Revista de Investigación Educativa
ISSN: 0212-4068
rie@um.es
Asociación Interuniversitaria de Investigación Pedagógica
España

Recamán Payo, Adriana; Nieto Martín, Santiago

PROVALIS RESEARCH, SOFTWARE ESPECIALIZADO PARA EL ANÁLISIS DE TEXTOS EN LA INVESTIGACIÓN EDUCATIVA. APLICACIÓN OPERATIVA

Revista de Investigación Educativa, vol. 30, núm. 2, junio, 2012, pp. 397-422
Asociación Interuniversitaria de Investigación Pedagógica
Murcia, España

Recamán Payo, Adriana; Nieto Martín, Santiago (2012). Provalis Research, software especializado para el análisis de textos en la investigación educativa. Aplicación operativa. *Revista de Investigación Educativa, 30* (2), 397-422.

PROVALIS RESEARCH, SOFTWARE ESPECIALIZADO PARA EL ANÁLISIS DE TEXTOS EN LA INVESTIGACIÓN EDUCATIVA. APLICACIÓN OPERATIVA[1]

Adriana Recamán Payo
Santiago Nieto Martín
Facultad de Educación - Universidad de Salamanca

RESUMEN

Las diversas alternativas que ofrecen las herramientas informáticas en el procesamiento de textos, posibilitan la descripción objetiva, sistemática y cuantitativa del contenido analizado. Y para ello ponemos a prueba el paquete Provalis Research *aplicado a unas cartas que valoran los comportamientos personales de los maestros.*

A través de software especializado en el análisis de documentos, el presente trabajo estudia la relevancia de las palabras frecuentes en la población objeto de estudio. Tratamos de esclarecer las posibilidades que ofrece la jerarquía terminológica y su contribución a la descripción de los hechos, *orientación de la acción investigadora y generación de nuevo conocimiento.*

Palabras clave: *Análisis de contenido; Investigación cualitativa; Frecuencia; Proximidad; Técnicas cluster.*

Correspondencia:
Adriana Recamán Payo (adrepa@usal.es). Universidad de Salamanca. Facultad de Educación. Edificio Europa. Dpto. de Didáctica, Organización y Métodos de Investigación. Seminario 10. Paseo Canalejas, 169. 37008, Salamanca. Telf.: 923 294 630 Ext. 3339.
Santiago Nieto Martín (snietom@usal.es).

1 Subprograma de *ayudas para becas y contratos de Formación de Profesorado Universitario del Programa Nacional de Formación de Recursos Humanos de Investigación, en el marco del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008-2011.*

PROVALIS RESEARCH, SPECIALIZED SOFTWARE FOR TEXT ANALYSIS IN EDUCATIONAL RESEARCH. OPERATIONAL APPLICATION

ABSTRACT

The different alternatives offered by computer software for text processing make it possible to describe content in an objective, systematic and qualitative way. In this vein, we test the Provalis Research *package software applied to letters evaluating the personal behaviour of teachers.*

By means of specialized software for document analysis, this study focuses on the relevance of frequent words in the population studied, in an attempt to shed light on the possibilities offered by a terminological hierarchy and its contribution to the description of facts*, the orientation of research activity and the generation of new knowledge.*

Keywords: *Content analysis; Qualitative research; Frequency; Proximity; Cluster techniques.*

1. INTRODUCCIÓN

El análisis de contenido suele enmarcase en una metodología de carácter cualitativo, aunque su interpretación requiere, con frecuencia, de una adecuada valoración cuantitativa. La información textual y gráfica que se aborda en este tipo de análisis es material documental y, en consecuencia, objeto de una observación indirecta de la realidad.

León et al (2011, 13) señalan que "en los últimos años, el estudio de las inferencias ha adquirido tanta relevancia que actualmente se consideran el núcleo de la comprensión e interpretación de la realidad y, por tanto, uno de los pilares de la cognición humana".

El análisis de pequeños o grandes volúmenes de información y la necesidad de extraer y descubrir regularidades, patrones y/o relaciones en su contenido es demandado desde muchos ámbitos de conocimiento como la medicina, la psicología, la sociólogía, la política, la etnografía… Holsti (1968) afirma que el contenido de cualquier tipo de comunicación (lingüístico, oral, icónico, gestual…) es susceptible de ser analizado pormenorizadamente, cualquiera que sea el número de personas implicadas en la comunicación o el instrumento que reúne los datos (agendas, diarios, cartas, cuestionarios, encuestas, tests, libros, anuncios, entrevistas, radio, prensa, televisión, literatura, ciencia...).

Pero al igual que la sociedad experimenta una progresiva evolución, el análisis de contenido también ha ampliado su campo de estudio. Podemos encontrarlo aplicado en el análisis de sitios *web*, en el descubrimiento y la extracción de información en informes, en quejas y opiniones de clientes (alumnos, usuarios…), en mensajes de foros, chats, sms y blogs, en etiquetado automático y clasificación de documentos, en desarrollo y validación de taxonomías, en detección de fraudes, en atribución de autoría o en análisis de patentes… (Llisterri, 2003).

2. OBJETIVOS

Los objetivos planteados en el desarrollo de este trabajo son:

- Evidenciar las posibilidades de análisis y clasificación de contenido educativo a través de software especializado en la minería de datos y de texto.
- Profundizar en el contenido textual de un libro de cartas de alumnos para esclarecer las relaciones y dependencias entre palabras clave, tópicos y casos.
- Avanzar en el proceso de clarificación de colecciones de naturaleza pedagógica a través del análisis de contenido y destacar la necesidad de estudios fundamentados a través del establecimiento de diccionarios de categorización.

3. METODOLOGÍA: ANÁLISIS DE CONTENIDO

A pesar de la ausencia de estándares y modelos analíticos, los datos derivados de estudios cualitativos deben ser consistentes y fiables, con registros sistemáticos y descripciones cuidadas.

La categorización es determinante en este proceso dado que dota de estructura al estudio y proporciona una base sólida sobre la que argumentar y a partir de la cual establecer conexiones y dependencias.

El investigador es el principal protagonista en la recogida de datos. Su concepción de la realidad, estudiada como un todo distinto de la suma de las partes que lo componen, permite explorar todas las posibilidades y relaciones que ofrece el contenido. En ello es relevante el análisis inductivo que el investigador aplica y que conlleva "una primera descripción de las situaciones de cada uno de los casos o eventos estudiados, con el fin de detectar progresivamente la existencia de unas regularidades entre ellos que constituyen la base o germen de una futura teoría adecuada a las condiciones y valores locales" (Anguera, 2008, 143).

Erickson (1986) señala los marcos de exigencia que le dan legitimidad metodológica a los estudios cualitativos y que deben cumplirse para alcanzar verdaderas cotas de calidad en los mismos: validez semántica, validez hermenéutica, validez pragmática.

Pero a pesar del cumplimiento de estas premisas, la metodología cualitativa, se ve aquejada de cierta carencia de objetividad que disminuye con la irrupción tecnológica. El software especializado contribuye, por un lado, a certificar los procesos de fiabilidad y validez del proceso, y por otro, proporciona herramientas informáticas capaces de transformar los datos cualitativos en información de naturaleza cuantitativa.

A partir de este momento hablamos de una complementariedad de la metodología cualitativa con la cuantitativa para el análisis de contenido, pues en palabras de Anguera (2008, 150) "si la metodología cualitativa nos ayudó en la obtención de datos que aportan una gran riqueza informativa, la cuantitativa nos suministra los recursos para su análisis más conveniente". Con ello nos desmarcamos de una elección paradigmática excluyente en favor de una integración de técnicas y convivencia de planteamientos.

Con el fin de acercar ambas posiciones metodológicas aparece el software especializado para análisis de datos cualitativos asistido por ordenador, el denominado

CAQDAS[2]; herramientas pragmáticas que respaldan la investigación cualitativa y posibilitan un manejo sistemático de los datos a través de búsquedas, registros y selecciones mecanizadas que ordenan y sintetizan la información.

El desarrollo de herramientas de investigación, estudio y análisis de información estructurada y no estructurada en documentos, encuestas, audio, vídeo e imágenes, apuesta por la combinación de técnicas cualitativas y cuantitativas. Para ello se implementan paquetes de software para el análisis de datos (*mixed-model qualitative data analysis*), que posibilitan el estudio manual y automatizado de grandes colecciones multimedia.

Así, el conjunto de programas de *Provalis Research*, integrado por los módulos *WordStat*, *QDA Miner* y *SimStat*, constituye uno de estos paquetes, a partir del cual se desarrolla el presente tabajo:

FIGURA 1
HERRAMIENTAS DE PROVALIS RESEARCH PARA EL ANÁLISIS DE TEXTO

Las razones que argumentaron la elección de este software se resumen en los siguientes términos:

- Provalis Research proporciona herramientas fáciles y potentes para el análisis de texto asistido por ordenador.
- Su concepción trimodular (minería de texto, minería de datos y análisis estadístico) permite realizar análisis profundos de contenido sin necesidad de recurrir a otros programas complementarios.
- Esta posibilidad de imbricación de módulos (ausente en *Atlas Ti, Nvivo*...) potencia los análisis y resultados, a la vez que reduce el tiempo de ejecución, sin la pérdida de datos ni los problemas de compatibilidad que se pueden generar con la exportación de información.
- Intuitivo y fácil de utilizar, *Provalis Research* integra herramientas innovadoras para el análisis de contenido multimedia, la identificación de patrones o la

2 Computer Assisted Qualitative Data Analysis Software.

configuración multiusuario, a lo que se añade un coste de adquisición inferior al de otros programas de similares características en el mercado (*SPSS Text Analytics*).

3.1. La minería de datos con QDA Miner

QDA Miner es un programa de análisis cualitativo de datos que permite anotar, codificar y recuperar documentos gráficos y textuales (minería de datos). Con ello estructuramos la información y clasificamos el contenido para análisis más profundos y sistemáticos. Dotado de herramientas exploratorias, posibilita la identificación de patrones y la relación entre los códigos asignados. El potencial del programa se incrementa a través de su integración con herramientas avanzadas de análisis estadístico (*SimStat*) y minería de texto (*WordStat*).

QDA Miner maneja proyectos complejos sobre los que gran cantidad de documentos se combinan con información categorial y numérica. Los casos, las variables y los códigos son los elementos estructurales básicos del programa. "Posee varias herramientas para asistir en la tarea de codificación y realizar análisis descriptivos, comparativos y exploratorios. Estas herramientas pueden usarse para sistematizar la codificación de documentos, asegurar la consistencia de la codificación, identificar regularidades y patrones y descubrir relaciones ocultas entre los códigos y otras propiedades de los casos." (Cisneros, 2009, 89).

Las herramientas estadísticas y de visualización que tiene integradas, así como las de agrupamiento (clustering), las de escalamiento multidimensional, los mapas de calor, los análisis de correspondencias y los análisis de secuencias, permiten identificar rápidamente tendencias, explorar los datos, describirlos, establecer comparaciones y probar hipótesis.

Una herramienta de generación de informes permite a los investigadores almacenar, en un sólo lugar, las consultas y análisis de resultados, tablas y gráficos, notas de investigación y citas. Posibilita crear esquemas para organizar y estructurar la información, los resultados, las interpretaciones, el seguimiento de casos y el trabajo de los diversos miembros del equipo.

QDA Miner posee un registro de comandos que realiza un seguimiento de cada acceso al proyecto, cada operación de codificación, cada cambio, consulta o análisis realizado. Con ello proporcionamos confiabilidad al análisis, ya que la herramienta posibilita documentar el proceso y supervisar el trabajo individual o colectivo, así como recordar y repetir las consultas y análisis efectuados previamente o deshacer algunas operaciones realizadas con anterioridad.

Junto con el administrador de informes, este registro de comandos proporciona referencias muy valiosas para la creación de una pista de auditoría detallada, de enorme utilidad para contribuir a garantizar la transparencia del proceso de investigación y la credibilidad del estudio.

La función principal de *QDA Miner* es asignar códigos a segmentos seleccionados de texto para luego analizar esos códigos y establecer relaciones entre ellos. Los códigos se agrupan formando categorías que poseen una estructura en árbol. En la estructura las categorías son los nodos y bajo ellas están los códigos asociados. Cuando se crea

un proyecto, el libro de códigos está vacío y comienza un proceso hermenéutico controlado que se mueve entre la creatividad y la desocultación.

La flexibilidad que ofrece la herramienta permite la división y fusión de código, la redimensión de segmentos de código y la búsqueda y reemplazo interactivo del mismo, además de su codificación y recodificación automática.

QDA Miner posee también importantes herramientas de recuperación de texto, como un motor de búsqueda de gran alcance fundamentado en *operadores booleanos* y consultas a tesauros o un buscador de palabras clave basado en diccionarios específicos para el análisis de contenido.

3.2. La minería de texto con WordStat

WordStat es un módulo de minería de textos para la extracción rápida de patrones o tendencias, así como el seguimiento cuidadoso y preciso de los métodos cualitativos de análisis de contenido sobre información no estructurada.

El módulo constituye una extensión de *QDA Miner* y *SimStat*, programas con los que se complementa para proporcionar una combinación flexible y dilatada de métodos cualitativos y cuantitativos en el análisis de contenido.

El análisis más básico que realiza consiste en un recuento de las palabras frecuentes, un simple análisis descriptivo sin relacionarlo con otras variables y sin ningún diccionario de categorización activo.

En el momento de procesar el texto se escogen el número máximo de ítems que queremos que aparezcan listados y se pueden agrupar en función de la frecuencia, la ocurrencia de caso, orden alfabético, o TF*IDF[3]. A partir de aquí, se pueden extraer temas automáticamente utilizando técnicas jerárquicas de cluster.

Wordstat ofrece la posibilidad de identificar, a través de dendrogramas, las coocurrencias de palabras. La agrupación de palabras revela tópicos en la colección de textos y el estudio en profundidad del dendrograma permite identificarlos rápidamente y recuperar los segmentos de texto asociados a cada tópico.

La herramienta de escala multidimensional (mapa 2D) se puede utilizar para representar la proximidad entre las palabras y los tópicos. A ello se suma el gráfico de burbujas, que añade información sobre la frecuencia de cada ítem a partir del diámetro de las esferas.

El gráfico de proximidad por su parte, representa la distancia entre una palabra clave (*keyword*) y todas las demás. De este modo podemos comparar coocurrencias. También podemos visualizar los párrafos asociados a tales relaciones y calcular estadísticos específicos para evaluar la fuerza de la asociación (Chi-cuadrado, Pearson, Student, Spearman…).

El análisis de correspondencias es también un modo muy eficiente de identificar patrones entre palabras en el documento y valores de otra variable. Esto permitiría,

3 TF*IDF: frecuencia del término ponderada por la inversa de la frecuencia del documento. Esta ponderación se basa en la suposición de que cuanto más a menudo un término aparece en un documento, más representativo de su contenido, sin embargo, en cuantos más documentos aparece el término, menos discriminante es.

por ejemplo, agrupar juntos aquellos documentos que emplean las mismas palabras y situar aparte a aquellos otros que utilizan diferentes.

En lo relativo a los diccionarios en *WordStat*, observamos que el programa permite medir conceptos específicos usando diccionarios de categorización. Estos diccionarios pueden utilizarse para asignar varias palabras o frases a las categorías de contenido.

Un *diccionario de categorización* es una estructura jerárquica donde cada categoría puede ser medida usando palabras, patrones, frases o reglas de proximidad.

El diccionario que utiliza el programa, *Regessive Imagery Dictionary (RID)*[4], se compone de cerca de 3200 palabras y raíces asignadas a 29 categorías de los procesos primarios de cognición, 7 categorías de los procesos secundarios de cognición y 7 categorías de emociones.

El RID (Martindale, 1975, 1990) sigue un esquema de codificación de análisis de contenido diseñado para medir el pensamiento primario vs pensamiento conceptual. El pensamiento conceptual es abstracto, lógico, orientado a la realidad y dirigido a la solución de problemas. El pensamiento primordial, por su parte, es asociativo, concreto y tiene poco en cuenta la realidad. Es el tipo de pensamiento que se encuentra en la fantasía y los sueños.

En base al RID, el *Wordstat* realiza un recuento de las ocurrencias de cada categoría en los documentos y permite conocer la composición de los mismos así como el número de palabras que fueron asignados a cada categoría. Esta información puede ser sometida a análisis estadísticos. La razón de ser del diccionario es que los procesos psicológicos se reflejan en el contenido de un texto. Así, por ejemplo, a mayor pensamiento primario implicado en la producción de un texto, menos abstracto, más unidad y más palabras vinculadas a las sensaciones.

El RID parece proporcionar un índice válido del constructo de los contenidos primarios vs conceptuales de pensamiento y así lo certifican determinados estudios, como por ejemplo el número significativamente mayor de contenido primordial que se ha encontrado en las historias de fantasía de personas creativas frente a sujetos no creativos (Martindale y Dailey, 1996), o la presencia de contenido más primordial en las producciones verbales de niños pequeños en comparación con niños mayores (West, Martindale y Sutton-Smith, 1985) y de los sujetos con esquizofrenia en comparación con los sujetos control (West y Martindale, 1988).

WordStat también ofrece herramientas que ayudan a construir un diccionario para análisis de contenido propio y una lista de exclusión que contiene aquellas palabras que en el proceso de análisis serán ignoradas por el programa.

Para abordar toda esta vorágine informativa e instrumental es necesario no perder de vista el contexto y los objetivos de la investigación. La densidad documental, la concurrencia de técnicas y una información excesivamente segmentada pueden desviar la atención o menoscabar la interpretación.

4 Regessive Imagery Dictionary, Latin Version translated by Ron Newbold.

4. MATERIAL OBJETO DE ESTUDIO

El material objeto de análisis se compone de 180 cartas de alumnos sobre sus antiguos maestros recogidas en el libro *Profesores que dejan huella* (Gato, 2006). Estas cartas, noventa de ellas de carácter positivo y otras noventa de carácter negativo, constituyen la muestra de nuestro estudio y contribuyen a fijar criterios sobre *qué es y qué no puede ser un profesor*.

Las cartas publicadas fueron seleccionadas opináticamente, atendiendo fundamentalmente a su contenido, entre las recopiladas por una profesora en su asignatura de *Relación educativa y estilo docente*.

Durante 1995 y 2004, sugiere a sus diferentes alumnos, como actividad de aula, la redacción de dos cartas dirigidas a sus antiguos profesores: una al profesor que había dejado una huella positiva en su vida, y otra a aquél, cuyo sólo recuerdo despertaba en él vivencias negativas;

Estas producciones, esencialmente recogidas en el ámbito extremeño, constituyen "un relato de una parte sustancial de la vida académica del autor de la misma, que tiene que ver con un período concreto de su trayectoria vital: infancia, adolescencia, o juventud" y de su realidad escolar "que va, desde los Centros de Educación Infantil, hasta los Centros de Estudios Superiores o Universitarios" (Gato, 2006).

En palabras de la autora, estas cartas aportan información sobre la *relación profesor-alumno que se establece en el aula en todo proceso de enseñanza-aprendizaje*, ofrecen manifestaciones anímicas y revelan *nuevos e inesperados matices al retrato del profesor.*

El libro, con sus 180 cartas, con una extensión media de 222 palabras/carta y 17 líneas/carta, constituye la base del presente estudio y está estructurado en dos grandes bloques:

Las 90 primeras cartas son de agradecimiento y valoración a la labor docente. Ensalzan valores como la confianza, paciencia, admiración, gratitud, respeto, cariño… Se subdividen a su vez en tres grandes grupos:

1.- Educar por lo que se es /transmisión de valores
2.- Afecto/atención personal
3.- Motivación/estímulo

Las restantes 90 cartas constituyen el bloque, consideradas, de carácter negativo. Claramente diferenciadas de las anteriores en el *signo* de su contenido, reflejan dificultades y temores en la relación profesor-alumno así como falta de interacción o intereses divergentes.

De acuerdo al contenido dominante en cada una de ellas, las cartas negativas se agrupan en:

1.- Falta de profesionalidad/incompetencia
2.- Distancia/impersonalidad
3.- Amenaza/miedo

Nuestro texto lo forman, pues, las cartas, y su distribución en el libro constituye una primera clasificación de las mismas, que utilizaremos como unidades de contexto.

TABLA 1
UNIDADES DE CONTEXTO DE LA POBLACIÓN

CATEGORÍA	SUBCATEGORÍA	Nº
CARTAS POSITIVAS	Educar por lo que se es/Transmisión de valores	30
	Afectivo/Atención personal	30
	Motivación/Estímulo	30
CARTAS NEGATIVAS	Falta de profesionalidad/Incompetencia	30
	Distancia/Impersonalidad	30
	Amenaza/Miedo	30

5. PLANTEAMIENTO DEL PROBLEMA

El material de que disponemos admite gran cantidad de variantes para su organización, pero la estructura elegida por la autora responde, en su opinión, a la intención de subrayar determinadas constantes que se evidencian en ellos.

Su método es la simple lectura de las impresiones, sinceras y espontaneas, expresadas en los diversos documentos, de los que entresaca los seis grandes matices y los dos grandes bloques en que divide la colección.

Procede, en este punto, realizar algunas preguntas de interés, como… ¿avala el estudio de los datos esta clasificación?, ¿existe una mejor estructuración de los documentos en función del contenido de los mismos?, ¿hay temáticas relevantes que no han sido consideradas?; si tuviésemos que seguir segmentando las unidades temáticas, ¿cúales serían los tópicos a considerar?, ¿es suficiente un análisis de palabras frecuentes para resolver estos planteamientos?

6. ANÁLISIS DE DATOS Y RESULTADOS

A continuación realizamos el estudio atendiendo a tres técnicas diferenciadas. De un lado el análisis simple de frecuencias, de otro la aplicación de técnicas jerarquicas y finalmente el escalamiento multidimensional.

6.1. Palabras frecuentes con WordStat

Efectuamos el análisis del texto a través del recuento de las palabras frecuentes en las 180 cartas objeto de estudio. Para obtener resultados lo más reveladores posibles establecemos tres condiciones analíticas previas: *lematización*, *lista de exclusión* y *frecuencia del término ponderada*.

6.1.1. Lematización

La *lematización* es un proceso de eliminación automática de partes no esenciales de las palabras (sufijos, prefijos) para reducirlas a su parte original (lema). Con ello se

facilita la eficacia de la indización y la recuperación de información. Así por ejemplo, en la lematización, todos los plurales se transforman en formas singulares y todos los verbos en tiempo pasado son reemplazados por tiempo presente. Igualmente los nombres, verbos, adjetivos y adverbios derivados de la misma raíz se transforman en esta sola palabra.

En *WordStat*, la *lematización* es un algoritmo de sustitución de sufijo. Este automatismo puede generar alguna sustitución de palabra no válida, pero esos errores no tendrán consecuencias importantes sobre el resultado de un análisis, además de que pueden ser controlados a través de una lista de sustituciones personalizadas. En ella corregimos aquellas lematizaciones que no se adecúan a la lengua española en general, ni a nuestro objeto de estudio en particular.

Veamos así las 30 palabras más frecuentes en el conjunto de las 180 cartas sin y con *lematización* respectivamente:

TABLA 2
FRECUENCIA SIN LEMATIZACIÓN Y CON ELLA

	FRECUENCIA		FRECUENCIA
NO	765	SER	873
ES	267	NO	765
MÁS	253	HABER	484
ALUMNOS	181	HACER	329
CLASE	147	TENER	325
SER	145	ALUMNO	294
PROFESOR	135	SABER	274
ERA	130	CLASE	259
GRACIAS	127	PROFESOR	256
CLASES	112	MÁS	253
RECUERDO	112	QUERER	236
SIEMPRE	109	IR	192
YA	109	DAR	182
NUNCA	103	ESTAR	177
DÍA	95	PODER	172
HE	95	RECORDAR	171
SÓLO	89	DECIR	155
TAMBIÉN	88	ENSEÑAR	150
COSAS	87	BUEN	147
HA	84	GRACIAS	134
AÑOS	82	MAESTRO	117
CÓMO	81	COSA	115
BIEN	79	SENTIR	111
VIDA	79	DÍA	109
CARTA	73	SIEMPRE	109
MAESTRA	73	YA	109
MEJOR	73	BIEN	108
ALUMNA	72	APRENDER	106
MUY	70	NUNCA	103
TIEMPO	70	AÑO	101

Fijémonos como la *lematización* altera considerablemente el orden de términos frecuentes.

Sin aplicar el algoritmo y las correspondientes adaptaciones del mismo a nuestro trabajo, el adverbio NO aparece como la palabra de mayor frecuencia en el análisis (765 apariciones) seguido de la forma conjugada ES (267).

Sin embargo, tras aplicar el proceso de lematización, la jerarquía de términos se reorganiza considerablemente. Al recoger todas las formas del verbo SER (soy, es, fuimos...) en una sola, encontramos que aparece un total de 873 veces, posicionándose como el término más frecuente en las cartas superando al adverbio de negación. Por tanto, la lematización es un proceso importante en la minería de textos y altera considerablemente el orden de términos frecuentes en los textos analizados.

Asimismo, hemos agrupado las palabras alumno, alumna, alumnos, alumnas y alumnado en un único término ALUMNO, (Figura 2). La suma de frecuencias independientes de cada uno de ellos (28,72, 181, 9 y 4 respectivamente) constituye la frecuencia global del término ALUMNO (294).

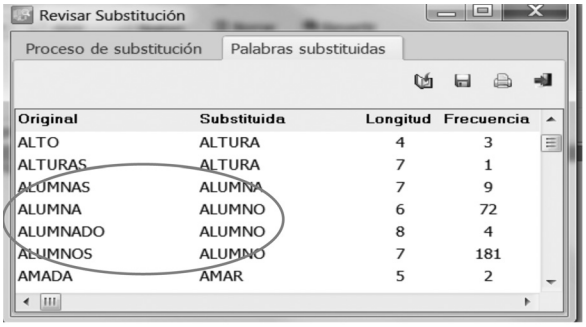


FIGURA 2
REVISIÓN DE LEMATIZACIÓN

En la *lematización* establecida hemos prescindido de las diferencias de género, la distinción entre singular o plural, la oposición individual/colectivo así como de la diversidad de tiempos y modos verbales dado que su estudio diverge de nuestros objetivos de análisis.

6.1.2. Lista de exclusión

La Tabla 2, frecuencia sin lematización y con ella, fue obtenida, además, como resultado de aplicar un diccionario de exclusión elaborado para eliminar del análisis aquellas palabras que no se adecúan a nuestros intereses.

La finalidad de su aplicación es excluir del estudio palabras con poco o ningún valor semántico, dado que basamos nuestro estudio sólo en palabras *llenas* (sustantivos, adjetivos, verbos y adverbios) con algunas salvedades, como veremos.

TABLA 3
FRECUENCIAS SIN Y CON LISTA DE EXCLUSIÓN

	RECUENCI
A	967
LA	824
EN	799
NO	771
EL	589
POR	494
ME	493
LO	453
CON	431
UN	395
LOS	370
SE	329
UNA	318
TE	304
PARA	286
COMO	282
ES	267
MÁS	254
SU	242
USTED	234
TU	231
NOS	229
LAS	228
PERO	196
TODO	193
MI	192
ALUMNOS	181
SUS	179
AL	170
SI	167

	FRECUENCIA
SER	873
NO	765
HABER	484
HACER	329
TENER	325
ALUMNO	294
SABER	274
CLASE	259
PROFESOR	256
MÁS	253
QUERER	236
IR	192
DAR	182
ESTAR	177
PODER	172
RECORDAR	171
DECIR	155
ENSEÑAR	150
BUEN	147
GRACIAS	134
MAESTRO	117
COSA	115
SENTIR	111
DÍA	109
SIEMPRE	109
YA	109
BIEN	108
APRENDER	106
NUNCA	103
AÑO	101

Así pues, conforman esta lista de exclusión términos que *en vez de significar*, señalan, indican a un elemento, persona, objeto (pronombres y artículos) o tienen función de nexo en la oración (preposiciones y conjunciones). A ello hemos añadido las locuciones conjuntivas y prepositivas, además de las letras del alfabeto, los números (cardinales y ordinales) y las contracciones.

6.1.3. Frecuencia del término ponderada

Partiendo de este inventario de términos frecuentes, *pulido* por la *lematización* y la *lista de exclusión*, obtenemos las palabras de mayor presencia en el texto de cartas. Sin embargo, con ello sólo obtenemos información sobre el número de ocurrencias de cada término. Es decir, sobre esta lista hemos aplicado una ponderación vinculada a la presencia o ausencia de términos en los diversos documentos analizados, lo que está directamente vinculado a procesos de *recuperación de información*.

Presumiblemente, un documento o zona del mismo donde se menciona una palabra con mucha frecuencia, está fuertemente vinculado a la temática asociada con dicha palabra. Por eso, un mecanismo justificado de medición consiste en asignar a cada término *(t)* un peso específico que dependerá del número de veces que aparece en del documento *(d)*: a igual número de ocurrencias de dos términos en el documento, igual peso para cada uno de ellos.

Este sistema de ponderación se llama *frecuencia del término* y se denota $\boldsymbol{tf_{t,d}}$.

Al mismo tiempo, definiremos la *frecuencia del documento* ($\boldsymbol{df_{d,t}}$) como el número de documentos *(d)* de la colección que contienen el término *(t)*.

Pero nos encontramos con que no todas las palabras de un documento, pese a su frecuencia, son igual de importantes. Efectivamente, ese es el motivo por el que hemos decidido no indexar todas las palabras y valernos de una lista de exclusión. Con ello hemos prescindido tanto de su recuperación como de su ponderación.

No obstante todavía se puede perfilar mucho más el proceso de análisis y para eso utilizamos la *frecuencia inversa del documento* (***idf***). Y es que además de los términos excluidos previamente al proceso, existen otros tantos que tienen poco o ningún poder para discriminar la relevancia temática del contenido analizado.

Así introducimos otro mecanismo que atenúa el efecto de las palabras que se producen con mucha frecuencia en nuestra colección de cartas *(N=180)*, para con ello lograr mayor significatividad en la determinación de la relevancia de los términos.

A tal fin se reduce el peso de un término a medida que aumenta su número de apariciones en el conjunto total de la colección, dado que cuanto más a menudo aparece un término en un documento, más representativo de su contenido, sin embargo, en cuantos más documentos aparece, menos discriminante es.

$$\mathbf{idf_t = \log \frac{N}{df_t}}$$

TABLA 4
FRECUENCIA DE TÉRMINOS Y DEL DOCUMENTO

t	$\mathbf{tf_{t,d}}$	$\mathbf{df_{d,t}}$	$\mathbf{idf_t}$
SER	873	174	$\text{idf}_{\text{SER}} = \log\frac{180}{174} = 0{,}01472$
MAESTRO	117	56	$\text{idf}_{\text{MAESTRO}} = \log\frac{180}{56} = 0{,}50708$

Resultado de combinar la *frecuencia de término* con *frecuencia inversa del documento*, se genera un peso para cada término *t* en cada documento *d*:

$$\mathbf{tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t}$$

La ocurrencia del término, ponderada por la inversa de la frecuencia del documento asigna un peso al término en el documento (Mannig et al, 2008), que es

— mayor cuando se produce muchas veces dentro de un pequeño número de documentos;
— menor cuando el término se produce menos veces en un documento, o se produce en muchos documentos;
— más bajo cuando el término se produce en casi todos los documentos.

TABLA 5
FRECUENCIA PONDERADA DE TÉRMINOS

t	$tf_{t,d.}$	$idf_{d,t}$	$tf\text{-}idf_{t,d}$
SER	873	0,01472	12,9
MAESTRO	117	0,50708	59,3

TABLA 6
PALABRAS FRECUENTES SIN Y CON PONDERACIÓN

	TÉRMINO	TF		TÉRMINO	TF * IDF
1	SER	873	1	**MAESTRO**	59,3
2	NO	765	2	**GRACIAS**	55,8
3	HABER	484	3	**ENSEÑAR**	47,4
4	HACER	329	4	DECIR	46,7
5	TENER	325	5	COSA	45,8
6	ALUMNO	294	6	BUEN	45,0
7	SABER	274	7	IR	44,9
8	CLASE	259	8	SENTIR	44,8
9	PROFESOR	256	9	PODER	44,7
10	MÁS	253	10	SIEMPRE	44,7
11	QUERER	236	11	NUNCA	44,2
12	IR	192	12	DÍA	44,0
13	DAR	182	13	ESTAR	43,7
14	ESTAR	177	14	MAL	42,8
15	PODER	172	15	MÁS	42,7
16	RECORDAR	171	16	VIDA	41,8
17	DECIR	155	17	SEGUIR	41,5
18	**ENSEÑAR**	150	18	YA	41,4
19	BUEN	147	19	DAR	41,1
20	**GRACIAS**	134	20	PERSONA	41,0
21	**MAESTRO**	117	21	CLASE	41,0
22	COSA	115	22	AÑO	40,8
23	SENTIR	111	23	VEZ	40,8
24	YA	109	24	SÓLO	40,6
25	DÍA	109	25	CÓMO	40,5
26	SIEMPRE	109	26	BIEN	40,4
27	BIEN	108	27	APRENDER	40,3
28	APRENDER	106	28	RECORDAR	40,0
29	NUNCA	103	29	TRABAJAR	39,8
30	VEZ	101	30	SABER	39,6

Así, por ejemplo (Tabla 6), el verbo SER y el adverbio de negación NO, son las dos palabras más frecuentes de nuestros documentos, pero, como observamos, también dos de los términos de menor relevancia. Igualmente, los términos ALUMNO y PROFESOR que aparecen en sexta y novena posición respecto a su frecuencia de aparición, *caen* en relevancia al trigésimo tercero y trigésimo octavo puesto respectivamente. Mientras tanto, MAESTRO pasa de ocupar un vigésimo primer lugar en el orden de frecuencias a posicionarse el primero en relevancia tras la ponderación.

6.2. Técnicas jerárquicas

A continuación utilizamos métodos jerárquicos de análisis cluster sobre las 500 palabras de mayor frecuencia basándonos en sus co-ocurrencias en los casos.

Con ello percibimos fácilmente palabras que con frecuencia aparecen próximas, de manera que los diversos grupos de palabras nos revelarán los diferentes temas y tópicos de las cartas.

Estas técnicas clúster nos permiten efectuar un proceso de aglomeración progresivo que minimiza la distancia entre las palabras frecuentes (o maximiza la similitud entre ellas).

También formamos los conglomerados por casos, es decir por documentos (cartas). Así comprobamos los clúster formados por las 180 cartas y comparamos esta agrupación con la correspondiente clasificación de la muestra establecida en la Tabla 1, unidades de contexto de la población.

El análisis de conglomerados y el escalamiento multidimensional de todas las palabras o categorías, permite obtener un árbol de clasificación (dendrograma), que nos muestra gráficamente el proceso de unión seguido y el nivel de fusión para cada situación.

6.2.1. Conglomerados por palabras clave (términos frecuentes)

Para nuestro análisis consideramos que dos términos coocurren cada vez que aparecen juntos en el mismo caso, dado que la extensión de las cartas es relativamente pequeña; si bien podríamos haber restringido su coocurrencia por párrafo, por oración o a un cierto rango de palabras preestablecido.

Además, la coocurrencia por caso es más apropiada para determinar con mayor facilidad los temas de cada carta.

De entre los diversos índices existentes (*Jaccard*, *Ochiai*, *Theta*…) para establecer la medida de similitud utilizada en la agrupación, utilizamos el *coeficiente de Sorensen*.

Este índice tiene en cuenta las ocurrencias de palabras pero no su frecuencia. Es similar al de *Jaccard* pero las correspondencias experimentan una doble ponderación:

$$\frac{2a}{2a + b + c}$$

Dónde *a* representa los casos en que ambas palabras ocurren, y *b* y *c* representan los casos en que sólo una de las palabras es encontrada.

En cuanto al tipo de agrupación escogemos el *clustering de segundo orden* que se basa en perfiles de coocurrencia en lugar de coocurrencia de palabras. Es decir, *dos palabras clave estarán próximas no necesariamente por su coocurrencia, sino porque aparecen en el similares contextos*. Esto favorece la agrupación de palabras que son sinónimos o formas alternativas de una misma palabra. Así, por ejemplo, las palabras colegio y escuela, no suelen aparecer juntas en el mismo documento, pero la agrupación de segundo orden es capaz de establecer una gran proximidad entre ellas dado que ambas coocurren con palabras como profesor o alumno (Grefenstette, 1994).

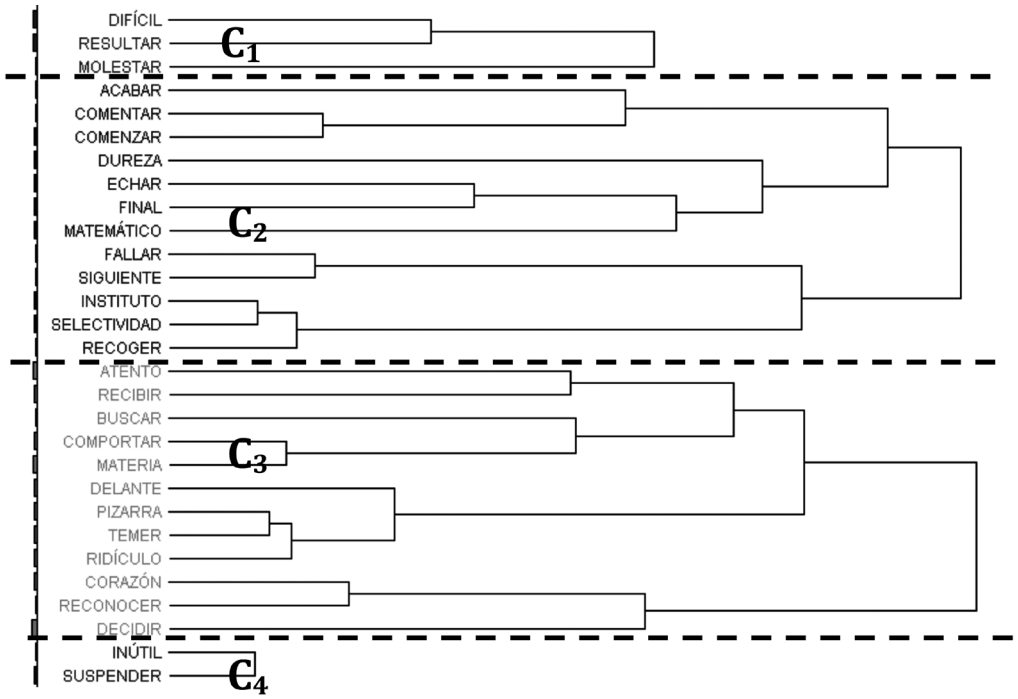


FIGURA 3
ORDEN DE AGLOMERACIÓN PARA K = 136

A partir del dendrograma obtenido con estas condiciones, a medida que aumentamos la distancia mínima requerida para agrupar documentos, aumentamos también el número de clusters formados. Es decir, la variación de *K* forma grupos adicionales de palabras. Cuanto menor es su valor, más amplios y generales son los grupos temáticos encontrados.

Para *k = 136*, hay 45 conglomerados de una sola palabra que eliminamos de la jerarquía obtenida por considerar que los elementos aislados no aportan información temática concluyente.

Entre los 91 clúster restantes encontramos la agrupación C_1, C_2, C_3 y C_4 de la Figura 3. Los cuatro subgrupos tienen un signo predominantemente negativo y una gran tendencia a constituir un único clúster a medida que *k* disminuye.

La recuperación de palabras clave (Figura 4) confirma que los subgrupos están conformados por cartas de signo predominantemente negativo:

C_1

aso	TOPICO	SIGNO	Variable
11	distancia/impersonalidad	Cartas negativas	CARTAS
44	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
64	amenaza/miedo	Cartas negativas	CARTAS
72	amenaza/miedo	Cartas negativas	CARTAS
80	amenaza/miedo	Cartas negativas	CARTAS
107	afecto/atención personal	Cartas positivas	CARTAS
120	afecto/atención personal	Cartas positivas	CARTAS

C_2

aso	TOPICO	SIGNO	Variable
10	distancia/impersonalidad	Cartas negativas	CARTAS
14	distancia/impersonalidad	Cartas negativas	CARTAS
21	distancia/impersonalidad	Cartas negativas	CARTAS
22	distancia/impersonalidad	Cartas negativas	CARTAS
34	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
39	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
44	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
49	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
61	amenaza/miedo	Cartas negativas	CARTAS
66	amenaza/miedo	Cartas negativas	CARTAS
76	amenaza/miedo	Cartas negativas	CARTAS
82	amenaza/miedo	Cartas negativas	CARTAS
84	amenaza/miedo	Cartas negativas	CARTAS
103	afecto/atención personal	Cartas positivas	CARTAS
104	afecto/atención personal	Cartas positivas	CARTAS
106	afecto/atención personal	Cartas positivas	CARTAS
109	afecto/atención personal	Cartas positivas	CARTAS
122	motivación/estímulo	Cartas positivas	CARTAS
128	motivación/estímulo	Cartas positivas	CARTAS
144	motivación/estímulo	Cartas positivas	CARTAS
164	transmisión de valores	Cartas positivas	CARTAS
172	transmisión de valores	Cartas positivas	CARTAS

C_3

aso	TOPICO	SIGNO	Variable
4	distancia/impersonalidad	Cartas negativas	CARTAS
6	distancia/impersonalidad	Cartas negativas	CARTAS
9	distancia/impersonalidad	Cartas negativas	CARTAS
11	distancia/impersonalidad	Cartas negativas	CARTAS
14	distancia/impersonalidad	Cartas negativas	CARTAS
30	distancia/impersonalidad	Cartas negativas	CARTAS
44	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
47	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
51	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
53	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
61	amenaza/miedo	Cartas negativas	CARTAS
63	amenaza/miedo	Cartas negativas	CARTAS
65	amenaza/miedo	Cartas negativas	CARTAS
66	amenaza/miedo	Cartas negativas	CARTAS
72	amenaza/miedo	Cartas negativas	CARTAS
75	amenaza/miedo	Cartas negativas	CARTAS
76	amenaza/miedo	Cartas negativas	CARTAS
78	amenaza/miedo	Cartas negativas	CARTAS
81	amenaza/miedo	Cartas negativas	CARTAS
82	amenaza/miedo	Cartas negativas	CARTAS
83	amenaza/miedo	Cartas negativas	CARTAS
84	amenaza/miedo	Cartas negativas	CARTAS
85	amenaza/miedo	Cartas negativas	CARTAS
88	amenaza/miedo	Cartas negativas	CARTAS
97	afecto/atención personal	Cartas positivas	CARTAS
98	afecto/atención personal	Cartas positivas	CARTAS
103	afecto/atención personal	Cartas positivas	CARTAS
107	afecto/atención personal	Cartas positivas	CARTAS
118	afecto/atención personal	Cartas positivas	CARTAS
124	motivación/estímulo	Cartas positivas	CARTAS

C_4

aso	TOPICO	SIGNO	Variable
2	distancia/impersonalidad	Cartas negativas	CARTAS
5	distancia/impersonalidad	Cartas negativas	CARTAS
39	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
44	falta de profesionalidad/incompetencia	Cartas negativas	CARTAS
82	amenaza/miedo	Cartas negativas	CARTAS

FIGURA 4
RECUPERACIÓN DE PALABRAS CLAVE EN CONTEXTO

El cluster C_3 por ejemplo, está formado por los términos *atento*, *recibir*, *buscar*, *comportar*, *materia*, *delante*, *pizarra*, *temer*, *ridículo*, *corazón*, *reconocer* y *decidir*.

Este tema aparece en 30 cartas diferentes, una sexta parte del total de la muestra, donde predomina el tópico *miedo.* Hace referencia al temor y sentimiento de ridículo que un alumno experimenta o le hace sentir su maestro cuando sale a la pizarra o expone la materia:

TABLA 7
RECUPERACIÓN DE PALABRAS CLAVE EN CONTEXTO

"engullidos por el ***miedo****, por el aceleramiento de* ***corazón*** *y ese sudor frío que sentía al salir a la* ***pizarra****" (caso 81, carta negativa, amenaza/miedo).*

"Tampoco puedo olvidar los exámenes orales a los que nos sometía todos los días y cómo disfrutaba poniéndonos en ***ridículo****" (caso 84, carta negativa, amenaza/miedo).*

"tenía ***temor*** *de equivocarme y de que usted, con su sorna característica, me pusiera en* ***ridículo****,* ***delante*** *de mis compañeras" (caso 63, carta negativa, amenaza/miedo).*

"pretendía siempre mostrar su superioridad y dejar en ***ridículo*** *a cuantos alumnos caían en sus manos. La* ***pizarra*** *era el escenario al que usted sacaba a sus alumnos, para mirarlos de arriba abajo y reírse de ellos" (caso 82, carta negativa, amenaza/miedo).*

"en vez de ***recibir*** *ayuda, por su parte, lo único que* ***recibió*** *fue un castigo" (caso 88, carta negativa, amenaza/miedo).*

6.2.2. Conglomerados por casos (documentos)

Vamos a agrupar las 180 cartas a través de métodos jerárquicos aglomerativos.

Para ello partimos del nivel *K= 0* con *n* grupos. La matriz de distancias utilizada para establecer las agrupaciones se fundamenta en los *coeficientes del coseno* calculados sobre la frecuencia relativa de las diferentes palabras clave en las cartas. Por tanto, cuanto más similar sea el contenido de dos cartas en términos de distribución de sus palabras clave, mayor será el coeficiente y la propensión de esas cartas a formar un clúster.

Supongamos que queremos dividir el conjunto de casos en dos grandes grupos. El establecimiento de *K*=2 no devuelve dos clúster de signos temáticos contrapuestos como cabría esperar (cartas *positivas* y cartas *negativas*). Asimismo, para un nivel *K=6*, se forman seis grupos muy dispares cuya distribución es la que se muestra en la figura siguiente, y que nada tiene que ver con la clasificación previa de las cartas (Tabla 1), que establecía seis grupos de igual tamaño ($n_1= n_2= n_3= n_4= n_5= n_6= 30$) y de *signo* diferenciado ($n_1= n_2= n_3=$ *Positivo*, $n_4= n_5= n_6=$ *Negativo*).

Hay que llegar a formar 26 o más grupos ($K \geq 26$) para empezar a disolver el *gran grupo* que concentra la mayor parte de las cartas y reducir su densidad a menos de 36 documentos.

Supongamos que $K = 26$ y que para simplificar ocultamos los grupos con $n_i \leq 3$ documentos, por considerarlos desprovistos de significatividad. Entonces, visualizamos 14 grupos diferenciados que ya ofrecen un perfil temático más definido en cuanto a categorías (cartas *positivas* vs cartas *negativas*), pero que no ofrecen información alguna

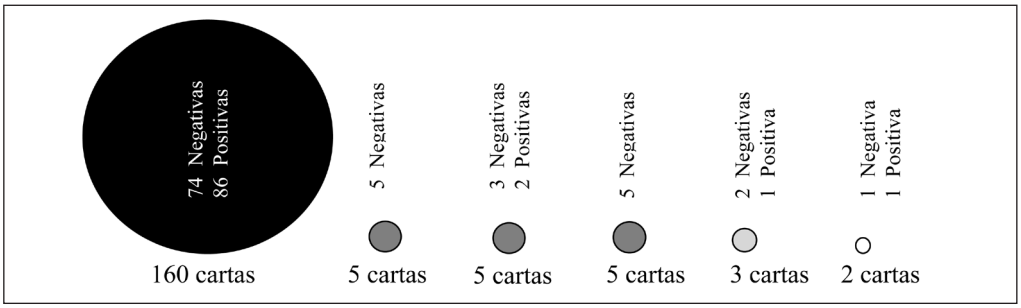


FIGURA 5
CONGLOMERADO PARA K= 6

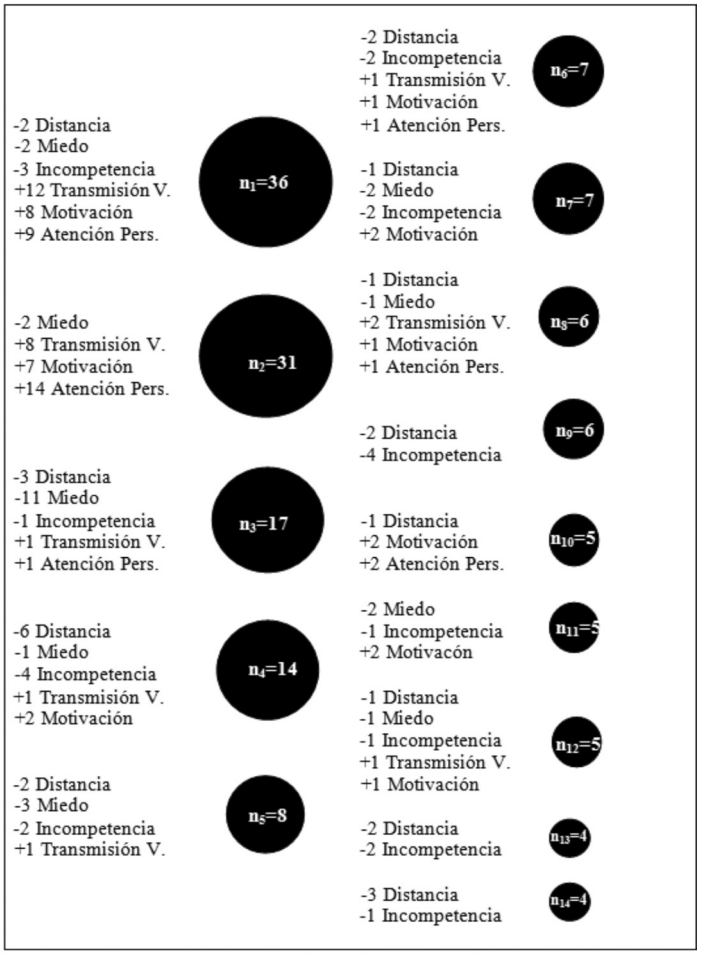


FIGURA 6
CONGLOMERADO PARA K=26

respecto a las subcategorías previamente establecidas (miedo, distancia, incompetencia, motivación, atención personal, transmisión de valores).

Si observamos con detenimiento esta división en grupos del conjunto de textos, veremos que los dos cluster más densos (C_1 y C_2) recogen buena parte de las cartas *positivas* de la distribución.

Por otro lado, las cartas *negativas* dominan en prácticamente todas las pequeñas agrupaciones restantes.

Independientemente del número de clusters determinado, el método no arroja las diferencias deseadas inter e intra grupos. La exploración general de las relaciones existentes entre cartas a través de procesos de aglomeración ascendentes basados en palabras clave no permite certificar, como una categorización posible de los documentos, el establecimiento de unidades de contexto expuestas en la Tabla 1, ni sugiere alguna otra alternativa clara.

6.3. Escalamiento Multidimensional

Esta "técnica multivariante de interdependencia trata de representar en un espacio geométrico de pocas dimensiones las proximidades existentes entre un conjunto de objetos", (Guerrero y Ramírez, 2009).

A través de ella representamos sobre un mapa de 2 dimensiones el conjunto de casos y palabras clave para estudiar su posición relativa. Sin embargo los resultados obtenidos en el proceso no son lo suficientemente claros como para hablar de un escalamiento multidimensional revelador. El gran volumen de elementos que forman parte del análisis no permite representar la proximidad entre palabras frecuentes con precisión.

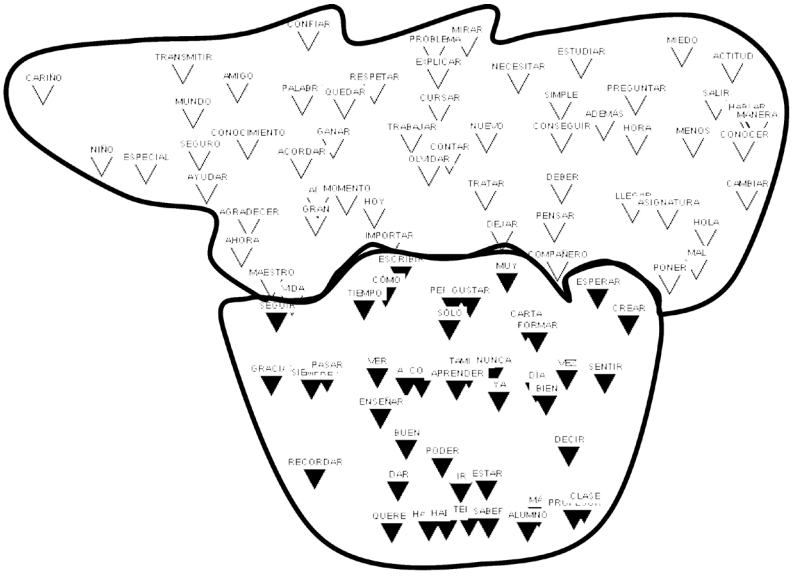


FIGURA 7
MAPA 2D PARA K = 2 Y N =100

Así pues, si rebajamos de 500 a 100 el número de términos basados en TF*IDF a representar, se produce una importante reducción del escenario gráfico, lo que mejora considerablemente la relación perceptible entre palabras clave.

Para *k = 2* grupos y *N= 100* términos *clave* obtenemos el gráfico de la Figura 7, con un valor $R^2 = 0{,}9054$ ($p<0{,}001$). En él, es difícil encontrar dos tópicos diferenciados con claridad. A medida que aumentemos el valor de *K* perdemos amplitud de conglomerados en favor de una mayor precisión temática intragrupos.

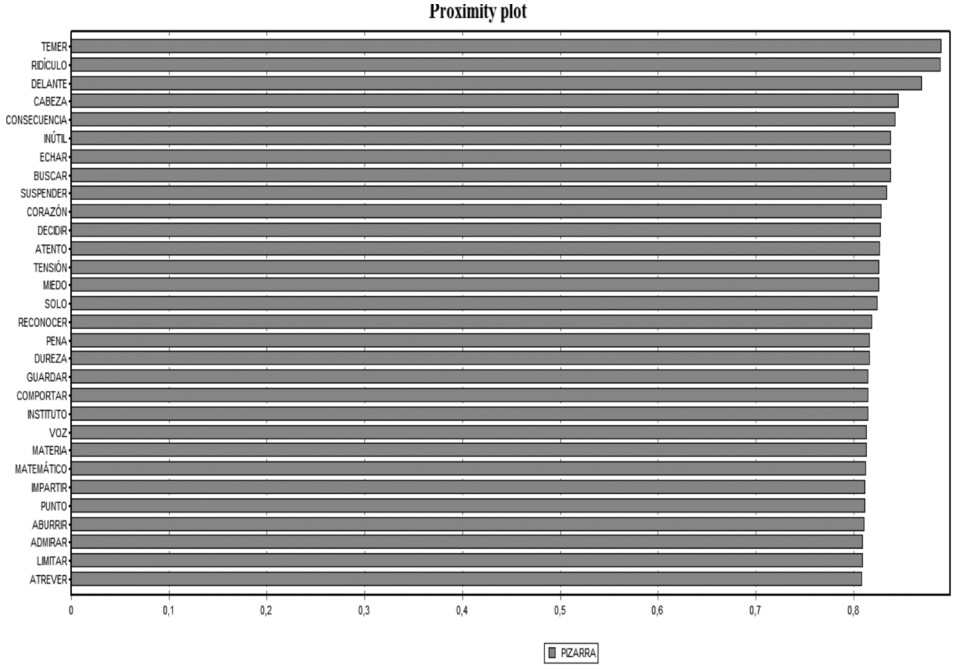
Si realizamos esta representación mediante un gráfico de burbujas, añadiremos, además, información relativa a la frecuencia de cada término en función del diámetro de cada círculo.

6.4. Gráfico de proximidad

Además del dendrograma, el gráfico de proximidades es otra salida del procedimiento clúster. Está basado en la matriz de distancias o similaridades entre casos que para nosotros se fundamenta en el *coeficiente de Sorensen*, medida de proximidad elegida.

Este gráfico representará la distancia entre una palabra clave y todas las demás. En general, lo que revela, es la escasa distancia existente entre la totalidad de los términos, lo cual respalda una vez más, la dificultad para constituir conglomerados bien diferenciados y unidades temáticas independientes.

Tomemos por ejemplo el término *pizarra*. El gráfico de proximidades resultante es el siguiente:



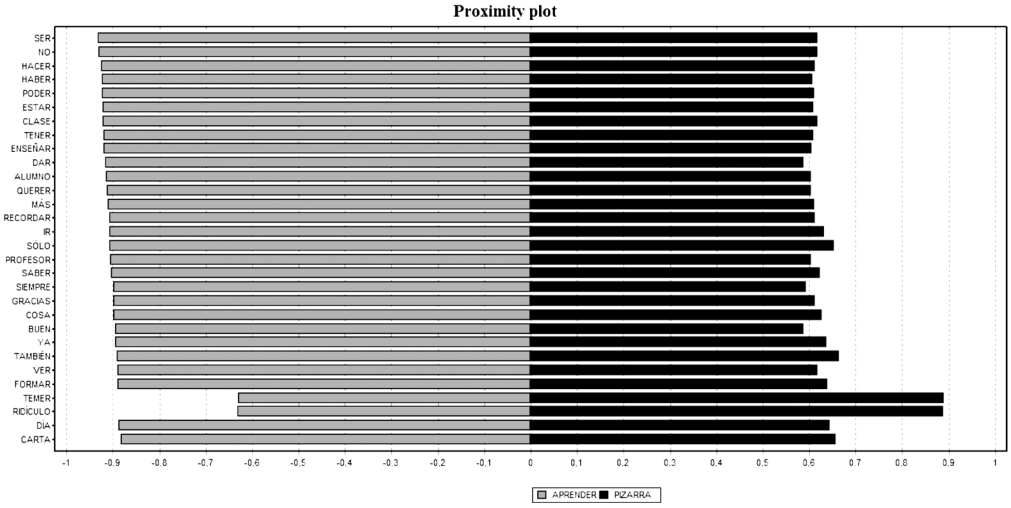
GRÁFICA 1
PROXIMIDADES DEL TÉRMINO PIZARRA

Observamos que son muchos los términos próximos a la palabra *pizarra*. Los tres más cercanos a ella son *temer*, *ridículo* y *delante.* Pero también superan el 0,8 de similaridad las palabras: *cabeza, consecuencia, inútil, echar, buscar, suspender, corazón, decidir, atento, tensión, miedo…*

TABLA 8
RECUPERACIÓN DE PALABRAS CLAVE EN CONTEXTO

***Miedo** a ser humillados **delante** de los compañeros, miedo a salir a la **pizarra** y sentimos **ridiculizados**,* (caso 75, carta negativa, amenaza/miedo).
*Nos hacía **temer** su **asignatura** y, sobre todo, a usted. Temíamos tremendamente hacer el **ridículo** en sus clases, algo, que por cierto, a usted le resultaba muy gracioso, y por eso **buscaba** la ocasión de dejamos malparados **delante** de los demás, sacándonos a la **pizarra**,* (caso 83, carta negativa, amenaza/miedo).

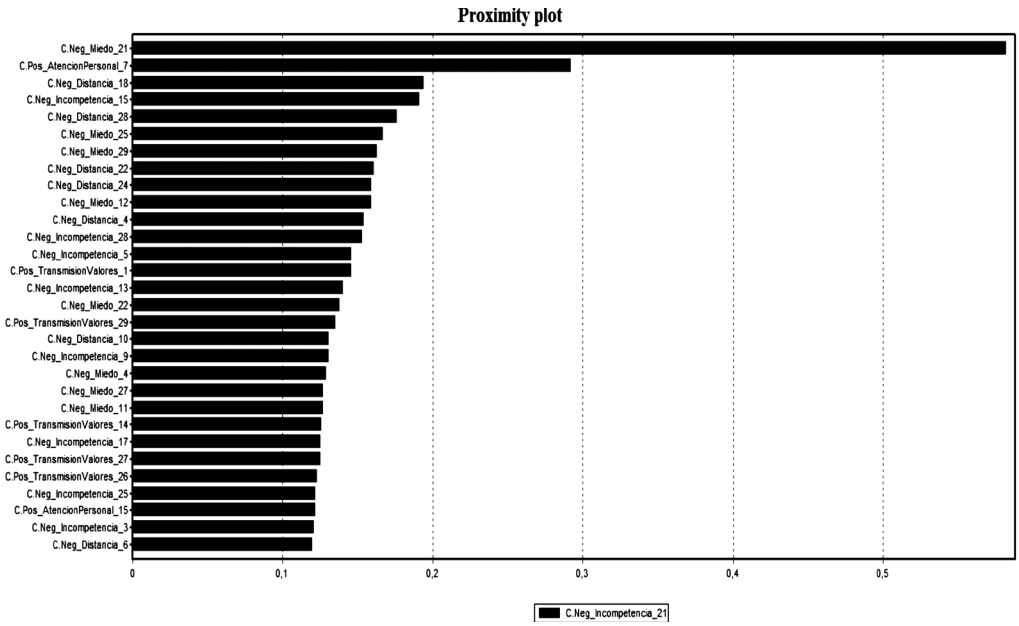
Realizamos, ahora, el cálculo de proximidades conjuntamente para los términos *pizarra* y *aprender.* De este modo podemos comparar las palabras que coocurren con uno y otro término.



GRÁFICA 2
PROXIMIDAD APRENDER - PIZARRA

Observamos que las palabras *temer* y *ridículo* tienen el mayor índice de proximidad con *pizarra* y sin embargo el menor índice con *aprender,* (Gráfica 2).

Si realizamos el gráfico de proximidades por casos en lugar de por términos clave, observamos que hay mayores valores de distancia entre unas cartas y otras que la encontrada entre palabras.



GRÁFICA 3
PROXIMIDAD DEL CASO 44

Vemos que la *Carta Negativa Incompetencia_21* (caso 44) tiene un índice de similaridad de *0,582* con la *Carta Negativa Miedo_21* (caso 74). Ello nos indica que abordan temáticas de suficiente afinidad como para conjeturar relaciones y ubicarlas con relativa proximidad.

TABLA 9
PROXIMIDAD DE CASOS

CASO 44	*CASO 74*
• *Sé que no soy buena para el dibujo,*	• *sigo sin saber dibujar, se me da fatal hacer un dibujo*
• *Recuerdo que una compañera habló con su madre , quien fue a quejarse*	• *Solías llamar a mi madre para enseñarle los dibujos tan espantosos que yo hacía*
• *si tú te hubieras molestado.*	• *¿Acaso alguien merece ese trato?*
• *alguien que tuvo la desgracia de ser alumna tuya.*	• *me siento incapaz de hacerlo. Todo eso te lo debo a ti.*

7. CONCLUSIONES

Verificar e inferir controlada, objetiva y sistemáticamente las producciones académicas a través de software especializado en el análisis de contenido, puede contribuir a una mayor detección de singularidades, a la mejor atención a la diversidad o a la eficacia comunicativa en el conjunto del centro y aula.

Sin embargo, la informática es una herramienta del proceso y no el todo de la investigación, es decir, entraña limitaciones y no aporta validez por sí sola aunque contribuye a la coherencia, agilidad y rigor en el registro y codificación de los datos.

Analizar solamente palabras frecuentes en los textos parece ser una condición necesaria, pero no suficiente, para llegar a profundizar en el contenido textual y esclarecer las relaciones y dependencias entre palabras clave, tópicos o casos.

Las listas de frecuencias desvelan repeticiones y evidencian pautas pero se requiere de una gran creatividad y trabajo teórico que lo cumplimente, ya que dichas listas por sí solas, no perfilan categorías ni confirman líneas temáticas precisas.

Para avanzar en el proceso de clarificación de datos y confirmar o desmentir si las unidades de contexto previamente establecidas (Tabla 1) pueden ser consideradas, o no, como un sistema de categorías, debemos, al menos, continuar el análisis a partir de los tres procesos siguientes:

1. Ampliar el análisis descriptivo de palabras a segmentos. Se ha percibido en las cartas un vocabulario común pero fuertemente acompañado de modificadores. Si bien las palabras no han discriminado suficientemente los tópicos, el aislamiento por segmentos nos dará mayor precisión sobre la opinión o actitud del hablante respecto al tema referido.
2. Comparar la lista de palabras frecuentes con dos importantes variables que contribuyen a describir nuestro texto:
   — La variable *signo*, nominal y que toma los valores *carta positiva* o *carta negativa.*
   — Y la variable *tópico*, también nominal y que comprende seis items distintos: *transmisión de valores, afecto-atención personal, motivación-estímulo, falta de profesionalidad-incompetencia, distancia-impersonalidad* y *amenaza-miedo.*
3. Utilizar el módulo *QDA Miner*, para etiquetar el texto en función a un sistema de categorías establecido. Torres y Perera (2009), construyeron uno de estos sistemas para foros de debate *online* con el fin de estudiar la comunicación asincrónica en la formación a través de Internet. Dado el contexto de los documentos que aquí se analizan, sería apropiado generar una categorización en función de valores y contravalores, y utilizarla para constituir un libro de código válido y fiable en la interpretación del contenido.
4. Finalmente, utilizar un diccionario de *WordStat* para *medir* conceptos específicos a través de un diccionario de categorización (latinRID[5]). Ello permitirá realizar un recuento de las ocurrencias de cada categoría en los documentos, así como conocer la composición de los mismos.

5 Latin RID es la versión en lengua hispana del Regressive Imagery Dictionary (RID).

En cualquier caso, el programa satisface la calidad investigadora que nos proponemos, con el añadido de proporcionarnos análisis cuantitativos gráficos que facilitan la interpretación adecuada para cada situación que se pueda plantear.

REFERENCIAS

Anguera, M. T. (2008). Metodologías cualitativas: Características, procesos y aplicaciones. En M. A. Verdugo, M. Crespo, M. Badía & B. Arias (Coord.), *Metodología en la investigación sobre discapacidad. Introducción al uso de las ecuaciones estructurales* (pp. 141-155). Salamanca, España: Instituto Universitario de Integración en la Comunidad, Universidad de Salamanca.

Bardin, L. (2002). *Análisis de contenido*. Madrid, España: Akal.

Cisneros Puebla, C. A. (2009). *QDA Miner. Software para Análisis Cualitativo de Datos. Guía del Usuario*. Canadá: Provalis Research. Recuperado de http://www.provalis-research.com /Documents/QDAMiner32ES.pdf

Clemente Díaz, M. (2003). *El analisis de contenido como tecnica de investigacion de la comunicacion social.* Recuperado de http://www.robertexto.com/archivo14/analisis.htm

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.) *Handbook of research on teaching* (119-161). New York, NY: McMillan.

García Guzmán, J. M. (1991). Los valores que promueve el Sistema Educativo, tal y como son percibidos por los agentes del mismo. En M. Muñoz-Repiso Izaguirre, J. M. Valle López, & J. L. Villalaín Benito (Comps.), *Educación y valores en España. Actas del seminario* (pp. 83-106). Madrid, España: MEC.

Gato, P. (2006). *Profesores que dejan huella*. Cáceres, España: Universidad de Extremadura.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Hingham, MA: Kluwer Academic Publishers.

Guerrero Casas, F. M., & Ramírez Hurtado, J. M. (2002). *El análisis de escalamiento multidimensional: una alternativa y un complemento a otras técnicas multivariantes*. Comunicación presentada a las X Jornadas de la Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa, Madrid (España). Recuperado de http://www.uv.es/asepuma/X/K11C.pdf

Holsti, O. R. (1968). Content analysis. En G. Lindzey, & E. Aronson (Eds.), *The handbook of social Psychology* [Vol. 2] (pp. 596-692). Reading, Inglaterra: Addison-Wesley.

León, J. A., Solari, M., Olmos, R., & Escudero, I. (2011). La generación de inferencias dentro de un contexto social. Un análisis de la comprensión lectora a través de protocolos verbales y una tarea de resumen oral. *Revista de Investigación Educativa, 29* (1), 13-42.

Llisterri Boix, J. (2003). Técnicas de procesamiento del lenguaje. En M. A. Martí Antonín (Coord.), *Tecnologías del lenguaje* (pp. 193-248) . Barcelona, España: UOC.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, England: Cambridge University Press.

Marín Ibáñez, R. (1993). *Los valores, un desafío permanente*. Madrid, España: Cincel.

Martindale, C., & Dailey, A. (1996). Creativity, primary process cognition, and personality. *Personality and Individual Differences, 20*, 409-414.

Miles, M. B., & Hubermas, A. M. (1984) *Qualitative data analysis: A source book*. Beverly Hills, CA: Sage

Torres, J. J., & Perera, V. H. (2009). Cálculo de la fiabilidad y concordancia entre codificadores de un sistema de categorías para el estudio del foro online en e-learning. *Revista de Investigación Educativa, 27* (1), 89-103.

West, A. N., & Martindale, C. (1988). Primary process content in paranoid schizophrenic speech. *Journal of Genetic Psychology, 149*, 547-553.

West, A. N., Martindale, C., & Sutton-Smith, B. (1985). Age trends in the content of children's spontaneous fantasy narratives. *Genetic, Social, and General Psychology Monographs*, *111*, 389-405.

Fecha de recepción: 4 de junio de 2011.
Fecha de revisión: 25 de julio de 2011.
Fecha de aceptación: 3 de febrero de 2012.