normas para el recuento.

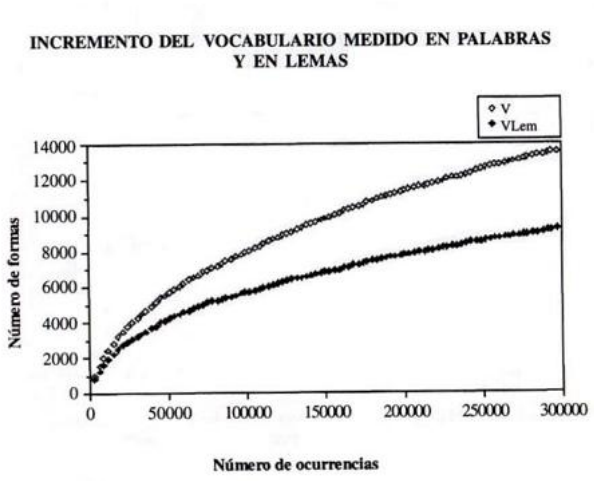

Figura 3.2: Ejemplo incremento del Vocabulario, presentado en el Libro *"Análisis estadístico de Textos"* Lebart, L. y otros, página 65.

Oraciones, secuencias y delimitadores

Para evitar realizar recuentos diferentes para las ocurrencias de segmentos que incluyan signos de puntuación y las de aquellos iguales a los anteriores pero sin signos de puntuación, en la práctica se buscan segmentos que no contengan palabras separadas por estos signos, ya sean signos débiles o fuertes. Se encuentran dos tipos de separadores que constituyen los delimitadores de secuencia:

- **Los separadores fuertes** o **separadores de oración:** aquí se encuentran los signos de puntuación tales como: el punto, punto de exclamación y punto de interrogación.
- **Los separadores débiles:** aquí se encuentran los signos de puntuación tales como: la coma, el punto y coma, dos puntos, guión, comillas y paréntesis.

Entonces, se puede decir que todas las sucesiones de ocurrencias no separadas por un delimitador de secuencia, son ocurrencias de segmentos.

LEYES EMPÍRICAS SOBRE EL VOCABULARIO

Si hablamos de leyes empíricas o distribuciones estadísticas relacionadas al ámbito lingüístico, la ley de Zipf figura entre las más nombradas.

Ley de Zipf

Cuando el corpus analizado comporta algunos miles de ocurrencias, se observa que la distribución de frecuencias adquiere ciertas características. G. K. Zipf llegó a la conclusión que en estos corpus de gran tamaño, los rangos de frecuencia de las unidades léxicas, ordenados en forma descendente, son inversamente proporcionales a la frecuencia correspondiente[15].

A veces se expresa esta relación de la siguiente manera: el producto "rango por frecuencia" es aproximadamente constante:

$$r\ f = c \tag{3.17}$$

siendo r el rango de la forma gráfica, f la frecuencia de la misma forma gráfica y c una constante (en general una décima parte del tamaño del corpus, T). Otra manera de expresar la misma relación surge al aplicar logaritmos:

$$log\ r + log\ f = log\ c \tag{3.18}$$

A modo de generalización de las fórmulas (3.17) y (3.18), se presentan las siguientes:

$$r^B\ f = c \tag{3.19}$$

$$B(log\ r) + log\ f = log\ c \tag{3.20}$$

siendo B un parámetro que ajusta la pendiente de la recta que surge en la relación 3.18. Esta última se conoce como "ley de Zipf generalizada".

Utilizando a $V(f)$, que simboliza el número de frecuencias mayor o igual a f y tomando una escala logarítmica, se puede representar la gama de frecuencias sobre un diagrama de Pareto. Este tipo de representación permite poner de manifiesto ciertos "accidentes" en la regularidad general de las distribuciones de frecuencia. Sobre todo, los puntos de la gama de frecuencias que corresponden a frecuencias muy altas o muy bajas, se apartan algo de la ley de Zipf.

[15]Si bien la ley de Zipf tiene un carácter general, la misma se originó en el estudio de textos en inglés.

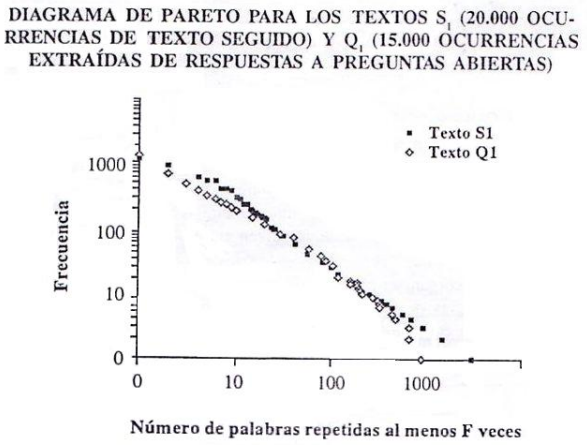

Figura 3.3: Ejemplo Diagrama de Pareto, presentado en el Libro *"Análisis estadístico de Textos"* Lebart, L. y otros, página 61.

Formas alternativas a la Ley de Zipf

Benoît Mandelbrot[16] presentó una alternativa razonable a la ecuación (3.17) en función de numerosos estudios avocados a la generalización de la ley de Zipf. Teniendo en cuenta la posibilidad de que el parámetro B en la fórmula (3.19) es en general diferente de 1 y que la relación (3.17) no se mantiene constante para valores bajos del rango r, introdujo la siguiente fórmula, generalización de la anterior:

$$(r+m)^B f = c \tag{3.21}$$

siendo m, B y c constantes que dependen del corpus estudiado (m es un "parámetro de ajuste" para valores bajos de r).

Otra interesante generalización es la relación rango-frecuencia de H. Edmundson, denominada "distribución del rango (de tres parámetros)":

$$f(r,c,b,a) = c(r+a)^{-b} \quad c>0, \;\; b>0, \;\; a \geq 0 \tag{3.22}$$

que es a su vez una generalización de la fórmula presentada por Mandelbrot.

La racionalidad detrás de la Ley de Zipf

Partiendo del estudio empírico realizado en la publicación de 1935 *The Psycho-Biology of Language*[17] y afirmando esta ley empírica desde un punto de vista filosófico en su último libro *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* en 1949, Zipf justifica

[16]Matemático polaco, creador –entre otros– de la denominada Geometría Fractal.

[17]En este libro Zipf muestra un diagrama $log(f)$ vs. $log(r)$ en donde se ajusta una curva rango-frecuencia de las palabras en Latín escritas por Platón en varios textos de su autoría.

gran parte de su visión respecto del comportamiento del ser humano: la lucha por minimizar el esfuerzo. Quizás lo más llamativo del caso es que, basándose en sólamente dos "modelos matemáticos" tan sencillos como (3.17) o (3.18), Zipf fue capaz de aplicar éstos en diversidad de aspectos.

No obstante lo anterior, muchos investigadores continuaron estudiando la racionalidad detrás de la ley de Zipf. Gran parte de los estudios realizados en este contexto se basaron en teoría de las probabilidades, partiendo de la probabilidad que una palabra ocurriera con cierta frecuencia en un corpus escogido arbitrariamente.

Uno de los más destacados es la distribución de probabilidad desarrollada por Gustav Herdan

$$p_f = \begin{cases} \frac{x-a}{x} & \text{si f=1} \\ (x-a)\frac{a(a+1)(a+2)...(a+f+2)}{x(x+1)(x+2)...(x+f-1)} & \text{si f=2,3,...} \end{cases} \tag{3.23}$$

siendo p_f la probabilidad que una forma gráfica aparezca con frecuencia f en un corpus de gran tamaño, a y x constantes que dependen del corpus, cumpliendo la relación $0 < a < x$. Esta distribución de probabilidad es conocida como la *fórmula de Waring-Herdan*.

Otros estudios del "comportamiento *rango-frecuencia*" en el lenguaje natural involucran otras distribuciones de probabilidad, como la log-normal o la Maxwell-Boltzmann (transformación de la distribución Gamma, muy usada en estadística mecánica y química), entre otras.

Modelización de distribuciones léxicas

Caso 1: Estimación del tamaño del vocabulario. Este es uno de los llamados "casos clásicos"de la lingüística cuantitativa. Se busca estimar la cantidad de diferentes ocurrencias que pertenecen a una cierta categoría dada una muestra de expresiones[18] de una categoría relevante.

Las estimaciones de curvas de incremento de vocabulario, así como también la interpolación usada para estimar estas curvas (en función de tamaños de muestra arbitrarios) y la extrapolación de estos métodos a muestras muy grandes a través de los modelos *LNRE*[19] son aplicaciones de uso corriente en la lingüística cuantitativa a nivel teórico.

Caso 2: Estimación del léxico. Las herramientas utilizadas en este caso son idénticas al ítem anterior, no obstante sus aplicaciones se centran a nivel práctico. En el campo de la lingüística computacional –rama definida en la sección 3.2.2– es menéster determinar cuán grande

[18] Traducción de *token*.

[19] *Large Number of Rare Events*: modelos estadísticos que permiten, por ejemplo, estimar el tamaño del vocabulario en estudios de atribuciones de autor, o extrapolar la proporción de hápaxes para así medir la productividad morfológica en la formación de palabras, entre otros.

tendría que ser el tamaño del vocabulario para que las unidades léxicas que no se encuentran en el mismo, denominadas *Out Of Vocabulary*[20] (OOV), se mantengan por debajo de un determinado umbral.

3.3.6. Tablas Léxicas y de Segmentos Repetidos

Las tablas léxicas y las de segmentos repetidos constituyen formas de reorganizar la información obtenidas por técnicas multivariadas de análisis.

- *Tabla Léxica*: tabla de doble entrada en la que las filas son unidades léxicas simples (forma gráfica o lema) y las columnas las partes del texto: locutores, autores, documentos, etc.. Esta tabla recoge la frecuencia k_{ij} de aparición de la unidad léxica situada en la fila i dentro de la j-ésima parte del texto. La denominación habitual para esta tabla es **T**.

- *Tabla Léxica Agregada*: matriz cuyo término general c_{ij} recoge el número de veces que la i-ésima unidad léxica simple ha sido utilizada en un conjunto de partes del texto, organizados bajo cierto criterio específico (ej. respuestas obtenidas de sujetos de un mismo grupo etáreo).

- *Tabla de Segmentos Repetidos*: de modo análogo a las anteriores, pueden construirse estas tablas, sustituyendo unidades léxicas simples por segmentos repetidos. Pueden ser tanto simples como agregadas.

3.3.7. Análisis de Correspondencias sobre Tablas Léxicas

El análisis de correspondencias es una técnica *descriptiva* de tablas de contingencia –o tablas cruzadas– y de ciertas tablas binarias –llamadas tablas de "presencia-ausencia"–. Dicha descripción se hace bajo la forma de una representación gráfica de las asociaciones entre filas y columnas. En este caso, la tabla usada es una *tabla léxica*, en la que se dispone de información sobre la frecuencia en que aparecen ciertas unidades léxicas en distintos textos o partes de un corpus textual. Esta información puede ser representada en un espacio de dimensiones reducidas conservando lo esencial de su estructura. Así como mencionan Lebart, Salem y Bécue[28] esta es una "simple extensión del dominio de aplicación del análisis de correspondencias, con procedimientos de cálculo y reglas de interpretación específicas."

Analisis de Correspondencia Simple

En general, el puntapié inicial de cada estudio se da con el procesamiento de las tablas *"variables por individuos"* a través de las técnicas antedichas.

[20]Por ejemplo, un programa para traducción de textos de un idioma a otro posée una base de datos compuesta por un vocabulario de tamaño V. Muchas veces ocurre que, por estar ciertas palabras del idioma de origen fuera de ese vocabulario, el software se ve imposibilitado de traducir las mismas al idioma de destino. Estas palabras se denominan *Out Of Vocabulary*.

Cabe recordar que la unidad considerada en este caso no es el individuo, sino la ocurrencia de una unidad léxica. Las filas constituyen así una partición del conjunto de las ocurrencias, mientras que las columnas constituyen una partición de los individuos entrevistados –para respuestas a preguntas abiertas, por ejemplo– o partes de un corpus –en el caso de textos literarios, discursos políticos, etc.–.

La comparación de dos perfiles-fila da cuenta de las asociaciones entre unidades léxicas y categorías de una variable, mientras que al comparar dos perfiles-columna se obtiene la proximidad existente entre las diferentes categorías de la variable seleccionada, en relación al vocabulario seleccionado.

El procedimiento en este caso es el siguiente: se elige una variable para construir la tabla léxica a ser analizada, y esta tabla será el insumo principal para el análisis de correspondencia simple, con todo lo que ello implica: obtención de perfiles-fila y perfiles-columna, construcción de los ejes de inercia respectivos y proyección de las nubes de perfiles fila y columna en los planos factoriales.

Análisis de Correspondencia Múltiple

La principal diferencia con el ACS es que la tabla a analizar no es una tabla léxica sino una tabla disyuntiva completa, o una tabla de Burt en su defecto.

Se puede decir entonces que es equivalente someter al análisis de correspondencias múltiple (ACM) a una tabla de contingencia **R**, cuyo elemento genérico muestra el valor que toma el individuo u observación i bajo la variable categórica k (para un total de j variables), o analizar una tabla binaria **Z** con i filas y tanta cantidad de columnas como la suma de las modalidades de las variables, expresada por k. El elemento genérico de esta tabla es z_{ik}:

$$z_{ik} = \begin{cases} 1 & \text{si el individuo i toma la modalidad k} \\ 0 & \text{en otro caso} \end{cases} \tag{3.24}$$

Si bien el análisis de esta última tabla Z es más "costoso" desde muchos puntos de vista (por ejemplo en recursos informáticos), es también más interesante ya que, basándose en esta tabla, el investigador puede construir otra tabla aún más informativa que ésta, la tabla de Burt, simétrica y compuesta de las *modalidades* de todas las variables usadas, la cual al ser dividida en bloques de matrices muestra como se comportan –a través de su frecuencia absoluta– los individuos respecto a las modalidades de cada variable y, al mismo tiempo, respecto de las modalidades cruzadas de otras variables. Es importante tener en cuenta que ambas tablas conducen a similares resultados. Ahora bien, la escala difiere ya que los valores propios de **B** son, como se desprende de la fórmula 3.14, el cuadrado de los valores hallados según la tabla **Z**.

A la hora del estudio, las variables activas seleccionadas tienen que estar relacionadas al enfoque deseado: éstas deben presentar cierto grado de homogeneidad.

Una vez construidos los ejes factoriales, se puede proyectar como unidades suplementarias a las restantes variables y/o elementos, para así observar estas unidades en los planos engendrados por las variables activas.

De este modo, las asociaciones entre ciertas palabras utilizadas y características de los individuos entrevistados se hacen visibles rápidamente. La proyección sistemática de variables suplementarias permite ahorrar tiempo y evitar los errores de interpretación que se producen en general al examinar, de forma secuencial, una serie de tablas de contingencia.

Si, por ejemplo, se construye una tabla en la que se establece la frecuencia en que cada unidad léxica aparece en el texto emitido por cierto individuo, es posible representar a los individuos en el espacio determinado por las unidades léxicas seleccionadas.

El examen de las palabras utilizadas tiene sumo interés a la hora de describir vocabularios característicos por ejemplo, en función de agrupamientos hechos a priori respecto a algunas de las variables cualitativas usadas. Es habitual escoger un umbral mínimo de referencia, ya que muchos reconocidos autores[21] coinciden en el hecho de que, en estudios estadísticos como estos, las palabras demasiado poco empleadas no tienen mucho sentido[13]. De todas formas se recomienda que aproximadamente el 75 % de la longitud total del corpus se conserve, además de corroborar la estabilidad de los resultados frente a pequeñas variaciones del umbral.

3.3.8. Análisis de Grupos sobre Tablas Léxicas

Estos métodos permiten representar las proximidades entre las filas o las columnas de una tabla léxica mediante la formación de grupos o clusters.

Se encuentran dos grandes familias dentro de estos métodos:

- **Los de clasificación jerárquica**, que permiten obtener un orden de clusters anidados a partir de un conjunto de elementos.
- **Los de clasificación directa**, que realizan una partición de la población en estudio en un número dado de grupos.

Los resultados obtenidos mediante los métodos de clasificación revelan ser, en la práctica, complementos indispensables de los resultados obtenidos a partir del análisis de correspondencias. Una de las razones mas importantes que motiva el empleo conjunto de los métodos factoriales y los de clasificación, es el enriquecimiento dimensional de las representaciones, ya que los reagrupamientos se realizan a partir de distancias calculadas en *todo el espacio* y no en el espacio reducido a los primeros planos factoriales.

[21] Bécue, Lebart y Salem, entre otros.

Clasificación Jerárquica

Este método se aplica a tablas cruzadas tales como las tablas léxicas, o las tablas léxicas agregadas. Se puede clasificar tanto las columnas de la tabla como el conjunto de las filas.

Se parte de un conjunto de *n* elementos, con pesos iguales o distintos, entre los cuales se han calculado las distancias dos a dos. En primer lugar, se juntan los dos elementos más próximos. Estos dos elementos forman ahora un nuevo elemento cuyo peso es la suma de los pesos de los dos elementos agrupados. Se vuelven a calcular las distancias entre este nuevo elemento y los restantes y se agregan de nuevo los dos elementos más próximos. Se reiterea el procedimiento hasta agotar el total de los elementos. La *(n-1)-ésima* operación reagrupa el conjunto de los elementos en una única clase.

El dendrograma

Esta forma de clasificación se puede representar de varias maneras, pero sin dudas el denominado árbol jerárquico o dendrograma constituye la representación más rica.

Los reagrupamientos formados a cada paso del algoritmo reúnen elementos cada vez más distantes a medida que se avanza en la construcción del árbol: el número de puntos ya agregados aumenta y la distancia mínima entre las clases que quedan por agrupar se incrementa.

Esta representación en forma de dendrograma de una clasificación jerárquica muestra claramente que los grupos formados a lo largo del proceso, constituyen una jerarquía anidada. La interpretación de ésta se basa en el análisis de las distancias entre elementos o grupos unidos en un mismo nodo.

Si el número de elementos por clasificar es importante, el dendrograma puede resultar dificil de estudiar. De esta manera, una solución práctica consiste en definir un *nivel de corte* que permita considerar sólo la parte superior del árbol. Se puede determinar de antemano el número de grupos entre los cuales se desea repartir el conjunto de elementos a clasificar, o el número de nodos superiores a retener para la clasificación.

Es posible repartir en el interior de los grupos de la jerarquía construida, un conjunto de elementos suplementarios. Para cada uno de estos elementos, el algoritmo de asignación procede de una forma muy simple: se comienza por buscar, entre los elementos terminales, el más próximo al elemento suplementario a clasificar y luego se agrega éste a la clase de la jerarquía que contiene al elemento terminal. El método permite clasificar un gran número de elementos suplementarios sin modificar la clasificación obtenida sobre el conjunto base.

3.3.9. Visualización de Datos Textuales

Para llevar a cabo el análisis de textos de cualquier naturaleza (corpus literarios, documentos o respuestas a preguntas abiertas en una encuesta, etc.) se utilizan las llamadas *tablas léxicas*, las cuales –tal como se mostrará a continuación– comportan un arreglo de doble entrada que, en general, es formado por unidades léxicas de cualquier índole (palabras, segmentos repetidos, lemas, etc.) y modalidades de una variable. A las mismas se las describirá utilizando la técnica de análisis de correspondencias afín de lograr, como se explicó en párrafos anteriores, una interpretación razonable de la estructura de datos con pérdidas mínimas de información.

En el análisis de respuestas a preguntas abiertas, que es el caso que nos concierne, se suele agrupar a las mismas en *textos artificiales* para llevar a cabo la caracterización de los individuos (por ejemplo, según los tramos etáreos de los encuestados, o también según su nivel educativo o de ingreso). Claro que, como suele suceder, los criterios de reagrupamiento más adecuados –o los más populares si se quiere– suelen ser desconocidos *a priori*. Es aconsejable[28] realizar las siguientes estrategias cuando ninguna forma de aunar los datos se impone respecto a las otras:

- Utilizar una partición de síntesis a través de técnicas de clasificación automática. Este punto se desarrollará a continuación en la sección "particiones en situaciones-tipo";
- Analizar directamente las respuestas no reagrupadas, punto desarrollado en la sección "análisis directo de respuestas".

ANÁLISIS DE CORRESPONDENCIA DE UNA TABLA LÉXICA

En secciones anteriores se vio cómo las respuestas se podían codificar de manera totalmente transparente para el usuario. El resultado de esta codificación puede tomar dos formas diferentes que se materializan en las tablas **R** y **T**.

Tablas Léxicas de Base

- Tabla **R**: tabla constituida por el total de respuestas consideradas en las filas –este número se suele denotar por *n*–, y en las columnas el número de palabras de la respuesta más larga de todo el corpus. En la práctica la tabla R no es rectangular: cada fila tiene longitud variable, debido a que la fila *i* de esta matriz contiene las palabras que componen la respuesta del individuo *i*.
- Tabla **T**: esta tabla posée la misma cantidad de filas que **R**, pero varía en la cantidad de columnas, dado que en la tabla **T** las columnas vienen dadas por la cantidad de palabras distintas que se encuentren en el corpus, es decir *V* columnas. La celda *(i,j)* contiene el número de veces que el individuo *i* utiliza la palabra *j* en su respuesta[22].

[22]Este es un tipo de *tabla dispersa*. Ver anexoA.

La tabla **T** se puede construir fácilmente a partir de la **R**; sin embargo la información en relación al orden de las palabras en cada respuesta se pierde en la matriz **T**.

Tablas Léxicas Agregadas

Como en la práctica es difícil poder encontrar respuestas lo suficientemente largas para poder aplicar sin restricciones determinados tratamientos estadísticos, es necesario reagrupar las mismas en categorías que tengan sentido. Como se vio anteriormente, la tabla disyuntiva completa **Z**, describe las respuestas de n individuos a una pregunta cerrada con p modalidades de respuesta. Cada pregunta cerrada contenida en esta matriz permite definir una partición de los individuos interrogados.

La tabla **C** por su parte tiene V filas y p columnas –número de modalidades de la pregunta cerrada considerada–. Su término general c_{ij} cuenta el número de veces que el conjunto de individuos que escogieron la modalidad j utilizan la palabra i. Esta matriz se obtiene mediante el siguiente producto:

$$\mathbf{C = T'Z} \tag{3.25}$$

Umbral de frecuencia de las palabras

La comparación de perfiles léxicos tiene sentido cuando las palabras poco frecuentes son eliminadas del análisis. Esto trae aparejado una brusca disminución del tamaño del vocabulario, T.

Construcción de la tabla léxica agregada

Como ya se mencionó líneas arriba, es necesario reagrupar respuestas en función de una variable categórica observada sobre los individuos. Se agregan las k filas de la tabla **T** y se obtiene la tabla de contingencia **C** "unidades léxicas por categorías de individuos". Esta matriz **C** permite la comparación de los perfiles léxicos de las categorías.

¿Que tabla analizar?

Escoger entre la tabla **T** o la **C**, es decir entre el análisis directo de las respuestas o el análisis de las respuestas reagrupadas no es tarea sencilla. En esta decisión intervienen numerosos aspectos. Si las respuestas individuales son cortas, es posible encontrar respuestas que coincidan en su contenido pero utilizando palabras totalmente distintas, dificultando así el reconocimiento de la proximidad existente entre ellas. En suma, todo esto dependerá del grado de refinamiento que se desea obtener y del tamaño de la parte del texto escogida (a mayor cantidad de elementos, se consolidan usos de ciertas locuciones o expresiones que difícilmente puedan ser reconocidos

en partes pequeñas y, más aún, se puede llegar a identificar las categorías de locutores que usan estos términos de manera significativa).

Lematización

Antes de lematizar[23], el investigador debe reconocer las locuciones y agrupar las palabras que forman una expresión en una única unidad léxica que conserva su forma inicial. Por ejemplo: *a_gusto*, *crema_de_enjuague*, *de_vez_en_cuando*, etc..

El análisis del corpus lematizado busca verificar si la estructura o patrón de los puntos-categoría observada en los planos factoriales proviene del vocabulario escogido. De ser así, tanto el análisis del corpus sin lematizar y el lematizado resultan similares en cuanto a la estructura de las categorías.

La lematización influye sobre el número de palabras retenidas, ya que los verbos, por ejemplo, serán eliminados en una cantidad menor respecto al caso donde no hay lematización del texto: las flexiones y formatos en masculino o femenino se reúnen en su verbo raíz, aumentando la frecuencia de éste en el total de palabras, a la vez que muchos verbos que por separado no caían en el umbral mínimo de frecuencias fijado anteriormente, ahora sí lo hacen. Incluso más: algunos de estos verbos aparecen entre las palabras más frecuentes.

El proceso de lematización incide directamente en una disminución del vocabulario, y por ende de la "riqueza sintáctica" del mismo, por ejemplo el uso de verbos en primera o tercera persona. Esto tiene como contrapartida análisis más robustos desde el punto de vista estadístico, ya que las frecuencias observadas son más altas que antes. Además, el lema ocupa en general una posición más próxima al centro de gravedad que cada una de sus correspondientes formas conjugadas o flexionadas[24].

Partición en Situaciones Tipo

Para lograr un análisis más riguroso es recomendable el reagrupamiento de individuos. Ahora bien, no es tarea sencilla elegir la "variable de reagrupamiento". Para escogerla de forma arbitraria, es necesario poseer un sólido conocimiento sobre el fenómeno observado. Si este no es el caso, es posible a través de un algoritmo de formación de grupos o de clasificación, "etiquetar" a los individuos bajo las características seleccionadas, sin privilegiar ninguna a priori.

ANÁLISIS DIRECTO DE LAS RESPUESTAS

Hasta ahora, se viene estudiando los perfiles de frecuencias de las partes de un corpus, cuando estas partes contienen textos relativamente importantes en lo que refiere a la longitud.

[23] Ver anexoA.
[24] Ver anexoA.

Para obtener dichas partes a partir de preguntas abiertas, se debió proceder a reagrupamientos *a priori*, según criterios escogidos también *a priori*.

Sin embargo, los métodos presentados en las secciones anteriores –Análisis de Correspondencias y Análisis de Clasificación– se pueden aplicar a las respuestas individuales. Este tratamiento, de las respuestas individuales, se recomienda en las siguientes situaciones:

1. Cuando las respuestas son ricas desde el punto de vista léxico, para que la comparación de los perfiles de frecuencia sea más provechosa.

2. Cuando se está buscando, con un trabajo de descripción preliminar, criterios para la creación de grupos.

En el primer caso una descripción directa de las preguntas es posible, dejando la puerta abierta para posteriores reagrupamientos si es que éstos pueden ayudar a la interpretación o permiten poner a prueba ciertas hipótesis. En el segundo caso la noción de perfil no tiene el mismo sentido: las respuestas se distinguen más por la presencia o ausencia de formas que por verdaderas variaciones entre perfiles de frecuencia.

En resumen, el análisis directo permitirá reagrupar las respuestas similares, dejando de lado las que se distinguen por la originalidad de su forma. Se trata entonces, de una ayuda a la postcodificación.

Análisis de T (tabla dispersa)

Para analizar directamente las respuestas, se somete a un análisis descriptivo a la tabla **T**, compuesta como ya se dijo líneas arriba por n filas y V columnas.

Algunas observaciones:

- La proximidad entre dos unidades léxicas (o sea, dos columnas de la matriz **T**) es tanto mayor cuanto más frecuente es la aparición de dichas unidades en las mismas respuestas (y no sólo en los mismos textos o agregados de respuestas), lo que permite captar mejor las proximidades sintagmáticas[25].

- Las fórmulas de transición presentadas en el cuadro de resumen3.1 de la sección3.1.2 permiten simplificar la interpretación de la tabla **T**: los n individuos estarán próximos a los centros de gravedad de las palabras que ellos emplean.

- Las n filas de **T** representan a los individuos entrevistados. Las respuestas a las preguntas cerradas del cuestionario pueden formar parte de las columnas de una nueva tabla T^+ y así posicionarse como elementos suplementarios sobre los planos factoriales obtenidos luego

[25]El *análisis sintagmático* de un texto abarca el estudio de su estructura y la relación entre sus partes. El estudio de las relaciones sintagmáticas revela las reglas de combinación fundamentales en la producción e interpretación de textos. Citado en Chandler, Daniel (2002) "Semiotics: The Basics" Edit. Routledge, Londres.

de un análisis de correspondencias de **T**, ayudando así a sugerir criterios para reagrupar las respuestas.

Validación externa: variables suplementarias y valores test

Las características de los individuos juegan un importante papel a la hora de describir y asignar a los mismos a ciertos grupos, configurados éstos de forma natural (por ejemplo, en el caso de variables como el nivel socio-económico) o artificial (cuando se construyen grupos de individuos según criterios de proximidad previamente definidos).

La valoración de las coordenadas de las variables suplementarias incluidas en la nueva tabla T^+ se realiza mediante el cálculo de *"valores-test"* que proporcionan una medida de su significación estadística.

Los cálculos de estos valores permiten una validación externa de los diferentes planos factoriales. Pueden incluso constituir un método para determinar el número de ejes significativos. Si una categoría suplementaria no es significativa sobre un eje, no será útil a la hora de validar un plano que lo contenga.

3.3.10. Elementos Característicos y Respuestas Modales

Las representaciones espaciales que resultan de las proyecciones en los planos factoriales se pueden enriquecer mediante resultados de naturaleza más probabilística: los *elementos característicos* o *unidades léxicas características*. Se trata de poner en evidencia las unidades léxicas presentes en cada parte del corpus con una frecuencia particularmente elevada comparada con la frecuencia global, o por el contrario una frecuencia particularmente baja.

Además, se pueden caracterizar los grupos de respuestas mediante las denominadas *respuestas modales*. Las respuestas modales contienen un gran número de palabras características de la parte del corpus a la que pertenecen.

ELEMENTOS CARACTERÍSTICOS

Se desea seleccionar elementos léxicos sobre o subutilizados en cada parte del corpus en comparación con la frecuencia global en todo el corpus.

Se supone que las palabras tienen una distribución hipergeométrica y de esta manera se compara la diferencia entre la frecuencia global de una palabra y su frecuencia en la parte estudiada. Aunque los índices calculados sirven como complemento para seleccionar los elementos característicos de cada una de las partes.

Cálculo de los elementos característicos

$k_{ij} \rightarrow$ subfrecuencia de la palabra i en la parte j del corpus
$k_{i.} \rightarrow$ frecuencia de la palabra i en la totalidad del corpus
$k_{.j} \rightarrow$ tamaño (número de ocurrencias) de la parte j del corpus
$k_{..} \rightarrow$ tamaño (número de ocurrencias) del corpus (o, simplemente, k)

Se tiene una población de k objetos, donde $k_{i.}$ objetos viene "marcados" con algún distintivo que los diferencia, o pueden ser ocurrencias de una misma palabra de frecuencia total $k_{i.}$. El número, por tanto, de objetos "no marcados" es igual a $k - k_{i.}$.

Entonces, mediante un procedimiento de extracción aleatoria sin reposición, se selecciona una muestra de $k_{.j}$ objetos. Luego, se calcula el número k_{ij} de objetos marcados que contiene la muestra.

Para cada muestra de tamaño $k_{.j}$, el número k_{ij} de objetos marcados puede tomar un valor entero entre 0 y $k_{i.}$, número total de objetos marcados. Para cada entero n comprendido entre 0 y $k_{i.}$, es posible hacer el recuento de muestras $(N(n))$ de tamaño $k_{.j}$ para las cuales k_{ij} es exactamente igual a n. La ley de probabilidad para una extracción sin reposición que se ajusta mejor a este experimento, es la *ley Hipergeométrica*.

$Prob(k, k_{i.}, k_{.j}, n)$ indica la probabilidad de obtener un número n de objetos marcados, al extraer una muestra sin reposición de tamaño $k_{.j}$, de un total de k objetos, de los cuales $k_{i.}$ son los objetos marcados.

La fórmula clásica de la ley Hipergeométrica es:

$$Prob(k, k_{i.}, k_{.j}, n) = \frac{\binom{k_{i.}}{n}\binom{k-k_{i.}}{k_{.j}-n}}{\binom{k}{k_{.j}}} \tag{3.26}$$

Se puede ahora utilizar ese modelo para emitir un juicio sobre la frecuencia absoluta k_{ij} observada en una muestra. Si el valor observado de k_{ij}, n, se encuentra muy próximo a la moda de la distribución, no se puede decir nada sobre el resultado observado. Si dicho valor es claramente superior a la moda, el investigador se interesará por la probabilidad $P_{sup}(k_{ij})$ de observar, bajo las anteriores hipótesis, un número de objetos marcados igual o superior a k_{ij} entre los $k_{.j}$ objetos seleccionados al azar. Si, por el contrario, dicho valor es claramente inferior a la moda, se calculará la probabilidad $P_{inf}(k_{ij})$ de observar un número de objetos marcados igual o inferior a k_{ij}.

$P_{sup}(k_{ij}) \rightarrow$ es la suma las probabilidades $Prob(k, k_{i.}, k_{.j}, n)$ para los valores de n comprendidos entre k_{ij} y $k_{i.}$.

$P_{inf}(k_{ij}) \rightarrow$ es la suma de las probabilidades $Prob(k, k_{i.}, k_{.j}, n)$ para los valores de n comprendidos entre 0 y k_{ij}.

Cálculo práctico de las unidades léxicas características

Las celdas de la tabla léxica agregada contienen las subfrecuencias k_{ij} de cada una de las palabras i en las distintas partes j. Para cada celda, se valoran las subfrecuencias respecto a los números k, $k_{i.}$, $k_{.j}$, mediante el modelo de probabilidad de Distribución Hipergeométrica.

Para seleccionar las probabilidades consideradas pequeñas –que señalan las palabras de uso diferenciado entre las partes–, se fija de manera arbitraria un *umbral*. En general, se escogen para el mismo valores habituales utilizados en los test estadísticos: 0,005; 0,01; 0,001; etc..

Unidades léxicas características negativas y positivas y unidades banales

La palabra i es una *palabra característica negativa* o *especificidad negativa* para la parte j del corpus, cuando dicha palabra está subrrepresentada en la parte j, es decir cuando $P_{inf}(k_{ij})$ es inferior al umbral.

Por el contrario, la palabra i es una *palabra característica positiva* o *especificidad positiva* para la parte j del corpus, cuando dicha palabra está sobrerrepresentada en la parte j, es decir cuando $P_{sup}(k_{ij})$ es inferior al umbral.

Cuando ninguna de las dos probabilidades $P_{sup}(k_{ij})$ y $P_{inf}(k_{ij})$, es inferior al umbral, se considera que la palabra i es banal[26] para todas las partes; dicha palabra pertenece al *vocabulario básico* del corpus.

Los diagnósticos obtenidos mediante este método de selección de los elementos característicos se puede presentar de diversas maneras:

- Listado de palabras en donde se indica si es un elemento característico positivo o negativo de las diferentes partes del corpus.
- Listado de las unidades características. Consiste en listados ordenados, para cada una de las partes. Se indican tanto las unidades sobrerrepresentadas como también las subrrepresentadas.

Valor-test

La probabilidad $P_{sup}(k_{ij})$ (respectivamente $P_{inf}(k_{ij})$) mide la desviación existente, por exceso (respectivamente por defecto) entre la frecuencia relativa de la palabra i en la parte j y su frecuencia relativa global $f_{i.} = k_{i.}/k$ en todo el corpus.

A esta probabilidad se le puede asociar el valor z_0 proveniente de una v.a. normal estándar tal que

[26] Ver Anexo A.

$$P\left(x \geq z_0\right) = P_{sup}\left(k_{ij}\right) \tag{3.27}$$

$$P\left(x \leq z_0\right) = P_{inf}\left(k_{ij}\right) \tag{3.28}$$

Este valor z_0 calculado para cada par (palabra *i*, parte *j*) recibe el nombre de *valor-test* y se nota *t(i,j)*. Bajo la hipótesis de distribución aleatoria de una palabra en las partes, el valor-test está comprendido en el intervalo [-1,96; 1,96] con una probabilidad del 95 %.

RESPUESTAS CARACTERÍSTICAS

Una técnica sencilla permitirá situar las palabras en su contexto inmediato. Esta técnica consiste en la selección automática de las *respuestas características* o *respuestas modales*. Éstas no son respuestas artificiales que dan un resumen de lo respondido por cada grupo, sino respuestas auténticas, seleccionadas en razón de su carácter representativo para una categoría de individuos.

Se ofrecen aquí dos maneras de seleccionar dichas respuestas características: la primera parte de las palabras características, mientras que la segunda se apoya en el cálculo de distancias según criterios geométricos simples –distancia de χ^2–.

Selección de las respuestas características utizando los elementos característicos

Consiste en buscar las respuestas que contienen, en la medida de lo posible, las palabras más características del grupo. Para dicha selección se utilizan los valores-test *t(i,j)*.

Para cada parte, se ordenan las palabras desde la más característica hasta la más anticaracterística. El rango de dicho orden es más pequeño cuanto más característica es la palabra. Se puede asociar a cada respuesta el rango medio de las palabras que contiene: un rango pequeño significa que la respuesta contiene sólo palabras muy características de la parte. Otra variante consiste en utilizar la media de los valores-test: al rango medio más pequeño corresponde el mayor valor-test medio.

Selección de las respuestas características utilizando la distancia de Chi-2

La tabla léxica entera, **T**, con k filas y V columnas, siendo V el número de palabras seleccionadas con el umbral de frecuencia escogido, representa las respuestas abiertas.

Una parte o grupo de respuestas es un conjunto de vectores-fila y el perfil léxico medio de dicho grupo se obtiene calculando el perfil medio de los vectores-fila de este conjunto. Si el reagrupamiento se efectúa según las modalidades de una pregunta cerrada cuyas respuestas vienen codificadas en una tabla **Z**, la tabla léxica agregada **C** se calcula mediante la fórmula

$$\mathbf{C} = \mathbf{T'Z} \tag{3.29}$$

Por tanto, se pueden calcular distancias entre respuestas o grupos de respuestas. Las respuestas (filas de **T**) y los grupos de respuestas (filas de **C'**) vienen representados por vectores en el espacio referenciado por las palabras, de dimensión V, siendo la distancia usada la χ^2.

Dichas distancias expresan la desviación entre el perfil de una respuesta y el perfil medio del grupo al cual pertenecen. La distancia entre un punto-fila i de **T** y un punto-fila m de **C'** viene dada por:

$$d^2(i,m) = \sum_{j=1}^{V} \frac{t_{..}}{t_{.j}} \left(\frac{t_{ij}}{t_{i.}} - \frac{c_{jm}}{c_{.m}} \right)^2 \tag{3.30}$$

Para cada parte, se puede ordenar las distancias en forma creciente y así seleccionar las respuestas más representativas en función de sus perfiles léxicos, es decir las respuestas correspondientes a las distancias más pequeñas.

Comparación de Criterios

El primer criterio favorece a las respuestas breves. Cuando existan respuestas muy cortas formadas por palabras muy características serán estas las primeras seleccionadas. Evidentemente, cuando esto ocurra, se obtendrá una visión reducida y muy parcial del contenido de las respuestas del grupo.

El segundo criterio, por el contrario, privilegia los perfiles suaves y por tanto a las respuestas largas. Cuando existan respuestas muy largas pueden resultar seleccionadas aunque no sean muy representativas del grupo.

De hecho, se debe consultar un número suficiente de respuestas características. Se suelen leer primero las que se obtienen según el primer criterio, pues constituyen un breve resumen pero claro y fácil de relacionar con las palabras características. Después, se leen las respuestas obtenidas con el segundo criterio, lo que permite captar la diversidad de contextos en los cuales aparecen las palabras características y ver con que palabras se asocian.

Comentarios

- Sin ninguna codificación ni interpretación previa del texto, se puede obtener una visualización de las proximidades entre palabras por una parte, y categorías por otra.
- Las respuestas presentadas son las respuestas íntegras, es por tanto, una manera de recuperar las palabras muy poco frecuentes.

- Al conservar las primeras respuestas características en relación a ambos criterios, se espera encontrar algunas respuestas en común.

3.3.11. Otras Técnicas Situadas en el ADT

A continuación se exhiben un conjunto de técnicas enmarcadas en el análisis estadístico de textos que, por distintos motivos, no han sido utilizados en este informe de pasantía; pero se cree convieniente destacarlos.

ANÁLISIS DE CONTIGÜIDAD

Los textos agrupados en corpus proporcionan al investigador numerosa información externa como autor, género, fecha de redacción en el caso de textos comparados de diferentes autores, y categorías socioprofesionales, grupos etários, nivel de instrucción en el caso de colapsar respuestas a preguntas abiertas. Esta información nos permite además establecer relaciones tales como la procedencia de un mismo autor, fechas próximas, mismo género literario, etc..

Las 3 estructuras de base

El ejemplo citado por Lebart[27] es útil para entender el núcleo conceptual de la técnica que se describe a continuación. En el mismo se utiliza un corpus ficticio formado por editoriales de un mismo periódico. Al citado corpus se lo divide en nueve partes, correspondiente cada una a diferentes autores (simbolizados por las letras A, B y C) y fechas (números 1, 2 y 3).

En la figura 3.4 se observan tres posibles formaciones: La primer estructura, **S1**, se denomina "partición" ya que toma en cuenta la autoría del artículo. Así, considera contiguos a artículos de un mismo autor. La segunda, **S2**, es conocida por "cronología" pues en este caso se toman en cuenta la *contigüidad de tiempos*, sin tomar en cuenta al autor del artículo; finalmente **S3** recibe el nombre de *grafo no orientado*, debido a que en esta estructura se reúnen las partes del corpus que presentan alguna relación de contigüidad o proximidad, como por ejemplo las distintas temáticas tratadas, filiación política, documentos relativos a regiones limítrofes, etc..

Más aún, es posible distinguir dos grupos estructurales: uno constituido por **S1** y **S2**, denominado de *estructuras de contigüidad a priori*, debido a que puede establecerse su relación antes de efectuar un análisis, sólo conociendo nombres de autores y momentos de publicación de los artículos, y el otro de *estructuras de contigüidad a posteriori* pues los vínculos entre ellos son fabricados en función de relaciones no superficiales, como las citadas líneas arriba.

- **S1 – Estructuras de partición:** se analizan a través del ANOVA, técnica que permite testear la heterogeneidad de ciertos reagrupamientos de variables.
- **S2 – Estructuras de series cronológicas:** se utilizan las series de tiempo y los procesos estocásticos para su estudio.

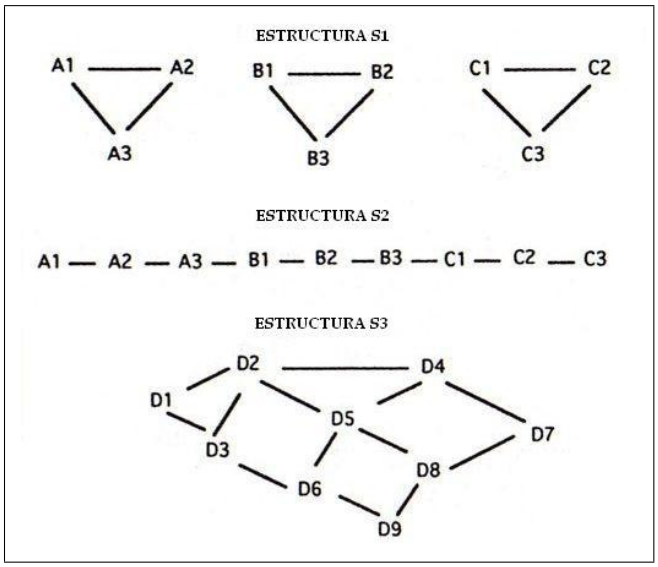

Figura 3.4: Estructuras de Contigüidad

- **S3 – Estructuras de grafos:** estructuras más generales que las anteriores pero relacionadas a técnicas menos populares que las dos precedentes, llamadas *análisis estadísticos de contigüidad*. Estas técnicas permiten extraer conclusiones más fácilmente en situaciones corrientes. En la medida que estos ordenamientos contienen como casos particulares a los dos anteriores (**S1** y **S2**), podemos decir que **S3** constituye una herramienta muy útil a la hora de tomar en cuenta la mayoría de las estructuras a priori susceptibles de ser observadas en la práctica en estadística textual.

Particularmente, en la fase de interpretación –delicada, por cierto– estos métodos son un buen complemento de los análisis puramente exploratorios, reduciendo así la parte de subjetividad inherente a todo comentario.

Homogeneidad de los valores de una variable respecto a una partición

En esta primera parte se estudiará el caso de una variable numérica que toma valores en cierto conjunto de individuos. Se busca estudiar con las definiciones precedentes, los factores sobre el conjunto de partes en que se dividió el corpus, creados a través del análisis de correspondencias de una tabla léxica.

En lo que sigue, J simboliza un conjunto de elementos y p al número de los mismos. Es posible definir sobre el conjunto $(J \times J)$ una relación denominada *relación de contigüidad*, *R(j,j')*. Esta es simétrica ($R(j,j') = R(j',j)$) pero no reflexiva (*R(j,j)* no es posible; es decir que cada parte j no es adyacente a sí misma).

Grafos asociados a estructuras de contigüidad

Hay varias formas para representar una relación de contigüidad. Una de ellas es a través de los denominados "grafos de contigüidad o adyacencia" designados con el símbolo G_i: G_1, G_2,...

Características de los G_i

- A cada elemento *j* del conjunto *J* le corresponde un *vértice* del grafo.
- Cada una de las parejas de vértices unidas por la relación de contigüidad *R(j,j')*, estan comunicadas entre sí por una *arista*, *m*.
- El número de aristas adyacentes a un mismo vértice *j*, es decir el número de aristas con una extremidad en *j*, se denomina el *grado* del vértice *j*-ésimo, m_j.
- Si todos los vértices están unidos por una arista, se dice que *G* es un *grafo completo*. Este tipo de grafos posee $p(p-1)$ aristas.
- Es usual distinguir las aristas según sus sentidos. De este modo, aquella arista que va desde *j* a *j'* es, en principio, distinta a aquella que lo hace de modo inverso.

Las relaciones a priori que interesan son aquellas relacionadas con los grafos simétricos (tales que *R(j,j') = R(j',j)*). Dichos grafos son conocidos en matemática discreta como "grafos no orientados" debido a que no existe relación de precedencia entre los elementos del mismo; es decir, no hay un único sentido vectorial entre elementos del conjunto *J*.

Matrices de contigüidad M

Las relaciones *R(j,j')* del párrafo precedente pueden ser representadas por medio de la denominada *matriz de contigüidad*, arreglo multidimensional, cuadrado y simétrico (pues establecimos más arriba que las relaciones de contigüidad estudiadas serán simétricas), compuesto de ceros y unos. Su término general es $m_{jj'}$, que vale:

$$m_{jj'} = \begin{cases} 1 & \text{si j contiguo j'} \\ 0 & \text{en otro caso} \end{cases} \tag{3.31}$$

Según la convención en la fórmula (3.31), los elementos de la diagonal principal son todos nulos.

La siguiente igualdad relaciona a las aristas *m*, los grados m_j y los elementos $m_{jj'}$ de la matriz **M**:

$$m = \sum_{j=1}^{p} m_j = \sum_{j=1}^{p} \sum_{j'=1}^{p} m_{jj'} \tag{3.32}$$

El coeficiente de contigüidad

Imaginemos ahora una variable aleatoria *Z*, cuyo soporte o dominio son los valores que toma ésta en cada uno de los vértices del grafo de contigüidad, z_j. Estos valores corresponden a medidas hechas al interior de cada fragmento del corpus, por ejemplo la longitud media de las frases en cada parte.

La media muestral de *Z* es:

Matriz de Contigüidad de la Estructura S1

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	0	0	0	1	0	0	1	0	0
B1	0	0	0	0	1	0	0	1	0
C1	0	0	0	0	0	1	0	0	1
A2	1	0	0	0	0	0	1	0	0
B2	0	1	0	0	0	0	0	1	0
C2	0	0	1	0	0	0	0	0	1
A3	1	0	0	1	0	0	0	0	0
B3	0	1	0	0	1	0	0	0	0
C3	0	0	1	0	0	1	0	0	0

Matriz de Contigüidad de la Estructura S2

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	0	1	0	0	0	0	0	0	0
B1	1	0	1	0	0	0	0	0	0
C1	0	1	0	1	0	0	0	0	0
A2	0	0	1	0	1	0	0	0	0
B2	0	0	0	1	0	1	0	0	0
C2	0	0	0	0	1	0	1	0	0
A3	0	0	0	0	0	1	0	1	0
B3	0	0	0	0	0	0	1	0	1
C3	0	0	0	0	0	0	0	1	0

Figura 3.5: Ejemplo de Tablas de Contigüidad para estructuras **S1** y **S2**, extraido de *Statistique Textuelle*; L. Lebart, A. Salem, pág. 201.

$$\bar{z} = \frac{1}{p}\sum_{j=1}^{p} z_j \tag{3.33}$$

La varianza global mide la dispersión de los valores z_j alrededor de su media, $\bar{z}$ respecto del conjunto de vértices del grafo G:

$$v(z) = \frac{1}{2p(p-1)}\sum_{j'=1}^{p}\sum_{j=1}^{p}(z_j - z_{j'})^2 = \frac{1}{p-1}\sum_{j=1}^{p}(z_j - \bar{z})^2 \tag{3.34}$$

La varianza local $v^*(z)$ por su parte mide la dispersión entre vértices contiguos en G, es decir calculada para aquellas parejas j y j' unidas por un vértice del grafo:

$$v^*(z) = \frac{1}{2m}\sum_{j'=1}^{p}\sum_{j=1}^{p} m_{jj'}(z_j - z_{j'})^2 \tag{3.35}$$

Coeficiente de contigüidad de Geary:

Se busca aquí ver si la variable z_j puede ser considerada independiente o no de la estructura del grafo G. La fórmula

$$c(z) = \frac{v^*(z)}{v(z)} \tag{3.36}$$

puede ser discutida según lo siguiente:

c(Z) $\rightarrow$ 1: ocurre si la varianza local está próxima de la global, lo cual se traduce en una distribución aleatoria de Z en el grafo (es decir, z_j es independiente del grafo G)

c(Z) $\rightarrow$ 0: aquí la varianza local es bastante menor que la global, lo que implica que Z depende de la forma del grafo (o sea, toma valores sobre vértices contiguos en lugar de hacerlo tomando los vértices 2 a 2 de manera cualquiera)

c(Z) >1: los valores de z_j son demasiado diferentes sobre los vértices vecinos.

Cálculo de coeficientes de asimetría y curtosis de $c(z)$

Tomando como base el supuesto de normalidad de las z_j (es decir, que cada valor z_j proviene de una muestra de variables aleatorias normales estándar e independientes), es posible calcular los momentos de tercer y cuarto orden de $c(Z)$ para luego efectuar el cálculo del coeficiente de asimetría y el de curtosis.

Coeficiente de contigüidad de Von Neumann

Este es un caso particular de las series de tiempo. Cuando la relación de contigüidad se reduce a una relación de consecutividad –como es el caso de la estructura S2– hay cambios en el numerador de $c(Z)$:

$$c = \frac{1}{2(p-1)} \frac{\sum_{j=2}^{p} (z_j - z_{j-1})^2}{v(Z)} \tag{3.37}$$

ya que el número total de conexiones de una estructura de consecutividad –como S2– entre p elementos es igual a $(p-1)$. Este coeficiente es un caso particular de $c(Z)$ de Geary.

Utilización del coeficiente c

Volviendo al ejemplo visto al comienzo de la presente sección, se supone una v.a. Z que toma valores sobre las 9 partes en las cuales está dividido el corpus. Se puede así introducir a Z como la frecuencia relativa de las ocurrencias de una forma gráfica, por ejemplo *'de'*. Primeramente, se calcula el coeficiente $c(Z)$ de Geary para las estructuras de contigüidad **S1** y **S2**, o sea las cantidades $c(Z|S1)$ y $c(Z|S2)$. Si $c(Z|S1)$ es cercano a 1, z_j no sufre variaciones marcadas debido a la diferencia de autor, mientras que si $c(Z|S2)$ es cercano a 1 se dice que z_j no sufre de efecto cronológico.

HOMOGENEIDAD DE LOS FACTORES EN FUNCIÓN DE UNA ESTRUCTURA A PRIORI

A la estructura S1 le corresponde la matriz de contigüidad **M**. Partiendo de la totalidad del corpus, se construyen las respectivas tablas léxicas (compuestas de n unidades y p particiones del texto), las que luego serán sometidas a un análisis de correspondencias para así extraer $(p-1)$ factores, denotados ellos como $F_\alpha(j)$, donde α indica el número de factor, $\alpha = 1, ..., (p-1)$ y j el número de partición, $j = 1, ..., p$.

Homogeneidad de un factor respecto a una estructura

El cálculo del coeficiente de contigüidad c respecto de los factores $F_\alpha(j)$ nos permite medir el vínculo existente entre cada factor y la estructura S1. Para cada factor este coeficiente

$$c(F_\alpha) = c_\alpha = \frac{(p-1)\sum^*(F_\alpha(j) - F_\alpha(j'))^2}{(2m)\sum^j(F_\alpha(j) - F_\alpha(.))^2} \tag{3.38}$$

siendo:

$F_\alpha(.) = \frac{1}{p}\sum_{j=1}^{p} F_\alpha(j)$ la media del factor α para cada valor que toma en las respectivos fragmentos del corpus. Dicha cantidad es nula si todas las partes tienen el mismo peso.

$\sum^*$ es una sumatoria correspondiente a aquellos vértices j y j' tales que los mismos sean adyacentes

$\sum^j$ es una sumatoria en todos los j.

Homogeneidad en el espacio formado por los k primeros factores

El análisis realizado para cada uno de los factores es tomado de forma aislada respecto a una estructura de contigüidad, para luego ser complementado por el estudio de las distancias entre los elementos del conjunto J y el espacio generado por los primeros k factores. Se presenta entonces la distancia chi-cuadrado entre 2 puntos j y j', escrito en función de los n_f factores obtenidos en el análisis de correspondencias:

$$d^2(j, j') = \sum_{\alpha=1}^{n_f}(F_\alpha(j) - F_\alpha(j'))^2 \tag{3.39}$$

En particular, la distancia al cuadrado de dos puntos j y j' en función del espacio generado por los k primeros factores es:

$$d_k^2(j, j') = \sum_{\alpha=1}^{k}(F_\alpha(j) - F_\alpha(j'))^2 \tag{3.40}$$

Así pues, la cantidad

$$G_k = \frac{p(p-1)\sum^* d_k^2(j, j')}{(m)\sum^{jj'} d_k^2(j, j')} \tag{3.41}$$

es un coeficiente c calculado a partir de las distancias $d_k^2(j, j')$ que toman ahora valores entre los subespacios generados por los k primeros factores.

En el caso particular cuando $k = 1$, ocurre que $G_1 = c(F_1)$. Cuando k iguala a el número total de factores hallados por análisis de correspondencias, o sea cuando $k = (p-1)$, G_k muestra el

cociente entre la varianza "dentro de la estructura" y la varianza total a través de las distancias χ^2 que fueron calculadas sobre la tabla de frecuencias. Los valores intermedios de k corresponden a aquellas distancias "filtradas" por los primeros factores.

En la medida que no todas las particiones del corpus tienen la misma importancia desde un punto de vista numérico, se calculará en aquellas partes del corpus que son poco semejantes entre sí la proporción

$$G_k = \frac{p(p-1)\sum^{*} p_{.j}p_{.j'}d_k^2(j,j')}{(m)\sum^{jj'} p_{.j}p_{.j'}d_k^2(j,j')} \tag{3.42}$$

en donde las distancias cuadráticas mostradas son ponderadas por el peso relativo de cada una de las particiones.

ANÁLISIS DISCRIMINANTE TEXTUAL

Los métodos precedentes tenían un objeto meramente descriptivo. Las técnicas de métodos exploratorios, más dinámicas y con un mayor nivel de interacción que la simple descripción, recurren a la estadística multidimensional para obtener una clara visualización de elementos o reagrupamientos de los mismos, provenientes ya sea de textos enteros o descomposiciones de éstos. Esta es una búsqueda de organizaciones, de trazos estructurales, de resúmenes sugestivos.

Estos métodos "completan" una *panoplia*[27] de técnicas que son a menudo correspondidas con decisiones, aunque en realidad son más bien de ayuda a la hora de tomar decisiones. En el dominio de la estadística textual, esto concierne por ejemplo a la atribución de un texto a un autor.

Los procedimientos estadísticos que permiten realizar estas atribuciones, dependen del análisis discriminante. Apuntan básicamente a predecir la pertenencia de un "individuo" a una clase o una categoría, a partir de variables medidas sobre este individuo. Esta predicción es posible por una fase de aprendizaje realizada sobre un conjunto de observaciones para los cuales las variables y categorías son conocidas simultáneamente (*learning* o *training sample*). El recorte y a veces la elaboración de las unidades estadísticas constituye una fase importante de la búsqueda. El análisis discriminante textual será parte importante de estas etapas preparatorias.

Técnicas de discriminación

Las aplicaciones más corrientes del análisis discriminante textual son las siguientes:

Estilometría: es, a grandes rasgos, el análisis discriminante basado en el reconocimiento de patrones. Un caso puede ser la atribución de autores o fechas. Aquí se busca dejar de lado cierto "ruido"contenido en el texto para poder reconocer características de forma o estilo, a través de

[27] Armadura completa con todas las piezas

distribuciones de vocabulario, de índices o de *ratios*, como por ejemplo la cantidad de formas gráficas que aparecen en determinado texto[28]. A través de las citadas herramientas se logra determinar elementos "invariantes" a la naturaleza del autor, como por ejemplo modismos o frases típicas usadas por éste.

Discriminación Global: las aplicaciones realizadas en búsqueda documental, por ejemplo codificación automática o tratamiento de respuestas a preguntas abiertas, hacen referencia al contenido, a la sustancia de los textos. La forma en que una respuesta es formulada, importa menos que su clasificación en un grupo de documentos que presenta cierta homogeneidad respecto a su contenido.

SEMIOMETRÍA

Semiometría hace referencia a un término introducido por un escritor francés, Jean-François Steiner, que estaba interesado en investigaciones de mercado. En muchas encuestas relacionadas, es común incluir preguntas enfocadas a obtener información respecto a estilos de vida o valores de los encuestados. Estas preguntas consisten en general en la descripción de actitudes u opiniones basándose en palabras o frases. Desde un tiempo a esta parte, se vienen llevando a cabo este tipo de encuestas en distintos países con lenguas tan diferentes como el francés, el español, el italiano o el alemán, por citar algunas[21].

Aplicación a preguntas cerradas

Básicamente, en los cuestionarios de este tipo se inserta un listado con cantidad prefijada de palabras y los encuestados tienen que asignar un puntaje en una escala ascendente –desde el punto de vista del "agrado" del encuestado– de 7 puntos, siendo entonces 1 una marca asociada al "total desagrado" con la palabra mencionada, y 7 se asociará pues al "agrado completo"respecto a la misma. Los puntajes de la escala pueden variar también para hacer más comprensible al encuestado el "sistema de preferencias"al que es expuesto en el estudio, por ejemplo cambiando la escala a la siguiente: -3 como valor de "desagrado total" y +3 con valor de "agrado total", siendo el valor 0 el neutro.

Estas variables pueden ser analizadas mediante análisis de componentes principales. Según uno de los estudios publicados por Lebart, se obtienen ejes factoriales relativamente estables[21], y a su vez los resultados de usar las mismas palabras en diferentes países son bastante similares (incluso teniendo en cuenta el problema de la traducción, tanto desde un punto de vista semántico (significado) como también sintáctico).

[28] El índice *D* de Simpson es un buen ejemplo. Ver [27], sección 8.2.2.

Palabra1	-3	-2	-1	0	~~+1~~	+2	+3
Palabra2	-3	-2	-1	0	+1	+2	~~+3~~
Palabra3	-3	-2	~~-1~~	0	+1	+2	+3
Palabra4	-3	~~-2~~	-1	0	+1	+2	+3
Palabra5	-3	-2	-1	0	~~+1~~	+2	+3
Palabra6	~~-3~~	-2	-1	0	+1	+2	+3
Palabra7	-3	-2	-1	0	+1	~~+2~~	+3
Palabra8	~~-3~~	-2	-1	0	+1	+2	+3

Cuadro 3.4: Ejemplo

Aplicación a preguntas abiertas

En este caso, los entrevistados son invitados a mencionar cuál o cuáles son, en su parecer, las palabras que consideran "agradables" o "desagradables", sin disponer de una batería de posibles opciones. En este caso, se puede analizar el vocabulario a través del análisis de correspondencias de tablas léxicas como la **C**, la **T** o la T^+ y extraer conclusiones significativas al respecto.

A modo de conclusión, se puede decir que la Semiometría presenta las siguientes ventajas:

- Es una herramienta versátil que puede ser usada en diferentes ocasiones;
- Es posible hacer comparaciones internacionales en campos de investigación muchas veces considerados como "intra-culturales";
- Puede establecerse si existe evolución y "variación cronológica" desde muchos puntos de vista, gracias a la relativa permanencia y pertinencia de los ítems usados;
- Simplicidad y riqueza otorgada por los cuestionarios con preguntas abiertas.

RECUPERACIÓN DE LA INFORMACIÓN

Sistemas de Búsqueda de Información o simplemente *Recuperación de Información*, ciencia que se nutre de matemática, inteligencia artificial, lingüística, estadística, informática y física entre tantas otras, tiene como objeto central –como el nombre lo deja ver– la búsqueda de metadatos, de documentos o información dentro de ellos, al interior de una base de datos –como puede ser el archivo de volúmenes pertenecientes a una biblioteca–o, por ejemplo, una "gran base de datos", como la internet. Los sistemas diseñados para estos fines son capaces de escoger, de entre una gran cantidad de documentos o bases de datos, aquellos archivos cuyo contenido hace referencia a algún tema específico, seleccionado de antemano por la persona que así lo requiere, buscando así disminuir la llamada "sobrecarga de información".

Se busca representar, almacenar y organizar la información contenida en documentos para atender los requerimientos del usuario. Un sistema de información eficiente será aquel que logre construir el "tesaurus"[29] de las palabras (es decir, agrupar estas palabras en categorías pertinentes al ámbito de aplicación), de asignar palabras índices a los documentos y de extraer frases que permitan resumir el o los documentos en cuestión. El enfoque es, ante todo, pragmático: se trata de disponer de una herramienta que otorgue al investigador opciones, con tasas de error, costos y restricciones asociadas a determinadas selecciones.

Historia

Esta disciplina remonta sus orígenes hacia la década de 1940 cuando Vannevar Bush, ingeniero y científico estadounidense postula en el artículo *As we may think* en 1945, la idea de utilizar computadoras para buscar y encontrar fragmentos relevantes de información, a través de un dispositivo mecánico llamado *Memex* en el cual se almacenarían todo tipo de documentos. En los años siguientes, tanto las fuerzas armadas norteamericanas –en pleno contexto de la Guerra Fría– así como también la industria de la informática, comenzaron a trabajar en la creación de sistemas mecánicos para la búsqueda y detección de información, hasta la época contemporánea donde las herramientas de búsqueda en internet –denominadas *web search engines*– facilitan la búsqueda de información a todo nivel. Con sólo acceder una palabra o frase relacionada al tema de interés, se obtiene un listado de direcciones en donde se encuentran estas palabras o frases –o por qué no, temas relacionados a éstas–, imágenes, comunidades, etc. Este tipo de aplicaciones son las más visibles dentro de la Recuperación de Información[30].

Por otra parte, basados en la recuperación de la información como eje central, se han presentado en las recientes Jornadas de Análisis de Datos Textuales (JADT) temas muy diversos, como por ejemplo los siguientes:

- Estudio de avisos de trabajo en internet de empresas relacionadas con las tecnologías de la información y aplicación del Partial Textual Credit Model (PTCM) para ordenar los requerimientos de estas empresas, y compararlos de modo pertinente con los programas académicos actuales
- Clasificación de documentos multimedia a través de los n-*grams*[31].
- El estudio del manejo de la relación con el cliente en comercio electrónico (electronic Customer Relationship Management) a través de herramientas de minería de datos (técnica

[29] Ver anexoA.

[30] Entre los casos más notorios figuran Google, Yahoo!, Live Search entre otras.

[31] Técnica independiente del idioma o lenguaje usado, que es utilizada para descomponer imágenes, textos o estructuras musicales en n partes. Por ejemplo, si queremos descomponer a la palabra informática en 3-*grams*, obtenemos lo siguiente: inf, nfo, for, orm, rmá, mát, tic, ica.

mencionada a continuación) y Análisis Discriminante No Paramétrico en un sitio web, basados en *micro-data*[32].

Data Mining

El principio de la *minería de datos* está estrechamente relacionado con la búsqueda de información: se trata de extraer –a través de algoritmos sofisticados– información previamente desconocida, válida y útil a partir de bases de datos (en general muy grandes) y luego usar la misma para la toma de decisiones. Si bien el término es relativamente reciente, las técnicas y la tecnología –como pudo observarse líneas arriba– ya han forjado su historia.

El término *data mining* es usado en aplicaciones relacionadas a modelos predictivos o de reconocimiento de patrones, como es el caso de las redes neuronales, o bien en el llamado *knowledge discovery* en donde se obtiene información legible y fácilmente entendible para el usuario.

Text Mining

Minería de textos o también *minería de datos textuales* hacen referencia al proceso de extraer información de alta calidad desde un texto. Esta "alta calidad" es definida previamente a través de técnicas como el aprendizaje por patrones estadísticos[33].

Generalmente, la minería de textos apunta hacia el agrupamiento y categorización de textos, extracción de conceptos y resumen de documentos, por citar algunos. En la actualidad, la minería de datos textuales multilingüe crece en adeptos a nivel mundial, dada su capacidad de extraer información entre textos de diferentes fuentes y lenguas. Por cierto, también otras áreas utilizan de forma creciente al *text mining*: aplicaciones biomédicas, tecnologías de la información, etc. Una de las aplicaciones más comunes es en los filtros *anti-spam* de los correos electrónicos, para detectar posible material no deseado y clasificarlo como tal.

[32]Micro-data hace referencia a datos sobre objetos individuales (p. ej. personas, transacciones, compañías, etc.). Extraido de: ils.unc.edu/õhjs/stats/tutorial_BasicConcepts.html

[33]Traducido de *statistical pattern learning*: una aproximación a la inteligencia artificial basada en la modelización estadística de los datos, basada ésta a su vez en las teorías de la probabilidad y toma de decisiones.

Parte III

Resultados

Capítulo 4

ACM Preguntas Abiertas Post-Codificadas

En el presente capítulo se exponen los resultados de análisis de correspondencias múltiple y simple, para cumplir con uno de los objetivos planteados en la investigación: comparar los resultados obtenidos mediante técnicas de análisis estadístico de textos que se muestran en el siguiente capítulo, con los obtenidos luego de aplicar análisis de correspondencia a las respuestas abiertas poscodificadas.

Cabe destacar que en todos los gráficos presentados, las dos coordenadas de cada punto son transformadas en *rangos*. En cada eje, los valores numéricos n son clasificados y sustituidos por sus rangos. El valor más pequeño tiene rango 1 y el más grande, n. Así, la escala original es sustituida por una aritmética, para lograr una más clara visión de los elementos de un gráfico factorial.

4.1. Análisis de Correspondencia Múltiple

En una primera instancia se procedió a realizar un análisis de correspondencia tomando como activas a las siguientes variables: *Grado* al cual concurría el/la niño/a al momento de la encuesta, *Trabajo del padre*, *Trabajo de la madre*, *Quién contestó el formulario*, si el/la niño/a *Repitió* algún año, *Instrucción del padre*, *Instrucción de la madre*, si el/la niño/a es *Saludable*, *Síndrome Global* y la variable cruzada *Sexo-Edad*; y utilizando las *preguntas post-codificadas A y B* como suplementarias[1].

Se obtuvo entonces la figura 4.1 que representa al primer plano factorial.

Las variables relacionadas con los padres contribuyen en un 79,2 % a la formación del primer eje factorial[2]. Al no estar las mismas relacionadas con los niños *per sé*, se optó por dejarlas como

[1] Ver Anexo B.

[2] Ver salida en el Anexo D.

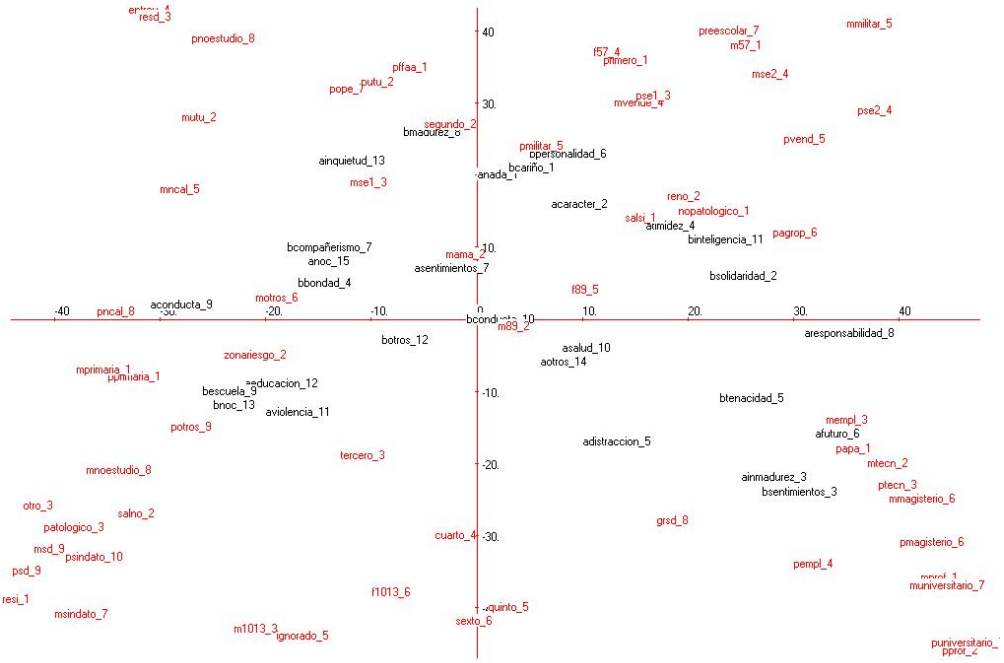

Figura 4.1: Primer plano factorial tomando como variables suplementarias las preguntas abiertas postcodificadas.

suplementarias en una segunda etapa.

De esta forma, se realizó otro ACM –teniendo en cuenta los comentarios mencionados líneas arriba– considerando a las variables *Grado*, *Quién contestó el formulario*, si el/la niño/a *Repitió* algún año, si el/la niño/a es *Saludable*, *Síndrome Global* y la variable cruzada *Sexo-Edad* como activas; y utilizando las *preguntas post-codificadas A y B*, *Trabajo del padre*, *Trabajo de la madre*, *Instrucción del padre* e *Instrucción de la madre* como suplementarias.

En este caso las variables *Sexo-Edad* y *Grado* son las que más contribuyen a la formación del primer eje factorial. Se puede observar que la contribución absoluta es casi el 80 % en ese eje factorial y poco más del 61 % en el segundo[3].

La figura 4.2 representa el primer plano factorial para este segundo ACM, considerando a la vez a las variables activas y suplementarias.

4.1.1. Descripción del Análisis

A continuación se presenta la descripción e interpretación de algunos de los principales ejes y planos factoriales[4]:

[3]Ver salidas en el Anexo D.

[4]Se presentan en el Anexo D las salidas correspondientes a los cinco primeros ejes factoriales.

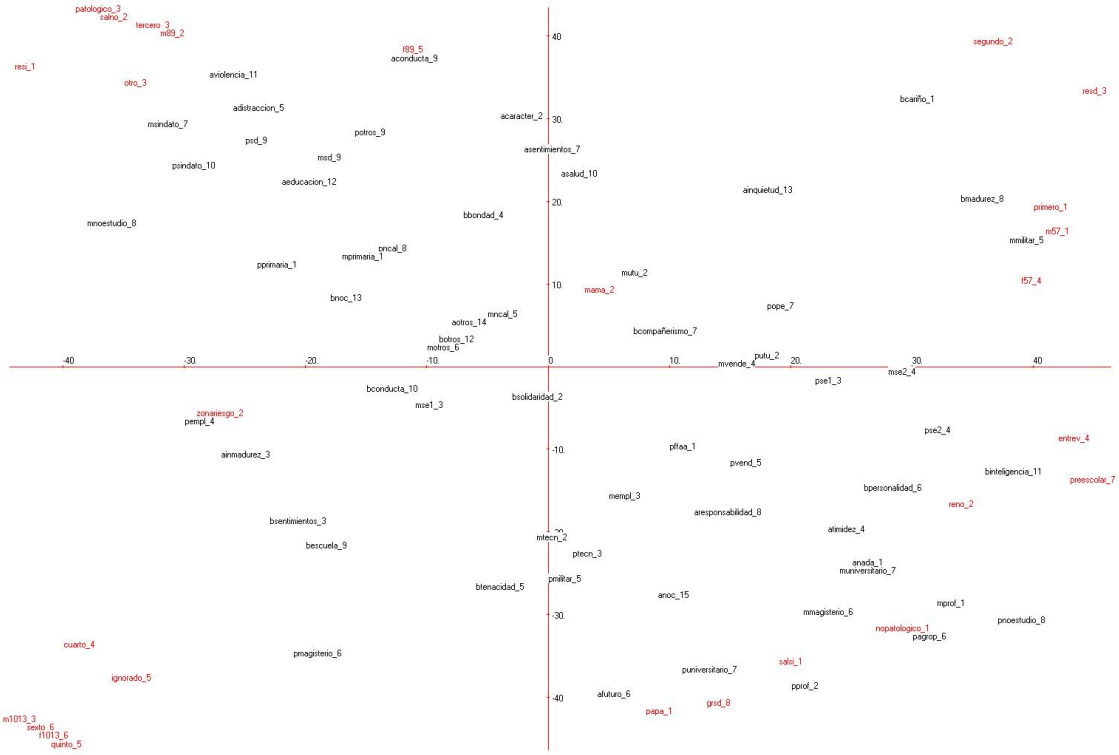

Figura 4.2: Primer plano factorial, tomando como variables suplementarias las preguntas abiertas postcodificadas, el trabajo y la instrucción de los padres.

Primer eje Las variables que más contribuyen a la formación de este factor son *Sexo-Edad* y *Grado* como fue destacado anteriormente. Además, alguna de las modalidades de estas variables están relativamente bien representadas[5]. Se puede considerar a este eje como un *"eje del desarrollo"*, en el sentido que, de derecha a izquierda se proyectan tanto las edades de los niños/as como el grado, de forma creciente. Para facilitar esta visión, en la figura 4.3 se trazaron líneas de colores uniendo los respectivos puntos.

Segundo eje *Síndrome global* y *Saludable* comienzan a contribuir más en la formación de este eje, aunque igualmente las variables *Sexo-Edad* y *Grado* continúan siendo las que más aportan a la construcción del mismo.

Tercer eje En este eje, la diferencia de contribuciones relativas entre *Sexo-Edad* y *Grado* por un lado y *Síndrome global* y *Saludable* por otro disminuye significativamente; estas dos últimas variables pasan a explicar más de un tercio de la formación del eje.

Cuarto y quinto eje Si bien en conjunto las variables más importantes siguen siendo *Sexo-Edad* y *Grado* en ambos casos, hay otra –*Quién respondió la encuesta*– que empieza a influir significativamente en la formación de estos ejes.

[5]Es de esperar que, al usar muchas modalidades en alguna de las variables –como es el presente caso–, la calidad de representación así como la inercia explicada por cada eje sean bajos.

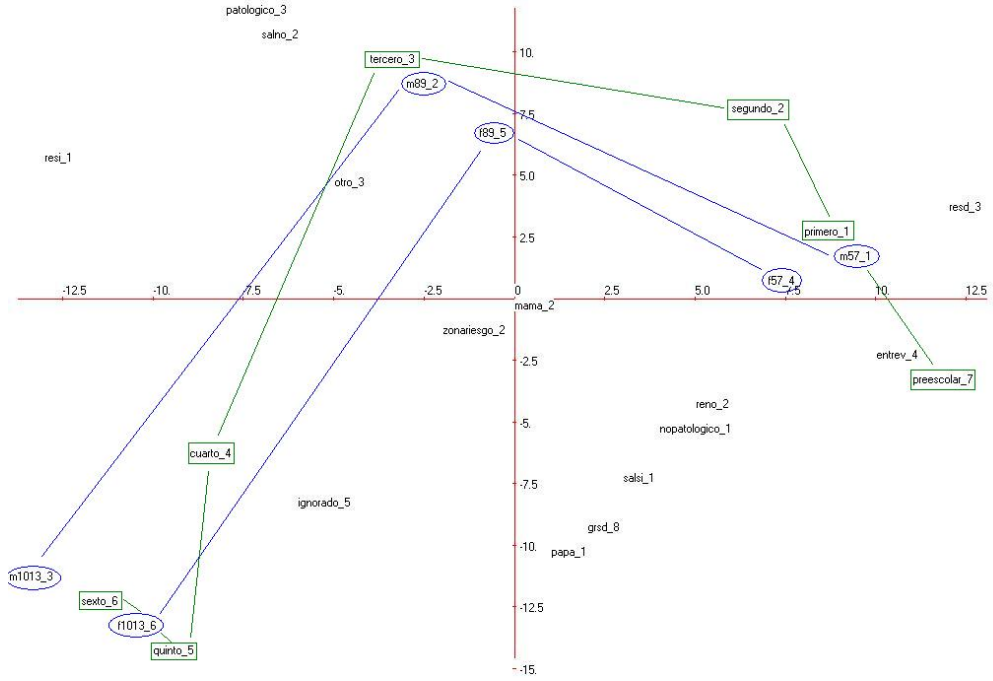

Figura 4.3: Variables activas proyectadas en el primer plano factorial

Primer plano factorial El plano formado por los dos primeros factores presenta ciertas peculiaridades. Por ejemplo, llama la atención el hecho que los varones se proyectan levemente por encima de las niñas de la misma edad, aunque siguiendo el mismo trayecto. Por otra parte, a pesar de ser ortogonales entre sí linealmente hablando, los primeros ejes factoriales parecen estar aproximadamente relacionados en un sentido no-lineal. Esto ocurre en general cuando las modalidades están ordenadas a priori –como es el caso de las variables *Grado* y *Sexo-Edad*– y su existencia hace que el segundo factor y los siguientes no puedan interpretarse por sí mismos, ya que son funciones polinómicas del primero. De esta manera, la información contenida en éste sigue apareciendo en los restantes. Esto pone de manifiesto la existencia del denominado *efecto Guttman* o *silla de montar*[6].

Plano formado por ejes 1 y 3 Tanto el segundo como el tercer eje hacen las veces de *"línea divisoria entre sanos y enfermos"*, pero es en este último donde se aprecian mejores contribuciones a su formación y calidad de representación de las modalidades de las variables *Síndrome Global* y *Saludable*. Esto se puede apreciar en la figura 4.4.

De esta forma, dado que el tercer eje factorial aporta más información que el segundo -a nivel del subespacio bidimensional- en cuanto a calidad de representación de las modalidades de las variables *Síndrome Global* y *Saludable*, se optó por el plano formado por los ejes 1 y 3. En la figura 4.5 en donde aparecen proyectadas como suplementarias las distintas modalidades de las preguntas A y B postcodificadas, se marcaron aquellas categorías cuyos valores-test eran sig-

[6]Citado en [15], pág. 226-232.

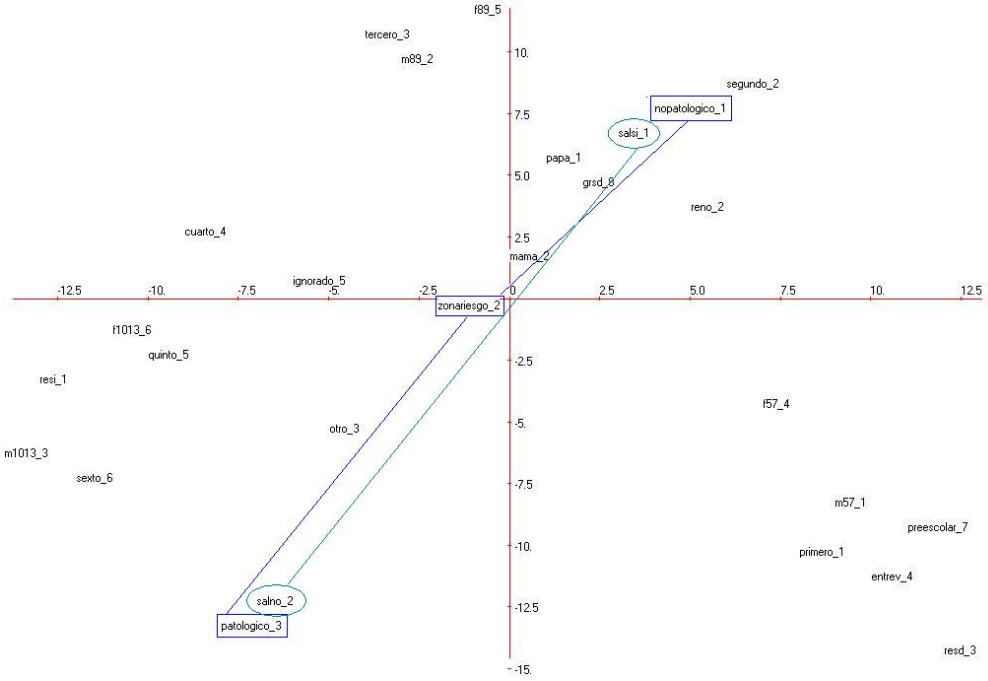

Figura 4.4: Plano formado por los ejes factoriales 1 y 3

nificativos. De esta forma, las modalidades que poséen valores-test que caen fuera del intervalo $[-1{,}96; 1{,}96]$ están encerradas en óvalos verdes para el eje 1 y en rectángulos de color azul para el eje 3.

Es de destacar, por un lado, la asociación de las modalidades *Nada* e *Inteligencia* –de las preguntas A y B, respectivamente– con las modalidades *No Patológico* y *Saludable* –de las variables *Síndrome Global* y *Saludable*, en forma respectiva–; y por otro lado Educación y Conducta –ambas modalidades de la pregunta A– con Patológico y No Saludable.

Por último, pero no menos importante, la inercia explicada por los cinco primeros ejes factoriales es de 36 %. Para obtener porcentajes "menos pesimistas" se usó como ponderadores los índices de Benzécri y Greenacre, obteniendo así porcentajes un tanto mayores (69 % y 40 %, respectivamente)[7]. Esto es esperable, ya que las variables usadas poséen, en mucho de los casos, un gran número de modalidades[8].

[7]El índice de Benzécri no considera aquellos $\lambda_i < \overline{\lambda}$ y se calcula como $\rho(\lambda_s) = \left[\frac{J}{J-1}\right]^2 \left[\lambda_s - \frac{1}{J}\right]^2$ y el índice de Greenacre –leve reformulación del anterior– se calcula como $\rho(\lambda_s) = \left[\frac{J}{J-1}\right]^2 \left[\sqrt{\lambda_s} - \frac{1}{J}\right]^2$.

[8]En general, es recomendable no sobrepasar las 8 modalidades[11].

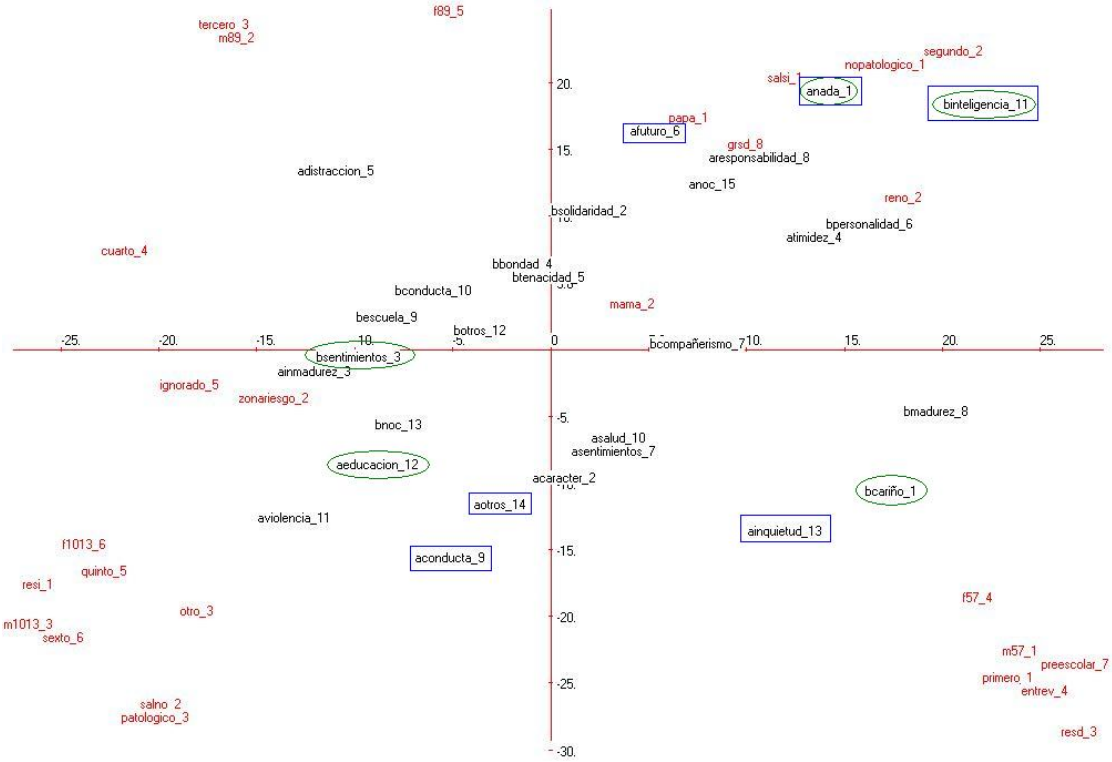

Figura 4.5: Proyección de las modalidades de las preguntas A y B en ejes 1 y 3

4.1.2. Análisis de Cluster

A partir del análisis factorial, el software DTM crea grupos utilizando el método jerárquico de vecinos recíprocos. Este algoritmo funciona de la siguiente manera:

Descripción del algoritmo:

Se define vecino recíproco de esta forma: dos elementos x_i y $x_{i'}$ son vecinos recíprocos si x_i es el elemento más cercano a $x_{i'}$ y viceversa.

1^a **etapa:** se crea una cadena de elementos sucesivos encontrando elementos próximos de forma sucesiva:

$x_1 \longrightarrow x_2 \longrightarrow x_3 \longrightarrow ... \longrightarrow x_{i-2} \longrightarrow x_{i-1} \longrightarrow x_i \longrightarrow ...$

Esta "cadena" se detiene necesariamente cuando dos elementos sucesivos son vecinos recíprocos:

$... \longrightarrow x_i \longrightarrow ... \longrightarrow x_{k-1} \leftrightarrow x_k$

De esta forma, la cadena se detendrá si x_{k-1} es también el vecino más próximo de x_k.

2^a **etapa:** para $k = 2$, la cadena comienza con un elemento formado por dos vecinos recíprocos,

$x_1 \leftrightarrow x_2$

Se escoge un nuevo elemento a partir del cual una cadena es construida, y la misma se detiene sobre nuevos vecinos recíprocos cuya unión forma un nodo.

3^a **etapa:** para $k > 2$ la búsqueda de vecinos recíprocos comienza en el elemento x_{k-2}. El algoritmo finaliza al crearse los $n - 1$ noods En general, la construcción de algoritmos de agregación genera un número importante de cálculos. Este algoritmo es efectivo en el sentido que reduce el número de operaciones de n^3 a n^2, siendo n el número de elementos a clasificar.

Descripción del análisis

En una primera instancia, al no poseer información *a priori* respecto a los posibles grupos a crear se optó por cinco clusters, luego descartado debido a la agrupación despareja (alguno de estos grupos se formaba tardíamente con pocos individuos). Sucesivamente, la cantidad de grupos deseados pasó de cuatro a tres, por similares motivos. En este último caso, se constató una formación relativamente homogénea en cuanto a la cantidad de individuos por grupo, además de valores-test significativos en todos los casos. La posición de los mismos se observa en la figura 4.6. En cuanto a la cantidad de ejes factoriales usados, se tomó en un principio 12 –cantidad asignada por defecto en el DTM–, para luego pasar a cinco ejes, sin verificarse cambios significativos en la formación de los grupos.

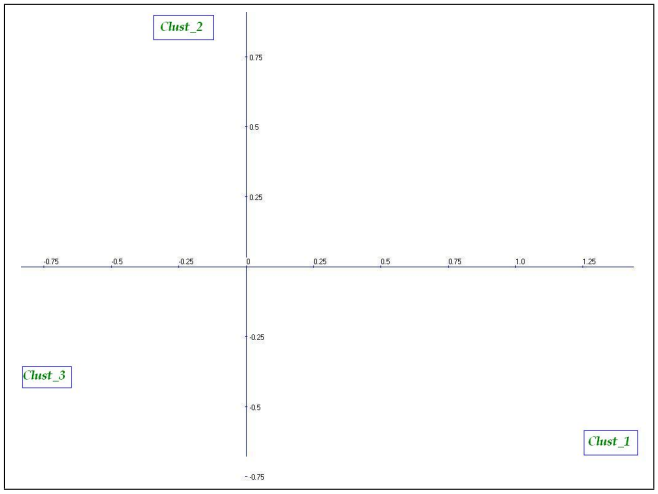

Figura 4.6: Grupos formados, proyectados en los ejes factoriales 1 y 3

De esta forma, los grupos presentan las siguientes características:

- **Primer cluster:** formado por 382 individuos, cuyos rasgos sobresalientes son *niños y niñas pequeños/as* (de entre cinco y siete años), en los *cursos iniciales* de educación básica (preescolares y primer año), *no repetidores*.
- **Segundo cluster:** se forma de 468 niños de ambos sexos, de *mediana edad* (de ocho a nueve años), cursando *segundo y tercer año* de escuela.
- **Tercer cluster:** aquí figuran 458 niñas y niños con edades más cercanas a la pubertad –de *10 a 13 años*–, en los cursos superiores dentro de la enseñanza primaria (4°, 5° *y* 6° *años*), *repetidores*.

Como se puede apreciar, la diferenciación entre grupos es claramente por tramo de edad, grado y, en menor medida, si son repetidores, no distinguiéndose entre sexos.

4.2. Análisis de Correspondencia Simple

Aquí se muestran algunos cruces de variables que, luego de realizar los ACM mencionados en la subsección anterior, parecieron interesantes.

Síndrome Global* con *Saludable : luego de obtener los ejes factoriales correspondientes, se procedió a proyectar las preguntas A y B para observar posibles asociaciones entre lo que los padres dicen de sus hijos y los resultados consecuencia de la aplicación del cuestionario CBCL. Si se observa la figura 4.7, donde aparece el plano factorial formado por los ejes 1 y 3, se aprecia una diferenciación entre los cuatro cuadrantes del gráfico. En el primero, la categoría que más aporta, la mejor representada y que tiene un valor-test significativo es *Patológico*. Siguiendo esta lógica, se observan las categorías *No Saludable*, *No Patológico* y *Saludable* en los cuadrantes 2, 3 y 4 respecivamente, con elevados valores-test y buena calidad de representación.

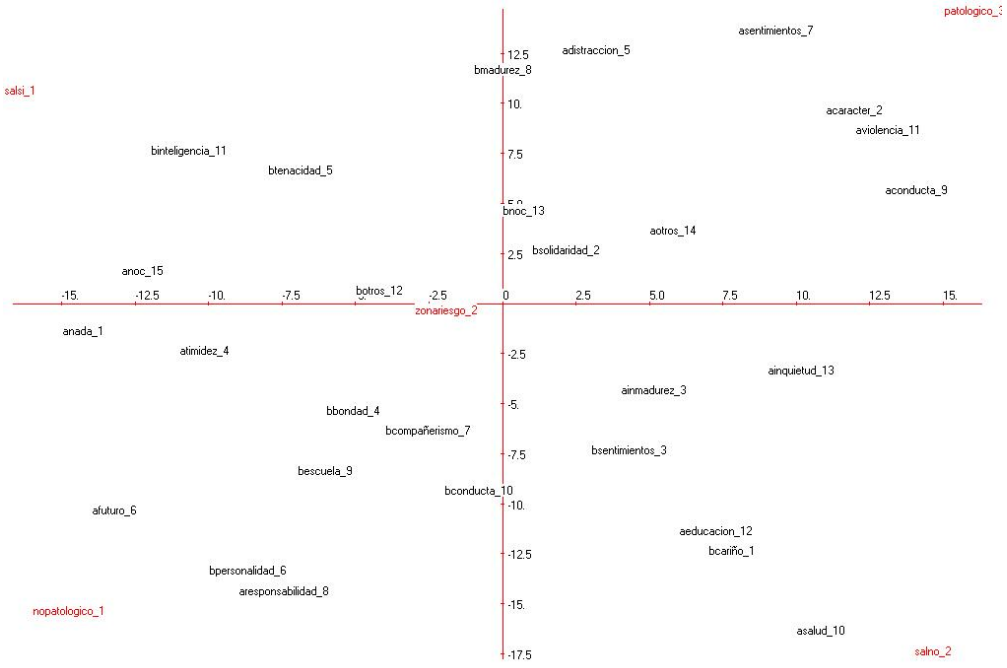

Figura 4.7: Plano formado por ejes factoriales 1 y 3

Si ahora se observan las proyecciones de las variables suplementarias en estos 4 cuadrantes se obtiene lo siguiente:

- *Cuadrante 1*: se observa las categorías *Carácter*, *Conducta*, *Violencia*, *Sentimientos* y *Distracción* pertenecientes a la pregunta A (*¿Qué es lo que más le preocupa acerca de su*

hijo/a?).

- *Cuadrante 2*: aquí las categorías significativas son *Salud*, *Inquietud* y *Madurez*, en este orden, correspondientes a la pregunta A y *Cariño* a la pregunta B (*¿Qué es lo mejor que le ve a su hijo/a?*).
- *Cuadrante 3*: *Futuro* y *Nada* de la pregunta A son las que sobresalen.
- *Cuadrante 4*: *No contesta* de pregunta A e *Inteligencia* de pregunta B se destacan en este cuadrante, por su significación en el primer eje factorial.

Síndrome Global* con *Repitió : Cruzando estas dos variables, y como se puede observar en el cuadro 4.1, la contribución de cada una a la construcción de los ejes factoriales, se puede decir que es indiferente escoger entre cualquier posible combinación de los ejes 1 a 4 para obtener y así extraer comentarios de un plano factorial. Sin embargo, si se observa la calidad de representación de cada una de las modalidades en los diferentes ejes factoriales, así como los valores-test asociados, el plano formado por los ejes 1 y 4 es el más adecuado en este cruce.

categories			coordinates				contributions				squared cosine				test-values			
iden - libelle	p.rel	disto	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
8 . repano_V#8																		
cat1 - resi_1	9.40	4.32	-1.62	.05	-.05	1.30	40.5	.1	.1	40.5	.61	.00	.00	.39	-28.2	.9	-.9	22.5
cat2 - reno_2	40.33	.24	.37	-.08	.08	-.30	9.2	.5	.5	9.2	.58	.02	.02	.37	27.6	-5.7	5.6	-22.0
catb - resd_3	.27	185.86	.76	9.72	-9.51	-.61	.3	49.5	49.5	.3	.00	.51	.49	.00	2.0	25.8	-25.2	-1.6
--- cumul. contribution =							50.0	50.0	50.0	50.0								
16 . sindglo1rec_V#16																		
cat1 - nopatologico_1	36.20	.38	.46	.13	.13	.37	12.6	1.2	1.2	12.6	.56	.05	.04	.35	26.9	7.8	7.6	21.5
cat2 - zonariesgo_2	5.70	7.78	-.67	-1.90	-1.85	-.54	4.2	40.1	40.1	4.2	.06	.46	.44	.04	-8.7	-24.6	-24.0	-7.0
cat3 - patologico_3	8.10	5.17	-1.58	.74	.72	-1.26	33.2	8.7	8.7	33.2	.48	.11	.10	.31	-25.1	11.8	11.5	-20.1
--- cumul. contribution =							50.0	50.0	50.0	50.0								

Cuadro 4.1: Salida DTM

En el presente caso, se repite nuevamente lo visto líneas arriba: la diferenciación en 4 cuadrantes, según la posición de las categorías activas *No Patológico*, *No Repitió*, *Patológico* y *Repitió* en los cuadrantes 1, 2, 3 y 4 respecivamente.

A continuación se presenta la figura 4.8, representando el plano formado por los ejes 1 y 4. Las variables suplementarias están proyectadas en este plano factorial. Las modalidades encerradas en óvalos de color verde son aquellas que presentan valores-test significativos en el eje 1, las encerradas en rectángulos azules presentan valores significativos en el eje 4 y las restantes no aparecen como significativas en este plano factorial.

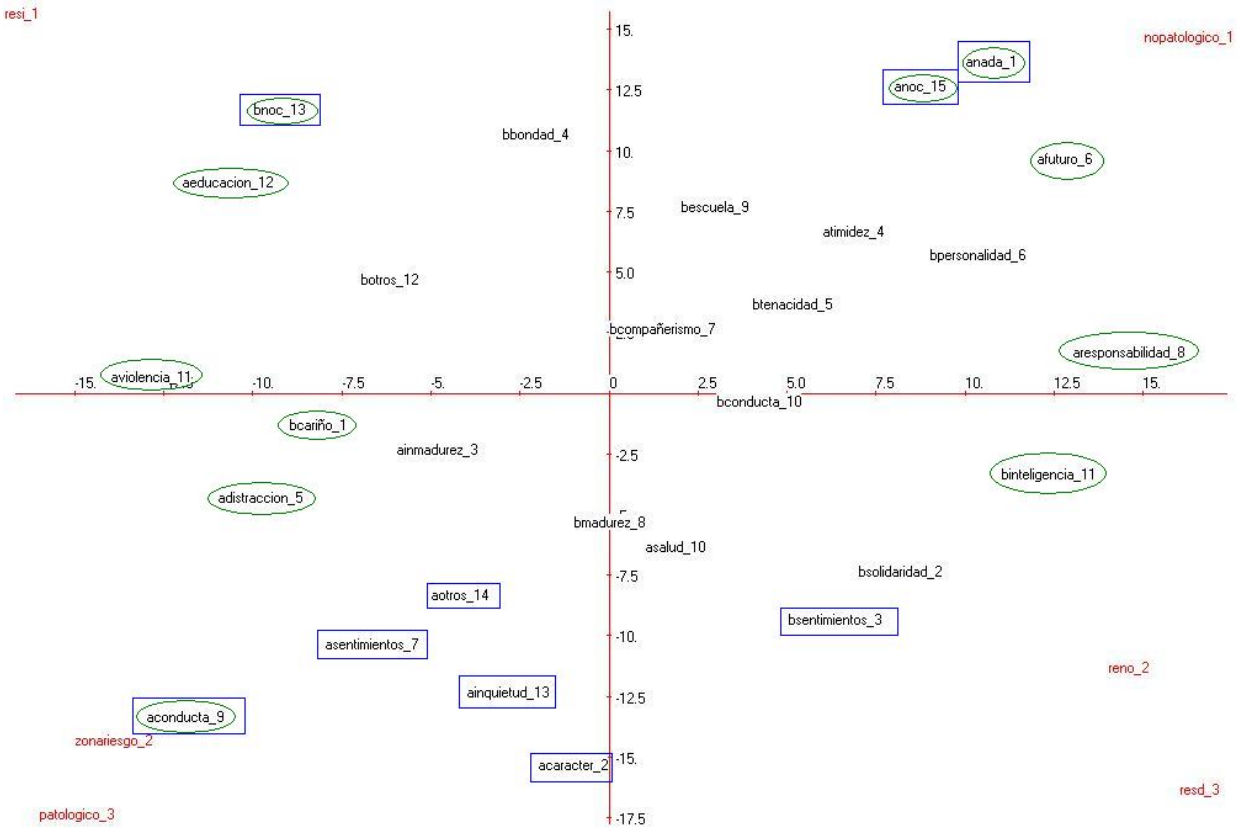

Figura 4.8: Plano factorial formado por los ejes 1 y 4

Capítulo 5

Análisis Textual de Preguntas Abiertas

El insumo principal de este capítulo es el corpus textual formado por las dos respuestas a preguntas abiertas existentes en el formulario CBCL, además de las variables cualitativas utilizadas en los capítulos precedentes. En las siguientes secciones, este texto es analizado de dos formas: como un solo corpus primero, y separando los subcorpus de las preguntas A y B después.

El número total de ocurrencias asciende a $T = 24543$, con un vocabulario –es decir, cantidad de palabras diferentes– $V = 2600$, con lo cual el número de palabras distintas asciende a $10,6\,\%$. A modo de filtro, se fijó un umbral de frecuencia 10 –o sea, aquellas formas gráficas que aparecían menos de diez veces en el total del corpus fueron descartadas–; así, el nuevo tamaño del corpus asciende a $T^* = 19828$ y la cantidad de palabras diferentes a $V^* = 299$.

Al observar por separado los subcorpus formados por las preguntas A y B, se encuentran algunas diferencias. Respecto al tamaño, el corpus B es mayor que el A, siendo los valores de T respectivamente $T_A = 11023$ y $T_B = 13520$. Si se observa la cantidad de palabras diferentes, el corpus A es más rico en vocabulario que el B: $V_A = 1771$ contra $V_B = 1541$, y de esta forma el porcentaje de palabras distintas es $16,1\,\%$ y $11,4\,\%$ para los corpus A y B, respectivamente. En una etapa posterior, al tomar el umbral de frecuencias mencionado, los tamaños de los corpus son $T_A^* = 7848$ y $T_B^* = 10766$, y sus respectivos vocabularios $V_A^* = 148$ y $V_B^* = 180$.

5.1. Visualización de tablas léxicas

En esta sección se realiza la descripción de las tablas léxicas teniendo en cuenta el corpus y los subcorpus arriba mencionados.

En primer lugar se obtuvo la tabla léxica del corpus total, conformada por 299 palabras y 1308 respuestas. Sobre esta tabla se realizó un análisis de correspondencia, obteniéndose 8 ejes factoriales –cantidad fijada por defecto–. Teniendo en cuenta los primeros 7 ejes y utilizando el método de vecinos recíprocos, fueron construidos 15 grupos. Luego de este paso, se construyó la

tabla léxica agregada, formada por las 299 formas gráficas mencionadas anteriormente y los 15 grupos obtenidos. Se obtuvieron así los elementos característicos y las respuestas modales para cada grupo.

Posteriormente, se repiten los pasos mencionados utilizando los subcorpus de las preguntas A y B, respectivamente.

Tabla léxica del corpus total: En el cuadro 5.1 se observan: el nombre de cada cluster, asignado por defecto por el DTM[1], la cantidad de individuos, las coordenadas de los clusters y sus valores-test. Se puede destacar al respecto que los valores-test de los grupos formados son, en la amplia mayoría de los casos, significativos en los cinco primeros ejes factoriales.

En cuanto a la cantidad de respuestas, los grupos más importantes son, respectivamente el n° 1 con 682, el n° 6 con 208, el n° 5 con 156 y el n° 4 con 153. Este ordenamiento casi coincide con la longitud del subcorpus generado para cada cluster; el orden en este caso es: $T^1_{A,B} = 13000$, $T^6_{A,B} = 2995$, $T^4_{A,B} = 1667$, $T^5_{A,B} = 1538$. La media global de palabras usadas por respuesta fue de 18,8; estando sólamente el grupo 1 por encima de esta media (23,9).

coordinates and test-values on axes 1 to 5

classes				coordinates					test-values				
iden - name	effec.	abs.w	disto	1	2	3	4	5	1	2	3	4	5
cut a of the tree into 15 classes													
a01a- class 1 / 15	682	682.00	.17	-.15	-.34	.12	.07	.02	-11.14	-27.19	10.27	5.85	1.85
a02a- class 2 / 15	10	10.00	9.94	.36	-.48	-2.15	-.41	2.11	2.20	-3.25	-15.07	-2.87	14.91
a03a- class 3 / 15	1	1.00	152.98	-2.41	1.49	-9.84	-4.44	-1.65	-4.64	3.15	-21.70	-9.84	-3.67
a04a- class 4 / 15	153	153.00	1.26	-.42	.12	-.59	-.35	-.70	-10.58	3.23	-17.21	-10.07	-20.62
a05a- class 5 / 15	156	156.00	1.73	-.75	.83	.46	-.23	.44	-19.17	23.32	13.60	-6.84	13.01
a06a- class 6 / 15	200	200.00	.66	.69	-.12	.27	-.15	-.06	20.53	-4.07	9.16	-5.16	-1.97
a07a- class 7 / 15	59	59.00	4.53	1.65	.40	.25	-.46	.07	24.98	6.65	4.39	-7.93	1.14
a08a- class 8 / 15	7	7.00	46.58	.97	.31	-1.12	-.10	-2.67	4.97	1.74	-6.57	-.60	-15.75
a09a- class 9 / 15	26	26.00	14.23	-.41	1.59	-1.10	1.99	-.31	-4.05	17.36	-12.54	22.73	-3.60
a10a- class 10 / 15	5	5.00	112.08	5.18	4.05	-2.18	.34	-1.44	22.31	19.23	-10.74	1.71	-7.20
a11a- class 11 / 15	1	1.00	772.33	10.53	9.36	-5.44	.60	-4.67	20.27	19.83	-12.00	1.33	-10.41
a12a- class 12 / 15	3	3.00	184.81	1.70	-.32	-8.63	-1.78	10.01	5.66	-1.19	-32.97	-6.82	38.66
a13a- class 13 / 15	1	1.00	907.49	2.72	-2.04	-19.68	-3.51	21.62	5.23	-4.31	-43.39	-7.79	48.17
a14a- class 14 / 15	3	3.00	215.32	1.35	4.22	-1.18	12.44	-.12	4.50	15.50	-4.51	47.81	-.47
a15a- class 15 / 15	1	1.00	926.77	3.33	7.12	-1.42	25.15	-.21	6.41	15.09	-3.13	55.75	-.46

Cuadro 5.1: Descripción de los clusters formados

A continuación se describe qué ocurre con algunos de estos grupos en cuanto a sus respuestas modales y los elementos característicos:

- *Grupo 1:* no hay similitudes claras en las primeras respuestas características, posiblemente por la existencia de artículos y pronombres como elementos característicos de este grupo.

[1]El DTM asigna por defecto la numeración *aXXa* para los grupos formados, siendo *XX* números que en el caso particular del cuadro 5.1 varían entre 01 y 15.

- *Grupos 2, 12 y 13:* podemos destacar para la pregunta A como para la B que tanto el contexto discursivo de las respuestas modales como los elementos característicos hacen referencia a la *inquietud*.
- *Grupos 4 y 5:* en ambos grupos se obtienen respuestas modales relacionadas con el *compañerismo*, diferenciándose éstos en que, en el primero los elementos característicos con especifidades positivas apuntan hacia los niños, y en el segundo hacia las niñas.
- *Grupo 7: voluntad*, *solidaridad* y *carácter* son sus elementos característicos, que aparecen además en las respuestas modales.
- *Grupo 8:* el *aprendizaje* aparece a la vez como elemento característico y se repite en las respuestas modales.
- *Grupo 9: nada* es parte de las respuestas modales y los elementos característicos al mismo tiempo.
- *Grupos 10, 14 y 15: inteligencia* y *madurez* se destacan en estos grupos.

Tabla léxica del subcorpus A: A continuación se presenta la figura 5.1, que muestra formas gráficas y los clusters formados a partir de éstas. Cabe destacar que, las formas gráficas *nada* e *inquietud* no están representadas en sus verdaderas coordenadas –están bastante más alejadas del baricentro de lo que se muestra aquí–, y lo mismo sucede con los clusters 3, 12, 13, 14 y 15.

En las siguientes líneas, se describen brevemente los grupos formados:

- *Grupo 2:* tomando en cuenta los elementos característicos junto con las respuestas modales, *falta de aprendizaje* puede tomarse como referente de este grupo.
- *Grupo 3 y 15: por ahora* y *nada* se destacan tanto en las respuestas modales como en los elementos característicos.
- *Grupo 4 y 5:* se nutren de comentarios de padres preocupados por el *futuro* de sus hijos.
- *Grupo 6:* aquí, la *educación* hacia el futuro es su característica más sobresaliente; pues si bien "la educación" aparece en las tres respuestas modales de la salida, las palabras *tenga*, *una* y *buena* figuran como elementos característicos en este grupo.
- *Grupo 7:* indudablemente, la *conducta* aparece destacada en este grupo.
- *Grupos 8, 11 y 12:* mientras que el cluster 8 refiere a las niñas (*muy* y *tímida* aparecen como elementos característicos, a la vez que en las respuestas modales), los clusters 11 y 12 hablan de la *timidez* en general.
- *Grupos 9, 13 y 14:* a la vez que el grupo 9 engloba frases respecto a los varones (*inquieto*, *nervioso* son elementos característicos), en los grupos 13 y 14 aparece el elemento característico *inquietud* con un sentido más general.
- *Grupo 10:* el *carácter*, tanto en las respuestas modales como en los elementos característicos, se destaca en este grupo.

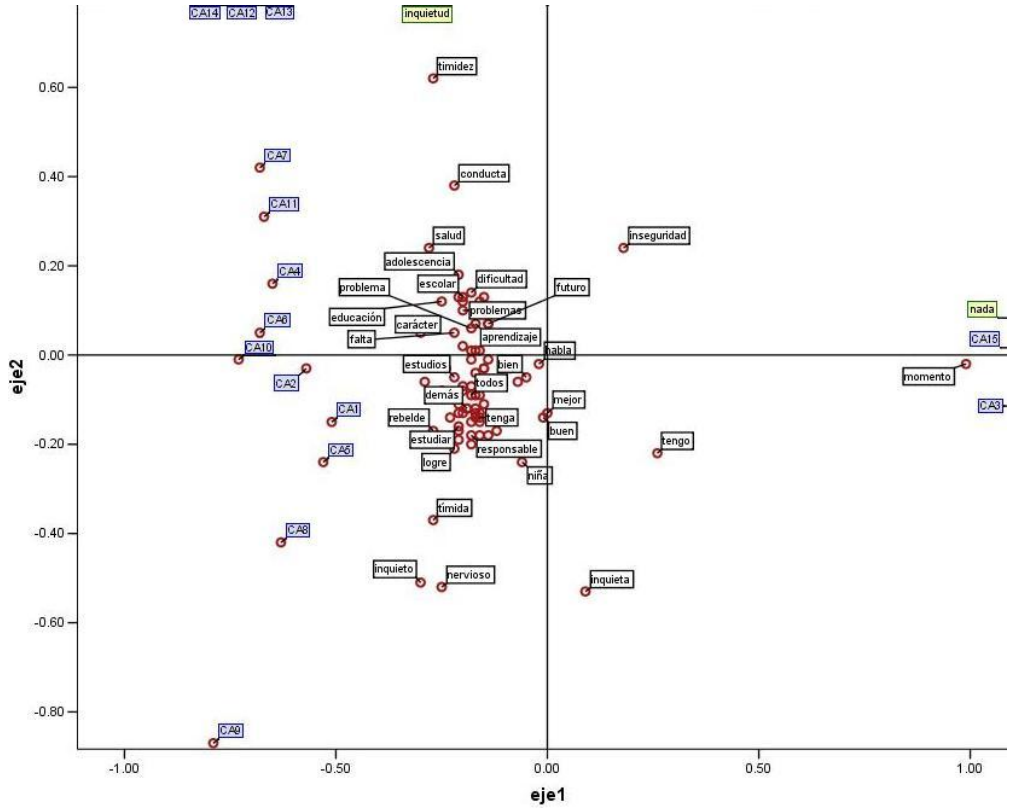

Figura 5.1: Formas gráficas y clusters del subcorpus de la pregunta A

Tabla léxica del subcorpus B: Se destaca en la figura 5.2 los clusters constituidos por las formas gráficas del subcorpus de la pregunta B, y a continuación, una breve descripción de los mismos. Ocurre algo similar a lo observado en el subcorpus de la pregunta A, en donde ciertos clusters –13, 14 y 15 en este caso– quedan fuera del área del gráfico.

- *Grupo 2: le gusta* y *escuela* aparecen tanto en las respuestas modales como en los elementos característicos.
- *Grupo 3: niño* y *cariñoso* sobresalen en respuestas y elementos característicos.
- *Grupo 4: compañero* es compartida tanto en las respuestas como en los elementos característicos; aunque en estos últimos se nombran otras características positivas de los niños respecto a sus pares (*compañerismo*, *solidaridad*, *respeto*).
- *Grupo 5:* en el mismo sentido que el cuarto cluster, *compañera* es el nexo común entre palabras características y sentencias modales, y otras como *cariñosa*, *buena*, *responsable* y *solidaria* se destacan entre las formas gráficas usadas.
- *Grupo 6:* aquí *buenos* y *sentimientos* son compartidos entre las palabras y las respuestas características, mientras que *sensibilidad* aparece destacada también en las formas representativas de este grupo.

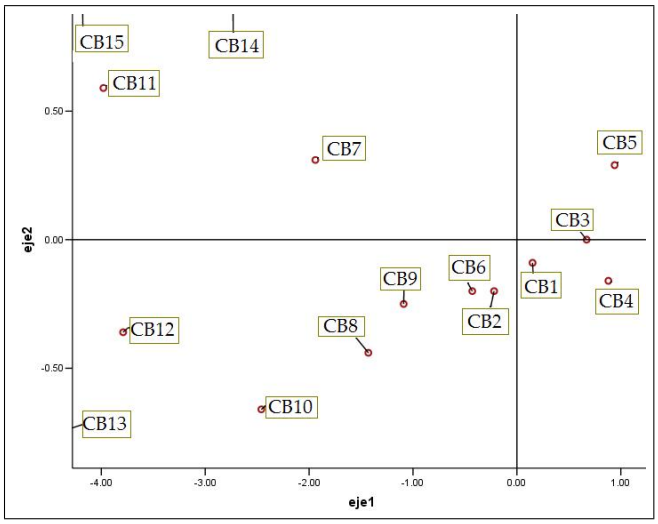

Figura 5.2: Grupos formados en el subcorpus B

- *Grupos 7 y 11: su*, *inteligencia*, *capacidad* sobresalen en palabras y respuestas características. Si bien en el grupo 7 *inteligencia* aparece acompañada de otras cualidades positivas en las respuestas modales, como *compañerismo* y *amor* por ejemplo, en el 11 *inteligencia* aparece sola.
- *Grupos 12 y 13: responsabilidad* se destaca en ambos casos.
- *Grupos 14 y 15: madurez* sobresale para preguntas y frases características.

5.2. Correspondencia Múltiple

Con el fin de comparar los resultados obtenidos en el capítulo 4, donde se optó por el análisis de correspondencias múltiple aplicado a las preguntas postcodificadas A y B, en el presente capítulo se muestra el estudio de estas mismas preguntas en su estado natural, es decir utilizando otra de las diferentes técnicas enmarcadas en el análisis estadístico de textos.

Para situarse aquí, es necesario tener en cuenta los insumos utilizados para el análisis de correspondencias de la sección 4. Estos son: las variables activas –*Grado*, *Quién contestó* el formulario, si *Repitió* algún año, si es *Saludable*, *Síndrome global* y *Sexo-Edad*–, los ejes factoriales 1 y 3 (y el plano formado por éstos) y por último los tres clusters descriptos en la sección mencionada.

A continuación se presentan los resultados para las preguntas A y B por separado. Se optó por trabajar de esta manera debido a que algunas formas gráficas, por más que aparezcan en una u otra pregunta, comportan significados diferentes: a modo de ejemplo, es distinto hablar de "no me preocupa *nada* de mi hijo" que de "no veo *nada* bueno en mi hijo".

Pregunta A

En las líneas siguientes, se presenta un gráfico con las palabras del subcorpus A proyectadas en los ejes 1 y 3, junto con las variables activas y los clusters formados a partir del análisis de correspondencias. Posteriormente, se describe lo que, *a priori*, es considerado más relevante para el presente análisis.

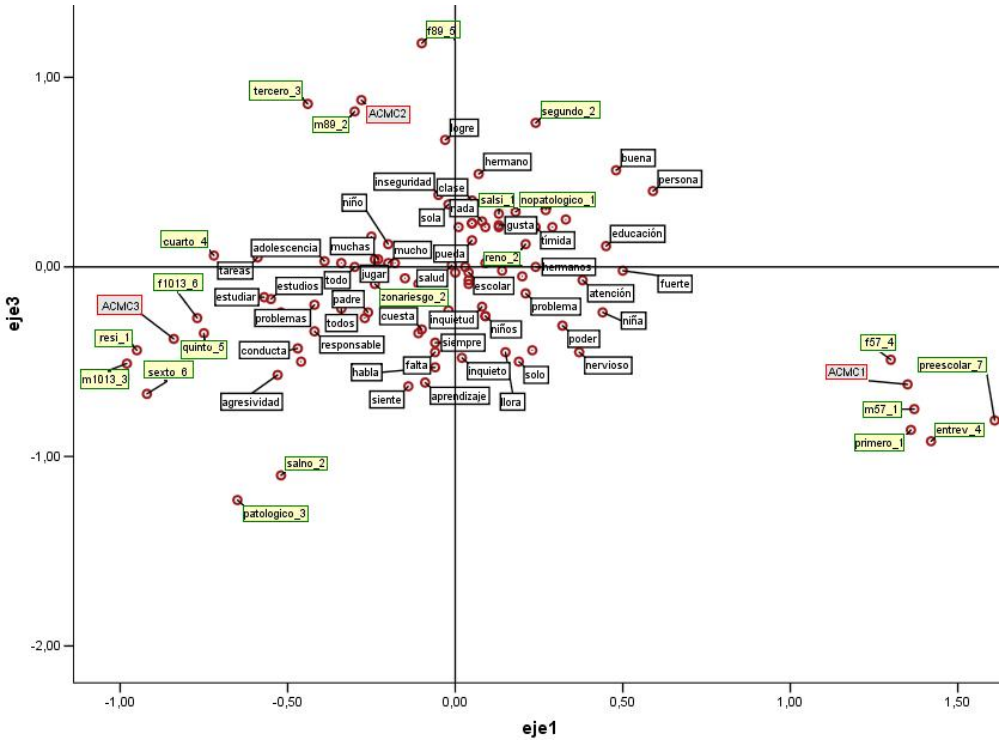

Figura 5.3: Palabras de la pregunta A, con variables activas y clusters

Como se observa en la figura 5.3, en el primer cuadrante se destaca la presencia de *nada*, asociada a las categorías *Saludable* y *No patológico* de las variables activas *Saludable* y *Sindrome global*, respectivamente; que son las variables que mejor explican el eje 3, donde *nada* posee un valor-test significativo.

El tercer cuadrante por su parte, muestra a la palabra *agresividad* cercana a las categorías *No saludable* y *Patológico* de las respectivas variables activas, seguida de cerca por la palabra *conducta*. La primera es significativa para ambos ejes factoriales, mientras que la segunda lo es sólo para el primer eje.

Para los restantes cuadrantes en cambio, no se encontraron formas gráficas cercanas validadas a través de valores test significativos. De todas formas, los gráficos pueden sugerir por donde continuar la investigación.

Pregunta B

En las líneas siguientes, se presenta un gráfico con las palabras del subcorpus B proyectadas en los ejes 1 y 3, junto con las variables activas y los clusters formados a partir del análisis de correspondencias. Posteriormente, se describe lo que, *a priori*, es considerado más relevante para el presente análisis.

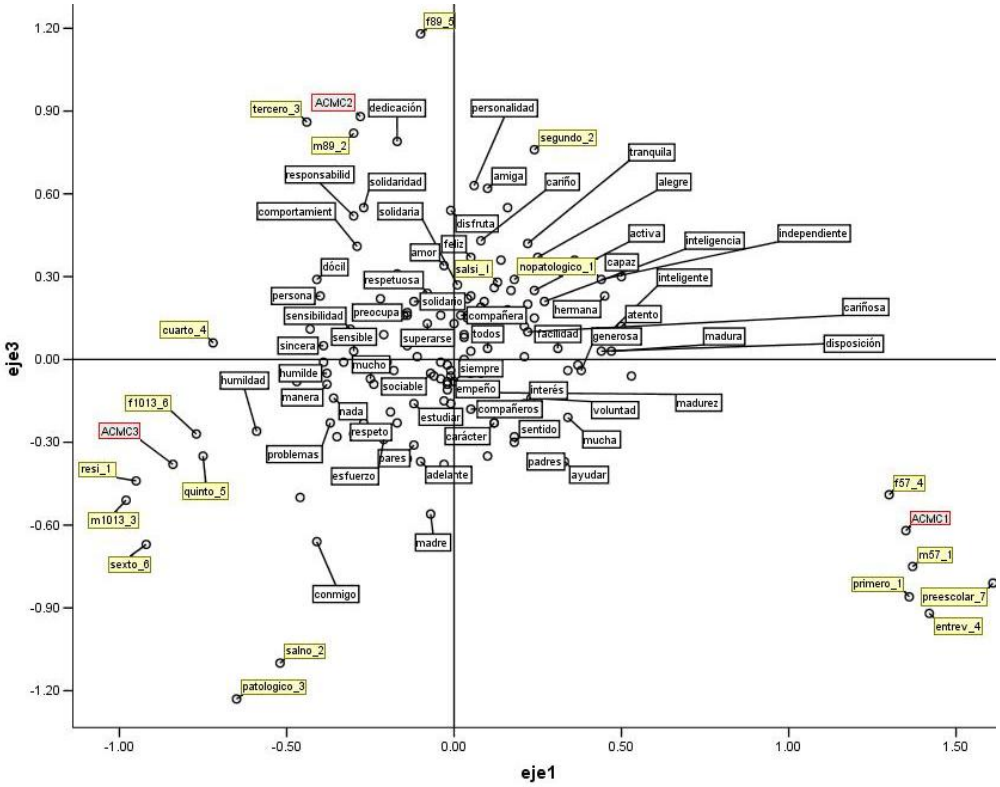

Figura 5.4: Palabras de la pregunta B, con variables activas y clusters

La figura 5.4 muestra por su parte una nube de puntos proyectada relativamente más concentrada que la observada en el apartado anterior. Esto ocurre no sólo con en este plano factorial, sino que si observa tanto el eje 2 como los ejes factoriales 4 y 5, las formas gráficas siguen siendo extremadamente baricéntricas.

Frente a esta situación se cree poco conveniente hablar de asociaciones entre las diferentes unidades léxicas y las varibales activas.

5.2.1. Comparación de las respuestas postcodificadas y textuales

Aquí se presenta una breve comparación de las preguntas A y B postcodificadas, con las mismas en su estado natural. Nuevamente, se presentan los resultados discriminando entre dichas preguntas.

Pregunta A

La figura 5.5 muestra la proyección de las palabras del subcorpus A y las categorías de la pregunta A poscodificada. En la misma se pueden observan algunas asociaciones que se mencionan a continuación.

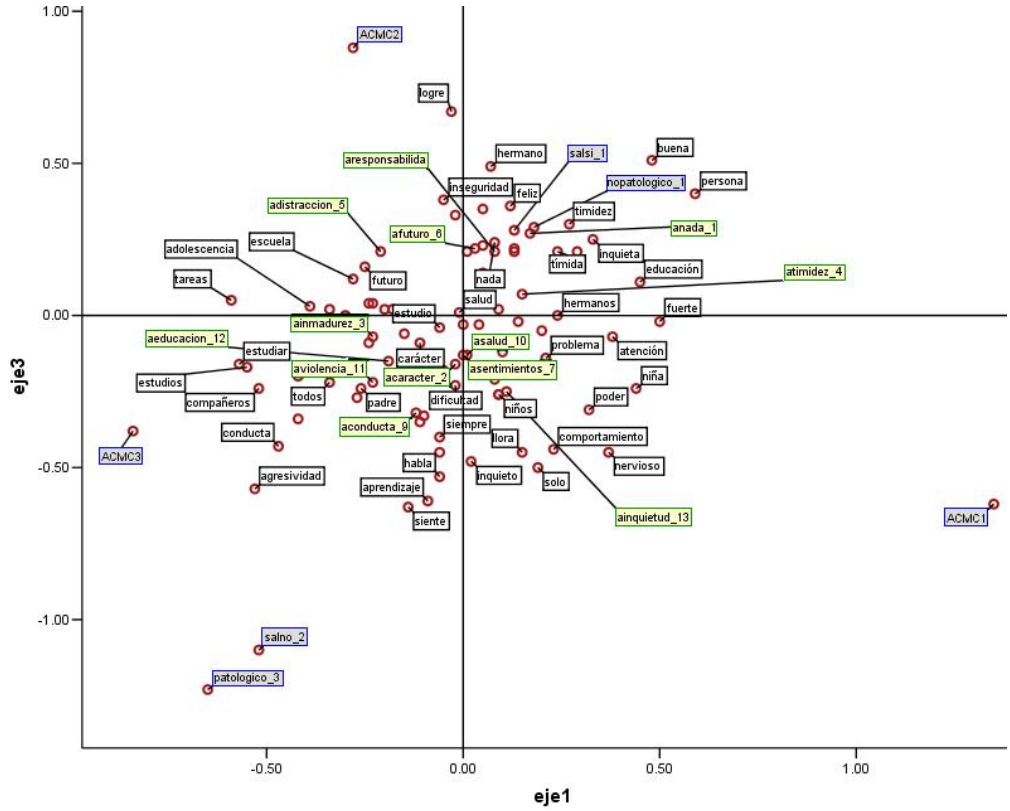

Figura 5.5: Palabras del subcorpus A y categorías de la pregunta A postcodificada proyectadas en plano factorial

No todas las categorías creadas para la pregunta abierta *"¿Que es lo que más le preocupa acerca de su hijo/a?"* tienen asociadas formas gráficas que puedan parecer, a priori, unidades de dicha categoría. Sin embargo, hay otras que llaman la atención del investigador.

A modo de ejemplo, se pueden mencionar las categorías *nada*, *conducta*, *inquietud* y *educación*. Para cada una de estas categorias se describen las formas gráficas relacionadas y proyectadas cerca de las mismas.

nada: es claramente pequeña la distancia entre la categoría *Nada* creada con la postcodificación y su forma gráfica homónima. Dicho de otro modo, tanto la categoría como la unidad *nada* son muy próximas entre sí.

conducta: la palabra homónima posee valores test significantes en ambos ejes factoriales, además de estar muy próxima en el gráfico a la categoría.

inquietud: las unidades léxicas *inquieto* y *comportamiento*, al igual que esta categoría, presentan valores test significativos para el eje 3. Del mismo modo, sus proyecciones en el plano factorial citado son cercanas.

educación: de forma similar al caso anterior, las palabras *estudiar* y *estudios* son cercanas a esta categoría, y además poseen valores test significativos en el mismo eje en el cual la categoría en cuestión.

Pregunta B

En el gráfico 5.6 se muestra la proyección de aquellas palabras del subcorpus B que superan el umbral fijado de frecuencias y la representación de las categorías de la pregunta *"¿Qué es lo mejor que le ve a su hijo/a?"*.

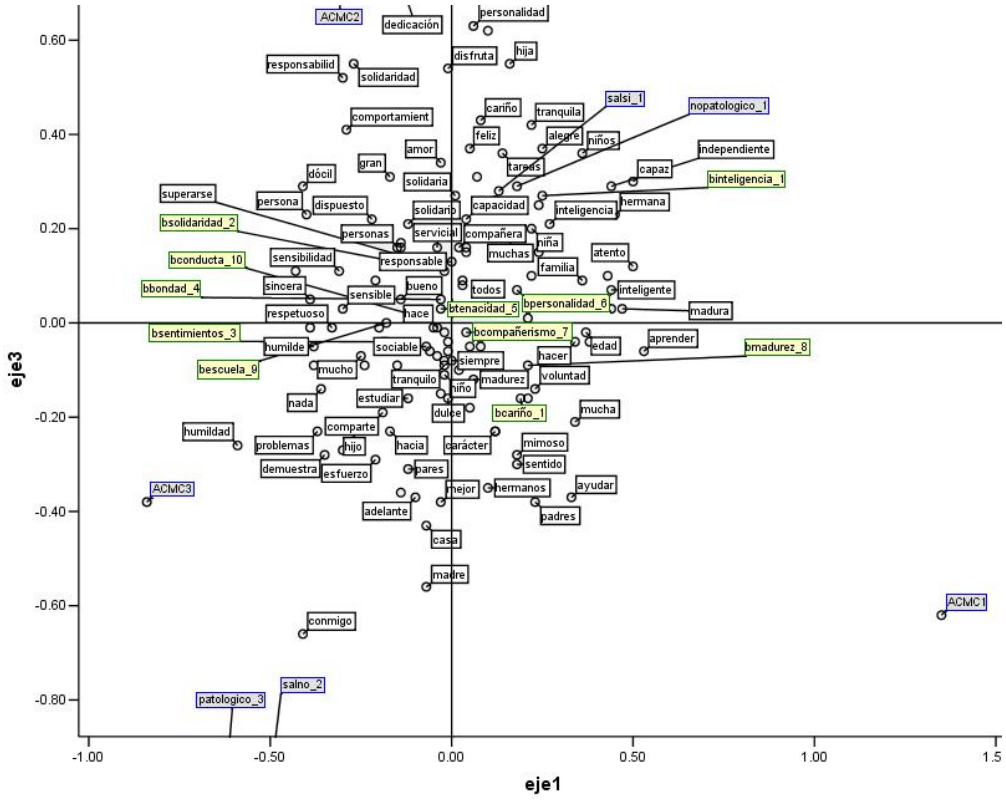

Figura 5.6: Palabras del subcorpus B y categorías de la pregunta B postcodificada proyectadas en plano factorial

En la figura anterior se puede apreciar, como ya se había mencionado, que la nube de las palabras del subcorpus B, así como las categorías de la misma pregunta, son baricéntricas.

A continuación se presentan aquellas categorías de la pregunta B que poseen, en por lo menos uno de los ejes de este plano factorial, valores test significativos y las formas gráficas que parecen estar asociadas a dichas categorías.

cariño: la forma gráfica *cariñoso* –con coordenadas (0.12;-0.23), dado que su etiqueta no aparece representada en la figura– está próxima a esta categoría.

sentimientos: en esta oportunidad las unidades *sensible* y *humildad* son las más próximas a la presente modalidad, además de presentar valores test significativos en el mismo eje factorial.

inteligencia: *inteligente* es una de las pocas formas gráficas que cumplen con los criterios mencionados anteriormente: valores test significativos y proyección cercana a la de la categoría en el plano factorial citado. Se puede mencionar además que la palabra *inteligencia*, a pesar de no poseer valores test significativos en los ejes que forman dicho plano, se proyecta muy cerca de la categoría del mismo nombre.

5.3. Concordancias

En el presente apartado se analizan las concordancias de algunas palabras escogidas, diferenciadas por pregunta.

Pregunta A

Nada: Como se observa en el cuadro 5.2, esta palabra aparece 161 veces en el subcorpus de la pregunta A. Aproximadamente en el 59 % de los casos esta forma gráfica aparece sola, es decir sin ningún contexto o, dicho de otra forma, la respuesta que dan los padres sobre qué les preocupa acerca de sus hijos es *nada*. En los restantes casos, esta unidad léxica posee contextos discursivos que, en la mayoría de las veces, van en el mismo sentido que el 59 % mencionado anteriormente: frases como "no me preocupa *nada*" o "*nada* en particular" reafirman esta cuestión. Sólo en unas pocas oportunidades –apenas el 8 %– el entorno de esta palabra tiene significados diferentes: "momentos como que no le importa *nada*", "no quiere *nada* con nadie", etc. En estos casos el sentido semántico se aparta de lo que podría denominarse "nada" como "exento de".

```
Concordance of words equivalent with:      nada
--------------------------------------
         frequency of repetition  161                                                                  response

                                                    nada                                              - 0008
 demasiado dada con todas las personas y no tiene miedo a nada                                        - 0010
                              por ahora no me preocupa nada                                           - 0014
                                       no me preocupa nada                                            - 0020
                                            no tengo nada                                             - 0021
                                           por ahora nada                                             - 0022
                                                    nada                                              - 0024
                                           por ahora nada                                             - 0029
                  me preocupa que no tiene temor a casi nada                                          - 0036
                                                    nada en particular
```

Cuadro 5.2: Concordancias de la palabra *Nada*

Escuela: Esta forma gráfica aparece 62 veces en este subcorpus, presentando las siguientes peculiaridades:

- en la mitad de los casos, los contextos discursivos hacen referencia al rendimiento en la escuela: frases como "que no rinde en la *escuela*" o "su lentitud en el aprendizaje en la *escuela*" son las más usadas;
- para alrededor del 14 % de estas frases, lo que está por venir aparece como preocupante para algunos padres: tal vez "que salga bien preparado de la *escuela*" o "el cambio de *escuela* al liceo" pueden ser clarificantes en este sentido;

Futuro: Esta unidad léxica tiene tan sólo 49 apariciones en este subcorpus. No obstante, los contextos discursivos de ésta parecen interesantes. Es posible efectuar una división en dos grupos:

- por un lado, aquellos respondentes que refieren al futuro sobre sus hijos, pero en un sentido general: "me preocupa su *futuro*", "el *futuro* que le espera" o "pienso en el *futuro*" son claros ejemplos;
- por otro, aquellos que están preocupados en temáticas particulares sobre sus vástagos: "que tiene muchas ganas de estudiar y no ve *futuro* en el uruguay" y "lo que más me preocupa es en el futuro su adolescencia y en lo laboral" corroboran lo anterior.

Pregunta B

Cariñoso: Es la palabra más frecuente –con un sentido semántico importante– con 134 apariciones. En 30 de los casos (22 %) esta palabra no tiene contexto, es decir aparece sola; y esto sumado a las 24 palabras que aparecen con adverbios superlativos ("que es muy *cariñoso*" por ejemplo) y a aquellas frases que imponen sólamente al adjetivo *cariñoso* ("que es un niño *cariñoso*", "también *cariñoso*", etc.), aumentan esta cifra al 56 %. El restante 44 % se divide entre aquellas frases en las cuales se emplean otros adjetivos además de la forma-polo citada (tales como "que es muy *cariñoso* y voluntarioso" o "generoso y súper *cariñoso*"),

Es importante destacar que esta unidad léxica –que es adjetivo masculino singular de la palabra *cariño*– fue tenida en cuenta a la vez con otras formas similares, tales como *cariñosa* o *cariñosos*. De todos modos la primera resultó escogida debido a la cantidad de apariciones en el texto de la pregunta B, largamente superior a las otras (87 apariciones para *cariñosa*, 14 para *cariño* y tan sólo una para *cariñosos*). Sin dudas que, en caso de lematizar el presente texto, el analista encontrará una frecuencia de aparición de la palabra *cariño* en el subcorpus tratado, superior a las 134 mencionadas líneas arriba.

Responsable: Tiene 88 apariciones en este subcorpus. En más de la mitad de las veces (48 para ser exactos), esta palabra se encuentra sola o con algún superlativo ("muy *responsable*", por ejemplo). En la restante mitad de los casos, ocurren dos cosas: por un lado, la palabra *responsable* aparece acompañada de otro adjetivo. Dentro de estas 23 frases, 10 de estos

adjetivos van por el lado del afecto (muy *responsable* y amigable, es cariñosa y *responsable*, es *responsable* y muy querida por sus compañeros), 10 con la capacidad en la escuela (es *responsable* y aplicada, es muy *responsable* y le pone empeño a las cosas) y las restantes 3 con conducta ejemplar (*responsable* y respetuoso, por ejemplo). En las otras 17, las oraciones utilizadas son más ricas desde el punto de vista del vocabulario –lenguaje más articulado y poco más extenso que en casos anteriores–, aunque desde el punto de vista semántico hay similitudes con los casos arriba mencionados, sobre todo con lo relacionado a los estudios (en 10 de los 17 casos se nombra a la *escuela* de algún modo), además de destacar responsabilidad en las tareas de la casa, entre otros.

Capítulo 6

Conclusiones

Se puede concluir que no se encontraron grandes diferencias en la utilización de las diferentes técnicas (ACM, ADT). A *priori*, esto puede ser consecuencia de la falta de conocimiento sobre el tema. Es importante recordar que existían variables, como la diferenciación en niveles de síndromes surgidas del mismo instrumento (CBCL), que no fueron utilizadas debido a que el estudio preliminar se encuentra en fase de validación externa, por lo que la Clínica de Psiquiatría Pediátrica del Hospital Pereira Rossell pidió que éstos no fueran utilizados hasta no terminar este proceso.

Uno de los objetivos de este trabajo era la introducción de algunas de las herramientas que facilitan la gestión y la descripción de corpus de gran tamaño y que a su vez permitan derivar información de ellos desde un punto de vista estadístico. Lo más destacable de estas herramientas es que permiten realizar un análisis más rico y con una menor pérdida de información. A modo de ejemplo, en el caso de análisis de preguntas abiertas, el analista no interviene sino hasta la interpretación final de los datos; es decir que no se presenta sesgo en la preparación de los mismos.

Otro de los objetivos del presente trabajo de pasantía consistía en la comparación del análisis de dos preguntas abiertas postcodificadas, con el análisis de las mismas sin ningún tipo de intervención previa, es decir las palabras y frases en su "estado natural", como se ha mencionado en algún pasaje de este informe. Para dicha comparación se procedió del siguiente modo:

1. Luego de aplicar la técnica de análisis de correspondencias múltiple a la tabla original de datos, considerando el plano formado por los ejes factoriales 1 y 3, se llegó a la conclusión de que el eje 1 muestra, siguiéndolo de derecha a izquierda, el *desarrollo en edad y grado* del niño/a; y por su parte el eje 3 hace las veces de *divisor entre niños sanos y enfermos*
2. Del punto anterior, se realizaron cruces de variables que parecieron interesantes, para luego aplicarles el método de análisis de correspondencia simple, y en cada caso se proyectaron las respuestas abiertas postcodificadas como variables suplementarias. Aquí se destacan como más importantes:

- Para el cruzamiento entre *Saludable* y *Síndrome global* el plano factorial (1,3) mostró una clara división entre niños *sanos* y *enfermos*. Es importante mencionar que la categoría *Nada* quedó proyectada del lado de los *sanos* y las categorías *violencia*, *carácter* y *conducta* del lado de los *enfermos*.
- Para la tabla que surgió del cruce entre *Síndrome global* con *Repitió*, nuevamente las categorías *Conducta* y *Nada* se asocian a las categorías *Patológico* y *No patológico* de la variable *Síndrome global*. Se reitera en este plano la división entre *sanos* y *enfermos*.

3. Al realizar un análisis de cluster sobre los ejes obtenidos según el análisis de correspondencias múltiple, se obtuvieron tres clusters, caracterizándose éstos según las variables *Sexo-Edad* y *Grado*.

4. Se realizó el armado y la visualización de tablas léxicas, teniendo en cuenta el corpus formado por las respuestas a las preguntas A y B y los subcorpus tomando estas respuestas por separado. En todos los casos se construyeron grupos los cuales fueron etiquetados según los elementos característicos y respuestas modales contenidos en ellos. Este análisis va en el mismo sentido que la post-codificación pero con una importante ventaja: *no* hay intervención subjetiva del analista al comienzo del mismo.

5. Al realizar el análisis de correspondencias múltiples proyectando como suplementarias las formas gráficas más importantes, se llegó a resultados similares que los obtenidos al proyectar las categorías de las preguntas postcodificadas.

6. En cuanto al estudio de las concordancias, se sometieron algunas de las palabras más frecuentes de los subcorpus A y B tomados por separado, analizando de este modo sus diferentes contextos discursivos.

6.1. Recomendaciones

Se presentan algunas sugerencias para futuras investigaciones:

- Considerando los problemas que se presentaron con la muestra mencionados en la sección 2.2.1, se sugiere realizar una reponderación de la muestra y reformulación del presente trabajo.
- Se cree conveniente la puesta en práctica de un enfoque multidisciplinario en la materia, debido a la falta de conocimiento por parte de los autores del presente informe en cuanto a la temática tratada –psiquiatría pediátrica– por un lado, y a la complejidad del tema por otro. Para ejemplificar esta cuestión, el criterio de elección de palabras para el análisis de concordancias hubiere sido diferente, teniendo un cabal conocimiento del tema, respecto del elegido en este trabajo. Siguiendo esta línea, es importante la utilización de todas las preguntas que se usan en un formulario o, dicho de otro modo, no incluir preguntas que luego no serán analizadas.

6.2. Temas pendientes

A continuación se presentan algunos de los temas que quedaron pendientes por distintas razones, como por ejemplo el software utilizado o aspectos surgidos a último momento.

- **Visualización de datos textuales a través de *Self Organization Maps* (SOM):** esta herramienta de corte gráfico, también conocida como Mapa de Kohonen (debido a su creador, el profesor finlandés Teuvo Kohonen) es un instrumento basado en las teorías de redes neuronales y análisis de contigüidad, útil para la visualización de datos multidimensionales en espacios reducidos, como puede ser un plano. Esta técnica es muy usada a nivel de análisis de textos a modo de encontrar, por ejemplo, asociaciones entre formas gráficas.

- **Estudio de segmentos repetidos:** con el propósito de lograr una más detallada descripción de los corpus, sería interesante realizar este tipo de estudio. Complementaría al estudio de otras tablas léxicas ya mencionadas.

- **Lematización:** realizar el análisis luego de aplicar esta técnica sería provechoso afín de encontrar posibles diferencias entre tomar el corpus total y el mismo corpus lematizado. En el software DTM –utilizado en el presente informe– existe la posibilidad de lematizar un corpus e incluso eliminar aquellas formas gráficas o, de forma más general, unidades léxicas que no aportan un significado importante a frases o contextos, tales como artículos, preposiciones y otros conectivos de uso corriente en el español. Hacerlo implica disponer de cierta dedicación, debido a su procedimiento enteramente manual[1].

- **Contigüidad:** una de las posibles aplicaciones de esta técnica, y al mismo tiempo, una de las vetas más interesantes consistiría en demostrar, si los corpus de las preguntas abiertas mencionadas, como fuente de información para las percepciones son diferentes o por el contrario estan próximos, según variables como el síndrome global, o sexo-edad por ejemplo, a lo que los psiquiatras toman como fuente de percepción de determinados problemas, para poder observar la utilidad y/o la posible reformulación de las preguntas abiertas del cuestionario.

- **Tablas yuxtapuestas:** estas tablas se forman como yuxtaposición –valga la redundancia– de diferentes tablas léxicas. Permiten observar si los perfiles de las formas gráficas varían al cambiar las particiones contra las que se comparan. En este caso sería interesante generar tablas yuxtapuestas para analizar cruzamientos entre formas gráficas y la variable sexo-edad, síndrome global y la instrucción de la madre, por ejemplo.

[1] Es de uso corriente en este tipo de estudios usar un software específico para lematizar el corpus a trabajar. El problema es que, en general, hay que disponer de los medios económicos para poder obtenerlo. Por información sobre posibles programas informáticos relacionados, visitar http://www.textanalysis.info

Bibliografía

[1] Achenbach, T.; Rescorla, L.: *Manual for the ASEBA School-Age Forms & Profiles*; University of Vermont, Research Center for Children, Youth and Families, 2001, pp. 22-23.

[2] Álvarez, R.: Las preguntas de respuesta abierta y cerrada en los cuestionarios. Análisis Estadístico de la Información, en *Metodología de Encuestas, Vol. 5, num. 1*; 2003, pp. 45-54.

[3] Aureli, E.; Fioredistella, D.: Recruitment via web and information technology: a model for ranking the competences in Italian job market; in *JADT 2006*, pp. 79-88.

[4] Baroni, M.; Evert, S.: *The zipfR package for lexical statistics: A tutorial introduction*, 2006. Disponible en http://www.r-project.org/

[5] Baroni, M.; Evert, S.: Chapter 38: Statistical methods for corpus exploitation, in *Corpus Linguistics. An International Handbook*; Lüdeling, A. and Kytö, M. (eds.), Mouton de Gruyter, Berlin, 2006.

[6] Biskri, I.; Rompré, L.; Laouamer, L.; Meunier, F.: Classification de documents Multimédia : vers une approche générale; in *JADT 2006*, pp. 189-200.

[7] Blanco, J.: *Introducción al Análisis Multivariado: Teoría y aplicaciones a la realidad latinoamericana*; Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, UDELAR, 2006.

[8] Camillo, F.; Liberati, C.: e-CRM, web semantic propensity models and micro-datamining: an application of Kernel Discriminant Analysis to the Glam on Web case. *JADT 2006*, pp. 235-243.

[9] Casella, G.; Berger, R.: *Statistical Inference*; Duxbury Advanced Series - Thompson Learning, 2004, pp. 621-627.

[10] Chateau, F; Lebart; L.: Assessing Sample Variability in the Visualization Techniques related to Principal Component Analysis: Bootstrap and Alternative Simulation Methods, in *COMPSTAT*; Physica Verlag ed., Heidelberg, 1996, pp. 205-210.

[11] Etxeberría, J.; García, E.; Gil, J.; Rodríguez, G.: *Análisis de datos y textos*, RA-MA editorial, Madrid, 1995.

[12] Evert, S.: How Random is a Corpus? The Library Methaphor, in *ZAA Vol. 54.2*; 2006, pp. 177-190.

[13] Fernández, K.: Análisis Textual: Generación y Aplicaciones, en *Metodología de Encuestas, Vol. 5, num. 1*; 2003, pp. 55-66.

[14] Gelbukh, A.; Sidorov, G.: Zipf and Heaps Laws'Coefficients Depend on Language, in *Conference on Intelligent Text Processing and Computational Linguistics*; Lecture Notes in Computer Science, 2004, pp. 332-335.

[15] Greenacre, M.: *Theory and Applications of Correspondance Analysis*, Academic Press, Londres, 1984.

[16] Härdle, W.; Simar, L.: *Applied Multivariate Statistical Analysis*; MD Tech, 2003, pp. 219-232.

[17] Hochsztain, E.; Tarsisto, A.: Inteligencia Empresarial: un tema imprescindible para los profesionales del año 2000, en *Segundas Jornadas Rioplatenses de Práctica Profesional*; Facultad de Ciencias Económicas y de Administración, UDELAR, Plural Ed., Montevideo, 1999, pp. 80-88.

[18] Instituto Nacional de Estadística (INE): *Manual Guía para la Codificación de Ocupaciones de Actividad; Clasificación Internacional Uniforme de Ocupaciones (CIUO-88), adaptada a Uruguay (CNUO-95)*; 1996.

[19] Labbé, D.; Hubert, P.: Vocabulary Richness, in *Colloque de l'ALCC-ACH*; Paris, 1994.

[20] Lebart, L.: Contiguity analysis and classification, in *Data Analysis Journal (corrected version)*; 2004, pp. 233-244.

[21] Lebart, L.: Semiometry: the use of words to describe lifestyles and values, in *JMRA Seminar*; 2002.

[22] Lebart, L.: Classification problems in text analysis and information retrieval, in *Advances in Data Science and Classification*; Rizzi, A. and coll., 1998, pp. 473-482.

[23] Lebart, L.: *Visualization of Textual Data: unfolding the Kohonen Maps*, (texto desconocido), 2005, pp. 73-80.

[24] Lebart, L. et coll.: *DTM Software: Exploratory statistical processing of complex data sets comprising both numerical and textual data*; disponible dans http://www.lebart.org/.

[25] Lebart, L.; Mirkin, B.G.: Correspondence analysis and classification, in *Seventh International Conference on Multivariate Analysis Barcelona Meeting* (corrected version); North Holland, 1992, pp. 341-357.

[26] Lebart, L.; Morineau, A.; Piron, M.: *Statistique exploratoire multidimensionnelle*; Dunod Ed., París, 1995, pp. 171-176.

[27] Lebart, L.; Salem, A.: *Statistique Textuelle*; Dunod ed., París, 1994.

[28] Lebart, L.; Salem, A.; Bécue, M.: *Análisis Estadístico de Textos*; Ed. Milenio, Lleida, 2000.

[29] Lelu, A.; Halleb, M.; Delprat, B. Recherche dínformation et Cartographie dans des corpus textuels a partir des fréquences de n-*grammes*. *JADT 1998*.

[30] Peña, D.: *Análisis de Datos Multivariantes*; McGraw Hill, Madrid, 2002.

[31] Terrádez Gurrea, M.: *Frecuencias Léxicas del Español coloquial: Análisis Cuantitativo y Cualitativo*; Facultat de Filología, Universitat de València, 2001, pp. 37-42.

[32] (Varios): *La Enciclopedia*; Salvat Editores, Mediasat Group, Madrid, 2004, pp. 7051-7054, 7733, 9137-9138, 14305-14308.

[33] Viola, L.; Garrido, G., Varela, A.: *Estudio Epidemiológico sobre la Salud Mental de los niños uruguayos*; Clínica de Psiquiatría Pediátrica, Facultad de Medicina, UDELAR, 2007.

[34] Wyllys, R.: Empirical and Theoretical Bases of Zipf's Law, in *Library Trends Vol. 30, num. 1*; 1981, pp. 53-64.

Parte IV

Anexos

Apéndice A

Glosario

El presente Glosario está basado en el homónimo que presentan L. Lebart y A. Salem[27], a modo de guía conceptual. Este ha sido modificado según los términos utilizados en este trabajo de pasantía.

Notas

- Los asteriscos indican otra definición en el presente glosario.
- Las siguientes abreviaciones que aparecen entre paréntesis refieren al dominio en el cual se aplica la correspondiente definición:
  - **(ac)** Análisis factorial de correspondencias
  - **(acm)** Análisis de correspondencias múltiples
  - **(cla)** Clasificación - clusters
  - **(sp)** Método de las Especificidades
  - **(sr)** Análisis de los segmentos repetidos
  - **(ling)** Lingüistica
  - **(stat)** Estadística
  - **(sa)** Segmentación automática

Glosario

algoritmo - conjunto de reglas operatorias propias de un cálculo.

caracter (sa) - signo tipográfico utilizado para la codificación del texto, sobre un soporte legible para el ordenador.

caracteres delimitadores / no delimitadores (sa) - distinción realizada sobre el conjunto de caracteres que entran en la composición del texto, que permite al software especializado segmentar al corpus en *ocurrencias** (serie de caracteres no delimitadores cuyas extremidades están limitadas por caracteres delimitadores). Se realizan las siguientes distinciones:

- los caracteres *delimitadores de ocurrencia* (conocidos también como *"delimitadores de forma"*) que son en general el espacio en blanco y los signos de puntuación usuales tales como el punto, el punto y coma, los dos puntos, etc.
- los caracteres *delimitadores de secuencia*: subconjunto de delimitadores de ocurrencia que corresponden, en general, a los separadores débiles y fuertes.
- los caracteres *separadores de frase*: subconjunto del delimitadores de secuencia que corresponde, en general, a los separadores fuertes.

concordancia (sa) - ocurrencias de una misma palabra (la denominada *forma-polo*) reagrupadas con fragmentos de su contexto más inmediato. La longitud de estos fragmentos varía según las necesidades del análisis.

construcción de contextos (sa) - reorganización de una secuencia textual, donde las ocurrencias de una forma gráfica son acompañados de un fragmento del contexto, pudiendo contener varias líneas de texto alrededor de la forma-polo. La longitud de este contexto se define por el número de casos antes y después de la forma-polo.

contexto (de una unidad léxica) (sa) - conjunto constituido por los términos vecinos de cada una de las ocurrencias que corresponden a la unidad léxica estudiada.

coocurrencia (sa) - determinación de las atracciones existentes entre pares de palabras en el interior de una unidad fijada (por ejemplo frase, párrafo, alrededores de una ocurrencia) de contexto.

corpus (ling) - conjunto limitado de elementos (o simplemente expresiones) en los cuales se basa el estudio de un fenómeno lingüístico.

- (lexicometría) conjunto de textos reunidos con el fin último de la comparación, siendo el corpus el insumo para un estudio cuantitativo.

delimitador de secuencia (sa) - subconjunto de caracteres *delimitadores** de *forma**, correspondiente a las signos de puntuación débiles y fuertes (en general el punto, los signos de interrogación y exclamación, la coma, el punto y coma, dos puntos, las comillas, los guiones y los paréntesis).

elemento característico - (de una partición) sinónimo de especificidad *positiva**.

elementos de un segmento (sr) - cada una de las unidades léxicas (en general formas gráficas) correspondiente a las ocurrencias que entran en su composición. Ej: A, B, C son respectivamente el primer, segundo y tercer elemento del segmento ABC.

enfoque paradigmático - aquellos métodos que parten del "despiece" del corpus en unidades mínimas (formas gráficas, por ejemplo). En este contexto se sitúan: el estudio de la riqueza del vocabulario, la contigüidad entre textos o los análisis de correspondencias aplicados a tablas léxicas, por citar algunos.

enfoque sintagmático - es aquel enfoque cuyo interés gira alrededor del contexto discursivo; es un enfoque secuencial del estudio del texto.

especificidad negativa (sp) - para un umbral fijado de especificidad, una unidad léxica i y una parte j dadas, la unidad i es denominada específica negativa de la parte j si su subfrecuencia es "anormalmente baja" en esta parte. Es decir, si la suma de las probabilidades calculadas a partir del modelo hipergeométrico para los valores iguales o inferiores a la subfrecuencia comprobada es inferior al umbral fijado al principio.

especificidad positiva (sp) - para un umbral fijado de especificidad, una unidad léxica i y una parte j dadas, la unidad i es llamada específica positiva de la parte j (o elemento característico* de esta parte) si su subfrecuencia es "anormalmente elevada" en esta parte. Es decir, si la suma de probabilidades calculadas a partir del modelo hipergeométrico para valores iguales o superiores a la subfrecuencia comprobada es inferior al umbral fijado al principio.

flexión - alteración morfológica (o de forma) de cierta palabra.

forma gráfica (sa) - secuencia de caracteres no delimitadores comprendida entre dos caracteres delimitadores.

forma común - forma que se repite en cada una de las particiones del corpus.

forma original - forma que encuentra todas sus casos en una sola partición del corpus.

frecuencia (sa) - número de ocurrencias de una unidad léxica en el corpus.

frecuencia de un segmento (sr) - es el número de ocurrencias de este segmento en todo el corpus.

frecuencia máxima (sa) - frecuencia de la forma más frecuente del corpus.

frecuencia relativa (sa) - la frecuencia de una unidad textual en el corpus o en una de sus particiones, respecto al tamaño del sub-corpus seleccionado.

gama de frecuencias (sa) - serie denotada por V_k, $k = 1, \ldots, F_{max}$, donde muestra la cantidad de formas de frecuencia k (como caso particular se tiene a las palabras de frecuencia 1, V_1, es decir los *hápax*).

glosario (sa) - lista constituida a partir de una reorganización de las unidades léxicas. Útil para una rápida localización de las ocurrencias de cada palabra.

gramática comparativa (ling) - aquella ciencia que estudia las relaciones, similitudes y diferencias entre dos o más idiomas.

hapax - del griego *hapax legomenon*, "cosa dicha una sola vez "

- (sa) forma cuya frecuencia es igual a uno en el corpus (hápax del corpus) o en una de sus particiones (hápax de la partición).

homonimia - cuando dos términos de diferente significado coinciden en sus significantes.

índice alfabético donde las unidades aparecen ordenadas alfabéticamente como su propio nombre lo indica.

índice jerárquico donde las unidades aparecen ordenadas por frecuencia decreciente, si dos palabras tienen la misma frecuencia se recurre al orden alfabético.

ítem de respuesta - (o modalidad de respuesta) elemento de respuesta establecido de antemano en una pregunta cerrada.

lematización - reagrupación de las ocurrencias del texto bajo una forma canónica (en general a partir de un diccionario). En español, se reagrupa según el siguiente criterio:

- las formas verbales se llevan al infinitivo,
- los sustantivos al singular,
- los adjetivos a masculino el singular.

A modo de ejemplo, si se tiene las formas *existirá, existió, existen*, al reagruparlas por el criterio de lematización expuesto líneas arriba, estas tres formas se reunirán en una sola, *existir*.

léxico (ling) - que concierne al léxico* o al vocabulario*.

- conjunto virtual de las palabras de una lengua.

lexicometría - conjunto de métodos que permiten reorganizar la secuencia textual y realizar análisis estadísticos sobre el vocabulario* de un corpus.

ocurrencia (sa) - serie de caracteres no - delimitadores delimitada en sus extremos por dos caracteres delimitadores*.

paradigma (ling) - conjunto de términos de una misma clase gramatical que pueden aparecer en un mismo contexto.

paradigmático/a (sa) - que concierne a la reagrupación en series de unidades léxicas*, independientemente de su orden (para que tenga sentido) en la cadena escrita.

parte - (de un corpus de textos) fragmento de texto correspondiente a las divisiones naturales del corpus o en una reagrupación de estas últimas.

partición - (de un corpus) división en partes constituidas por fragmentos consecutivos de texto, no teniendo intersección común y cuya unión es igual al corpus.

particiones en situaciones-tipo (cla) - clases de una partición* de un conjunto de observaciones, que sintetizan una batería de variables de interés (sexo, edad, profesión, etc).

polisemia - significado múltiple de una palabra.

post-codificación - operación manual que consiste en armar un listado de las principales categorías de respuestas libres sobre una submuestra de las mismas, para luego cerrar la pregunta abierta correspondiente, llevando todas las respuestas a estas categorías.

pregunta abierta - pregunta de respuesta libre; pueden ser numéricas (ej: ¿Cuales deben ser, según usted, los recursos mínimos de una familia que tiene tres niños menores de 16 años?), o textuales (ej: ¿puede justificar su elección?).

pregunta cerrada - pregunta cuyas respuestas posibles son propuestas explícitamente a la persona interrogada.

pronombre - clase de palabra que ejerce las mismas funciones gramaticales que el sustantivo.

rango - orden que ocupa determinada unidad léxica en un texto, según su frecuencia (por ejemplo, la unidad más frecuente del corpus tendrá rango $r = 1$)

respuesta característica o modal- (de una clase de individuos o de una parte* del corpus*) respuestas íntegras, seleccionadas en razón de su carácter representativo para una categoría de individuos, partiendo de los elementos característicos que ella contiene.

secuencia (sa) - serie de ocurrencias del corpus no separadas por un delimitador* de secuencia.

segmentación - operación que consiste en delimitar unidades léxicas en un texto.

segmentación automática - conjunto de operaciones realizadas por medio de procedimientos informáticos que segmentan, según reglas predefinidas, un texto almacenado en un soporte informático. Las diferentes unidades segmentadas se denominan unidades mínimas.

segmento (sr) - toda serie de ocurrencias consecutivas en el corpus y no separadas por un delimitador* de secuencia.

segmento repetido (sr) - serie de unidades léxicas* cuya frecuencia es superior o igual a 2 en el corpus.

semántica - estudio del significado de las palabras

separadores de frases (sa) - subconjunto de los caracteres delimitadores* de secuencia* correspondientes a las puntuaciones fuertes, en general: punto, signo de interrogación, signo de exclamación.

sintagma (ling) - conjunto de unidades en secuencia que forman una unidad dentro de la frase.

sintagmática (sa) - que concierne las reglas de combinación fundamentales en la producción e interpretación de la cadena de textos.

subfrecuencia (sa) - número de ocurrencias de una unidad léxica en una parte del corpus.

subsegmentos (sr) - para un segmento dado, todos los segmentos de tamaño inferior y comprendidos en el mismo. Ej: AB y BC son dos subsegmentos del segmento ABC.

sufijo - afijo propuesto con el cual se forman derivados léxicos (p. ej. en la palabra "casero", *-ero* es el sufijo).

sustantivo - parte de una oración cuya función es ser núcleo del sintagma nominal.

tabla de contingencia (stat) - sinónimo de tabla de frecuencias o de tabla cruzada: tabla cuyas filas y columnas representan respectivamente las modalidades de dos preguntas (o dos variables nominales), y el término general representa el número de individuos correspondiente a cada pareja de modalidades.

tabla de segmentos repetidos (TSR) - cuadro de doble entrada, cuyas filas las constituyen los segmentos repetidos encontrados en las distintas particiones del corpus. Las filas de la TSR son escogidas por orden lexicométrico* de los segmentos. (es decir, por orden de frecuencia).

tabla dispersa - matriz de grandes dimensiones en la que la mayoría de sus elementos valen cero.

tabla léxica (o tabla léxica amputada) - cuadro de doble entrada que surge de la supresión de ciertas filas de la TLE (las que por ejemplo corresponden a unidades que no alcanzan el umbral mínimo de frecuencia).

tabla léxica entera (TLE) - tabla de doble entrada cuyas filas son el total de unidades léxicas. El término genérico k_{ij} de la TLE es igual al total de veces que la unidad i aparece en la parte j del corpus.

tamaño (sa) - (de un corpus, de una parte de este corpus, de un fragmento de texto, etc.) número de ocurrencias contenidos en un corpus. Se denota T al tamaño del corpus.

tamaño de un segmento (sr) - número de ocurrencias que componen el segmento seleccionado.

tesaurus - palabra derivada del latín que significa *tesoro*; se refiere a un listado de palabras o términos empleados para representar conceptos, temas o contenidos de los documentos estudiados.

umbral (stat) - cantidad arbitraria fijada de antemano para determinar el número de elementos a retener.

unidad banal (sp) - para una partición dada del corpus, unidad léxica que no presenta ninguna especificidad, ni positiva ni negativa.

unidades léxicas - término general que refiere a la unidad de recuento en un corpus textual. Se puede hablar de unidades léxicas simples* y unidades léxicas complejas*.

unidades léxicas complejas - unidades compuestas, caracterizadas por los denominados segmentos repetidos* y cuasisegmentos*.

unidades léxicas simples - unidades básicas del léxico; se dividen en dos: formas gráficas* y lemas*.

valores-test (ac o acm) - cantidades que permiten apreciar la significación de la posición de un elemento suplementario sobre un eje factorial. Brevemente, si un valor-test sobrepasa a 2 en valor absoluto, tiene 95 % de posibilidades que la posición del elemento correspondiente no se deba al azar.

variables de tipo T - variables que crecen en forma aproximadamente proporcional al tamaño de un texto

variables de tipo V - variables que tienden a crecer cada vez menos a medida que aumenta el tamaño del texto.

vocabulario (sa) - conjunto de unidades léxicas* distintas en un corpus.

vocabulario de base (sp) - conjunto de formas del todo el corpus que no presentan ninguna especificidad (ni negativa ni positiva), para un umbral fijado (i.e. el conjunto de las unidades banales* para cada una de las particiones del corpus).

Apéndice B

Definición de las variables

1. **grado**: Grado escolar que esta cursando el niño/a al momento de realizar la entrevista
    - 1 - Primero - primero_1
    - 2 - Segundo - segundo_2
    - 3 - Tercero - tercero_3
    - 4 - Cuarto - cuarto_4
    - 5 - Quinto - quinto_5
    - 6 - Sexto - sexto_6
    - 7 - Preescolar - preescolar_7
    - 999 - Sin dato - grsd_8
2. **sexo**: Sexo del niño/a
    - 1 - Masculino - masculino_1
    - 2 - Femenino - femenino_2
3. **edadrec**: Edad del niño/a recodificada en tres tramos
    - 1 - de 5 a 7 años - 5a7_1
    - 2 - 8 y 9 años - 8y9_2
    - 3 - de 10 a 13 años - 10a13_3
4. **trabpadrec**: Trabajo del padre
    - 1 - Fuerzas Armadas - pffaa_1
    - 2 - Prof., Gerentes, Poder Ejecutivo - pprof_2
    - 3 - Técnicos - ptecn_3
    - 4 - Empleado de oficina - pempl_4

- 5 - Vendedor, comercio - pvend_5
- 6 - Agricultor, trab calif agrop y pesq - pagrop_6
- 7 - Operarios, artesanos, oficios - pope_7
- 8 - Trabajador no calificado - pncal_8
- 9 - Otros - potros_9
- 999 - Sin dato - psindato_10

5. **trabajomadrec**: Trabajo de la madre
   - 1 - Prof., Gerentes, Poder Ejecutivo - mprof_1
   - 2 - Técnicos - mtecn_2
   - 3 - Empleado de oficina - mempl_3
   - 4 - Vendedor, comercio - mvende_4
   - 5 - Trabajador no calificado - mncal_5
   - 6 - Otros - motros_6
   - 999 - Sin dato - msindato_7
6. **este**: ¿Quién contestó el formulario?
   - 1 - Padre - papa_1
   - 2 - Madre - mama_2
   - 3 - Otro familiar - otro_3
   - 4 - Entrevistador - entrev_4
   - 9 - Ignorado - ignorado_5
7. **clasesp**: ¿Está su hijo/a en una clase o escuela especial o recibe servicios especiales?
   - 1 - Si - clsi_1
   - 2 - No - clno_2
   - 3 - Sin dato - clsd_3
8. **repano**: ¿Ha repetido algún año?
   - 1 - Si - resi_1
   - 2 - No - reno_2
   - 3 - Sin dato - resd_3
9. **probacad**: ¿Ha tenido su hijo/a algún problema académico u otros problemas en la escuela?
   - 1 - Si - prsi_1

- 2 - No - prno_2
- 3 - Sin dato - prsd_3

10. **enfermed**: ¿Padece su hijo(a) de alguna enfermedad, incapacidad física o mental?
    - 1 - Si - ensi_1
    - 2 - No - enno_2
    - 3 - Sin dato - ensd_3
11. **instrucc**: Instrucción del padre
    - 1 - Primaria - pprimaria_1
    - 2 - UTU - putu_2
    - 3 - Secundaria 1er Ciclo - pse1_3
    - 4 - Secundaria 2do Ciclo - pse2_4
    - 5 - Escuela Militar o Policial - pmilitar_5
    - 6 - Magisterio o Profesorado - pmagisterio_6
    - 7 - Universitario o Terciario - puniversitario_7
    - 8 - Sin estudios - pnoestudio_8
    - 999 - Sin dato - psd_9
12. **instruc1**: Instrucción de la madre
    - 1 - Primaria - mprimaria_1
    - 2 - UTU - mutu_2
    - 3 - Secundaria 1er Ciclo - mse1_3
    - 4 - Secundaria 2do Ciclo - mse2_4
    - 5 - Escuela Militar o Policial - mmilitar_5
    - 6 - Magisterio o Profesorado - mmagisterio_6
    - 7 - Universitario o Terciario - muniversitario_7
    - 8 - Sin estudios - mnoestudio_8
    - 999 - Sin dato - msd_9
13. **nse**: Nivel Socioeconómico. Se desconocía como esta variable estaba construida.
    - 1 - Bajo - bajo_1
    - 2 - Medio - medio_2
    - 3 - Alto - alto_3
    - 4 - Sin dato - nodato_4

14. **saludabl**: Saludable. Representa, *a priori*, si el niño/a es saludable o no. Una de las preguntas realizadas a los padres sobre sus niños consistía en saber si éstos habían sido alguna vez referidos a profesionales de la salud mental. Cuando la respuesta era negativa, se podía decir que el niño era *saludable*.

    - 1 - Si - salsi_1
    - 2 - No - salno_2

15. **patolog**: Patológico. Representa si el niño/a es patológico o no.

    - 1 - Si - patsi_1
    - 2 - No - patno_2

16. **sindglo1rec**: Sindrome Global. Índice construido a partir de una serie de variables del cuestionario CBCL. Representa el estado de salud mental del niño/a luego de aplicado el test.

    - 1 - No patológico - nopatologico_1
    - 2 - Zona de riesgo - zonariesgo_2
    - 3 - Patológico - patologico_3

17. **sexoedadrec**: Sexo-Edad. Esta variable fue construida cruzando las variables sexo y edad.

    - 1 - niños de 5 a 7 años - m57_1
    - 2 - niños de 8 y 9 años - m89_2
    - 3 - niños de 10 a 13 años - m1013_3
    - 4 - niñas de 5 a 7 años - f57_4
    - 5 - niñas de 8 y 9 años - f89_5
    - 6 - niñas de 10 a 13 años - f1013_6

18. **pregArec**: ¿Qué es lo que más le preocupa acerca de su hijo/a? Esta pregunta estaba formulada como una pregunta de respuesta libre. Cabe destacar que en la base a la que se tuvo acceso no aparecía digitada en su totalidad. Luego de realizar la digitación total de las respuestas, se procedió a su postcodificación donde se obtuvieron, en un primer momento, 29 códigos que luego algunos de ellos fueron colapsados. Esto explica que los códigos presentados a continuación no sean consecutivos.

    - 1 - Nada - anada_1
    - 2 - Futuro - afuturo_6
    - 4 - Conducta - aconducta_9
    - 5 - Salud - asalud_10

- 6 - Violencia - aviolencia_11
- 7 - Educación - aeducacion_12
- 8 - Inquietud - ainquietud_13
- 10 - Carácter - acaracter_2
- 13 - Inmadurez - ainmadurez_3
- 15 - Timidez - atimidez_4
- 17 - Distracción - adistraccion_5
- 22 - Sentimientos - asentimientos_7
- 24 - Responsabilidad - aresponsabilidad_8
- 94 - Otros - aotros_14
- 97 - No contesta - anoc_15

19. **pregBrec**: ¿Qué es lo mejor que le ve a su hijo/a?. Idem anterior. En este caso los códigos obtenidos en una primera instancia fueron 33, que luego de colapsar alguno de ellos quedaron en los 13 que se presentan a continuación.

    - 1 - Cariño - bcariño_1
    - 2 - Compañerismo - bcompañerismo_7
    - 3 - Madurez - bmadurez_8
    - 4 - La escuela - bescuela_9
    - 5 - Conducta - bconducta_10
    - 6 - Inteligencia - binteligencia_11
    - 11 - Solidaridad - bsolidaridad_2
    - 14 - Sentimientos - bsentimientos_3
    - 17 - Bondad - bbondad_4
    - 18 - Tenacidad - btenacidad_5
    - 19 - Personalidad - bpersonalidad_6
    - 94 - Otros - botros_12
    - 97 - No contesta - bnoc_13

Por favor utilice letra de imprenta

CUESTIONARIO SOBRE EL COMPORTAMIENTO DE NIÑOS(AS) DE 6-18 AÑOS

Para completar en la oficina
ID #

NOMBRE COMPLETO DEL NIÑO(A): Primer Nombre Segundo Nombre Apellido

SEXO ☐ Masculino ☐ Femenino

EDAD

GRUPO ÉTNICO (blanco, negro, amerindio, asiático...)

FECHA DE HOY Mes ____ Dia ____ Año _____

FECHA DE NACIMIENTO Mes ____ Dia ____ Año _____

GRADO ESCOLAR _____

No va a la escuela: ☐

Por favor complete este cuestionario con su opinión sobre el comportamiento de su hijo(a). Hágalo aunque usted piensa que otras personas no están de acuerdo con su opinión. Siéntase en la libertad de escribir comentarios adicionales al final de cada frase y en el espacio que se provee en la página 2.

TRABAJO USUAL DE LOS PADRES, inclusive si ahora no está trabajando
(por favor especifique - por ejemplo: Mecánico, jardinero, maestro/a, ama de casa, albañil, policía, hace changas, vendedor/a ambulante, profesional).

TRABAJO DEL PADRE: _____

TRABAJO DE LA MADRE: _____

ESTE CUESTIONARIO FUE CONTESTADO POR:

☐ **Padre (Nombre y apellido)** _____

☐ **Madre (Nombre y apellido)** _____

☐ **Otra persona (Nombre y relación con el/la niño(a))** _____

I. ¿Cuáles son las actividades deportivas en las que más le gusta participar a su hijo(a)? Por ejemplo: natación, football, patinaje, bicicleta, basketball, karate, handball, pescar, etc.

☐ Ninguno

	En comparación con otros niños(as) de su edad, ¿cuánto tiempo le dedica a cada uno de estos deportes?				**En comparación con otros niños(as) de su edad, ¿qué tan bueno es él/ella en estos deportes?**			
	Menos que los demás	**Igual que los demás**	**Más que los demás**	**No lo sé**	**Peor que los demás**	**Igual que los demás**	**Mejor que los demás**	**No lo sé**
a. _____	☐	☐	☐	☐	☐	☐	☐	☐
b. _____	☐	☐	☐	☐	☐	☐	☐	☐
c. _____	☐	☐	☐	☐	☐	☐	☐	☐

II. ¿Cuáles son las actividades, juegos o pasatiempos favoritos de su hijo(a) además de los deportes? Por ejemplo, colección de figuritas, cartas; juegos de armar; jugar con muñecos/as, leer, tocar música, cantar, etc. **(No incluya** escuchar la radio o ver televisión).

☐ Ninguno

	En comparación con otros niños(as) de su edad, ¿cuánto tiempo le dedica a cada uno de estas actividades?				**En comparación con otros niños(as) de su edad, ¿cómo es él/ella en estas actividades?**			
	Menos que los demás	**Igual que los demás**	**Más que los demás**	**No lo sé**	**Peor que los demás**	**Igual que los demás**	**Mejor que los demás**	**No lo sé**
a. _____	☐	☐	☐	☐	☐	☐	☐	☐
b. _____	☐	☐	☐	☐	☐	☐	☐	☐
c. _____	☐	☐	☐	☐	☐	☐	☐	☐

III. ¿Cuáles son las organizaciones, equipos, clubes o grupos a los que pertenece su hijo(a)?

☐ Ninguno

	En comparación con otros niños(as) de su edad, ¿qué tan activo(a) es en cada uno de los grupos?			
	Menos que los demás	**Igual que los demás**	**Más que los demás**	**No lo sé**
a. _____	☐	☐	☐	☐
b. _____	☐	☐	☐	☐
c. _____	☐	☐	☐	☐

IV. ¿Qué trabajos o tareas hace su hijo(a)? Por ejemplo: cuidar de otros niños, hacer la cama, trabajar en una tienda, hacer mandados, vender en los ómnibus, etc. (Incluya tareas o trabajos pagados y no pagados.)

☐ Ninguno

	En comparación con otros niños(as) de su edad, ¿cómo lleva a cabo estas tareas?			
	Peor que los demás	**Igual que los demás**	**Mejor que los demás**	**No lo sé**
a. _____	☐	☐	☐	☐
b. _____	☐	☐	☐	☐
c. _____	☐	☐	☐	☐

Asegúrese que contestó todas las preguntas.

©Derechos Reservados 2001 T.M. Achenbach, U. of Vermont
1 S. Prospect St., Burlington, VT 05401 **LA LEY PROHIBE LA REPRODUCCION NO AUTORIZADA**
www.ASEBA.org

Por favor utilizar letra de imprenta. Asegúrese que contestó todas las preguntas.

V. 1. ¿Cuántos amigos o amigas íntimos(as) tiene su hijo(a)? (No incluya a sus hermanos o hermanas.)
❒ **Ninguno** ❒ **1** ❒ **2 ó 3** ❒ **4 o más**

2. Sin contar las horas en que está en la escuela, ¿cuántas veces a la semana participa su hijo(a) en actividades con sus amigos(as)? ❒ **Menos de 1** ❒ **1 ó 2** ❒ **3 o más**

VI. En comparación con otros niños o niñas de la misma edad, ¿cómo . . .

	¿Peor que los demás?	¿Igual que los demás?	¿Mejor que los demás?	
a. se lleva con sus hermanos y hermanas?	❒	❒	❒	❒ No tiene hermanos o hermanas
b. se lleva con otros niños y niñas?	❒	❒	❒	
c. se comporta con su papá y mamá?	❒	❒	❒	
d. juega solo(a) y hace sus tareas solo(a)?	❒	❒	❒	

VII. 1. Desempeño escolar. ❒ **Si su hijo(a) no está en la escuela, por favor escriba la razón.** _____

	Marque una respuesta para cada materia.	**Fue reprobado**	**Por debajo del promedio**	**Promedio**	**Más alto que el promedio**
	a. Lectura, Español o Literatura	❒	❒	❒	❒
	b. Historia o Estudios sociales	❒	❒	❒	❒
Otras materias, como por ejemplo, idiomas, cursos de computadoras, comercio, etc. **No incluya** cursos como educación física, artes industriales, etc.	c. Matemáticas o Aritmética	❒	❒	❒	❒
	d. Ciencias	❒	❒	❒	❒
	e. _____	❒	❒	❒	❒
	f. _____	❒	❒	❒	❒
	g. _____	❒	❒	❒	❒

2. ¿Está su hijo(a) en una clase o escuela especial o recibe servicios especiales? ❒ **No** ❒ **Sí—¿En qué tipo de clase o escuela especial está? (Especifique):**

3. ¿Ha repetido algún año? ❒ **No** ❒ **Sí—¿Qué año o años y por qué?**

4. ¿Ha tenido su hijo(a) algún problema académico u otros problemas en la escuela? ❒ **No** ❒ **Sí—por favor describa:**

¿Cuándo empezaron estos problemas?

¿Han terminado estos problemas? ❒ **No** ❒ **Sí—¿Cuándo terminaron?**

¿Padece su hijo(a) de alguna enfermedad, incapacidad física o mental? ❒ **No** ❒ **Sí—por favor describa el problema:**

¿Qué es lo que más le preocupa acerca de su hijo(a)?

¿Qué es lo mejor que le ve a su hijo(a)? Por favor describa:

A continuación hay una lista de frases que describen a los(las) niños(as) y jóvenes. Para cada frase que describa cómo es su hijo(a) ***ahora o durante los últimos seis meses*** haga un círculo en el número ***2*** si la frase describe a su hijo(a) ***muy a menudo***. Haga un círculo en el número ***1*** si la frase describe a su hijo(a) ***en cierta manera o algunas veces***. Haga un círculo en el ***0*** si la descripción con respecto a su hijo(a) ***no es cierta***. Por favor conteste todas las frases de la mejor manera posible inclusive si algunas de ellas parecen no describir a su hijo(a). ***Por favor escriba en letra de imprenta. Asegúrese que contestó todas las preguntas.***

0 = No es cierto (que sepa usted) **1 = En cierta manera, algunas veces** **2 = Muy cierto o cierto a menudo**

0 1 2 1. Se comporta como si tuviese menos edad
0 1 2 2. Toma bebidas alcohólicas sin permiso de los padres (describa):_____
_____
0 1 2 3. Discute, protesta mucho
0 1 2 4. Deja sin terminar lo que él/ella empieza
0 1 2 5. Disfruta de muy pocas cosas
0 1 2 6. Se hace caca encima o en lugares inadecuados
0 1 2 7. Es engreído, se manda la parte
0 1 2 8. No puede concentrarse o prestar atención por mucho tiempo
0 1 2 9. No puede sacarse algunos pensamientos de la cabeza. Ideas fijas, obsesiones (describa):_____
0 1 2 10. No puede quedarse quieto(a); es inquieto(a) o hiperactivo(a)
0 1 2 11. Es demasiado dependiente o apegado(a) a los adultos
0 1 2 12. Se queja de que se siente solo(a)
0 1 2 13. Está confundido(a) o embarullado
0 1 2 14. Llora mucho
0 1 2 15. Es cruel con los animales
0 1 2 16. Es cruel, amenaza o hace maldades a otros
0 1 2 17. Sueña despierto, se pierde en sus propios pensamientos
0 1 2 18. Se hace daño a sí mismo(a) eliberadamente o ha intentado suicidarse
0 1 2 19. Exige mucha atención
0 1 2 20. Destruye sus propias cosas
0 1 2 21. Destruye las pertenencias de sus familiares o de otras personas
0 1 2 22. Desobedece en casa
0 1 2 23. Desobedece en la escuela
0 1 2 24. No come bien
0 1 2 25. No se lleva bien con otros niños(as)/jóvenes
0 1 2 26. No parece sentirse culpable después de portarse mal
0 1 2 27. Se pone celoso(a) fácilmente
0 1 2 28. No respeta las reglas en casa, en la escuela o en otro lugar
0 1 2 29. Tiene miedo de ciertas situaciones, animales o lugares (no incluya la escuela) (describa):
_____
0 1 2 30. Le da miedo ir a la escuela
0 1 2 31. Tiene miedo de que pueda pensar o hacer algo malo
0 1 2 32. Siente que tiene que ser perfecto/a
0 1 2 33. Siente o se queja de que nadie lo/la quiere
0 1 2 34. Siente que los demás lo/la quieren perjudicar
0 1 2 35. Se siente inferior o cree que no vale nada
0 1 2 36. Se lastima accidentalmente con mucha frecuencia, propenso a accidentes
0 1 2 37. Se mete mucho en peleas
0 1 2 38. Los demás se burlan de él/ella a menudo
0 1 2 39. Se junta con niños(as)/jóvenes que se meten en problemas
0 1 2 40. Oye sonidos o voces que no existen (describa):
_____
_____
0 1 2 41. Impulsivo; actúa sin pensar
0 1 2 42. Prefiere más estar solo que con otras personas
0 1 2 43. Dice mentiras o hace trampas
0 1 2 44. Se muerde las uñas
0 1 2 45. Nervioso(a), tenso(a)
0 1 2 46. Tiene movimientos rápidos, sacudidas, tics **describa:**_____
_____
0 1 2 47. Tiene pesadillas
0 1 2 48. No les cae bien a otros niños(as)/jóvenes
0 1 2 49. Padece de estreñimiento
0 1 2 50. Demasiado ansioso(a) o miedoso(a)
0 1 2 51. Se queja de mareos
0 1 2 52. Se siente demasiado culpable
0 1 2 53. Come demasiado
0 1 2 54. Se siente demasiado cansado sin razón para estarlo
0 1 2 55. Tiene sobrepeso
56. Problemas físicos ***sin causa médica conocida***
0 1 2 a. Dolores o molestias (sin que sean del estómago o dolores de cabeza)
0 1 2 b. Dolores de cabeza
0 1 2 c. Náuseas, ganas de vomitar
0 1 2 d. Problemas con los ojos (si no usa lentes) (describa): _____
_____
0 1 2 e. Sarpullido o irritación en la piel
0 1 2 f. Dolores de estómago
0 1 2 g. Vómitos
0 1 2 h. Otros (describa): _____

Asegúrese que contestó todas las preguntas.
Por favor pase a la página siguiente.

Por favor escriba en letra de imprenta. Asegúrese que contestó todas las preguntas.

0 = No es cierto (que sepa usted) **1 = En cierta manera, algunas veces** **2 = Muy cierto o cierto a menudo**

0 1 2 57. Ataca a la gente físicamente
0 1 2 58. Mete el dedo en la nariz, se araña la piel u otras partes del cuerpo (describa): _____
_____
0 1 2 59. Se toca sus genitales en público
0 1 2 60. Se toca demasiado sus genitales
0 1 2 61. Tiene bajo rendimiento en la escuela
0 1 2 62. Tiene mala coordinación o torpeza
0 1 2 63. Prefiere estar con niños(as)/jóvenes mayores
0 1 2 64. Prefiere estar con niños(as)/jóvenes menores
0 1 2 65. Se rehusa a hablar
0 1 2 66. Repite ciertas acciones una y otra vez, "tiene manías" (describa): _____
_____
0 1 2 67. Se fuga de la casa
0 1 2 68. Grita mucho
0 1 2 69. Reservado(a); se calla todo
0 1 2 70. Ve cosas que no existen (describa): ___
_____
_____
0 1 2 71. Se cohibe o se avergüenza con facilidad
0 1 2 72. Prende fuego
0 1 2 73. Problemas sexuales (describa): _____
_____
_____
0 1 2 74. Le gusta llamar la atención o hacerse el/la payaso(a)
0 1 2 75. Es demasiado tímido(a)
0 1 2 76. Duerme menos que la mayoría de los/las niños(as)/jóvenes
0 1 2 77. Duerme más que la mayoría de los/las niños(as)/jóvenes durante el día y/o la noche (describa): _____
_____
0 1 2 78. No presta atención o se distrae facilmente
0 1 2 79. Tiene problemas en el habla (describa): _____
_____
0 1 2 80. Se queda con la mirada fija, mirando el vacío
0 1 2 81. Roba en casa
0 1 2 82. Roba fuera de casa
0 1 2 83. Almacena demasiadas cosas que no necesita (describa): _____
_____

0 1 2 84. Comportamiento raro (describa): _____
_____
0 1 2 85. Ideas raras (describa): _____
_____
0 1 2 86. Obstinado(a), malhumorado(a), irritable
0 1 2 87. Súbitos cambios de humor o sentimientos
0 1 2 88. Queda contrariado, pone mala cara, "entrompado"
0 1 2 89. Desconfiado(a), receloso(a)
0 1 2 90. Dice groserías, usa lenguaje obsceno
0 1 2 91. Habla de querer matarse
0 1 2 92. Habla o camina cuando está dormido(a) (describa): _____
_____
0 1 2 93. Habla demasiado
0 1 2 94. Se burla mucho de los demás
0 1 2 95. Le dan rabietas o tiene mal genio
0 1 2 96. Parece preocupado por temas sexuales
0 1 2 97. Amenaza a otros
0 1 2 98. Se chupa el dedo
0 1 2 99. Fuma, masca o inhala tabaco
0 1 2 100. No duerme bien (describa): _____
_____
0 1 2 101. Se hace la rabona a la escuela
0 1 2 102. Poco activo(a), lento(a), o le falta energía
0 1 2 103. Infeliz, triste, o deprimido(a)
0 1 2 104. Más ruidoso(a) de lo común
0 1 2 105. Usa drogas sin motivo médico (no incluya alcohol o tabaco) (describa): _____
_____
0 1 2 106. Comete actos de vandalismo, como romper ventanas u otras cosas
0 1 2 107. Se orina en la ropa durante el día
0 1 2 108. Se orina en la cama
0 1 2 109. Es quejoso, se lamenta
0 1 2 110. Desea ser del sexo opuesto
0 1 2 111. Se aisla, no se relaciona con los demás
0 1 2 112. Se preocupa mucho
113. Por favor anote cualquier otro problema que su niño(a) tenga y que no está incluído en esta lista
0 1 2 _____
0 1 2 _____
0 1 2 _____

POR FAVOR ASEGÚRESE QUE CONTESTÓ TODAS LAS PREGUNTAS **SUBRAYE LA PREGUNTA(S) QUE LE PREOCUPE(N)**

Apéndice C

Análisis descriptivo de las variables

Distributions of categorical variables

	bases			weight			
	absolu	%/total	%/expr.	absolu	%/total	%/expr.	histogram of weight
1 . grado_V#1							
cat1 - primero_1	220	16.82	16.82	220.00	16.82	16.82	********
cat2 - segundo_2	227	17.35	17.35	227.00	17.35	17.35	********
cat3 - tercero_3	208	15.90	15.90	208.00	15.90	15.90	********
cat4 - cuarto_4	256	19.57	19.57	256.00	19.57	19.57	**********
cat5 - quinto_5	195	14.91	14.91	195.00	14.91	14.91	*******
cat6 - sexto_6	77	5.89	5.89	77.00	5.89	5.89	***
cat7 - preescolar_7	84	6.42	6.42	84.00	6.42	6.42	***
catb - grsd_8	41	3.13	3.13	41.00	3.13	3.13	**
together	1308	100.00	100.00	1308.00	100.00	100.00	
2 . sexo_V#2							
cat1 - masculino_1	643	49.16	49.16	643.00	49.16	49.16	**********
cat2 - femenino_2	665	50.84	50.84	665.00	50.84	50.84	**********
together	1308	100.00	100.00	1308.00	100.00	100.00	
3 . edadrec_V#3							
cat1 - 5a7_1	385	29.43	29.43	385.00	29.43	29.43	**********
cat2 - 8y9_2	430	32.87	32.87	430.00	32.87	32.87	**********
cat3 - 10a13_3	493	37.69	37.69	493.00	37.69	37.69	**********
together	1308	100.00	100.00	1308.00	100.00	100.00	
4 . trabpadrec_V#4							
cat1 - pffaa_1	50	3.82	3.82	50.00	3.82	3.82	**
cat2 - pprof_2	70	5.35	5.35	70.00	5.35	5.35	***
cat3 - ptecn_3	38	2.91	2.91	38.00	2.91	2.91	**
cat4 - pempl_4	74	5.66	5.66	74.00	5.66	5.66	***
cat5 - pvend_5	241	18.43	18.43	241.00	18.43	18.43	*********
cat6 - pagrop_6	43	3.29	3.29	43.00	3.29	3.29	**
cat7 - pope_7	352	26.91	26.91	352.00	26.91	26.91	**********
cat8 - pncal_8	222	16.97	16.97	222.00	16.97	16.97	********
cat9 - potros_9	62	4.74	4.74	62.00	4.74	4.74	***
catb - psindato_10	156	11.93	11.93	156.00	11.93	11.93	******
together	1308	100.00	100.00	1308.00	100.00	100.00	
5 . trabajomadrec_V#5							
cat1 - mprof_1	120	9.17	9.17	120.00	9.17	9.17	*****
cat2 - mtecn_2	55	4.20	4.20	55.00	4.20	4.20	**
cat3 - mempl_3	134	10.24	10.24	134.00	10.24	10.24	*****
cat4 - mvende_4	154	11.77	11.77	154.00	11.77	11.77	******
cat5 - mncal_5	722	55.20	55.20	722.00	55.20	55.20	**********
cat6 - motros_6	50	3.82	3.82	50.00	3.82	3.82	**

```
catb - msindato_7              73    5.58    5.58        73.00    5.58    5.58    ***
       together              1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
   6 . este_V#6

cat1 - papa_1                  184   14.07   14.07       184.00   14.07   14.07    *******
cat2 - mama_2                 1013   77.45   77.45      1013.00   77.45   77.45    ************************************
cat3 - otro_3                   77    5.89    5.89        77.00    5.89    5.89    ***
cat4 - entrev_4                  5     .38     .38         5.00     .38     .38    *
cat9 - ignorado_5               29    2.22    2.22        29.00    2.22    2.22    **
       together              1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------

                        ------   bases   -------     -------- weight---------
                         absolu  %/total  %/expr.     absolu  %/total  %/expr.      histogram   of weights
-------------------------------------------------------------------------------------------------------------------------------
   7 . clasesp_V#7

cat1 - clsi_1                  48    3.67    3.67        48.00    3.67    3.67    **
cat2 - clno_2                1244   95.11   95.11      1244.00   95.11   95.11    ********************************************
catb - clsd_3                  16    1.22    1.22        16.00    1.22    1.22    *
       together              1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
   8 . repano_V#8

cat1 - resi_1                 246   18.81   18.81       246.00   18.81   18.81    *********
cat2 - reno_2                1055   80.66   80.66      1055.00   80.66   80.66    ************************************
catb - resd_3                   7     .54     .54         7.00     .54     .54    *
       together              1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
   9 . probacad_V#9

cat1 - prsi_1                 160   12.23   12.23       160.00   12.23   12.23    ******
cat2 - prno_2                1136   86.85   86.85      1136.00   86.85   86.85    ****************************************
catb - prsd_3                  12     .92     .92        12.00     .92     .92    *
       together              1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  10 . enfermed_V#10

cat1 - ensi_1                 121    9.25    9.25       121.00    9.25    9.25    *****
cat2 - enno_2                1177   89.98   89.98      1177.00   89.98   89.98    ******************************************
catb - ensd_3                  10     .76     .76        10.00     .76     .76    *
       together              1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  11 . instrucc_V#11

ca10 - pprimaria_1             384   29.36   29.36       384.00   29.36   29.36    **************
ca20 - putu_2                  251   19.19   19.19       251.00   19.19   19.19    *********
ca30 - pse1_3                  205   15.67   15.67       205.00   15.67   15.67    ********
ca40 - pse2_4                  147   11.24   11.24       147.00   11.24   11.24    ******
ca50 - pmilitar_5               41    3.13    3.13        41.00    3.13    3.13    **
ca60 - pmagisterio_6             6     .46     .46         6.00     .46     .46    *
ca70 - puniversitario_7        136   10.40   10.40       136.00   10.40   10.40    *****
ca80 - pnoestudio_8              4     .31     .31         4.00     .31     .31    *
catb - psd_9                   134   10.24   10.24       134.00   10.24   10.24    *****
       together              1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  12 . instruc1_V#12

ca10 - mprimaria_1             365   27.91   27.91       365.00   27.91   27.91    *************
ca20 - mutu_2                  162   12.39   12.39       162.00   12.39   12.39    ******
ca30 - mse1_3                  247   18.88   18.88       247.00   18.88   18.88    *********
ca40 - mse2_4                  222   16.97   16.97       222.00   16.97   16.97    ********
ca50 - mmilitar_5                7     .54     .54         7.00     .54     .54    *
ca60 - mmagisterio_6            50    3.82    3.82        50.00    3.82    3.82    **
ca70 - muniversitario_7        182   13.91   13.91       182.00   13.91   13.91    *******
ca80 - mnoestudio_8              3     .23     .23         3.00     .23     .23    *
catb - msd_9                    70    5.35    5.35        70.00    5.35    5.35    ***
       together              1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  13 . nse_V#13

cat1 - bajo_1                  856   65.44   65.44       856.00   65.44   65.44    *******************************
cat2 - medio_2                 241   18.43   18.43       241.00   18.43   18.43    *********
cat3 - alto_3                  191   14.60   14.60       191.00   14.60   14.60    *******
catb - nodato_4                 20    1.53    1.53        20.00    1.53    1.53    *
       together              1308  100.00  100.00      1308.00  100.00  100.00
```

```
                        ------   bases   -------     -------- weight---------
                         absolu  %/total  %/expr.     absolu  %/total  %/expr.      histogram   of weights
-------------------------------------------------------------------------------------------------------------------------------
  14 . saludabl_V#14

cat1 - salsi_1             1044   79.82   79.82      1044.00   79.82   79.82    ************************************
cat2 - salno_2              264   20.18   20.18       264.00   20.18   20.18    **********
       together            1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  15 . patolog_V#15

cat1 - patsi_1              266   20.34   20.34       266.00   20.34   20.34    **********
cat2 - patno_2             1042   79.66   79.66      1042.00   79.66   79.66    ************************************
       together            1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  16 . sindglo1rec_V#16

cat1 - nopatologico_1       947   72.40   72.40       947.00   72.40   72.40    *********************************
cat2 - zonariesgo_2         149   11.39   11.39       149.00   11.39   11.39    ******
cat3 - patologico_3         212   16.21   16.21       212.00   16.21   16.21    ********
       together            1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  17 . sexoedadrec_V#17

cat1 - m57_1                188   14.37   14.37       188.00   14.37   14.37    *******
cat2 - m89_2                206   15.75   15.75       206.00   15.75   15.75    ********
cat3 - m1013_3              249   19.04   19.04       249.00   19.04   19.04    *********
cat4 - f57_4                197   15.06   15.06       197.00   15.06   15.06    *******
cat5 - f89_5                224   17.13   17.13       224.00   17.13   17.13    ********
cat6 - f1013_6              244   18.65   18.65       244.00   18.65   18.65    *********
       together            1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  18 . pregArec_V#18

cat1 - anada_1              184   14.07   14.07       184.00   14.07   14.07    *******
ca10 - acaracter_2           58    4.43    4.43        58.00    4.43    4.43    ***
ca13 - ainmadurez_3          36    2.75    2.75        36.00    2.75    2.75    **
ca15 - atimidez_4            73    5.58    5.58        73.00    5.58    5.58    ***
ca17 - adistraccion_5        45    3.44    3.44        45.00    3.44    3.44    **
cat2 - afuturo_6            105    8.03    8.03       105.00    8.03    8.03    ****
ca22 - asentimientos_7       47    3.59    3.59        47.00    3.59    3.59    **
ca24 - aresponsabilidad_8    29    2.22    2.22        29.00    2.22    2.22    **
cat4 - aconducta_9           59    4.51    4.51        59.00    4.51    4.51    ***
cat5 - asalud_10             72    5.50    5.50        72.00    5.50    5.50    ***
cat6 - aviolencia_11         42    3.21    3.21        42.00    3.21    3.21    **
cat7 - aeducacion_12        150   11.47   11.47       150.00   11.47   11.47    ******
cat8 - ainquietud_13         77    5.89    5.89        77.00    5.89    5.89    ***
ca94 - aotros_14            124    9.48    9.48       124.00    9.48    9.48    *****
ca97 - anoc_15              207   15.83   15.83       207.00   15.83   15.83    ********
       together            1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
  19 . pregBrec_V#19

cat1 - bcariño_1            135   10.32   10.32       135.00   10.32   10.32    *****
ca11 - bsolidaridad_2        88    6.73    6.73        88.00    6.73    6.73    ****
ca14 - bsentimientos_3      119    9.10    9.10       119.00    9.10    9.10    *****
ca17 - bbondad_4            109    8.33    8.33       109.00    8.33    8.33    ****
ca18 - btenacidad_5         107    8.18    8.18       107.00    8.18    8.18    ****
ca19 - bpersonalidad_6       66    5.05    5.05        66.00    5.05    5.05    ***
cat2 - bcompañerismo_7      193   14.76   14.76       193.00   14.76   14.76    *******
cat3 - bmadurez_8            42    3.21    3.21        42.00    3.21    3.21    **
cat4 - bescuela_9            80    6.12    6.12        80.00    6.12    6.12    ***
cat5 - bconducta_10          68    5.20    5.20        68.00    5.20    5.20    ***
cat6 - binteligencia_11      77    5.89    5.89        77.00    5.89    5.89    ***
ca94 - botros_12            127    9.71    9.71       127.00    9.71    9.71    *****
ca97 - bnoc_13               97    7.42    7.42        97.00    7.42    7.42    ****
       together            1308  100.00  100.00      1308.00  100.00  100.00
-------------------------------------------------------------------------------------------------------------------------------
```

Apéndice D

Salidas

D.1. Análisis Correspondencias

D.1.1. Primer análisis de corresponcencia múltiple

A continución se presenta la salida para el análisis de correspondencia múltiple considerando las variables: Grado al cual concurría el niño/a al momento de la encuesta, Trabajo del padre, Trabajo de la madre, Quién contestó el formulario, Si el niño/a repitió algún año, Instrucción del padre, Instrucción de la madre, Si el niño es saludable, Síndrome Global recodificado y la variable cruzada Sexo-Edad; y utilizando las preguntas post-codificadas A y B como suplementarias.

```
 ====================================
       10 questions  (active)
       62 Associated categories
 ====================================
 ------------------------------------------------------------------------------------------
    1 . grado_V#1                                          (   8 categories )
    4 . trabpadrec_V#4                                     (  10 categories )
    5 . trabajomadrec_V#5                                  (   7 categories )
    6 . este_V#6                                           (   5 categories )
    8 . repano_V#8                                         (   3 categories )
   11 . instrucc_V#11                                      (   9 categories )
   12 . instruc1_V#12                                      (   9 categories )
   14 . saludabl_V#14                                      (   2 categories )
   16 . sindglo1rec_V#16                                   (   3 categories )
   17 . sexoedadrec_V#17                                   (   6 categories )
 ------------------------------------------------------------------------------------------

 ====================================
        2 questions  (supplementary)
       28 Associated categories
 ====================================
 ------------------------------------------------------------------------------------------
   18 . pregArec_V#18                                      (  15 categories )
   19 . pregBrec_V#19                                      (  13 categories )
 ------------------------------------------------------------------------------------------


Accuracy of computation :                trace before diagonalisation       5.2000
 ----------------------------------       sum of eigenvalues                 5.2000

 histogram of the 52 first eigenvalues
 -----------------------------------------

 +--------+------------+----------+----------+-------------------------------------------------------------------------------------+
 ! number !    Eigen   ! percent. ! cumulat. !                                                                                     !
 !        !    value   !          ! percent. !                                                                                     !
```

```
+--------+-----------+---------+---------+-------------------------------------------------------------------------------+
!    1   !    .2911   !    5.60 !    5.60 ! ****************************************************************************** !
!    2   !    .2134   !    4.10 !    9.70 ! ********************************************************                       !
!    3   !    .1841   !    3.54 !   13.24 ! ************************************************                               !
!    4   !    .1770   !    3.40 !   16.64 ! ***********************************************                                !
!    5   !    .1659   !    3.19 !   19.84 ! *********************************************                                 !
!    6   !    .1543   !    2.97 !   22.80 ! *******************************************                                   !
!    7   !    .1475   !    2.84 !   25.64 ! ******************************************                                    !
!    8   !    .1395   !    2.68 !   28.32 ! ******************************************                                    !
!    9   !    .1374   !    2.64 !   30.97 ! *****************************************                                     !
!   10   !    .1349   !    2.60 !   33.56 ! *****************************************                                     !
!   11   !    .1262   !    2.43 !   35.99 ! ****************************************                                      !
!   12   !    .1253   !    2.41 !   38.40 ! ****************************************                                      !
!   13   !    .1221   !    2.35 !   40.74 ! ***************************************                                       !
!   14   !    .1169   !    2.25 !   42.99 ! **************************************                                        !
!   15   !    .1146   !    2.20 !   45.20 ! **************************************                                        !
!   16   !    .1124   !    2.16 !   47.36 ! *************************************                                         !
!   17   !    .1109   !    2.13 !   49.49 ! *************************************                                         !
!   18   !    .1074   !    2.07 !   51.56 ! ************************************                                          !
!   19   !    .1068   !    2.05 !   53.61 ! ************************************                                          !
+--------+-----------+---------+---------+-------------------------------------------------------------------------------+
1-------------------------------------------------------------------------------------------------------------------------
      coordinates, contributions and squared cosine of active categories on axes  1 to  5
 -------------------------------------------------------------------------------------------------------------------------

 +-----------------------------------------+-----------------------------+-------------------------+-------------------------+
 !               categories                !          coordinates        !      contributions      !      squared cosine     !
 !-----------------------------------------+-----------------------------+-------------------------+-------------------------!
 ! iden - libelle             p.rel  disto !   1     2     3     4     5 !   1    2    3    4    5 !   1    2    3    4    5 !
 +-----------------------------------------+-----------------------------+-------------------------+-------------------------+
 !                                                                                                                           !
 !   1 . grado_V#1                                                                                                           !
 !                                                                                                                           !
 ! cat1 - primero_1            1.68   4.95 !  .12  1.00  1.00   .72  -.06 !  .1  8.0  9.2  5.0   .0 ! .00  .20  .20  .11  .00 !
 ! cat2 - segundo_2            1.74   4.76 ! -.06   .36   .01  -.75  -.59 !  .0  1.0   .0  5.5  3.6 ! .00  .03  .00  .12  .07 !
 ! cat3 - tercero_3            1.59   5.29 ! -.13  -.24  -.06 -1.18  -.74 !  .1   .4   .0 12.6  5.2 ! .00  .01  .00  .26  .10 !
 ! cat4 - cuarto_4             1.96   4.11 ! -.06  -.54  -.60   .12  -.03 !  .0  2.7  3.8   .2   .0 ! .00  .07  .09  .00  .00 !
 ! cat5 - quinto_5             1.49   5.71 !  .00  -.75  -.48   .50  1.17 !  .0  4.0  1.9  2.1 12.4 ! .00  .10  .04  .04  .24 !
 ! cat6 - sexto_6               .59  15.99 ! -.04  -.88  -.45   .72   .69 !  .0  2.2   .6  1.7  1.7 ! .00  .05  .01  .03  .03 !
 ! cat7 - preescolar_7          .64  14.57 !  .29  1.42   .71   .69   .38 !  .2  6.1  1.7  1.7   .6 ! .01  .14  .03  .03  .01 !
 ! catb - grsd_8                .31  30.90 !  .18  -.44   .28   .37  -.19 !  .0   .3   .1   .2   .1 ! .00  .01  .00  .00  .00 !
 +-----------------------------------------+--------- cumul.  contribution =   .4 24.6 17.4 29.0 23.6 +-------------------------+
 !                                                                                                                           !
 !   4 . trabpadrec_V#4                                                                                                      !
 !                                                                                                                           !
 ! cat1 - pffaa_1               .38  25.16 ! -.08   .47  -.75   .12   .29 !  .0   .4  1.2   .0   .2 ! .00  .01  .02  .00  .00 !
 ! cat2 - pprof_2               .54  17.69 ! 2.28 -1.30   .74   .65 -1.05 ! 9.5  4.2  1.6  1.3  3.5 ! .29  .09  .03  .02  .06 !
 ! cat3 - ptecn_3               .29  33.42 !  .92  -.31  -.32  -.75   .15 !  .8   .1   .2   .9   .0 ! .03  .00  .00  .02  .00 !
 ! cat4 - pempl_4               .57  16.68 !  .54  -.60  -.11  -.21   .17 !  .6  1.0   .0   .1   .1 ! .02  .02  .00  .00  .00 !
 ! cat5 - pvend_5              1.84   4.43 !  .39   .30  -.30  -.28   .38 ! 1.0   .8   .9   .8  1.6 ! .03  .02  .02  .02  .03 !
 ! cat6 - pagrop_6              .33  29.42 !  .38   .09   .14   .51   .16 !  .2   .0   .0   .5   .0 ! .01  .00  .00  .01  .00 !
 ! cat7 - pope_7                2.69   2.72 ! -.13   .43  -.12   .03  -.04 !  .2  2.3   .2   .0   .0 ! .01  .07  .01  .00  .00 !
 ! cat8 - pncal_8               1.70   4.89 ! -.69   .00  -.34   .56  -.61 ! 2.8   .0  1.1  3.0  3.9 ! .10  .00  .02  .06  .08 !
 ! cat9 - potros_9               .47  20.10 ! -.40  -.16   .14  -.28  -.31 !  .3   .1   .0   .2   .3 ! .01  .00  .00  .00  .00 !
 ! catb - psindato_10           1.19   7.38 ! -.75  -.60  1.16  -.51   .71 ! 2.3  2.0  8.7  1.8  3.6 ! .08  .05  .18  .04  .07 !
 +-----------------------------------------+--------- cumul.  contribution = 17.6 10.8 13.9  8.6 13.3 +-------------------------+
 !                                                                                                                           !
 !   5 . trabajomadrec_V#5                                                                                                   !
 !                                                                                                                           !
 ! cat1 - mprof_1               .92   9.90 !  1.81  -.68   .49   .45  -.64 ! 10.3  2.0  1.2  1.0  2.2 ! .33  .05  .02  .02  .04 !
 ! cat2 - mtecn_2               .42  22.78 !   .88  -.24  -.18  -.12   .14 !  1.1   .1   .1   .0   .1 ! .03  .00  .00  .00  .00 !
 ! cat3 - mempl_3              1.02   8.76 !   .69  -.15   .08  -.42   .50 !  1.7   .1   .0  1.0  1.5 ! .05  .00  .00  .02  .03 !
 ! cat4 - mvende_4             1.18   7.49 !   .15   .41  -.13  -.44   .38 !   .1   .9   .1  1.3  1.1 ! .00  .02  .00  .03  .02 !
 ! cat5 - mncal_5              5.52    .81 !  -.43   .16  -.16   .14  -.12 !  3.4   .7   .8   .6   .4 ! .22  .03  .03  .02  .02 !
 ! cat6 - motros_6              .38  25.16 !  -.27   .00  -.16   .05  -.02 !   .1   .0   .1   .0   .0 ! .00  .00  .00  .00  .00 !
 ! catb - msindato_7            .56  16.92 !  -.81  -.87  1.18  -.36   .37 !  1.2  2.0  4.2   .4   .5 ! .04  .04  .08  .01  .01 !
 +-----------------------------------------+--------- cumul.  contribution = 18.0  5.7  6.5  4.4  5.8 +-------------------------+
 !                                                                                                                           !
 !   6 . este_V#6                                                                                                            !
 !                                                                                                                           !
 ! cat1 - papa_1               1.41   6.11 !   .69  -.23  -.20   .13  -.02 !  2.3   .3   .3   .1   .0 ! .08  .01  .01  .00  .00 !
 ! cat2 - mama_2               7.74    .29 !  -.05   .09  -.02  -.02  -.03 !   .1   .3   .0   .0   .0 ! .01  .03  .00  .00  .00 !
 ! cat3 - otro_3                .59  15.99 !  -.84  -.35   .35   .07  -.04 !  1.4   .3   .4   .0   .0 ! .04  .01  .01  .00  .00 !
 ! cat4 - entrev_4              .04 260.60 !  -.57  1.76  -.29  1.68  -.34 !   .0   .6   .0   .6   .0 ! .00  .01  .00  .01  .00 !
 ! cat9 - ignorado_5            .22  44.10 !  -.23  -.97  1.16  -.69  1.41 !   .0  1.0  1.6   .6  2.7 ! .00  .02  .03  .01  .05 !
 +-----------------------------------------+--------- cumul.  contribution =  3.9  2.5  2.4  1.4  2.7 +-------------------------+
 !                                                                                                                           !
```

```
!   8 . repano_V#8                                                                                                  !
!                                                                                                                   !
! cat1 - resi_1               1.88   4.32 !  -.93  -.73   .04   .01  -.37 !  5.5  4.6   .0   .0  1.6 !  .20  .12  .00  .00  .03 !
! cat2 - reno_2               8.07    .24 !   .22   .16  -.02  -.01   .09 !  1.3   .9   .0   .0   .4 !  .20  .10  .00  .00  .03 !
! catb - resd_3                .05 185.86 !  -.56  1.73  1.50  1.65   .28 !   .1   .8   .7   .8   .0 !  .00  .02  .01  .01  .00 !
+-----------------------------------------+-------- cumul.  contribution =  6.9  6.3   .7   .8  1.9 +-------------------------+
!                                                                                                                   !
!  11 . instrucc_V#11                                                                                               !
!                                                                                                                   !
! ca10 - pprimaria_1          2.94   2.41 !  -.67  -.11  -.26   .57  -.51 !  4.5   .2  1.1  5.4  4.6 !  .19  .01  .03  .14  .11 !
! ca20 - putu_2               1.92   4.21 !  -.12   .46  -.27  -.09   .03 !   .1  1.9   .8   .1   .0 !  .00  .05  .02  .00  .00 !
! ca30 - pse1_3               1.57   5.38 !   .16   .42  -.31  -.34   .23 !   .1  1.3   .8  1.0   .5 !  .00  .03  .02  .02  .01 !
! ca40 - pse2_4               1.12   7.90 !   .75   .39  -.15  -.61   .51 !  2.2   .8   .1  2.3  1.8 !  .07  .02  .00  .05  .03 !
! ca50 - pmilitar_5            .31  30.90 !   .02   .26  -.62   .19   .59 !   .0   .1   .6   .1   .7 !  .00  .00  .01  .00  .01 !
! ca60 - pmagisterio_6         .05 217.00 !  1.53  -.54 -1.35 -1.30  -.01 !   .4   .1   .5   .4   .0 !  .01  .00  .01  .01  .00 !
! ca70 - puniversitario_7     1.04   8.62 !  1.84 -1.08   .57   .37  -.61 ! 12.1  5.7  1.8   .8  2.3 !  .39  .14  .04  .02  .04 !
! ca80 - pnoestudio_8          .03 326.00 !  -.37  1.16  -.15   .52   .68 !   .0   .2   .0   .0   .1 !  .00  .00  .00  .00  .00 !
! catb - psd_9                1.02   8.76 !  -.86  -.60  1.59  -.68   .93 !  2.6  1.7 14.1  2.6  5.3 !  .08  .04  .29  .05  .10 !
+-----------------------------------------+-------- cumul.  contribution = 21.9 11.9 19.9 12.8 15.3 +-------------------------+
1
 +-----------------------------------------+-----------------------------+-------------------------+-------------------------+
 !               categories                !          coordinates        !      contributions      !      squared cosine     !
 !-----------------------------------------+-----------------------------+-------------------------+-------------------------!
 ! iden - libelle             p.rel  disto !   1     2     3     4     5  !  1   2   3   4   5  !  1   2   3   4   5  !
 +-----------------------------------------+-----------------------------+-------------------------+-------------------------+
 !                                                                                                                 !
 !  12 . instruc1_V#12                                                                                             !
 !                                                                                                                 !
 ! ca10 - mprimaria_1          2.79   2.58 !  -.74  -.08  -.16   .59  -.49 !  5.2   .1   .4  5.4  4.1 !  .21  .00  .01  .13  .09 !
 ! ca20 - mutu_2               1.24   7.07 !  -.40   .36  -.07  -.09  -.03 !   .7   .8   .0   .1   .0 !  .02  .02  .00  .00  .00 !
 ! ca30 - mse1_3               1.89   4.30 !  -.12   .18  -.51  -.29   .35 !   .1   .3  2.6   .9  1.4 !  .00  .01  .06  .02  .03 !
 ! ca40 - mse2_4               1.70   4.89 !   .33   .46  -.06  -.65   .53 !   .6  1.7   .0  4.1  2.9 !  .02  .04  .00  .09  .06 !
 ! ca50 - mmilitar_5            .05 185.86 !   .73  1.45  -.08   .29  -.08 !   .1   .5   .0   .0   .0 !  .00  .01  .00  .00  .00 !
 ! ca60 - mmagisterio_6         .38  25.16 !  1.24  -.33   .31   .40  -.19 !  2.0   .2   .2   .4   .1 !  .06  .00  .00  .01  .00 !
 ! ca70 - muniversitario_7     1.39   6.19 !  1.56  -.69   .42   .18  -.42 ! 11.6  3.1  1.3   .2  1.5 !  .39  .08  .03  .01  .03 !
 ! ca80 - mnoestudio_8          .02 435.00 !  -.73  -.24  -.62  1.17  -.39 !   .0   .0   .0   .2   .0 !  .00  .00  .00  .00  .00 !
 ! catb - msd_9                 .54  17.69 !  -.82  -.59  1.73  -.60  1.00 !  1.2   .9  8.7  1.1  3.2 !  .04  .02  .17  .02  .06 !
 +-----------------------------------------+-------- cumul.  contribution = 21.7  7.5 13.4 12.3 13.2 +-------------------------+
 !                                                                                                                 !
 !  14 . saludabl_V#14                                                                                             !
 !                                                                                                                 !
 ! cat1 - salsi_1              7.98    .25 !   .15   .10  -.17   .04   .09 !   .6   .4  1.3   .1   .4 !  .09  .04  .12  .01  .04 !
 ! cat2 - salno_2              2.02   3.95 !  -.60  -.40   .68  -.15  -.37 !  2.5  1.5  5.1   .2  1.7 !  .09  .04  .12  .01  .04 !
 +-----------------------------------------+-------- cumul.  contribution =  3.1  1.9  6.4   .3  2.1 +-------------------------+
 !                                                                                                                 !
 !  16 . sindglo1rec_V#16                                                                                          !
 !                                                                                                                 !
 ! cat1 - nopatologico_1       7.24    .38 !   .23   .11  -.13   .00   .08 !  1.3   .4   .6   .0   .3 !  .14  .03  .04  .00  .02 !
 ! cat2 - zonariesgo_2         1.14   7.78 !  -.30  -.06  -.08  -.03   .22 !   .3   .0   .0   .0   .3 !  .01  .00  .00  .00  .01 !
 ! cat3 - patologico_3         1.62   5.17 !  -.82  -.46   .62   .01  -.53 !  3.7  1.6  3.4   .0  2.7 !  .13  .04  .07  .00  .05 !
 +-----------------------------------------+-------- cumul.  contribution =  5.4  2.1  4.1   .0  3.3 +-------------------------+
 !                                                                                                                 !
 !  17 . sexoedadrec_V#17                                                                                          !
 !                                                                                                                 !
 ! cat1 - m57_1                1.44   5.96 !   .32  1.07   .89   .58   .03 !   .5  7.7  6.2  2.7   .0 !  .02  .19  .13  .06  .00 !
 ! cat2 - m89_2                1.57   5.35 !   .00  -.02  -.16 -1.15  -.68 !   .0   .0   .2 11.9  4.4 !  .00  .00  .00  .25  .09 !
 ! cat3 - m1013_3              1.90   4.25 !  -.28  -.90  -.40   .46   .35 !   .5  7.2  1.6  2.3  1.4 !  .02  .19  .04  .05  .03 !
 ! cat4 - f57_4                1.51   5.64 !   .11  1.01   .73   .59  -.05 !   .1  7.3  4.3  2.9   .0 !  .00  .18  .09  .06  .00 !
 ! cat5 - f89_5                1.71   4.84 !   .05   .02  -.27  -.94  -.67 !   .0   .0   .7  8.5  4.7 !  .00  .00  .01  .18  .09 !
 ! cat6 - f1013_6              1.87   4.36 !  -.10  -.72  -.49   .45   .85 !   .1  4.5  2.4  2.1  8.1 !  .00  .12  .05  .05  .17 !
 +-----------------------------------------+-------- cumul.  contribution =  1.2 26.7 15.4 30.4 18.7 +-------------------------+
1-------------------------------------------------------------------------------------------------------------------------------
      coordinates and test values, of categories on axes       1 a  5
 ---------------------------------------------------------------------------------------------------------------------------

 +-----------------------------------------------------+-----------------------------------+-------------------------------+
 !                    categories                       !            coordinates            !          test-values          !
 !-----------------------------------------------------+-----------------------------------+-------------------------------!
 ! iden - title                eff.    p.abs     disto !    1      2      3      4      5   !  1     2     3     4     5   !
 +-----------------------------------------------------+-----------------------------------+-------------------------------+
 !                                                                                                                   !
 !    1 . grado_V#1                                                                                                  !
 !                                                                                                                   !
 ! cat1 - primero_1            220    220.00      4.95 !   .12   1.00   1.00    .72   -.06 !  1.9  16.3  16.3  11.7   -.9 !
 ! cat2 - segundo_2            227    227.00      4.76 !  -.06    .36    .01   -.75   -.59 ! -1.0   5.9    .1 -12.4  -9.7 !
 ! cat3 - tercero_3            208    208.00      5.29 !  -.13   -.24   -.06  -1.18   -.74 ! -2.0  -3.7   -.9 -18.6 -11.6 !
 ! cat4 - cuarto_4             256    256.00      4.11 !  -.06   -.54   -.60    .12   -.03 ! -1.0  -9.7 -10.6   2.2   -.5 !
```

```
! cat5 - quinto_5              195    195.00      5.71 !    .00   -.75   -.48    .50   1.17 !    .0 -11.4  -7.3   7.5  17.8 !
! cat6 - sexto_6                77     77.00     15.99 !   -.04   -.88   -.45    .72    .69 !   -.3  -8.0  -4.0   6.5   6.3 !
! cat7 - preescolar_7           84     84.00     14.57 !    .29   1.42    .71    .69    .38 !   2.8  13.4   6.7   6.5   3.6 !
! catb - grsd_8                 41     41.00     30.90 !    .18   -.44    .28    .37   -.19 !   1.2  -2.8   1.8   2.4  -1.2 !
+-----------------------------------------------------+-----------------------------------+-------------------------------+
!                                                                                                                           !
!    4 . trabpadrec_V#4                                                                                                     !
!                                                                                                                           !
! cat1 - pffaa_1                 50     50.00     25.16 !   -.08    .47   -.75    .12    .29 !   -.6   3.4  -5.4    .9   2.1 !
! cat2 - pprof_2                 70     70.00     17.69 !   2.28  -1.30    .74    .65  -1.05 !  19.6 -11.1   6.4   5.6  -9.0 !
! cat3 - ptecn_3                 38     38.00     33.42 !    .92   -.31   -.32   -.75    .15 !   5.8  -1.9  -2.0  -4.7    .9 !
! cat4 - pempl_4                 74     74.00     16.68 !    .54   -.60   -.11   -.21    .17 !   4.8  -5.3   -.9  -1.9   1.5 !
! cat5 - pvend_5                241    241.00      4.43 !    .39    .30   -.30   -.28    .38 !   6.7   5.1  -5.1  -4.7   6.6 !
! cat6 - pagrop_6                43     43.00     29.42 !    .38    .09    .14    .51    .16 !   2.6    .6    .9   3.4   1.1 !
! cat7 - pope_7                 352    352.00      2.72 !   -.13    .43   -.12    .03   -.04 !  -2.8   9.4  -2.7    .7   -.9 !
! cat8 - pncal_8                222    222.00      4.89 !   -.69    .00   -.34    .56   -.61 ! -11.3    .0  -5.5   9.1 -10.0 !
! cat9 - potros_9                62     62.00     20.10 !   -.40   -.16    .14   -.28   -.31 !  -3.3  -1.3   1.1  -2.2  -2.5 !
! catb - psindato_10            156    156.00      7.38 !   -.75   -.60   1.16   -.51    .71 !  -9.9  -7.9  15.4  -6.8   9.4 !
+-----------------------------------------------------+-----------------------------------+-------------------------------+
!                                                                                                                           !
!    5 . trabajomadrec_V#5                                                                                                  !
!                                                                                                                           !
! cat1 - mprof_1                 120    120.00      9.90 !   1.81   -.68    .49    .45   -.64 !  20.8  -7.8   5.6   5.2  -7.3 !
! cat2 - mtecn_2                  55     55.00     22.78 !    .88   -.24   -.18   -.12    .14 !   6.6  -1.8  -1.4   -.9   1.1 !
! cat3 - mempl_3                 134    134.00      8.76 !    .69   -.15    .08   -.42    .50 !   8.4  -1.9   1.0  -5.1   6.1 !
! cat4 - mvende_4                154    154.00      7.49 !    .15    .41   -.13   -.44    .38 !   1.9   5.4  -1.8  -5.9   5.1 !
! cat5 - mncal_5                 722    722.00       .81 !   -.43    .16   -.16    .14   -.12 ! -17.1   6.4  -6.6   5.6  -4.6 !
! cat6 - motros_6                 50     50.00     25.16 !   -.27    .00   -.16    .05   -.02 !  -1.9    .0  -1.1    .4   -.1 !
! catb - msindato_7               73     73.00     16.92 !   -.81   -.87   1.18   -.36    .37 !  -7.1  -7.6  10.4  -3.2   3.3 !
+-----------------------------------------------------+-----------------------------------+-------------------------------+
!                                                                                                                           !
!    6 . este_V#6                                                                                                           !
!                                                                                                                           !
! cat1 - papa_1                  184    184.00      6.11 !    .69   -.23   -.20    .13   -.02 !  10.1  -3.4  -3.0   1.8   -.3 !
! cat2 - mama_2                 1013   1013.00       .29 !   -.05    .09   -.02   -.02   -.03 !  -3.5   5.8  -1.4  -1.1  -2.1 !
! cat3 - otro_3                  77     77.00     15.99 !   -.84   -.35    .35    .07   -.04 !  -7.6  -3.1   3.1    .7   -.3 !
! cat4 - entrev_4                 5      5.00    260.60 !   -.57   1.76   -.29   1.68   -.34 !  -1.3   4.0   -.6   3.8   -.8 !
! cat9 - ignorado_5              29     29.00     44.10 !   -.23   -.97   1.16   -.69   1.41 !  -1.2  -5.3   6.3  -3.8   7.7 !
+-----------------------------------------------------+-----------------------------------+-------------------------------+
!                                                                                                                           !
!    8 . repano_V#8                                                                                                         !
!                                                                                                                           !
! cat1 - resi_1                  246    246.00      4.32 !   -.93   -.73    .04    .01   -.37 ! -16.1 -12.6    .7    .1  -6.5 !
! cat2 - reno_2                 1055   1055.00       .24 !    .22    .16   -.02   -.01    .09 !  16.2  11.6  -1.4   -.9   6.3 !
! catb - resd_3                    7      7.00    185.86 !   -.56   1.73   1.50   1.65    .28 !  -1.5   4.6   4.0   4.4    .7 !
+-----------------------------------------------------+-----------------------------------+-------------------------------+
!                                                                                                                           !
!   11 . instrucc_V#11                                                                                                      !
!                                                                                                                           !
! ca10 - pprimaria_1             384    384.00      2.41 !   -.67   -.11   -.26    .57   -.51 ! -15.6  -2.6  -6.1  13.3 -11.9 !
! ca20 - putu_2                  251    251.00      4.21 !   -.12    .46   -.27   -.09    .03 !  -2.1   8.0  -4.8  -1.5    .5 !
! ca30 - pse1_3                  205    205.00      5.38 !    .16    .42   -.31   -.34    .23 !   2.5   6.6  -4.9  -5.3   3.5 !
! ca40 - pse2_4                  147    147.00      7.90 !    .75    .39   -.15   -.61    .51 !   9.7   5.0  -2.0  -7.8   6.6 !
! ca50 - pmilitar_5               41     41.00     30.90 !    .02    .26   -.62    .19    .59 !    .1   1.7  -4.0   1.2   3.8 !
! ca60 - pmagisterio_6             6      6.00    217.00 !   1.53   -.54  -1.35  -1.30   -.01 !   3.8  -1.3  -3.3  -3.2    .0 !
! ca70 - puniversitario_7        136    136.00      8.62 !   1.84  -1.08    .57    .37   -.61 !  22.6 -13.3   7.0   4.5  -7.5 !
! ca80 - pnoestudio_8              4      4.00    326.00 !   -.37   1.16   -.15    .52    .68 !   -.8   2.3   -.3   1.0   1.4 !
! catb - psd_9                   134    134.00      8.76 !   -.86   -.60   1.59   -.68    .93 ! -10.4  -7.3  19.4  -8.3  11.3 !
+-----------------------------------------------------+-----------------------------------+-------------------------------+
1
 +-----------------------------------------------------+-----------------------------------+-------------------------------+
 !                    categories                       !            coordinates            !          test-values          !
 !-----------------------------------------------------+-----------------------------------+-------------------------------!
 ! iden - title                 eff.    p.abs     disto !    1      2      3      4      5   !   1     2     3     4     5   !
 +-----------------------------------------------------+-----------------------------------+-------------------------------+
 !                                                                                                                           !
 !   12 . instruc1_V#12                                                                                                      !
 !                                                                                                                           !
 ! ca10 - mprimaria_1             365    365.00      2.58 !   -.74   -.08   -.16    .59   -.49 ! -16.6  -1.8  -3.7  13.2 -11.1 !
 ! ca20 - mutu_2                  162    162.00      7.07 !   -.40    .36   -.07   -.09   -.03 !  -5.4   4.9  -1.0  -1.2   -.4 !
 ! ca30 - mse1_3                  247    247.00      4.30 !   -.12    .18   -.51   -.29    .35 !  -2.1   3.1  -8.8  -5.0   6.1 !
 ! ca40 - mse2_4                  222    222.00      4.89 !    .33    .46   -.06   -.65    .53 !   5.4   7.5  -1.1 -10.6   8.7 !
 ! ca50 - mmilitar_5                7      7.00    185.86 !    .73   1.45   -.08    .29   -.08 !   1.9   3.9   -.2    .8   -.2 !
 ! ca60 - mmagisterio_6            50     50.00     25.16 !   1.24   -.33    .31    .40   -.19 !   9.0  -2.4   2.2   2.9  -1.3 !
 ! ca70 - muniversitario_7        182    182.00      6.19 !   1.56   -.69    .42    .18   -.42 !  22.6 -10.1   6.1   2.6  -6.2 !
 ! ca80 - mnoestudio_8              3      3.00    435.00 !   -.73   -.24   -.62   1.17   -.39 !  -1.3   -.4  -1.1   2.0   -.7 !
 ! catb - msd_9                    70     70.00     17.69 !   -.82   -.59   1.73   -.60   1.00 !  -7.1  -5.0  14.9  -5.1   8.6 !
```

14 . saludabl_V#14														
cat1 - salsi_1	1044	1044.00	.25	.15	.10	-.17	.04	.09	10.8	7.2	-12.4	2.6	6.8
cat2 - salno_2	264	264.00	3.95	-.60	-.40	.68	-.15	-.37	-10.8	-7.2	12.4	-2.6	-6.8
16 . sindglo1rec_V#16														
cat1 - nopatologico_1	947	947.00	.38	.23	.11	-.13	.00	.08	13.5	6.6	-7.4	.2	4.9
cat2 - zonariesgo_2	149	149.00	7.78	-.30	-.06	-.08	-.03	.22	-3.8	-.8	-1.0	-.4	2.8
cat3 - patologico_3	212	212.00	5.17	-.82	-.46	.62	.01	-.53	-13.0	-7.3	9.9	.1	-8.4
17 . sexoedadrec_V#17														
cat1 - m57_1	188	188.00	5.96	.32	1.07	.89	.58	.03	4.7	15.8	13.2	8.6	.5
cat2 - m89_2	206	206.00	5.35	.00	-.02	-.16	-1.15	-.68	.0	-.4	-2.5	-18.0	-10.7
cat3 - m1013_3	249	249.00	4.25	-.28	-.90	-.40	.46	.35	-5.0	-15.8	-7.0	8.0	6.1
cat4 - f57_4	197	197.00	5.64	.11	1.01	.73	.59	-.05	1.7	15.5	11.1	8.9	-.7
cat5 - f89_5	224	224.00	4.84	.05	.02	-.27	-.94	-.67	.9	.3	-4.4	-15.4	-11.1
cat6 - f1013_6	244	244.00	4.36	-.10	-.72	-.49	.45	.85	-1.7	-12.5	-8.4	7.7	14.7
18 . pregArec_V#18														
cat1 - anada_1	184	184.00	6.11	.00	.19	-.19	.07	.00	-.1	2.7	-2.8	1.0	.0
ca10 - acaracter_2	58	58.00	21.55	.04	.11	.11	-.25	.10	.3	.9	.9	-2.0	.8
ca13 - ainmadurez_3	36	36.00	35.33	.33	-.30	-.07	.01	.07	2.0	-1.9	-.4	.0	.4
ca15 - atimidez_4	73	73.00	16.92	.17	.09	.08	-.07	.26	1.5	.8	.7	-.6	2.3
ca17 - adistraccion_5	45	45.00	28.07	.07	-.19	.08	-.36	-.17	.4	-1.3	.5	-2.5	-1.2
cat2 - afuturo_6	105	105.00	11.46	.65	-.18	-.10	.04	.17	6.9	-1.9	-1.1	.5	1.8
ca22 - asentimientos_7	47	47.00	26.83	-.06	.03	.13	-.06	-.05	-.4	.2	.9	-.4	-.4
ca24 - aresponsabilidad_8	29	29.00	44.10	.63	-.03	.04	-.19	.01	3.4	-.2	.2	-1.0	.1
cat4 - aconducta_9	59	59.00	21.17	-.46	.00	.31	-.17	-.21	-3.6	.0	2.4	-1.3	-1.7
cat5 - asalud_10	72	72.00	17.17	.05	-.07	.17	-.06	-.07	.4	-.6	1.5	-.5	-.6
cat6 - aviolencia_11	42	42.00	30.14	-.25	-.14	.16	-.14	-.26	-1.6	-.9	1.0	-.9	-1.7
cat7 - aeducacion_12	150	150.00	7.72	-.27	-.11	.01	.01	-.10	-3.5	-1.5	.1	.1	-1.3
cat8 - ainquietud_13	77	77.00	15.99	-.14	.20	.05	.07	-.03	-1.2	1.8	.5	.7	-.3
ca94 - aotros_14	124	124.00	9.55	.04	-.08	.01	.11	-.10	.5	-.9	.1	1.3	-1.2
ca97 - anoc_15	207	207.00	5.32	-.15	.06	-.09	.12	.11	-2.4	1.0	-1.4	2.0	1.7
19 . pregBrec_V#19														
cat1 - bcariño_1	135	135.00	8.69	.02	.19	.22	-.07	-.17	.2	2.4	2.7	-.8	-2.1
ca11 - bsolidaridad_2	88	88.00	13.86	.29	.02	-.12	-.20	.06	2.8	.2	-1.2	-1.9	.6
ca14 - bsentimientos_3	119	119.00	9.99	.37	-.31	-.05	.01	.10	4.3	-3.6	-.5	.1	1.1
ca17 - bbondad_4	109	109.00	11.00	-.16	.02	-.08	.00	-.19	-1.7	.2	-.9	.0	-2.1
ca18 - btenacidad_5	107	107.00	11.22	.30	-.13	.03	.03	.15	3.3	-1.4	.3	.4	1.7
ca19 - bpersonalidad_6	66	66.00	18.82	.04	.22	-.06	.08	.07	.3	1.8	-.5	.7	.5
cat2 - bcompañerismo_7	193	193.00	5.78	-.16	.09	.02	.00	.03	-2.4	1.3	.3	.0	.5
cat3 - bmadurez_8	42	42.00	30.14	-.06	.31	.12	-.09	.02	-.4	2.1	.8	-.6	.1
cat4 - bescuela_9	80	80.00	15.35	-.37	-.12	-.17	.22	.03	-3.4	-1.1	-1.6	2.0	.2
cat5 - bconducta_10	68	68.00	18.24	-.04	-.02	-.25	.06	-.06	-.3	-.1	-2.1	.5	-.5
cat6 - binteligencia_11	77	77.00	15.99	.25	.09	.07	-.15	.10	2.3	.8	.7	-1.3	.9
ca94 - botros_12	127	127.00	9.30	-.09	-.05	.03	.02	-.02	-1.1	-.6	.3	.2	-.3
ca97 - bnoc_13	97	97.00	12.48	-.34	-.14	.10	.08	-.04	-3.5	-1.4	1.1	.8	-.4

D.1.2. Segundo análisis de corresponcencia múltiple

La siguiente salida corresponde al segundo ACM, en donde las variables trabajo e instrucción de los padres dejaron de ser activas y pasaron a ser variables suplementarias conjuntamente con las preguntas A y B.

```
-------------------------------------------------------------------------------------------------------------------
      Eigenvalues
-------------------------------------------------------------------------------------------------------------------
```

```
 Accuracy of computation :                trace before diagonalisation       3.5000
 ----------------------------------       sum of eigenvalues                 3.5000

 histogram of the 21 first eigenvalues
 -------------------------------------------

 +--------+------------+----------+----------+-------------------------------------------------------------------------------------------+
 ! number !    Eigen   ! percent. ! cumulat. !                                                                                           !
 !        !    value   !          ! percent. !                                                                                           !
 +--------+------------+----------+----------+-------------------------------------------------------------------------------------------+
 !    1   !    .3355   !    9.58  !    9.58  ! ************************************************************************************** !
 !    2   !    .2878   !    8.22  !   17.81  ! ********************************************************************                   !
 !    3   !    .2768   !    7.91  !   25.72  ! ******************************************************************                     !
 !    4   !    .1904   !    5.44  !   31.16  ! ********************************************                                           !
 !    5   !    .1864   !    5.33  !   36.48  ! *******************************************                                            !
 !    6   !    .1822   !    5.20  !   41.69  ! ******************************************                                             !
 !    7   !    .1790   !    5.11  !   46.80  ! *****************************************                                              !
 !    8   !    .1763   !    5.04  !   51.84  ! *****************************************                                              !
 !    9   !    .1738   !    4.97  !   56.81  ! ****************************************                                               !
 !   10   !    .1687   !    4.82  !   61.62  ! ***************************************                                                !
 !   11   !    .1647   !    4.71  !   66.33  ! **************************************                                                 !
 !   12   !    .1598   !    4.57  !   70.90  ! *************************************                                                  !
 !   13   !    .1571   !    4.49  !   75.39  ! *************************************                                                  !
 !   14   !    .1539   !    4.40  !   79.78  ! ************************************                                                   !
 !   15   !    .1520   !    4.34  !   84.13  ! ************************************                                                   !
 !   16   !    .1473   !    4.21  !   88.34  ! ***********************************                                                    !
 !   17   !    .1416   !    4.05  !   92.38  ! **********************************                                                     !
 !   18   !    .1256   !    3.59  !   95.97  ! *****************************                                                          !
 !   19   !    .0755   !    2.16  !   98.13  ! ******************                                                                     !
 !   20   !    .0497   !    1.42  !   99.55  ! ************                                                                           !
 +--------+------------+----------+----------+-------------------------------------------------------------------------------------------+
1-------------------------------------------------------------------------------------------------------------------------------------------
      coordinates, contributions and squared cosine of active categories on axes  1 to  5
 -------------------------------------------------------------------------------------------------------------------------------------------

 +----------------------------------------+-----------------------------+----------------------+-------------------------+
 !               categories               !          coordinates        !     contributions    !     squared cosine      !
 !----------------------------------------+-----------------------------+----------------------+-------------------------!
 ! iden - libelle             p.rel  disto !   1     2     3     4     5 !  1   2   3   4   5 !  1   2   3   4   5  !
 +----------------------------------------+-----------------------------+----------------------+-------------------------+
 !                                                                                                                         !
 !   1 . grado_V#1                                                                                                         !
 !                                                                                                                         !
 ! cat1 - primero_1            2.80   4.95 !  1.36   .14  -.86   .09  -.08 ! 15.4   .2  7.4   .1   .1 !  .37  .00  .15  .00  .00 !
 ! cat2 - segundo_2            2.89   4.76 !   .24   .74   .76  -.23  -.82 !   .5  5.5  6.1   .8 10.6 !  .01  .11  .12  .01  .14 !
 ! cat3 - tercero_3            2.65   5.29 !  -.44  1.04   .86  -.29   .75 !  1.6 10.0  7.1  1.2  7.9 !  .04  .21  .14  .02  .11 !
 ! cat4 - cuarto_4             3.26   4.11 !  -.72  -.25   .06  1.04   .03 !  5.0   .7   .0 18.6   .0 !  .12  .02  .00  .26  .00 !
 ! cat5 - quinto_5             2.48   5.71 !  -.75 -1.27  -.35  -.41  -.20 !  4.2 13.9  1.1  2.2   .6 !  .10  .28  .02  .03  .01 !
 ! cat6 - sexto_6                .98  15.99 !  -.92  -.99  -.67 -1.31   .59 !  2.5  3.4  1.6  8.9  1.8 !  .05  .06  .03  .11  .02 !
 ! cat7 - preescolar_7         1.07  14.57 !  1.61  -.12  -.81   .04   .73 !  8.3   .0  2.5   .0  3.0 !  .18  .00  .05  .00  .04 !
 ! catb - grsd_8                .52  30.90 !   .09  -.39   .21   .14  -.58 !   .0   .3   .1   .1  1.0 !  .00  .00  .00  .00  .01 !
 +----------------------------------------+-------- cumul.  contribution = 37.5 34.0 25.9 31.9 25.0 +-------------------------+
 !                                                                                                                         !
 !   6 . este_V#6                                                                                                          !
 !                                                                                                                         !
 ! cat1 - papa_1               2.34   6.11 !   .05  -.44   .23   .69   .55 !   .0  1.6   .5  5.9  3.8 !  .00  .03  .01  .08  .05 !
 ! cat2 - mama_2              12.91    .29 !   .03   .06   .00  -.11   .04 !   .0   .2   .0   .8   .1 !  .00  .01  .00  .04  .01 !
 ! cat3 - otro_3                .98  15.99 !  -.46   .41  -.50  1.01 -1.65 !   .6   .6   .9  5.3 14.3 !  .01  .01  .02  .06  .17 !
 ! cat4 - entrev_4              .06 260.60 !  1.42  -.09  -.92 -1.31 -6.50 !   .4   .0   .2   .6 14.4 !  .01  .00  .00  .01  .16 !
 ! cat9 - ignorado_5            .37  44.10 !  -.47  -.32  -.08 -3.13   .50 !   .2   .1   .0 19.0   .5 !  .01  .00  .00  .22  .01 !
 +----------------------------------------+-------- cumul.  contribution =  1.3  2.5  1.6 31.5 33.1 +-------------------------+
 !                                                                                                                         !
 !   8 . repano_V#8                                                                                                        !
 !                                                                                                                         !
 ! cat1 - resi_1               3.13   4.32 !  -.95   .56  -.44   .33  -.30 !  8.5  3.4  2.2  1.8  1.5 !  .21  .07  .05  .03  .02 !
 ! cat2 - reno_2              13.44    .24 !   .21  -.13   .12  -.10   .05 !  1.8   .8   .7   .6   .2 !  .18  .07  .06  .04  .01 !
 ! catb - resd_3                .09 185.86 !  1.97   .32 -1.86  2.69  2.67 !  1.0   .0  1.1  3.4  3.4 !  .02  .00  .02  .04  .04 !
 +----------------------------------------+-------- cumul.  contribution = 11.3  4.3  4.0  5.9  5.1 +-------------------------+
 !                                                                                                                         !
 !  14 . saludabl_V#14                                                                                                     !
 !                                                                                                                         !
 ! cat1 - salsi_1             13.30    .25 !   .13  -.27   .28   .05  -.01 !   .7  3.4  3.7   .2   .0 !  .07  .29  .31  .01  .00 !
 ! cat2 - salno_2               3.36   3.95 !  -.52  1.07 -1.10  -.19   .05 !  2.7 13.4 14.7   .6   .0 !  .07  .29  .31  .01  .00 !
 +----------------------------------------+-------- cumul.  contribution =  3.4 16.8 18.5   .8   .1 +-------------------------+
 !                                                                                                                         !
 !  16 . sindglo1rec_V#16                                                                                                  !
```

```
!                                                                                                                           !
! cat1 - nopatologico_1       12.07     .38 !   .18  -.25   .29   .10   .03 !  1.2  2.6  3.7   .7   .1 !  .09  .16  .22  .03  .00 !
! cat2 - zonariesgo_2          1.90    7.78 !  -.24  -.05  -.09  -.68  -.33 !   .3   .0   .1  4.7  1.1 !  .01  .00  .00  .06  .01 !
! cat3 - patologico_3          2.70    5.17 !  -.65  1.15 -1.23   .01   .10 !  3.5 12.4 14.8   .0   .1 !  .08  .25  .29  .00  .00 !
+-----------------------------------------+---------- cumul.  contribution =  5.0 15.0 18.5  5.4  1.3 +--------------------------+
!                                                                                                                           !
!  17 . sexoedadrec_V#17                                                                                                    !
!                                                                                                                           !
! cat1 - m57_1                 2.40    5.96 !  1.37   .10  -.75  -.01  1.03 ! 13.5   .1  4.9   .0 13.5 !  .32  .00  .09  .00  .18 !
! cat2 - m89_2                 2.62    5.35 !  -.30  1.02   .82  -.69   .22 !   .7  9.5  6.3  6.5   .7 !  .02  .19  .12  .09  .01 !
! cat3 - m1013_3               3.17    4.25 !  -.98  -.53  -.51   .69   .29 !  9.1  3.1  3.0  8.0  1.5 !  .23  .07  .06  .11  .02 !
! cat4 - f57_4                 2.51    5.64 !  1.30   .06  -.49   .01 -1.17 ! 12.7   .0  2.2   .0 18.3 !  .30  .00  .04  .00  .24 !
! cat5 - f89_5                 2.85    4.84 !  -.10   .61  1.18   .52  -.05 !   .1  3.7 14.3  4.1   .0 !  .00  .08  .29  .06  .00 !
! cat6 - f1013_6               3.11    4.36 !  -.77 -1.01  -.27  -.61  -.29 !  5.5 11.0   .8  6.1  1.4 !  .14  .23  .02  .09  .02 !
+-----------------------------------------+---------- cumul.  contribution = 41.5 27.4 31.5 24.6 35.4 +--------------------------+
1--------------------------------------------------------------------------------------------------------------------------------
      coordinates and test values, of categories on axes       1 a  5
 --------------------------------------------------------------------------------------------------------------------------------

+-------------------------------------------------------+------------------------------------+--------------------------------+
!                    categories                         !              coordinates           !          test-values           !
!-------------------------------------------------------+------------------------------------+--------------------------------!
! iden - title                  eff.     p.abs    disto !   1      2      3      4      5    !  1     2     3     4     5     !
+-------------------------------------------------------+------------------------------------+--------------------------------+
!                                                                                                                             !
!    1 . grado_V#1                                                                                                            !
!                                                                                                                             !
! cat1 - primero_1               220    220.00     4.95 !  1.36    .14   -.86    .09   -.08 !  22.1   2.2 -13.9   1.4  -1.4 !
! cat2 - segundo_2               227    227.00     4.76 !   .24    .74    .76   -.23   -.82 !   4.0  12.2  12.6  -3.8 -13.7 !
! cat3 - tercero_3               208    208.00     5.29 !  -.44   1.04    .86   -.29    .75 !  -7.0  16.4  13.5  -4.6  11.7 !
! cat4 - cuarto_4                256    256.00     4.11 !  -.72   -.25    .06   1.04    .03 ! -12.8  -4.5   1.1  18.6    .5 !
! cat5 - quinto_5                195    195.00     5.71 !  -.75  -1.27   -.35   -.41   -.20 ! -11.4 -19.2  -5.4  -6.2  -3.1 !
! cat6 - sexto_6                  77     77.00    15.99 !  -.92   -.99   -.67  -1.31    .59 !  -8.3  -9.0  -6.0 -11.9   5.4 !
! cat7 - preescolar_7             84     84.00    14.57 !  1.61   -.12   -.81    .04    .73 !  15.3  -1.1  -7.7    .4   6.9 !
! catb - grsd_8                   41     41.00    30.90 !   .09   -.39    .21    .14   -.58 !    .6  -2.5   1.4    .9  -3.8 !
+-------------------------------------------------------+------------------------------------+--------------------------------+
!                                                                                                                             !
!    6 . este_V#6                                                                                                             !
!                                                                                                                             !
! cat1 - papa_1                  184    184.00     6.11 !   .05   -.44    .23    .69    .55 !    .8  -6.5   3.4  10.1   8.0 !
! cat2 - mama_2                 1013   1013.00      .29 !   .03    .06    .00   -.11    .04 !   2.1   3.9    .2  -7.1   2.9 !
! cat3 - otro_3                   77     77.00    15.99 !  -.46    .41   -.50   1.01  -1.65 !  -4.2   3.7  -4.6   9.2 -14.9 !
! cat4 - entrev_4                  5      5.00   260.60 !  1.42   -.09   -.92  -1.31  -6.50 !   3.2   -.2  -2.1  -2.9 -14.6 !
! cat9 - ignorado_5               29     29.00    44.10 !  -.47   -.32   -.08  -3.13    .50 !  -2.6  -1.8   -.4 -17.0   2.7 !
+-------------------------------------------------------+------------------------------------+--------------------------------+
!                                                                                                                             !
!    8 . repano_V#8                                                                                                           !
!                                                                                                                             !
! cat1 - resi_1                  246    246.00     4.32 !  -.95    .56   -.44    .33   -.30 ! -16.6   9.8  -7.7   5.8  -5.1 !
! cat2 - reno_2                 1055   1055.00      .24 !   .21   -.13    .12   -.10    .05 !  15.5  -9.8   8.6  -7.1   3.8 !
! catb - resd_3                    7      7.00   185.86 !  1.97    .32  -1.86   2.69   2.67 !   5.2    .8  -4.9   7.1   7.1 !
+-------------------------------------------------------+------------------------------------+--------------------------------+
!                                                                                                                             !
!   14 . saludabl_V#14                                                                                                        !
!                                                                                                                             !
! cat1 - salsi_1                1044   1044.00      .25 !   .13   -.27    .28    .05   -.01 !   9.5 -19.5  20.0   3.4   -.9 !
! cat2 - salno_2                 264    264.00     3.95 !  -.52   1.07  -1.10   -.19    .05 !  -9.5  19.5 -20.0  -3.4    .9 !
+-------------------------------------------------------+------------------------------------+--------------------------------+
!                                                                                                                             !
!   16 . sindglo1rec_V#16                                                                                                     !
!                                                                                                                             !
! cat1 - nopatologico_1          947    947.00      .38 !   .18   -.25    .29    .10    .03 !  10.7 -14.6  17.0   6.1   1.8 !
! cat2 - zonariesgo_2            149    149.00     7.78 !  -.24   -.05   -.09   -.68   -.33 !  -3.0   -.6  -1.2  -8.9  -4.3 !
! cat3 - patologico_3            212    212.00     5.17 !  -.65   1.15  -1.23    .01    .10 ! -10.4  18.3 -19.6    .2   1.5 !
+-------------------------------------------------------+------------------------------------+--------------------------------+
!                                                                                                                             !
!   17 . sexoedadrec_V#17                                                                                                     !
!                                                                                                                             !
! cat1 - m57_1                   188    188.00     5.96 !  1.37    .10   -.75   -.01   1.03 !  20.3   1.5 -11.1   -.1  15.2 !
! cat2 - m89_2                   206    206.00     5.35 !  -.30   1.02    .82   -.69    .22 !  -4.7  16.0  12.7 -10.8   3.5 !
! cat3 - m1013_3                 249    249.00     4.25 !  -.98   -.53   -.51    .69    .29 ! -17.2  -9.3  -8.9  12.1   5.2 !
! cat4 - f57_4                   197    197.00     5.64 !  1.30    .06   -.49    .01  -1.17 !  19.8    .9  -7.5    .2 -17.8 !
! cat5 - f89_5                   224    224.00     4.84 !  -.10    .61   1.18    .52   -.05 !  -1.6  10.0  19.3   8.6   -.9 !
! cat6 - f1013_6                 244    244.00     4.36 !  -.77  -1.01   -.27   -.61   -.29 ! -13.3 -17.5  -4.7 -10.5  -5.0 !
+-------------------------------------------------------+------------------------------------+--------------------------------+
!                                                                                                                             !
!    4 . trabpadrec_V#4                                                                                                       !
```

```
!                                                                                                          !
! cat1 - pffaa_1             50      50.00     25.16 !   .08  -.10   .11   .14  -.24 !   .6  -.7   .8  1.0 -1.8 !
! cat2 - pprof_2             70      70.00     17.69 !   .13  -.34   .27   .04   .31 !  1.2 -2.9  2.3   .4  2.7 !
! cat3 - ptecn_3             38      38.00     33.42 !   .02  -.18   .42   .13   .08 !   .1 -1.1  2.7   .8   .5 !
! cat4 - pempl_4             74      74.00     16.68 !  -.25  -.06  -.05   .01   .09 ! -2.2  -.6  -.4   .1   .8 !
! cat5 - pvend_5            241     241.00      4.43 !   .11  -.11   .13  -.08   .15 !  1.8 -2.0  2.2 -1.4  2.5 !
! cat6 - pagrop_6            43      43.00     29.42 !   .20  -.25   .01   .34   .17 !  1.3 -1.7   .1  2.3  1.1 !
! cat7 - pope_7             352     352.00      2.72 !   .13   .05  -.04  -.04  -.04 !  2.8  1.1 -1.0  -.9  -.8 !
! cat8 - pncal_8            222     222.00      4.89 !  -.12   .08  -.09   .13  -.13 ! -2.0  1.4 -1.4  2.1 -2.1 !
! cat9 - potros_9            62      62.00     20.10 !  -.13   .23  -.02   .10  -.11 ! -1.1  1.9  -.2   .8  -.9 !
! catb - psindato_10        156     156.00      7.38 !  -.25   .18  -.21  -.20  -.09 ! -3.4  2.4 -2.8 -2.7 -1.2 !
+----------------------------------------------------+---------------------------------+----------------------------+
1
+----------------------------------------------------+---------------------------------+----------------------------+
!                   categories                       !           coordinates           !        test-values         !
!----------------------------------------------------+---------------------------------+----------------------------!
! iden - title             eff.    p.abs     disto !    1     2     3     4     5  !   1    2    3    4    5  !
+----------------------------------------------------+---------------------------------+----------------------------+
!                                                                                                          !
!    5 . trabajomadrec_V#5                                                                                 !
!                                                                                                          !
! cat1 - mprof_1            120     120.00      9.90 !   .21  -.23   .21   .10   .31 !  2.4 -2.7  2.4  1.1  3.6 !
! cat2 - mtecn_2             55      55.00     22.78 !   .00  -.18  -.03   .00   .05 !   .0 -1.3  -.2   .0   .4 !
! cat3 - mempl_3            134     134.00      8.76 !   .03  -.13   .05   .01   .14 !   .4 -1.6   .5   .2  1.7 !
! cat4 - mvende_4           154     154.00      7.49 !   .10   .01   .04  -.12   .12 !  1.3   .1   .5 -1.6  1.6 !
! cat5 - mncal_5            722     722.00       .81 !  -.02   .05  -.03   .00  -.09 ! -1.0  2.0 -1.1  -.2 -3.7 !
! cat6 - motros_6            50      50.00     25.16 !  -.09   .02  -.10   .26  -.07 !  -.6   .1  -.7  1.8  -.5 !
! catb - msindato_7          73      73.00     16.92 !  -.31   .24  -.17  -.06  -.12 ! -2.7  2.1 -1.5  -.5 -1.0 !
+----------------------------------------------------+---------------------------------+----------------------------+
!                                                                                                          !
!   11 . instrucc_V#11                                                                                     !
!                                                                                                          !
! ca10 - pprimaria_1        384     384.00      2.41 !  -.19   .08  -.18   .13  -.10 ! -4.3  1.9 -4.3  2.9 -2.4 !
! ca20 - putu_2             251     251.00      4.21 !   .12   .01   .04  -.11   .09 !  2.1   .2   .7 -1.9  1.6 !
! ca30 - pse1_3             205     205.00      5.38 !   .14  -.01   .16   .03   .00 !  2.2  -.1  2.4   .5   .0 !
! ca40 - pse2_4             147     147.00      7.90 !   .20  -.09   .21   .02   .07 !  2.6 -1.1  2.8   .2   .9 !
! ca50 - pmilitar_5          41      41.00     30.90 !   .00  -.20   .04   .09  -.22 !   .0 -1.3   .3   .6 -1.4 !
! ca60 - pmagisterio_6        6       6.00    217.00 !  -.18  -.27   .90   .03   .23 !  -.5  -.7  2.2   .1   .6 !
! ca70 - puniversitario_7   136     136.00      8.62 !   .08  -.29   .20  -.04   .21 !  1.0 -3.6  2.5  -.5  2.6 !
! ca80 - pnoestudio_8         4       4.00    326.00 !   .54  -.25  -.04  -.55   .83 !  1.1  -.5  -.1 -1.1  1.7 !
! catb - psd_9              134     134.00      8.76 !  -.21   .22  -.27  -.20  -.14 ! -2.6  2.7 -3.3 -2.5 -1.7 !
+----------------------------------------------------+---------------------------------+----------------------------+
!                                                                                                          !
!   12 . instruc1_V#12                                                                                     !
!                                                                                                          !
! ca10 - mprimaria_1        365     365.00      2.58 !  -.14   .08  -.17   .07  -.17 ! -3.2  1.9 -3.8  1.7 -3.8 !
! ca20 - mutu_2             162     162.00      7.07 !   .04   .08   .00  -.05   .05 !   .5  1.1   .0  -.7   .7 !
! ca30 - mse1_3             247     247.00      4.30 !  -.10  -.04   .05  -.04   .13 ! -1.7  -.7   .8  -.7  2.3 !
! ca40 - mse2_4             222     222.00      4.89 !   .19   .00   .16  -.05   .01 !  3.0   .0  2.5  -.9   .1 !
! ca50 - mmilitar_5           7       7.00    185.86 !   .88   .10   .12   .66   .56 !  2.3   .3   .3  1.8  1.5 !
! ca60 - mmagisterio_6       50      50.00     25.16 !   .13  -.23   .09  -.09   .09 !  1.0 -1.7   .6  -.6   .7 !
! ca70 - muniversitario_7   182     182.00      6.19 !   .15  -.19   .18   .03   .16 !  2.2 -2.8  2.6   .4  2.4 !
! ca80 - mnoestudio_8         3       3.00    435.00 !  -.54   .10  -.91   .58 -2.61 !  -.9   .2 -1.6  1.0 -4.5 !
! catb - msd_9               70      70.00     17.69 !  -.15   .18  -.29  -.05  -.16 ! -1.3  1.5 -2.5  -.4 -1.3 !
+----------------------------------------------------+---------------------------------+----------------------------+
!                                                                                                          !
!   18 . pregArec_V#18                                                                                     !
!                                                                                                          !
! cat1 - anada_1            184     184.00      6.11 !   .17  -.19   .27   .05  -.12 !  2.4 -2.7  4.0   .8 -1.7 !
! ca10 - acaracter_2         58      58.00     21.55 !  -.02   .25  -.16  -.08   .15 !  -.1  2.0 -1.2  -.7  1.1 !
! ca13 - ainmadurez_3        36      36.00     35.33 !  -.23  -.11  -.07  -.18   .15 ! -1.4  -.7  -.5 -1.1   .9 !
! ca15 - atimidez_4          73      73.00     16.92 !   .15  -.18   .07  -.14  -.12 !  1.3 -1.5   .6 -1.3 -1.1 !
! ca17 - adistraccion_5      45      45.00     28.07 !  -.21   .29   .21  -.06   .08 ! -1.4  2.0  1.4  -.4   .5 !
! cat2 - afuturo_6          105     105.00     11.46 !   .03  -.38   .22   .04   .08 !   .4 -4.1  2.3   .4   .8 !
! ca22 - asentimientos_7     47      47.00     26.83 !   .00   .18  -.13   .18   .02 !   .0  1.3  -.9  1.3   .1 !
! ca24 - aresponsabilidad_8  29      29.00     44.10 !   .08  -.15   .21  -.24   .00 !   .4  -.8  1.1 -1.3   .0 !
! cat4 - aconducta_9         59      59.00     21.17 !  -.12   .56  -.32  -.01  -.12 !  -.9  4.4 -2.5  -.1  -.9 !
! cat5 - asalud_10           72      72.00     17.17 !   .01   .18  -.13   .09   .05 !   .1  1.5 -1.1   .8   .4 !
! cat6 - aviolencia_11       42      42.00     30.14 !  -.23   .43  -.22  -.30   .13 ! -1.5  2.8 -1.4 -2.0   .8 !
! cat7 - aeducacion_12      150     150.00      7.72 !  -.19   .17  -.15   .13  -.02 ! -2.4  2.2 -1.9  1.6  -.2 !
! cat8 - ainquietud_13       77      77.00     15.99 !   .11   .15  -.25  -.13   .04 !  1.0  1.3 -2.3 -1.2   .4 !
! ca94 - aotros_14          124     124.00      9.55 !  -.05   .04  -.18  -.03  -.03 !  -.5   .5 -2.1  -.3  -.4 !
! ca97 - anoc_15            207     207.00      5.32 !   .06  -.22   .12   .05   .03 !   .9 -3.5  1.9   .8   .5 !
+----------------------------------------------------+---------------------------------+----------------------------+
!                                                                                                          !
!   19 . pregBrec_V#19                                                                                     !
!                                                                                                          !
```

```
! cat1 - bcariño_1            135    135.00      8.69 !    .19    .31   -.16   -.02    .09 !   2.4   3.9  -1.9   -.2   1.1 !
! ca11 - bsolidaridad_2        88     88.00     13.86 !   -.02   -.02    .11    .16   -.23 !   -.2   -.2   1.1   1.6  -2.3 !
! ca14 - bsentimientos_3      119    119.00      9.99 !   -.19   -.17   -.04   -.06    .38 !  -2.1  -2.0   -.5   -.7   4.3 !
! ca17 - bbondad_4            109    109.00     11.00 !   -.03    .13    .05    .04   -.16 !   -.4   1.4    .6    .5  -1.7 !
! ca18 - btenacidad_5         107    107.00     11.22 !   -.03   -.20    .03   -.02   -.04 !   -.3  -2.1    .3   -.2   -.4 !
! ca19 - bpersonalidad_6       66     66.00     18.82 !    .18   -.12    .07    .23   -.09 !   1.5  -1.0    .6   1.9   -.7 !
! cat2 - bcompañerismo_7      193    193.00      5.78 !    .04    .02   -.02    .02   -.10 !    .6    .4   -.3    .3  -1.6 !
! cat3 - bmadurez_8            42     42.00     30.14 !    .21    .14   -.09   -.04    .09 !   1.4    .9   -.6   -.3    .6 !
! cat4 - bescuela_9            80     80.00     15.35 !   -.18   -.18    .00   -.03   -.09 !  -1.6  -1.6    .0   -.3   -.8 !
! cat5 - bconducta_10          68     68.00     18.24 !   -.12   -.01    .02    .06   -.09 !  -1.1   -.1    .2    .5   -.8 !
! cat6 - binteligencia_11      77     77.00     15.99 !    .25   -.11    .27   -.05    .09 !   2.2  -1.0   2.4   -.4    .8 !
! ca94 - botros_12            127    127.00      9.30 !   -.05    .02   -.01   -.13    .01 !   -.6    .3   -.1  -1.5    .2 !
! ca97 - bnoc_13               97     97.00     12.48 !   -.15    .05   -.09   -.06    .12 !  -1.5    .6  -1.0   -.6   1.3 !
+-----------------------------------------------------+-----------------------------------+-------------------------------+
```

D.1.3. Correspondencias Simples

Parte de la salida del cruzamiento de las variables *Sindrome Global* con *Saludable*

```
-------------------------------------------------------------------------------------------------------------------------------
      Eigenvalues
-------------------------------------------------------------------------------------------------------------------------------

 Accuracy of computation :               trace before diagonalisation       1.5000
 ----------------------------------      sum of eigenvalues                 1.5000

 histogram of the  3 first eigenvalues
 ---------------------------------------

 +--------+------------+----------+----------+-----------------------------------------------------------------------------+
 ! number !    Eigen   ! percent. ! cumulat. !                                                                             !
 !        !    value   !          ! percent. !                                                                             !
 +--------+------------+----------+----------+-----------------------------------------------------------------------------+
 !    1   !    .7675   !   51.16   !   51.16  ! ********************************************************************************* !
 !    2   !    .5000   !   33.33   !   84.50  ! **************************************************                              !
 !    3   !    .2325   !   15.50   !  100.00  ! ************************                                                     !
 +--------+------------+----------+----------+-----------------------------------------------------------------------------+
1-------------------------------------------------------------------------------------------------------------------------------
      coordinates, contributions and squared cosine of active categories on axes  1 to  3
 -------------------------------------------------------------------------------------------------------------------------------

 +-----------------------------------------+-----------------------------+-------------------------+-------------------------+
 !               categories                !         coordinates         !      contributions      !      squared cosine     !
 !-----------------------------------------+-----------------------------+-------------------------+-------------------------!
 ! iden - libelle             p.rel  disto !   1     2     3     0     0 !   1    2    3    0    0 !   1    2    3    0    0 !
 +-----------------------------------------+-----------------------------+-------------------------+-------------------------+
 !                                                                                                                             !
 !  14 . saludabl_V#14                                                                                                         !
 !                                                                                                                             !
 ! cat1 - salsi_1             39.91    .25 !  -.44   .00   .24   .00   .00 ! 10.1   .0 10.1   .0   .0 !  .77  .00  .23  .00  .00 !
 ! cat2 - salno_2             10.09   3.95 !  1.74   .00  -.96   .00   .00 ! 39.9   .0 39.9   .0   .0 !  .77  .00  .23  .00  .00 !
 +-----------------------------------------+-------- cumul.  contribution = 50.0   .0 50.0   .0   .0 +-------------------------+
 !                                                                                                                             !
 !  16 . sindglo1rec_V#16                                                                                                      !
 !                                                                                                                             !
 ! cat1 - nopatologico_1      36.20    .38 !  -.44  -.36  -.24   .00   .00 !  9.1  9.3  9.1   .0   .0 !  .51  .34  .15  .00  .00 !
 ! cat2 - zonariesgo_2         5.70   7.78 !   .00  2.79   .00   .00   .00 !   .0 88.6   .0   .0   .0 !  .00 1.00  .00  .00  .00 !
 ! cat3 - patologico_3         8.10   5.17 !  1.97  -.36  1.08   .00   .00 ! 40.9  2.1 40.9   .0   .0 !  .75  .02  .23  .00  .00 !
 +-----------------------------------------+-------- cumul.  contribution = 50.0100.0 50.0   .0   .0 +-------------------------+
1-------------------------------------------------------------------------------------------------------------------------------
      coordinates and test values, of categories on axes       1 a  3
 -------------------------------------------------------------------------------------------------------------------------------

 +-----------------------------------------------------+-----------------------------------+-------------------------------+
 !                    categories                       !            coordinates            !          test-values          !
 !-----------------------------------------------------+-----------------------------------+-------------------------------!
 ! iden - title                eff.    p.abs     disto !    1     2     3     0     0   !   1     2     3     0     0   !
 +-----------------------------------------------------+-----------------------------------+-------------------------------+
 !                                                                                                                           !
 !   14 . saludabl_V#14                                                                                                      !
 !                                                                                                                           !
 ! cat1 - salsi_1             1044   1044.00       .25 !  -.44   .00   .24   .00   .00 ! -31.7   .0  17.4   .0   .0 !
 ! cat2 - salno_2              264    264.00      3.95 !  1.74   .00  -.96   .00   .00 !  31.7   .0 -17.4   .0   .0 !
```

16 . sindglo1rec_V#16													
cat1 - nopatologico_1	947	947.00	.38	-.44	-.36	-.24	.00	.00	-25.8	-21.0	-14.2	.0	.0
cat2 - zonariesgo_2	149	149.00	7.78	.00	2.79	.00	.00	.00	.0	36.2	.0	.0	.0
cat3 - patologico_3	212	212.00	5.17	1.97	-.36	1.08	.00	.00	31.3	-5.7	17.2	.0	.0
18 . pregArec_V#18													
cat1 - anada_1	184	184.00	6.11	-.39	-.26	-.01	.00	.00	-5.8	-3.7	-.2	.0	.0
ca10 - acaracter_2	58	58.00	21.55	.39	.08	.15	.00	.00	3.1	.6	1.2	.0	.0
ca13 - ainmadurez_3	36	36.00	35.33	.13	.08	-.02	.00	.00	.8	.5	-.1	.0	.0
ca15 - atimidez_4	73	73.00	16.92	-.19	.03	-.02	.00	.00	-1.7	.3	-.1	.0	.0
ca17 - adistraccion_5	45	45.00	28.07	.07	.48	.25	.00	.00	.5	3.3	1.7	.0	.0
cat2 - afuturo_6	105	105.00	11.46	-.39	-.06	-.07	.00	.00	-4.2	-.6	-.7	.0	.0
ca22 - asentimientos_7	47	47.00	26.83	.25	-.09	.29	.00	.00	1.7	-.6	2.0	.0	.0
ca24 - aresponsabilidad_8	29	29.00	44.10	-.16	-.25	-.13	.00	.00	-.9	-1.4	-.7	.0	.0
cat4 - aconducta_9	59	59.00	21.17	.63	.12	.09	.00	.00	5.0	1.0	.7	.0	.0
cat5 - asalud_10	72	72.00	17.17	.27	.21	-.40	.00	.00	2.4	1.8	-3.5	.0	.0
cat6 - aviolencia_11	42	42.00	30.14	.45	.32	.13	.00	.00	2.9	2.1	.9	.0	.0
cat7 - aeducacion_12	150	150.00	7.72	.22	.02	-.07	.00	.00	2.9	.3	-.9	.0	.0
cat8 - ainquietud_13	77	77.00	15.99	.27	.17	-.02	.00	.00	2.4	1.6	-.2	.0	.0
ca94 - aotros_14	124	124.00	9.55	.17	.02	.06	.00	.00	2.0	.3	.7	.0	.0
ca97 - anoc_15	207	207.00	5.32	-.29	-.10	.01	.00	.00	-4.6	-1.6	.2	.0	.0
19 . pregBrec_V#19													
cat1 - bcariño_1	135	135.00	8.69	.22	.08	-.08	.00	.00	2.8	1.0	-1.0	.0	.0
ca11 - bsolidaridad_2	88	88.00	13.86	.05	-.18	.02	.00	.00	.5	-1.7	.2	.0	.0
ca14 - bsentimientos_3	119	119.00	9.99	.09	-.01	-.05	.00	.00	1.1	-.2	-.5	.0	.0
ca17 - bbondad_4	109	109.00	11.00	-.02	-.10	-.03	.00	.00	-.2	-1.1	-.4	.0	.0
ca18 - btenacidad_5	107	107.00	11.22	-.12	-.01	.10	.00	.00	-1.3	-.1	1.0	.0	.0
ca19 - bpersonalidad_6	66	66.00	18.82	-.19	-.07	-.08	.00	.00	-1.6	-.6	-.7	.0	.0
cat2 - bcompañerismo_7	193	193.00	5.78	.00	.08	-.03	.00	.00	-.1	1.2	-.5	.0	.0
cat3 - bmadurez_8	42	42.00	30.14	.03	.02	.24	.00	.00	.2	.1	1.6	.0	.0
cat4 - bescuela_9	80	80.00	15.35	-.07	-.08	-.05	.00	.00	-.6	-.8	-.4	.0	.0
cat5 - bconducta_10	68	68.00	18.24	.03	.01	-.05	.00	.00	.2	.1	-.4	.0	.0
cat6 - binteligencia_11	77	77.00	15.99	-.25	-.20	.13	.00	.00	-2.2	-1.8	1.1	.0	.0
ca94 - botros_12	127	127.00	9.30	-.01	.16	.00	.00	.00	-.1	1.9	.0	.0	.0
ca97 - bnoc_13	97	97.00	12.48	.05	.06	.07	.00	.00	.5	.6	.7	.0	.0

D.1.4. Clusters

Coordenadas de los tres clusters formados con los primeros 5 ejes factoriales de la correspondencia múltiple.

```
-------------------------------------------------------------------------------------------------------------------------------
     consolidating the partition around the  3 clusters centroids
     through  10 k-means-like iterations
-------------------------------------------------------------------------------------------------------------------------------

  Increase  of inter-classes variance
--------------------------------------

                          +-----------+----------+----------+----------+
                          ! iteration ! v.total  ! v.inter ! ratio    !
                          +-----------+----------+----------+----------+
                          !     0     ! 1.276871 !  .501722 !   .3929  !
                          !     1     ! 1.276871 !  .533094 !   .4175  !
                          !     2     ! 1.276871 !  .545017 !   .4268  !
                          !     3     ! 1.276871 !  .545091 !   .4269  !
                          !     4     ! 1.276871 !  .545091 !   .4269  !
                          +-----------+----------+----------+----------+
             stop  after  iteration   4 : Increase of inter-classes variance
             during the previous iteraion is only   .000 %.
```

decomposition of variance computed on 5 axes

	variances before	variances after	bases before	bases after	weight before	weight after	distances before	distances after
variance inter-classes	.5017	.5451						
variance intra-class								
class 1 / 3	.2207	.2221	380	382	380.00	382.00	.7104	.7181
class 2 / 3	.4087	.2559	540	468	540.00	468.00	.3168	.4371
class 3 / 3	.1458	.2538	388	458	388.00	458.00	.5547	.5112
total variance	1.2769	1.2769						

ratio (variance inter/total variance) : before... .3929
after ... .4269

coordinates and test-values on axes 1 to 5

classes: iden - name	effec.	abs.w	disto	coordinates 1	2	3	4	5	test-values 1	2	3	4	5
cut a of the tree into 3 classes													
aa1a- class 1 / 3	382	382.00	.72	1.35	.07	-.62	.03	-.09	31.25	1.71	-14.46	.61	-2.00
aa2a- class 2 / 3	468	468.00	.44	-.28	.83	.88	-.04	.09	-7.51	22.29	23.66	-1.09	2.51
aa3a- class 3 / 3	458	458.00	.51	-.84	-.91	-.38	.02	-.02	-22.25	-24.03	-9.99	.51	-.62

D.2. Análisis estadístico de textos

D.2.1. Tablas léxicas

Aquí se presenta parte de una tabla léxica agregada como ejemplo.

table Words - Texts

		a01a	a02a	a03a	a04a	a05a	a06a	a07a	a08a	a09a	a10a	a11a	a12a	a13a	a14a	a15a
a	i	322.	2.	0.	10.	12.	76.	5.	0.	0.	0.	0.	0.	0.	0.	0.
actitud	i	6.	0.	0.	0.	0.	7.	0.	0.	0.	0.	0.	0.	0.	0.	0.
activa	i	9.	0.	0.	0.	2.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.
actividades	i	6.	0.	0.	0.	0.	5.	0.	0.	0.	0.	0.	0.	0.	0.	0.
adelante	i	19.	0.	0.	0.	0.	3.	0.	0.	0.	0.	0.	0.	0.	0.	0.
además	i	8.	0.	0.	1.	0.	1.	1.	0.	0.	0.	0.	0.	0.	0.	0.
adolescencia	i	3.	0.	0.	0.	0.	7.	1.	0.	0.	0.	0.	0.	0.	0.	0.
agresividad	i	9.	0.	0.	0.	1.	3.	0.	0.	0.	0.	0.	0.	0.	0.	0.
ahora	i	19.	0.	0.	0.	0.	1.	1.	0.	2.	0.	0.	0.	0.	0.	0.
al	i	45.	1.	0.	2.	2.	7.	2.	0.	0.	0.	0.	0.	0.	0.	0.
alegre	i	17.	0.	0.	5.	7.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.
alegría	i	1.	0.	0.	0.	0.	7.	6.	0.	0.	0.	0.	0.	0.	0.	0.
algo	i	43.	0.	0.	0.	0.	3.	0.	0.	0.	0.	0.	0.	0.	0.	0.
alguien	i	9.	0.	0.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
algunas	i	6.	0.	0.	0.	0.	5.	1.	0.	0.	0.	0.	0.	0.	0.	0.
alumna	i	4.	0.	0.	0.	6.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
amable	i	7.	0.	0.	3.	2.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.
amiga	i	4.	0.	0.	0.	9.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
amigo	i	11.	0.	0.	6.	0.	2.	0.	0.	0.	0.	0.	0.	0.	0.	0.
amigos	i	34.	0.	0.	2.	1.	7.	0.	0.	0.	0.	0.	0.	0.	0.	0.
amistad	i	5.	0.	0.	0.	0.	6.	0.	0.	0.	0.	0.	0.	0.	0.	0.
amor	i	4.	0.	0.	0.	0.	7.	3.	0.	0.	0.	0.	0.	0.	0.	0.

```
aprende              i  12.   0.   0.   1.   2.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
aprender             i  17.   1.   0.   3.   1.  19.   1.   0.   0.   0.   0.   0.   0.   0.   0.
aprendizaje          i   8.   0.   0.   3.   0.   8.   0.   7.   0.   0.   0.   0.   0.   0.   0.
así                  i  11.   0.   0.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
atención             i  18.   0.   0.   1.   0.   5.   0.   0.   0.   0.   0.   0.   0.   0.   0.
atento               i  10.   1.   0.   4.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
aunque               i  15.   0.   0.   0.   0.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
ayuda                i  17.   0.   0.   0.   1.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
ayudar               i  17.   0.   0.   2.   2.   6.   0.   0.   0.   0.   0.   0.   0.   0.   0.
año                  i  15.   0.   0.   0.   0.   3.   0.   0.   0.   0.   0.   0.   0.   0.   0.
años                 i  30.   0.   0.   1.   0.   3.   0.   0.   0.   0.   0.   0.   0.   0.   0.
bastante             i   8.   0.   0.   3.   2.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
bien                 i  53.   1.   0.   7.   5.   2.   0.   0.   0.   0.   0.   0.   0.   0.   0.
bondad               i   0.   0.   0.   1.   0.   9.   0.   0.   0.   0.   0.   0.   0.   0.   0.
buen                 i  44.   1.   0.  49.   3.  17.   3.   1.   2.   0.   0.   0.   0.   0.   0.
buena                i  68.   0.   0.   2.  71.  18.   1.   0.   0.   0.   0.   0.   0.   0.   0.
bueno                i  27.   0.   0.  16.   1.   3.   0.   0.   0.   0.   0.   0.   0.   0.   0.
buenos               i   9.   0.   0.   4.   1.   6.   0.   0.   0.   0.   0.   0.   0.   0.   0.
cada                 i   8.   0.   0.   0.   0.   8.   0.   0.   0.   0.   0.   0.   0.   0.   0.
cambios              i   6.   0.   0.   0.   0.   4.   0.   0.   0.   0.   0.   0.   0.   0.   0.
capacidad            i   9.   1.   0.   0.   0.  33.   3.   0.   0.   1.   0.   0.   0.   0.   0.
capaz                i  16.   0.   0.   4.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
cariño               i   7.   1.   0.   0.   0.   7.   2.   0.   0.   0.   0.   0.   0.   0.   0.
cariñosa             i  49.   0.   0.   0.  34.   4.   0.   0.   0.   0.   0.   0.   0.   0.   0.
cariñoso             i  79.   1.   0.  52.   0.   2.   0.   0.   2.   0.   0.   0.   0.   0.   0.
carácter             i  23.   0.   0.  10.   3.  14.   6.   0.   1.   0.   0.   0.   0.   0.   0.
casa                 i  38.   0.   0.   5.   1.   9.   0.   0.   0.   0.   0.   0.   0.   0.   0.
clase                i  24.   0.   0.   0.   3.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
como                 i  61.   1.   0.   1.   7.  18.   1.   0.   0.   0.   0.   0.   0.   0.   0.
comparte             i   9.   0.   0.   2.   2.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.
compartir            i  19.   0.   0.   0.   3.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
compañera            i  50.   0.   0.   0.  77.   3.   0.   0.   0.   0.   0.   0.   0.   0.   0.
compañerismo         i   3.   1.   0.   0.   0.  22.  19.   0.   0.   1.   0.   0.   0.   0.   0.
compañero            i  33.   0.   0.  72.   1.   1.   1.   1.   3.   0.   0.   0.   0.   0.   0.
compañeros           i  24.   0.   0.   0.   1.   4.   1.   0.   0.   0.   0.   0.   0.   0.   0.
comportamiento       i  17.   0.   0.   4.   1.  12.   6.   0.   0.   0.   0.   0.   0.   0.   0.
con                  i 299.   1.   0.  22.  21.  64.   5.   0.   0.   0.   0.   0.   0.   0.   0.
conducta             i  18.   0.   0.   7.   4.   9.   2.   0.   0.   0.   0.   0.   0.   0.   0.
conmigo              i   5.   0.   0.   2.   5.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
corazón              i  18.   1.   0.  11.   2.  11.   0.   0.   0.   0.   0.   0.   0.   0.   0.
cosas                i  82.   1.   0.   3.   2.  12.   0.   0.   0.   0.   0.   0.   0.   0.   0.
cualquier            i   6.   0.   0.   3.   0.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
cuando               i  88.   0.   0.   5.   1.  10.   0.   0.   0.   0.   0.   0.   0.   0.   0.
cuesta               i  20.   0.   0.   1.   4.   2.   0.   0.   0.   0.   0.   0.   0.   0.   0.
da                   i   9.   0.   0.   1.   0.   1.   0.   0.   0.   0.   0.   0.   0.   0.   0.
de                   i 386.   5.   0.  40.  19. 176.   4.   4.   0.   0.   0.   0.   0.   0.   0.
dedicación           i   3.   0.   0.   0.   0.   1.   8.   0.   0.   0.   0.   0.   0.   0.   0.
del                  i  15.   0.   0.   0.   1.   4.   0.   0.   0.   0.   0.   0.   0.   0.   0.
demasiado            i  20.   0.   0.   1.   4.   2.   1.   0.   0.   0.   0.   0.   0.   0.   0.
demuestra            i  12.   0.   0.   0.   0.   2.   0.   0.   0.   0.   0.   0.   0.   0.   0.
-------------------------------------------------------------------------------------------------------------------------------------------
```

Análisis de cluster

Se presentan las coordenadas de los 15 clusters formados a partir del análisis de correspondencias sobre la tabla léxica, considerando el corpus total, es decir las preguntas A y B.

```
-------------------------------------------------------------------------------------------------------------------------------------------
     consolidating the partition around the 15 clusters centroids
     through  10 k-means-like iterations
-------------------------------------------------------------------------------------------------------------------------------------------

  Increase  of inter-classes variance
 ---------------------------------------

                          +-----------+----------+----------+----------+
                          ! iteration ! v.total  !  v.inter ! ratio   !
                          +-----------+----------+----------+----------+
                          !     0     ! 5.855031 ! 4.702869 !   .8032  !
                          !     1     ! 5.855031 ! 4.808125 !   .8212  !
                          !     2     ! 5.855031 ! 4.811449 !   .8218  !
```

```
                    !     3     ! 5.855031 ! 4.812643 !   .8220  !
                    +-----------+----------+----------+----------+
               stop  after  iteration   3 : Increase of inter-classes variance
               during the previous iteraion is only   .025 %.


decomposition of variance computed on    7 axes
-------------------------------------------------

                +------------------------+-----------------+-------------+-------------------+-----------------+
                !                        !    variances    !   bases     !       weight      !    distances    !
                !                        ! before   after  ! before after!  before    after  ! before   after  !
                +------------------------+-----------------+-------------+-------------------+-----------------+
                !                        !                 !             !                   !                 !
                ! variance inter-classes ! 4.7029   4.8126 !             !                   !                 !
                !                        !                 !             !                   !                 !
                ! variance intra-class   !                 !             !                   !                 !
                !                        !                 !             !                   !                 !
                !        class   1 / 15  !  .3542    .2409 !  782    682 !   782.00   682.00 !  .1233    .1721 !
                !        class   2 / 15  !  .0218    .0188 !   11     10 !    11.00    10.00 ! 8.9889   9.9446 !
                !        class   3 / 15  !  .0000    .0000 !    1      1 !     1.00     1.00 !******* 152.9839 !
                !        class   4 / 15  !  .0824    .1355 !  103    153 !   103.00   153.00 ! 1.5836   1.2571 !
                !        class   5 / 15  !  .0975    .1283 !  112    156 !   112.00   156.00 ! 2.2485   1.7294 !
                !        class   6 / 15  !  .2167    .1810 !  196    200 !   196.00   200.00 !  .6438    .6624 !
                !        class   7 / 15  !  .0911    .0936 !   53     59 !    53.00    59.00 ! 4.6675   4.5281 !
                !        class   8 / 15  !  .0891    .0792 !    8      7 !     8.00     7.00 !40.5026  46.5820 !
                !        class   9 / 15  !  .1579    .0927 !   29     26 !    29.00    26.00 !12.9449  14.2279 !
                !        class  10 / 15  !  .0233    .0233 !    5      5 !     5.00     5.00 !******* 112.0761 !
                !        class  11 / 15  !  .0000    .0000 !    1      1 !     1.00     1.00 !******* 772.3338 !
                !        class  12 / 15  !  .0102    .0102 !    3      3 !     3.00     3.00 !******* 184.8067 !
                !        class  13 / 15  !  .0000    .0000 !    1      1 !     1.00     1.00 !******* 907.4863 !
                !        class  14 / 15  !  .0079    .0390 !    2      3 !     2.00     3.00 !******* 215.3181 !
                !        class  15 / 15  !  .0000    .0000 !    1      1 !     1.00     1.00 !******* 926.7738 !
                !                        !                 !             !                   !                 !
                ! total variance         ! 5.8550   5.8550 !             !                   !                 !
                +------------------------+-----------------+-------------+-------------------+-----------------+

 ratio   (variance inter/total variance) :  before...  .8032
-----------------------------------------   after ...  .8220

coordinates and test-values on axes 1 to   5
-------------------------------------------------
+-------------------------------------------------+-----------------------------------+-----------------------------------+
!                  classes                        !            coordinates            !            test-values            !
+-------------------------------------------------+-----------------------------------+-----------------------------------+
! iden - name               effec.    abs.w  disto  !   1     2     3     4     5     !   1      2      3      4      5     !
+-------------------------------------------------+-----------------------------------+-----------------------------------+
!                                                                                                                         !
!     cut a of the tree into 15 classes                                                                                   !
!                                                                                                                         !
! a01a- class   1 / 15         682    682.00    .17  !  -.15   -.34    .12    .07    .02  ! -11.14 -27.19  10.27   5.85   1.85  !
! a02a- class   2 / 15          10     10.00   9.94  !   .36   -.48  -2.15   -.41   2.11  !   2.20  -3.25 -15.07  -2.87  14.91  !
! a03a- class   3 / 15           1      1.00 152.98  ! -2.41   1.49  -9.84  -4.44  -1.65  !  -4.64   3.15 -21.70  -9.84  -3.67  !
! a04a- class   4 / 15         153    153.00   1.26  !  -.42    .12   -.59   -.35   -.70  ! -10.58   3.23 -17.21 -10.07 -20.62  !
! a05a- class   5 / 15         156    156.00   1.73  !  -.75    .83    .46   -.23    .44  ! -19.17  23.32  13.60  -6.84  13.01  !
! a06a- class   6 / 15         200    200.00    .66  !   .69   -.12    .27   -.15   -.06  !  20.53  -4.07   9.16  -5.16  -1.97  !
! a07a- class   7 / 15          59     59.00   4.53  !  1.65    .40    .25   -.46    .07  !  24.98   6.65   4.39  -7.93   1.14  !
! a08a- class   8 / 15           7      7.00  46.58  !   .97    .31  -1.12   -.10  -2.67  !   4.97   1.74  -6.57   -.60 -15.75  !
! a09a- class   9 / 15          26     26.00  14.23  !  -.41   1.59  -1.10   1.99   -.31  !  -4.05  17.36 -12.54  22.73  -3.60  !
! a10a- class  10 / 15           5      5.00 112.08  !  5.18   4.05  -2.18    .34  -1.44  !  22.31  19.23 -10.74   1.71  -7.20  !
! a11a- class  11 / 15           1      1.00 772.33  ! 10.53   9.36  -5.44    .60  -4.67  !  20.27  19.83 -12.00   1.33 -10.41  !
! a12a- class  12 / 15           3      3.00 184.81  !  1.70   -.32  -8.63  -1.78  10.01  !   5.66  -1.19 -32.97  -6.82  38.66  !
! a13a- class  13 / 15           1      1.00 907.49  !  2.72 -2.04 -19.68  -3.51  21.62  !   5.23  -4.31 -43.39  -7.79  48.17  !
! a14a- class  14 / 15           3      3.00 215.32  !  1.35   4.22  -1.18  12.44   -.12  !   4.50  15.50  -4.51  47.81   -.47  !
! a15a- class  15 / 15           1      1.00 926.77  !  3.33   7.12  -1.42  25.15   -.21  !   6.41  15.09  -3.13  55.75   -.46  !
+-------------------------------------------------+-----------------------------------+-----------------------------------+
```

D.2.2. Análisis de Correspondencia Múltiple

Parte de un Glosario ordenado alfabeticamente

```
 Summary of results
-------------------
          total number of responses =    1308
              total number of words =   24540
          number of distinct words  =    2595
          percent.of distinct words =      10.6

selection of words
------------------
               frequency threshold  =       9
                        kept words  =   19828
                 distinct kept word =     299

!-----------------------------------!
!      words (alphabetical order)   !
!-------!---------------------!-----!
!  num. !    used words       ! freq.!
!-------!---------------------!-----!
!     1 ! a                   !  427 !
!     2 ! actitud             !   13 !
!     3 ! activa              !   12 !
!     4 ! actividades         !   11 !
!     5 ! adelante            !   22 !
!     6 ! además              !   11 !
!     7 ! adolescencia        !   11 !
!     8 ! agresividad         !   13 !
!     9 ! ahora               !   23 !
!    10 ! al                  !   59 !
!    11 ! alegre              !   30 !
!    12 ! alegría             !   14 !
!    13 ! algo                !   46 !
!    14 ! alguien             !   10 !
!    15 ! algunas             !   12 !
!    16 ! alumna              !   10 !
!    17 ! amable              !   13 !
!    18 ! amiga               !   13 !
!    19 ! amigo               !   19 !
!    20 ! amigos              !   44 !
!    21 ! amistad             !   11 !
!    22 ! amor                !   14 !
!    23 ! aprende             !   16 !
!    24 ! aprender            !   42 !
!    25 ! aprendizaje         !   26 !
!    26 ! así                 !   12 !
!    27 ! atención            !   24 !
!    28 ! atento              !   15 !
!    29 ! aunque              !   16 !
!    30 ! ayuda               !   19 !
!    31 ! ayudar              !   27 !
!    32 ! año                 !   18 !
!    33 ! años                !   34 !
!    34 ! bastante            !   14 !
!    35 ! bien                !   68 !
!    36 ! bondad              !   10 !
!    37 ! buen                !  120 !
!    38 ! buena               !  160 !
!    39 ! bueno               !   47 !
!    40 ! buenos              !   20 !
!    41 ! cada                !   16 !
!    42 ! cambios             !   10 !
!    43 ! capacidad           !   47 !
!    44 ! capaz               !   21 !
!    45 ! cariño              !   17 !
!    46 ! cariñosa            !   87 !
!    47 ! cariñoso            !  136 !
!    48 ! carácter            !   57 !
!    49 ! casa                !   53 !
!    50 ! clase               !   28 !
!    51 ! como                !   89 !
!    52 ! comparte            !   13 !
!    53 ! compartir           !   23 !
!    54 ! compañera           !  130 !
!    55 ! compañerismo        !   46 !
!    56 ! compañero           !  112 !
!    57 ! compañeros          !   30 !
!    58 ! comportamiento      !   40 !
!    59 ! con                 !  412 !
```

```
!    60 ! conducta            !   40 !
!    61 ! conmigo             !   13 !
```

Parte de un Glosario ordenado por frecuencia

```
!-----------------------------------!
!      words (frequency order)      !
!-------!--------------------!------!
!  num. !    used words      ! freq.!
!-------!--------------------!------!
!   220 ! que                ! 1367 !
!    95 ! es                 ! 1159 !
!   296 ! y                  ! 1084 !
!   180 ! muy                !  697 !
!   263 ! su                 !  682 !
!    68 ! de                 !  634 !
!   153 ! la                 !  592 !
!    93 ! en                 !  496 !
!   189 ! no                 !  437 !
!     1 ! a                  !  427 !
!    59 ! con                !  412 !
!   158 ! lo                 !  402 !
!   238 ! se                 !  356 !
!    90 ! el                 !  327 !
!   155 ! le                 !  311 !
!   160 ! los                !  304 !
!   212 ! por                !  271 !
!   286 ! un                 !  240 !
!   265 ! sus                !  226 !
!   199 ! para               !  217 !
!   275 ! tiene              !  197 !
!   183 ! nada               !  173 !
!   287 ! una                !  172 !
!   169 ! me                 !  171 !
!   214 ! preocupa           !  162 !
!    38 ! buena              !  160 !
!   279 ! todo               !  153 !
!   154 ! las                !  147 !
!    47 ! cariñoso           !  136 !
!    54 ! compañera          !  130 !
!    73 ! demás              !  125 !
!    37 ! buen               !  120 !
!    97 ! escuela            !  118 !
!    56 ! compañero          !  112 !
!   187 ! niño               !  111 !
!    65 ! cuando             !  104 !
!   125 ! gusta              !  103 !
!   231 ! responsable        !  102 !
!    63 ! cosas              !  100 !
!   179 ! mucho              !   98 !
!   249 ! siempre            !   94 !
!   291 ! veces              !   89 !
!    51 ! como               !   89 !
!   186 ! niña               !   88 !
!    46 ! cariñosa           !   87 !
!   128 ! hace               !   83 !
!   193 ! o                  !   81 !
!   109 ! está               !   77 !
!   181 ! más                !   75 !
!   246 ! ser                !   73 !
!   146 ! inteligente        !   72 !
!   222 ! quiere             !   71 !
!    35 ! bien               !   68 !
!   203 ! pero               !   67 !
!   299 ! él                 !   64 !
!    85 ! edad               !   64 !
!   239 ! sea                !   59 !
!    10 ! al                 !   59 !
!    91 ! ella               !   59 !
!   170 ! mejor              !   59 !
!   129 ! hacer              !   58 !
!   242 ! sensible           !   57 !
!    48 ! carácter           !   57 !
```

49	casa	53
119	futuro	51
217	problemas	48
39	bueno	47
43	capacidad	47
280	todos	47
188	niños	47
13	algo	46
55	compañerismo	46
248	si	45
209	poco	45
258	solidaria	45
20	amigos	44
62	corazón	43
171	mi	43
198	padres	42
24	aprender	42
60	conducta	40
105	estudiar	40
58	comportamiento	40
114	familia	37
244	sentimientos	36
216	problema	36
229	respetuoso	35
115	feliz	35
113	falta	34
33	años	34
292	veo	34
218	pueda	34
81	dulce	34
204	persona	32
137	hijo	31
132	hay	31
135	hermanos	31
11	alegre	30
133	hermana	30
57	compañeros	30
197	padre	30

D.2.3. Ejemplos de concordancias

Concordancias de *cariñoso*

```
Concordance of words equivalent with:      cariñoso
--------------------------------------
          frequency of repetition  134                                                                  response

                                                        cariñoso                                      - 0012
                                             que es muy cariñoso y siempre esta dispuesto hacer todo lo que se le pida   - 0015
    él a pesar de sus cambios de actitud es un niño muy cariñoso y respetuoso con las personas       - 0016
                                                 es muy cariñoso                                      - 0034
                                                        cariñoso y demostrativo tanto cuando esta bien como cuando esta  - 0041
                                             que es muy cariñoso y voluntarioso                       - 0049
                                             que es muy cariñoso con los hermanos y los padres        - 0050
                                                        cariñoso                                      - 0057
                                                 es muy cariñoso                                      - 0061
                                              educado y cariñoso                                      - 0068
                                         es un niño muy cariñoso                                      - 0080
                                                   y es cariñoso                                      - 0085
                                                        cariñoso                                      - 0093
       lo mejor es que es un buen compañero y es lo mas cariñoso que hay                              - 0100
                                                     lo cariñoso que es en su casa                    - 0110
                                                    muy cariñoso y servicial                          - 0119
                                                     lo cariñoso                                      - 0129
                                                 es muy cariñoso y le encanta jugar con los compañeros y compartir       - 0131
                                                 que es cariñoso y prolijo                            - 0135
                                                 que es cariñoso                                      - 0141
                                                        cariñoso y que tiene mucha confianza en mi    - 0161
                                             que es muy cariñoso                                      - 0183
                                                    muy cariñoso con quien se lo brinda               - 0199
                                                        cariñoso                                      - 0208
                                                        cariñoso y dispuesto a lo que se le proponga  - 0224
```

Concordancias de *responsable*

```
Concordance of words equivalent with:      responsable
--------------------------------------
          frequency of repetition   88                                                                                     response

                                                   responsable                                                           - 0058
                                               muy responsable y amigable                                                - 0104
                                     es cariñosa y responsable                                                           - 0165
                                            es muy responsable                                                           - 0171
                         es un niño sano e inteligente y responsable                                                     - 0176
                                                   responsable y aprende rápido                                          - 0204
                                                   responsable                                                           - 0218
                                                es responsable                                                           - 0223
                                                   responsable                                                           - 0227
                                                es responsable                                                           - 0230
                                                   responsable                                                           - 0239
                                        que es muy responsable con todo lo referente a la escuela                        - 0266
  es una niña normal qe disfruta todo lo que hace y es responsable y estudiosa                                           - 0281
                                            es muy responsable                                                           - 0289
                                               muy responsable                                                           - 0292
                                                   responsable                                                           - 0310
                                                   responsable                                                           - 0320
                                                   responsable                                                           - 0332
                                            es muy responsable y emprendedora                                            - 0375
 que hace lo emprende con muchas ganas y además es muy responsable                                                       - 0402
                                                   responsable                                                           - 0424
                                                es responsable                                                           - 0431
                   que es una niña muy tranquila y muy responsable por sus estudios                                      - 0433
                                                es responsable en sus deberes y que realmente hace todo sola             - 0490
                                                   responsable y respetuoso                                              - 0500
                                               muy responsable tanto para las tareas de la casa como para la escuela - 0512
                                                   responsable                                                           - 0548
                                                es responsable                                                           - 0557
                                es muy compañera y responsable                                                           - 0563
                                        que es muy responsable y estudiosa                                               - 0565
                                                   responsable                                                           - 0566
                                            es muy responsable tanto en la escuela como en sus cosas personales          - 0573
```

Concordancias de *futuro*

```
Concordance of words equivalent with:      futuro
--------------------------------------
          frequency of repetition   49                                                                                     response

                           me preocupa la conducta para el futuro                                                        - 0006
                                                        su futuro                                                        - 0052
                     que tenga una buena educación para el futuro                                                        - 0077
                                 que logre en la vida un futuro sin tener necesidades                                    - 0083
                                            pienso en el futuro                                                          - 0097
                                                        el futuro                                                        - 0139
                                                        su futuro                                                        - 0144
                                lo mas preocupante es el futuro que les espera                                           - 0150
                                          me preocupa su futuro                                                          - 0169
                                                        el futuro                                                        - 0202
enfrentarán los desafíos que tendrán que enfrentar en el futuro                                                          - 0223
              que tiene muchas ganas de estudiar y no ve futuro en el uruguay                                            - 0232
          en general que tenemos los padres en cuanto al futuro de nuestros hijos                                        - 0233
                                                        su futuro                                                        - 0242
                                               que en el futuro pueda traumarse por tener polidactilia                   - 0246
                                                        su futuro                                                        - 0248
                                                        su futuro sin duda                                               - 0262
                                                        el futuro                                                        - 0285
                                                        su futuro                                                        - 0422
                                          me preocupa su futuro                                                          - 0431
                                                        el futuro                                                        - 0485
                       lo que más me preocupa es en el futuro su adolescencia y en lo laboral                            - 0491
                                                   creo el futuro                                                        - 0496
```

```
                          educación buena para su futuro como persona y como estudiante          - 0500
                                               su futuro por su dificultad auditiva              - 0508
                                               su futuro a nivel educacional                     - 0540
          que pueda tener un buen estudio para en un futuro pueda tener un buen trabajo          - 0559
la presión de grupo a la que se pueda enfrentar en el futuro                                     - 0588
                                               el futuro que le espera por vivir                 - 0663
                                               el futuro                                         - 0664
                                               el futuro que le espera                           - 0665
                                               su futuro                                         - 0673
                                 por ahí pensar en el futuro                                     - 0705
                                    que tenga un buen futuro                                     - 0727
                                               su futuro                                         - 0748
                                               su futuro                                         - 0762
```

Índice alfabético

análisis de clusters, 19
análisis de correspondencia, 16
análisis factorial, 14

concordancia, 31
corpus, 4, 29
Cuasisegmento, 28

discriminación global, 56

elementos característicos, 44
equivalencia distribucional, 18
especificidad negativa, 46
estilometría, 55

Forma Gráfica, 27
formas-polo, 31

gramática, 19
Gramática Generativo-Transformacional, 21

hápax, 4, 30

identificación, 27
Inventarios de Segmentos Repetidos, 31

Lema, 27
ley de Zipf, 4
ley Hipergeométrica, 45
leyes empíricas, 23

palabra característica negativa, 46
perfil medio, 17

respuestas características, 47
respuestas modales, 44, 47

segmentación del texto, 27
Segmento Repetido, 28

Tabla de Segmentos Repetidos, 36
Tabla Léxica, 36
Tabla Léxica Agregada, 36

unidades léxicas, 29
unidades léxicas características, 44

Índice de figuras

3.1. Concordancias de la palabra *A* 31
3.2. Ejemplo incremento del Vocabulario, presentado en el Libro *"Análisis estadístico de Textos"* Lebart, L. y otros, página 65. 32
3.3. Ejemplo Diagrama de Pareto, presentado en el Libro *"Análisis estadístico de Textos"* Lebart, L. y otros, página 61. 34
3.4. Estructuras de Contigüidad 50
3.5. Ejemplo de Tablas de Contigüidad para estructuras **S1** y **S2**, extraido de *Statistique Textuelle*; L. Lebart, A. Salem, pág. 201. 52

4.1. Primer plano factorial tomando como variables suplementarias las preguntas abiertas postcodificadas. 62
4.2. Primer plano factorial, tomando como variables suplementarias las preguntas abiertas postcodificadas, el trabajo y la instrucción de los padres. 63
4.3. Variables activas proyectadas en el primer plano factorial 64
4.4. Plano formado por los ejes factoriales 1 y 3 65
4.5. Proyección de las modalidades de las preguntas A y B en ejes 1 y 3 66
4.6. Grupos formados, proyectados en los ejes factoriales 1 y 3 67
4.7. Plano formado por ejes factoriales 1 y 3 68
4.8. Plano factorial formado por los ejes 1 y 4 70

5.1. Formas gráficas y clusters del subcorpus de la pregunta A 74
5.2. Grupos formados en el subcorpus B 75
5.3. Palabras de la pregunta A, con variables activas y clusters 76
5.4. Palabras de la pregunta B, con variables activas y clusters 77
5.5. Palabras del subcorpus A y categorías de la pregunta A postcodificada proyectadas en plano factorial 78
5.6. Palabras del subcorpus B y categorías de la pregunta B postcodificada proyectadas en plano factorial 79

Índice de cuadros

3.1. Relaciones entre la nube fila y la nube columna 17
3.2. Cadena del tratamiento estadístico 24
3.3. Ejemplo de un Corpus 29
3.4. Ejemplo 57

4.1. Salida DTM 69

5.1. Descripción de los clusters formados 72
5.2. Concordancias de la palabra *Nada* 80