ESTUDIOS FILOLÓGICOS
DOI: 10.4067/s0071-17132018000200333

Un Enfoque de Reconocimiento de Patrones para el Análisis de Disponibilidad Léxica en Estudiantes de Pedagogía en Matemática

A Pattern Recognition Approach for the Analysis of Lexical Availability in Math Teaching Program Students

DARÍO ROJAS DÍAZ
CAROLINA ZAMBRANO MATAMALA
PEDRO SALCEDO LAGOS
MIGUEL FRIZ CARRILLO

Universidad de Concepción, Facultad de Educación, Chile.
Correo electrónico: dariorojas@udec.cl

Universidad de Concepción, Facultad de Educación, Chile.
Correo electrónico: carozambrano@udec.cl

Universidad de Concepción, Facultad de Educación, Chile.
Correo electrónico: psalcedo@udec.cl

Universidad del Bío Bío, Departamento de Ciencias de la Educación.
Correo electrónico: mfriz@ubiobio.cl

Los estudios de disponibilidad léxica permiten la caracterización y análisis del léxico disponible de un grupo de individuos. Los enfoques de análisis actuales principalmente se centran en la descripción mediante índices que resumen características grupales de los vocablos, sin ofrecer un análisis enfocado en los individuos. El objetivo del presente trabajo es proponer un modelo espacio-vectorial para la representación y análisis de los lexicones obtenidos de encuestas de disponibilidad léxica, permitiendo así, complementar estos estudios mediante un modelo matricial del corpus léxico basado en los modelos tradicionales de recuperación de información. La viabilidad del modelo propuesto se ha demostrado mediante la utilización de técnicas de reconocimiento de patrones como máquinas de soporte de vectores y clustering en un estudio de caso con alumnos de pedagogía de dos universidades chilenas. En esta investigación se concluye que es posible la utilización de estas técnicas para la visualización, agrupación y clasificación de estudiantes según su léxico disponible.

Palabras claves: Disponibilidad Léxica, Recuperación de Información, k-means, SVM, Educación

The lexical availability studies allow the characterization and analysis of the available lexicon of a group of individuals. Current analysis approaches focus mainly on index descriptions that summarize group characteristics of the words, without offering an analysis focused on individuals. The objective of the present work is to propose a space-vector model for the representation and analysis of the lexicons obtained from lexical availability surveys, allowing to complement the lexical availability studies, through a model of lexical corpus-based representation based on the traditional information retrieval models. The feasibility of the proposed model has been demonstrated through the use of pattern recognition techniques such as vector support and clustering machines in a case study of lexical availability in teaching students from two Chilean universities. In this research it is concluded that it is possible to use these techniques for the visualization, grouping and classification of students according to their available lexicon.

Key words: Lexical Availability, Clustering, k-means, SVM, Education

1. Introducción

En cualquier ambiente de enseñanza-aprendizaje se lleva a cabo un proceso de comunicación que involucra tanto a los individuos como a los instrumentos utilizados. Este proceso ocurre comúnmente durante la comunicación alumno-profesor, a través de la interacción pedagógica, permitiendo al docente actuar como elemento mediador clave y aclaratorio del aprendizaje (Casado 2002). En una interacción comunicativa de este tipo, los estudiantes utilizan de forma natural cierto conjunto de palabras o 'vocabulario' que les permite comunicarse con sus pares y profesores. Sin embargo, debido a que la interacción pedagógica ocurre en un contexto específico y especializado de comunicación, es necesario que el léxico del alumno sea acorde a la situación para que, mediante la estructura conceptual que estos representan, puedan realizar una reelaboración progresiva de su significado (Osses & Jaramillo 2008). Por lo tanto, si el 'vocabulario' utilizado por los actores del proceso de comunicación es deficiente o no es concordante, la interacción pedagógica podría llevarse a cabo de forma deficiente.

La caracterización del léxico disponible, que posee un individuo o grupo en una situación o contexto específico, se puede lograr a través del análisis de la disponibilidad léxica (Ferreira *et al*. 2014). Este tipo de análisis permite describir el léxico disponible de un grupo de individuos, principalmente mediante el cálculo de estadísticos grupales que indican, entre otras cosas, cuáles son los vocablos más disponibles para el grupo en torno a un centro de interés (contexto). No obstante, a pesar de que la mayor parte de este análisis se realiza sobre la base de la obtención del lexicón de cada individuo, los resultados que se logran están principalmente orientados a la caracterización de los grupos y los vocablos en

su conjunto. Esto último provoca que el proceso de caracterización, de nivel individual, se vea dificultado cuando existe una gran cantidad de sujetos en el análisis, producto de que no hay herramientas de procesamiento masivo que estén enfocadas al análisis del lexicón individual, pudiendo ser este a veces tan o más importante para estudios como los que tratan sobre la interacción pedagógica.

El presente trabajo propone un modelo de análisis espacio-vectorial para representar la información recogida mediante encuestas de disponibilidad léxica en torno a uno o más centros de interés. El enfoque propuesto permite realizar, en forma eficiente, análisis cuantitativos mediante técnicas de reconocimiento de patrones, que permiten abordar el lexicón individual y sus vocablos bajo una misma representación.

La estructura del artículo es la siguiente: primero se realiza una breve introducción al análisis de disponibilidad léxica y los índices de caracterización grupal más utilizados, en seguida se hace una formulación del modelo vectorial propuesto, luego se desarrolla un estudio de caso con el objetivo de mostrar nuevos enfoques de análisis habilitados por el modelo, finalizando con las conclusiones y alcances del trabajo realizado.

2. El análisis de disponibilidad léxica

Los estudios de la disponibilidad léxica nacen de un proyecto de la UNESCO en la década de los cincuenta, en el cual se perseguía facilitar el aprendizaje de la lengua francesa a los habitantes no nativos de Francia, buscando así, facilitar la integración social a partir de la simplificación de una lengua base en común (Michea 1953). La disponibilidad léxica refleja el caudal léxico utilizado en una situación comunicativa, donde aparecen ciertas palabras que, siendo muy usadas, están estrechamente relacionadas con conceptos aparecidos en tales interacciones de comunicación (López 1993).

Para obtener el léxico disponible de un grupo se realizan pruebas de asociación con centros de interés, las que dan como resultado el vocabulario activo de los sujetos en torno a esos contextos. Bajo estas premisas, este tipo de estudios supone que son más disponibles aquellas palabras que primero se recuperan de la memoria ante un estímulo. Así, una de las formas más comunes de obtener el léxico disponible de un conjunto de personas respecto de un centro de interés, involucra solicitar a los participantes escribir, de forma ordenada, las palabras que primero asocien con un centro de interés, considerando comúnmente un tiempo límite para realizar tal tarea.

El diccionario mental obtenido por tal procedimiento puede considerarse parte del lexicón mental del individuo (Ferreira *et al*. 2014). Este lexicón es permeable al entorno de desarrollo de las personas y, por lo tanto, puede variar en el tiempo. Sin embargo, este tipo de encuestas requieren de recursos materiales y tiempo de post-procesamiento considerables, lo que ha conducido a otros autores a proponer softwares especializados como LexMath (Salcedo *et al*. 2015; Del Valle *et al*. 2016), que permiten, utilizando Internet, controlar los tiempos exactos de la prueba y realizar el pre-procesamiento necesario para el análisis de las

respuestas, incluyendo la corrección ortografía y el cálculo automático de los estadígrafos léxicos más utilizados.

Un enfoque distinto de análisis de disponibilidad léxica es propuesto por (Echeverría *et al*. 2008). En este se presenta un programa computacional que caracteriza las encuestas de disponibilidad léxica mediante grafos, representando las palabras con nodos y simbolizando a través de aristas las relaciones entre ellas. De esta forma, los grafos pueden ser interpretados como redes semánticas que expresan las relaciones semánticas subyacentes en el corpus, lo que permite obtener una caracterización conexionista de los vocablos pertenecientes al léxico disponible de un grupo de individuos.

Respecto a los resultados obtenidos en los análisis de disponibilidad léxica, principalmente en Latinoamérica y España (López 1993; Germany & Cartes 2000; Urzúa *et al*. 2006; Valencia 2010; Ferreira *et al*. 2014), es común la utilización de cinco índices que permiten determinar la riqueza léxica de los grupos de sujetos y disponibilidad de los vocablos. Estos son:

- **Número de Lexicones (*N*):** corresponde a la cantidad de individuos considerados en un grupo, por lo que también representa la cantidad de lexicones en un centro de interés.
- **Promedio de Respuestas (*XR*):** indica cuántos vocablos en promedio ha respondido un grupo de individuos ante un estímulo o centro de interés específico.
- **Numero de Palabras Diferentes (*NPD*):** es un índice que da cuenta del total de vocablos que produce un grupo de sujetos para algún centro de interés, considerando cada palabra una única vez (a pesar de, posiblemente, haber sido indicada por más de un individuo).
- Índice de Cohesión (***IC***)**:** es un indicador del grado de coincidencia en las respuestas de los individuos. Este permite analizar comparativamente el léxico de un centro de interés en distintos grupos de sujetos, estableciendo cuan similares son las respuestas en cada uno de ellos. Se puede considerar como el grado de homogeneidad de un grupo respecto a su léxico disponible. El índice IC se obtiene del cociente entre el promedio de respuestas del grupo dividido por el total de palabras diferentes (XR/ NPD).
- Índice de Disponibilidad Léxica (***IDL*)**: corresponde al grado de disponibilidad de un vocablo en el grupo, ante un estímulo o centro de interés. La forma más común de calcular este índice es utilizando la siguiente ecuación:

$$idl(v_i) = \frac{f_1\lambda^0 + f_2\lambda^1 + f_3\lambda^2 + \cdots + f_p\lambda^{p-1} + \cdots + f_t\lambda^{t-1}}{N}$$

donde f_p es la frecuencia de aparición del vocablo v_i en la posición p, siendo $p = 0$ cuando el vocablo es indicado en primera posición y $p = t$ cuando el vocablo es indicado en la última posición t, con $t > 0$. La expresión λ^{p-1} es denominada tasa de

sustitución o factor de ponderación a la posición (Salcedo *et al*. 2013), y su valor va decreciendo a medida que la posición es mayor.

Por consiguiente, generalmente las investigaciones sobre análisis de disponibilidad léxica incluyen estadígrafos con los índices indicados anteriormente, incluyendo además listados con las palabras ordenadas según su IDL para cada centro de interés.

Según lo indicado anteriormente y considerando que cada individuo posee un sistema interno de representación distinto, el cual es influenciado por algunos factores internos y externos que no son controlables por las pruebas de disponibilidad léxica, es posible que el computo de los índices grupales presentados pueda hacer que se pierda información cuantitativa producto de la naturaleza resumida de estos. De esta forma, una caracterización vectorial del léxico disponible que, sin resumir la información, permitiera su comparación y agrupación, podría ser de gran ayuda en este sentido. Además, una caracterización de este tipo podría soportar en forma conjunta la representación de vocablos y lexicones, permitiendo medir cuantitativamente las relaciones entre estos.

Sobre la base de lo expuesto, el objetivo del modelo propuesto en este artículo es permitir la representación de la disponibilidad léxica para uno o varios centros de interés, con tal de facilitar la comparación de vocablos y lexicones en forma cuantitativa, otorgando adicionalmente un método de comparación objetivo entre individuos y grupos, el cual puede servir de complemento a los índices grupales existentes.

3. Modelo espacio-vectorial propuesto

Como ya se ha mencionado, la principal forma de analizar la disponibilidad léxica de los individuos es a través de los índices *IDL*, *N, NPD, XR* e *IC*. El presente trabajo propone un modelo espacio-vectorial para los lexicones basado en las matrices de frecuencia término-documento utilizadas comúnmente para representar palabras y documentos en los sistemas de recuperación de información (Salton *et al*. 1975; Soumen 2002). Este enfoque es la base para caracterizar corpus textuales de gran tamaño, incorporando en su formulación no sólo la frecuencia de palabras, sino que también un esquema de ponderación que permite considerar las particularidades de la información en su conjunto.

Al igual que los modelos de representación textual, el modelo que se propone está basado en la creación de una estructura de corpus léxico para modelar a los centros de interés. Cada centro está compuesto por un conjunto de vectores que caracterizan a los lexicones de cada individuo, los que a su vez están formados por los vocablos producidos por cada sujeto en la prueba de disponibilidad léxica.

En nuestro enfoque, un centro de interés se define como una matriz pesada M, donde las columnas representan los lexicones y las filas todos los vocablos producidos. En este corpus, la representación incluye un esquema de pesos asignados a cada elemento de la matriz que caracteriza el par vocablo-lexicón de acuerdo a la siguiente función:

$f(i,j) = h(i,j) * g(i)$, donde $h(i,j)$ es una función de peso local para el vocablo en el lexicón l_j y $g(i)$ una función de peso global para el vocablo v_i en el corpus.

Por un lado, el esquema de pesos de esta matriz define una función de peso local que permite establecer ponderaciones basadas en la frecuencia de las palabras dentro de un documento. Así, generalmente se tiende a definir funciones que otorgan pesos más grandes a las palabras más frecuentes, como una forma de diferenciar la importancia que tienen las palabras para ser consideradas representantes de dicha información textual. Debido a la naturaleza de los lexicones, estos no tienen palabras repetidas (si existen, son eliminadas en el post-procesado de la encuesta), ya que están compuestos por vocablos. Por ello, los únicos valores de frecuencia que puede tener un vocablo en un lexicón son: 0 cuando el vocablo no está presente, y 1 cuando forma parte de este. En consecuencia, siempre será un vector con valores binarios, por lo que una función de peso basada en la frecuencia local de los vocablos no tiene utilidad en nuestro modelo, lo que nos lleva a definirla como una función constante $h(i,j) = 1$.

Por otro lado, las funciones de peso global permiten asignar ponderaciones a las palabras considerando su contexto en el corpus completo, es decir, se considera la contabilización de las palabras según los documentos en que aparecen. En este sentido, estas funciones asignan pesos mayores a las palabras poco frecuentes y pesos menores a las palabras muy frecuentes con tal de equilibrar las ponderaciones locales con las globales y obtener una representación balanceada en los valores de la matriz. El objetivo de esto último es tener en cuenta las palabras que, a nivel global, son poco frecuentes, ya que la presencia de estas en ciertos documentos debe ser considerada de mayor importancia (con una alta ponderación), a causa de que su baja frecuencia permite distinguir y representar de mejor forma a los textos que pertenece. Así mismo, los modelos vectoriales tradicionales incorporan regularmente alguna variante de la función denominada "función de frecuencia inversa de documento" o IDF (por las siglas en inglés de Inverse Document Frecuency) (Chen *et al*. 2016). Esta función, definida como $IDF(p_i) = \log(N/f_{pi})$, está definida sobre la base del cociente entre la cantidad de documentos (N) y la frecuencia f_{pi}, de una palabra cualquiera p_i en el corpus total, contabilizándola solamente una vez por documento. De esta forma, cuanto mayor sea la presencia de la palabra en los distintos documentos, menor será el valor de IDF. Por el contrario, si la palabra aparece en pocos de ellos, entonces mayor será el valor de IDF, indicando que la palabra es una mejor representante a nivel global de dichos textos.

Las propiedades de la función de peso global resultan útiles para el modelo propuesto. En forma análoga, los vocablos menos frecuentes en un centro de interés pueden representar de mejor forma a los lexicones que pertenecen. Si además de la frecuencia, se considera a la posición en la que la palabra es producida globalmente, entonces el índice IDL puede ser considerado una función de peso global que incorpora esos dos criterios para ponderar la importancia de los vocablos en el contexto del centro de interés.

En este mismo ámbito, se puede definir una función de peso global basada en el IDL de los vocablos, siendo esta función la que incorpora la información referente al orden

de aparición de los vocablos al modelo propuesto. Así, la función global de peso utilizada es definida como:

$$g(v_i, l_j) = \begin{cases} 0 & , si\ v_i \notin l_j \\ 1 - idl(v_i) & , si\ v_i \in l_j \end{cases}$$

donde $g(v_i, l_j)$, es el peso del vocablo v_i en el lexicón l_j, siendo este igual a cero cuando no está contenido en el lexicón l_j. La Figura 1, muestra un esquema de la conformación de la matriz según la definición dada del corpus léxico *M*.

Lexicones por individuo

M= *Todos los vocablos*

	$\boldsymbol{l_1}$	$\boldsymbol{l_2}$	$\boldsymbol{l_j}$	⋯	$\boldsymbol{l_N}$
$\boldsymbol{v_1}$	$f(v_1, l_1)$	$f(v_1, l_2)$	$f(v_1, l_j)$	⋯	$f(v_1, l_N)$
$\boldsymbol{v_2}$	$f(v_2, l_1)$	$f(v_2, l_2)$	$f(v_2, l_j)$	⋯	$f(v_2, l_N)$
$\boldsymbol{v_i}$	$f(v_i, l_1)$	$f(v_i, l_2)$	$f(v_i, l_j)$	⋯	$f(v_i, l_N)$
⋮	⋮	⋮	⋮	⋱	⋮
$\boldsymbol{v_{NPD}}$	$f(v_{NPD}, l_1)$	$f(v_{NPD}, l_2)$	$f(v_{NPD}, l_j)$	⋯	$f(v_{NPD}, l_N)$

Figura 1. Representación matricial de Lexicones

Para ejemplificar lo anterior, se puede considerar el conjunto de lexicones (artificiales) pertenecientes al centro de interés "colores" que son expuestos en el Cuadro 1.

Cuadro 1. Centro de interés "colores" de ejemplo.

Lexicón	Vocablos
l_1	rojo, azul, verde, amarillo, naranja, café
l_2	azul, rojo, verde, lila, negro
l_3	rojo, rosado, azul, blanco, plateado
l_4	blanco, rojo, violeta, amarillo

El IDL para cada vocablo del Cuadro 1 se presenta en el Cuadro 2, con el cual se ha obtenido la matriz del corpus mediante el esquema de pesos propuesto. El Cuadro 3 muestra cómo los vocablos menos disponibles tienden a tener un peso global más alto, mientras que los más disponibles tienen un peso global más bajo, aunque se debe considerar que estos últimos tienen mayor presencia en varios lexicones a la vez. Si se realiza sobre la matriz del Cuadro 3 una reducción de dimensiones, mediante un análisis de componentes principales, el resultado de este proceso se puede apreciar gráficamente en la Figura 2.

Cuadro 2. IDL para cada vocablo del ejemplo.

Vocablo	IDL
rojo	.95
azul	.68
blanco	.43
verde	.41
amarillo	.36
rosado	.23
violeta	.20
lila	.18
naranja	.16
negro	.16
plateado	.16
café	.15

Cuadro 3. Matriz pesada para el centro de interés “colores”

Vocablo	$\overline{l_1}$	$\overline{l_2}$	$\overline{l_3}$	$\overline{l_4}$
rojo	.05	.05	.05	.05
azul	.32	.32	.32	0
verde	.60	.60	0	0
amarillo	.64	0	0	.64
naranja	.84	0	0	0

café	.85	0	0	0
lila	0	.82	0	0
negro	0	.84	0	0
rosado	0	0	.78	0
blanco	0	0	.57	.57
plateado	0	0	.84	0
violeta	0	0	0	.80

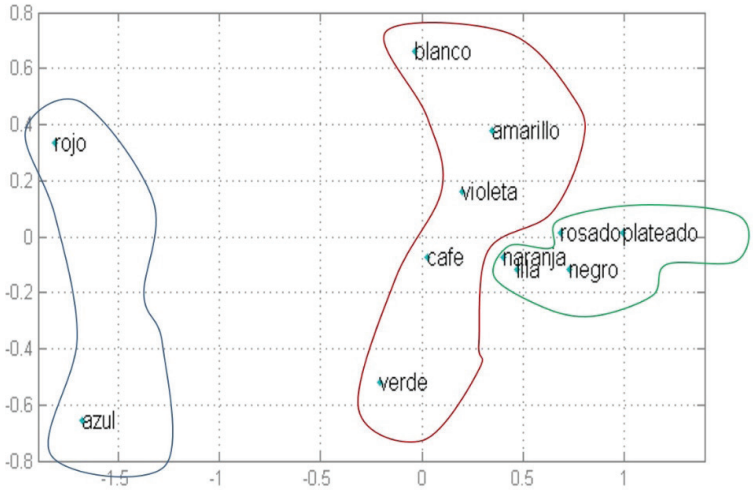

Figura 2. Representación gráfica del centro de interés "colores" obtenida mediante el modelo propuesto y la agrupación generada por el algoritmo de clustering.

Otra característica del modelo vectorial es que, debido a que la representación se produce en un espacio vectorial multidimensional, es posible aplicar técnicas de reconocimiento de patrones que tengan en cuenta la disposición espacial de los datos. Un ejemplo de estas técnicas corresponde a los algoritmos de clustering, los que permiten, en forma no supervisada, encontrar agrupaciones de los datos. Así, mediante la aplicación del algoritmo de clustering k-medias sobre la matriz original (no reducida), se puede determinar la existencia de tres grupos (k=3): {rojo, azul}, {verde, amarillo, café, blanco, violeta} y {naranja, lila, negro, rosado, plateado}. Estos grupos son mostrados en la Figura 2 a través de los contornos cerrados azul, rojo y verde respectivamente.

Las técnicas de clustering pueden resultar útiles en el análisis de grandes volúmenes de información y, por lo tanto, en muestras grandes de lexicones, permitiendo encontrar agrupaciones de acuerdo con la naturaleza del corpus y teniendo en cuenta la disponibilidad de los vocablos. En este mismo contexto, la evaluación de la calidad de la agrupación, en el sentido de cuán similares son los miembros en grupo (alta cohesión) y considerando a la vez cuán diferentes son los grupos entre sí, puede ser medida por los "índices de calidad de clustering", los cuales otorgan una valorización relativa a la calidad de las agrupaciones, permitiendo, por ejemplo, comparar la calidad de las agrupaciones de los vocablos o lexicones como una medida complementaria a los índices de cohesión. En este sentido, evaluar la calidad de la agrupación para distintas configuraciones de grupos en un mismo centro de interés, permite, de forma experimental, encontrar objetivamente cual es la cantidad óptima de subgrupos de elementos que se encuentran presentes en el corpus modelado.

Otra característica del modelo propuesto consiste en la posibilidad de utilizar técnicas de clasificación supervisadas como las máquinas de soporte de vector SVM (por las siglas de su nombre en inglés Support Vector Machine). La clasificación de patrones trata de resolver cómo clasificar algún objeto en una categoría definida a priori. Por lo tanto, el clasificador es un algoritmo que, a través de la definición de un vector de características para cada elemento, aprende a clasificar en base a ejemplos dados, decidiendo en forma automática a que clase o categoría pertenece dicho vector. Por ejemplo, se puede crear un modelo vectorial del léxico disponible sobre un centro de interés y determinar el género del individuo que produjo cada lexicón. De esta forma, dando una serie de ejemplos al clasificador con lexicones y el género del individuo que lo produjo, este puede aprender a clasificar un lexicón en una categoría de género, incluso para casos desconocidos de lexicones que no se le han enseñado. Aunque las técnicas y métodos para alcanzar dicho aprendizaje pueden ser abordados mediante múltiples enfoques, nuestro trabajo utiliza las SVM debido a la particularidad de ser una de las técnicas de mayor utilización en la clasificación de documentos (Martínez *et al*. 2011).

En forma resumida, el modelo propuesto permite complementar los análisis de disponibilidad léxica a través de las siguientes características:

- ***Contiene los índices más comúnmente utilizados***: los índices *IDL, N, NPD, XR* e *IC,* pueden ser calculados directamente desde la matriz.
- ***Contiene una representación de los vocablos y los lexicones***: Se puede utilizar para analizar a los vocablos y a los individuos de forma indistinta según las necesidades.
- ***Permite la comparación entre centros de interés***: es posible combinar las matrices para generar una representación conjunta de más de un centro de interés a la vez.
- ***Se puede incorporar información adicional a los individuos o vocablos***: se puede incorporar a la matriz información adicional como elementos demográficos o resultados de otros estudios para generar una representación

más completa de los individuos al utilizar técnicas de análisis multidimensional.
- ***Puede ser visualizada en forma espacial***: Se puede obtener una proyección en distintas dimensiones que permite visualizar tanto a las palabras como a los lexicones espacialmente.
- ***Permite el análisis directamente a través de técnicas de reconocimiento de patrones***: Se pueden utilizar técnicas derivadas del reconocimiento de patrones como clasificadores, algoritmos de clustering, y cualquier otra técnica que pueda operar con una representación vectorial de datos.

Con el objetivo de comprobar la utilidad del modelo propuesto, se presenta a continuación un caso de estudio de disponibilidad léxica en estudiantes de dos universidades chilenas.

4. Estudio de caso

4.1. Objetivo

El objetivo del estudio de caso es demostrar algunas de las posibilidades de análisis que ofrece el reconocimiento de patrones mediante el modelo propuesto. Precisamente por ello, no se busca un análisis del léxico disponible en detalle, como tradicionalmente se realizaría al intentar concluir o formular hipótesis de su conformación, sino que más bien, se busca presentar distintos tipos de resultados que pueden complementar los análisis tradicionales de disponibilidad léxica existentes.

4.2. Objetivos específicos

Mediante el modelo propuesto se plantean los siguientes objetivos específicos:

- Realizar una representación espacial de lexicones y vocablos con tal de poder visualizar sus similitudes y distribución.
- Demostrar la utilización de técnicas de reconocimiento de patrones para agrupar los lexicones según su similitud en forma automática, permitiendo la caracterización de subgrupos.
- Demostrar la utilización de técnicas de reconocimiento de patrones para clasificar y potencialmente predecir características de los alumnos a partir de sus lexicones.

4.3. Encuesta de disponibilidad léxica

Para nuestro estudio de caso se realizó la prueba de disponibilidad léxica para los alumnos de dos universidades chilenas que pertenecen a la carrera de Pedagogía en

Matemática y Computación de la Universidad de Concepción (UdeC), y a la carrera de Pedagogía en Matemática de la Universidad del Bío-Bío (UBB), abarcando en ambos casos alumnos de primero a cuarto año. El instrumento de recolección de datos utilizado corresponde a la prueba de disponibilidad léxica empleada por (Valencia & Echeverría 1999), y los centros de interés seleccionados corresponden a las cinco agrupaciones temáticas de los 21 estándares disciplinares que determina el Ministerio de Educación de Chile para la enseñanza de las matemáticas en educación media (MINEDUC Chile 2012): Datos y Azar, Sistemas Numéricos y Algebra, Cálculo, Estructuras Algebraicas, y Geometría.

Se debe puntualizar que los vocablos han sido transformados en el proceso de transcripción digital, eliminando tildes intencionalmente, evitando así, que el procesamiento computacional determine diferencias artificiales entre estos, producto de errores ortográficos de los sujetos en sus respuestas. De esta forma, en las secciones donde se presentan los resultados del estudio de caso, los vocablos son presentados en cursiva y sin tildes.

4.4. Estadígrafos

El Cuadro 4 muestra los estadígrafos de disponibilidad léxica por cada centro de interés. Como se puede apreciar, el centro de interés Estructuras Algebraicas es el que tiene una menor cantidad de vocablos por lexicón en promedio. Al contrario, el centro de interés Geometría es el que más vocablos por respuesta posee, lo que se ve reflejado en la cantidad de palabras distintas que este centro tiene. Por otro lado, comparando Datos y Azar con Geometría, se puede ver que a pesar de que Datos y Azar tiene una cantidad considerablemente mayor de palabras distintas, Geometría igualmente lo supera en el promedio de vocablos por respuesta, lo que influye en su mayor índice de cohesión, indicando a la vez, una mayor homogeneidad en el léxico disponible de los alumnos respecto a este último tema.

Cuadro 4. Estadígrafos por centro de interés

Centro de Interés	N	XR	NPD	IC
Datos y Azar	126	12.77	483	.0264
Cálculo	126	13.78	529	.0260
Estructuras Algebraicas	123	10.16	372	.0273
Geometría	126	19.94	423	.0471
Sistemas Numéricos	126	13.05	455	.0287

4.5. Representación espacial de los centros de interés

Como ya se ha mencionado, una de las ventajas de obtener una representación vectorial de los lexicones, es que estos pueden ser visualizados espacialmente. Con tal de observar los centros de interés en estudio, se obtuvo una versión reducida a tres dimensiones, mediante un análisis de componentes principales (PCA) de un único corpus léxico compuesto por todos los centros de interés.

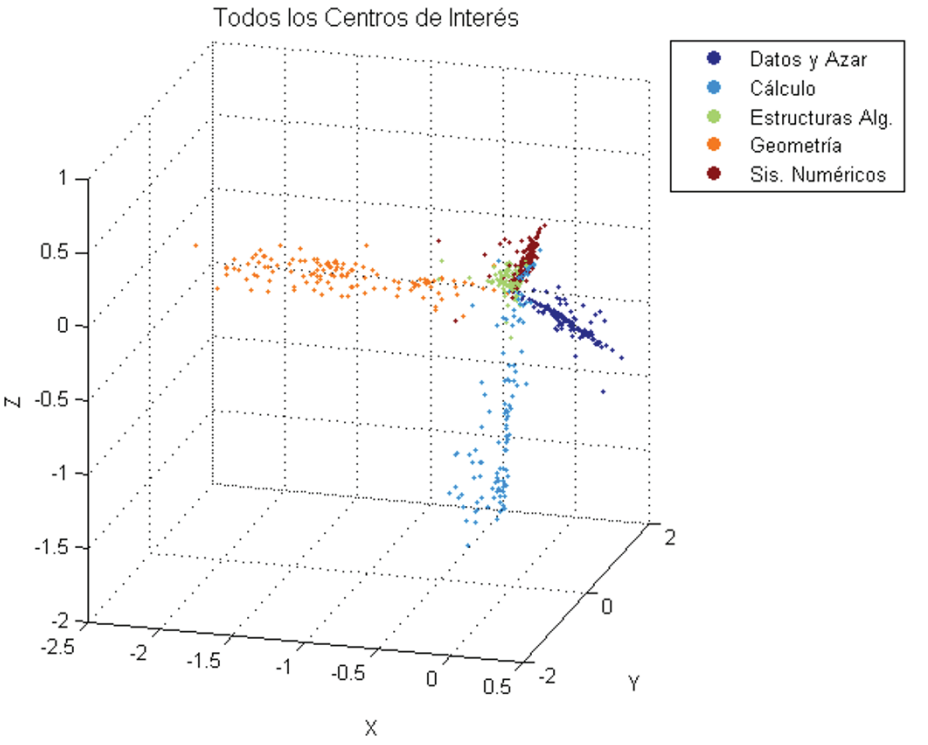

Figura 3. Representación espacial de los individuos de todos los centros de interés

En la Figura 3 se puede observar la agrupación espacial de los lexicones según su centro de interés, siendo Datos y Azar junto a Geometría los más distinguibles, mientras que Estructuras Algebraicas y Sistemas Numéricos están más cercanos espacialmente entre sí, tendiendo sus elementos a ubicarse en la misma distribución. La mayor similitud en estos centros de interés se debe principalmente al subconjunto de vocablos que poseen en común, lo que produce que los vectores que representan a estos lexicones se asemejen numéricamente.

4.6. Representación espacial de los vocablos e índices de disponibilidad léxica

Los estudios de disponibilidad léxica generalmente presentan tablas con el listado de los vocablos con más alto índice de disponibilidad. Complementariamente a esto, a través del modelo propuesto, se puede obtener la transpuesta de la matriz original del corpus léxico. De esta forma, y reduciendo a dos dimensiones el resultado de la transposición, se puede obtener una representación espacial de los vocablos de cada centro de interés junto a su índice de disponibilidad. En este mismo contexto, la Figura 4 muestra la distribución espacial para cuatro de los cinco centros de interés y sus 25 vocablos de mayor disponibilidad según su *IDL,* indicando su orden como superíndice.

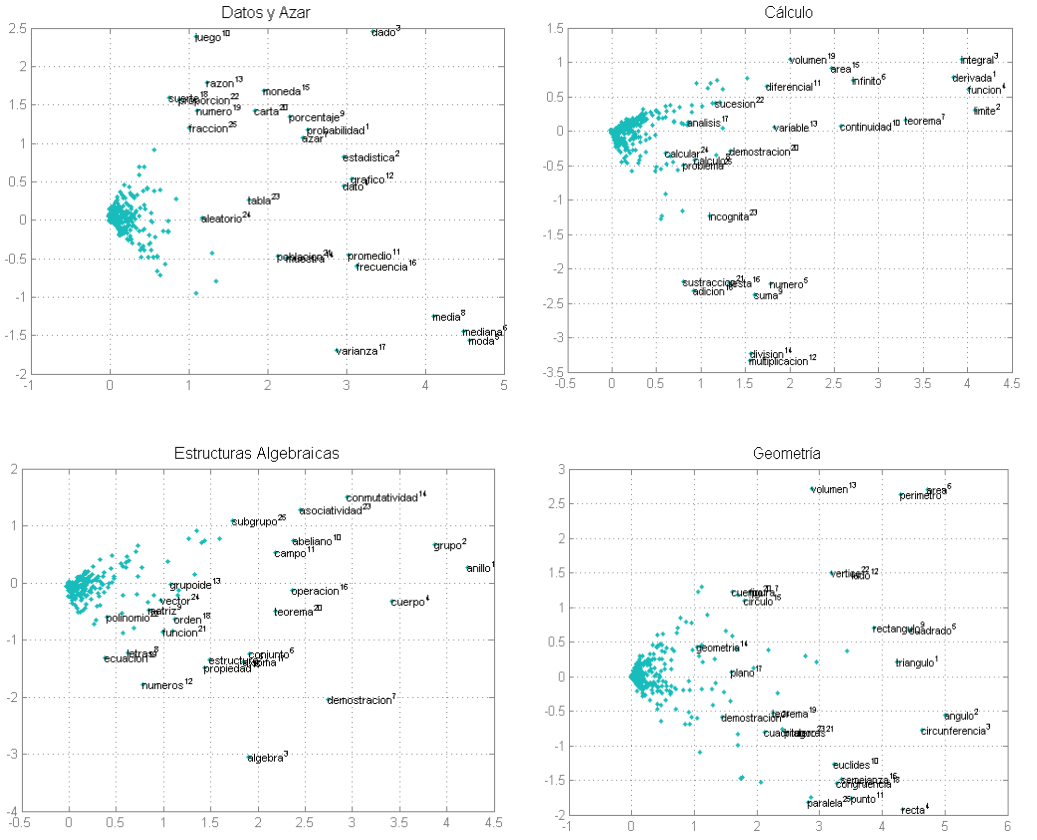

Figura 4. Distribución espacial de los vocablos para los centros de interés Datos y Azar, Cálculo, Estructuras Algebraicas y Geometría.

Por una parte, se puede ver en la Figura 4 que en el centro de interés Datos y Azar, existen vocablos de alta disponibilidad como *carta, dado, juego* y *suerte.* Estos vocablos pueden asociarse más a juegos de azar que al lenguaje técnico propio de dicha temática. Por otra parte, en los demás centros de interés, los vocablos son más propios de cada área temática y es común ver al menos una de las palabras que componen el nombre del centro de interés, tales como *azar*, *calculo*, *algebra*, *geometria* y *numero*. Adicionalmente, aparecen vocablos como *probabilidad, funcion, anillo y reales* que tienen directa relación con los contenidos y el vocabulario específico de cada área.

La Figura 4 también permite apreciar la cercanía espacial que poseen ciertos vocablos de una misma temática. Por ejemplo, para el centro de interés Cálculo, los vocablos *suma, resta, division* y *multiplicacion,* aparecen aglomerados al igual que *volumen* y *area* (a pesar de que su orden de disponibilidad no es consecutiva). Así mismo, en el centro de interés Sistemas Numéricos, también se puede observar la agrupación de los mismos vocablos *suma, resta, division* y *multiplicacion*, demostrando la falta de especificidad del léxico disponible de los alumnos para diferenciar lo que son el uso de las operaciones elementales en la práctica del cálculo y la definición propia de estas operaciones en la teoría. Por último, en el centro de interés Geometría, también contiene algunas agrupaciones de vocablos representativos a través de los grupos conformados por *volumen, perimetro* y *area*, y por el grupo *rectangulo* y *cuadrado*.

4.7. Representación espacial de los lexicones

El modelo propuesto permite asociar el lexicón de los estudiantes con otros datos. En este aspecto, la Figura 5 muestra algunas de estas asociaciones mediante la etiquetación de las respuestas de los alumnos. Hay que puntualizar que, con tal de no introducir un sesgo en la distribución espacial de los lexicones, no se ha incorporado información adicional a las matrices, ya que de haberlo hecho la representación resultante tendería a separar los elementos de acuerdo con la característica introducida, lo que podría ser útil para otros tipos de análisis. Sin embargo, lo que se busca es obtener una representación "sólo" de los lexicones y poder observar, según su propia conformación, cómo estos se distribuyen espacialmente según el modelo. Para estos efectos, en la generación de la Figura 5, sólo se han utilizado las características adicionales para etiquetar las respuestas, no considerándolas para la transformación de la matriz.

Como se puede observar en la Figura 5, en el centro de interés Datos y Azar se han etiquetado los estudiantes provenientes de la Universidad del Bío-Bío (UBB) y los estudiantes de la Universidad de Concepción (UdeC). En esta se puede ver que existe un nivel más alto de agrupación en los alumnos de la UBB, mientras que los estudiantes de la UdeC presentan una mayor dispersión en sus lexicones, lo que indica una mayor homogeneidad en las respuestas de la UBB con respecto a la UdeC desde el punto de vista espacial (similitud de las respuestas según el modelo). Para el centro de interés Cálculo se ha etiquetado el nivel de los estudiantes, correspondiente a la cantidad de años que

ha permanecido en la universidad. Se puede notar que, a medida que aumentan los años en la universidad, los lexicones disminuyen su dispersión. Por ejemplo, los años Tres y Cuatro claramente están agrupados en el área superior derecha, mientras que los años Uno y Dos están más dispersos por todo el espacio con una tendencia hacia el área inferior izquierda. Esto último puede ser causado porque el avance en sus estudios hace que su léxico disponible sea más homogéneo, a causa del natural aprendizaje de las materias propias del área del Cálculo.

Para el centro de interés Estructuras Algebraicas, la Figura 5 etiqueta cada respuesta con el género del estudiante. En esta se puede apreciar que no existe una relación directa del género del estudiante y su distribución espacial, por lo que no se advierte un patrón de agrupamiento visual con esta dimensionalidad. Por último, en el centro de interés

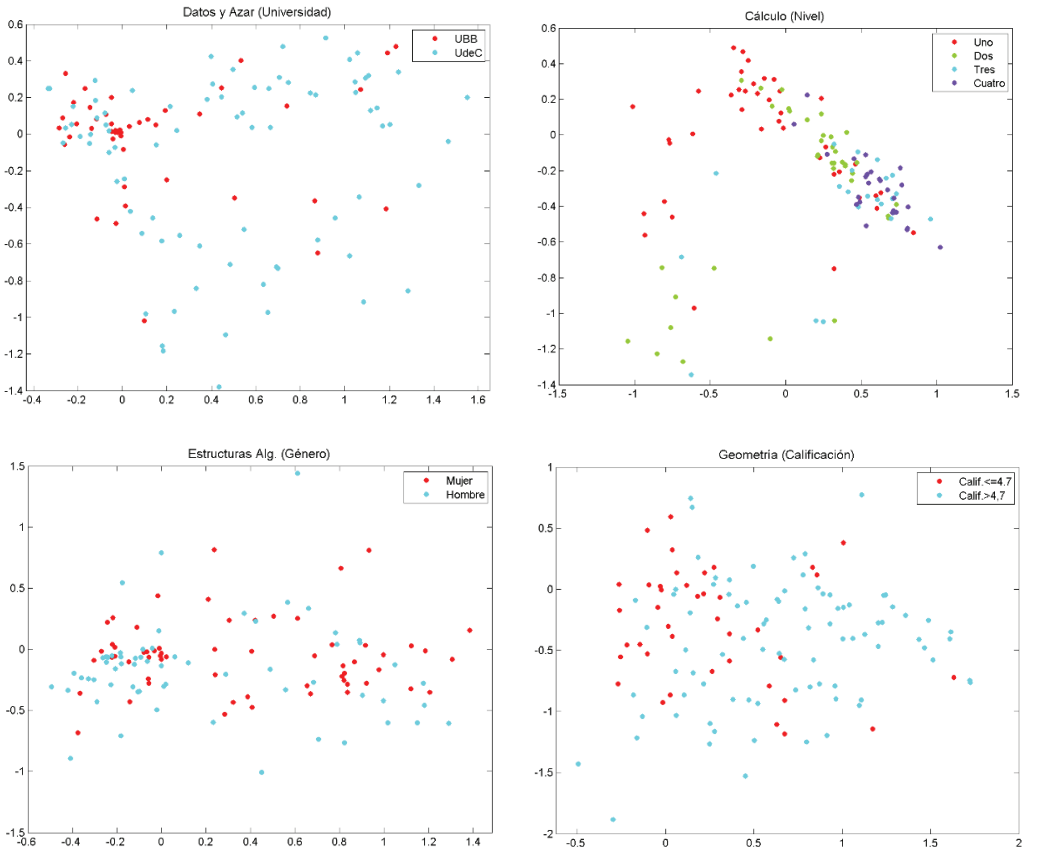

Figura 5. Representación espacial de algunos centros de interés y sus lexicones asociados a características adicionales.

Geometría, se han etiquetado los lexicones según dos rangos de calificaciones: de *1* a *4.7* y de *4.8* a *7*. Esta calificación corresponde al promedio de todas las calificaciones finales obtenidas por los alumnos en asignaturas relacionadas con los centros de interés estudiados. Como se puede observar, mayoritariamente las calificaciones inferiores o iguales a *4.7* se distribuyen más cercanas al origen de las coordenadas, lo que indicaría entre otras cosas que sus lexicones son menos homogéneos y menos disponibles en comparación a los individuos con calificaciones superiores a *4.7*.

4.8. Agrupación espacial de los lexicones

La representación mostrada anteriormente es realizada en dos dimensiones, por lo que es muy posible que, en una dimensionalidad más alta, puedan existir relaciones espaciales que no son detectables en forma visual. Para complementar el análisis, es posible realizar un análisis de clúster por cada centro de interés. En cada corpus léxico se aplicó el algoritmo k-means (MacQueen 1967; Jain 2010). Producto de que k-means no puede determinar por sí mismo la cantidad de clústeres que existen en la muestra, se ha aplicado este proceso configurando la técnica para que obtenga de 2 a 5 agrupaciones, utilizando en cada uno de los casos el índice de validación de clustering PBM (Pakhira *et al*. 2004) para determinar cuáles de estas configuraciones es la óptima.

Los resultados de este procedimiento se muestran en el Cuadro 5, donde para cada centro de interés se asocia: el valor alcanzado por el índice de validación de clustering (PBM), la cantidad de clústeres óptimos según el criterio definido (Cantidad Clústeres) y el detalle de los estadígrafos asociados a cada clúster obtenido. Se ha decidido no utilizar el índice de cohesión para determinar la cantidad de grupos óptimos, debido a que este es muy sensible a la cantidad de elementos, lo que no permite discriminar una configuración de otra sin poder evitar tal situación. Por el contrario, el índice de validación PBM tiende a valorar más las agrupaciones con una alta separabilidad entre-clústeres y una alta cercanía de los elementos intra-clúster, esto, según su formulación matemática, es menos sensible a la diferencia de cantidad de elementos en los grupos.

Cuadro 5. Mejores configuraciones de clústeres según índice PBM

Centro de Interés	PBM	Cantidad Clústeres	Grupo	Cantidad Individuos	NPD	XR	IC
Datos y Azar	.0007	3	1	36	196	15.1	.077
			2	21	162	18.3	.113
			3	69	300	9.8	.033
Cálculo	.0495	4	1	68	315	10.9	.035
			2	13	151	21.4	.142
			3	24	132	14.0	.106
			4	21	189	18.1	.096
Estructuras Alg.	.2080	2	1	79	269	7.5	.028
			2	44	196	14.9	.076
Geometría	.0384	2	1	47	257	25.6	.100
			2	79	320	16.6	.052
Sistemas Numéricos	.2108	2	1	36	184	14.5	.079
			2	90	392	12.5	.032

Como se puede apreciar en el Cuadro 5, en tres de los cinco corpus léxicos se encontró que, según el índice PBM, la separabilidad óptima corresponde a dos grupos. En cambio, para los centros de interés Datos y Azar, y Cálculo, son tres y cuatro respectivamente la cantidad de agrupaciones óptimas. Comparando estos resultados con el Cuadro 4, el centro de interés Datos y Azar tiene el menor índice de cohesión y la segunda cantidad mayor de vocablos. Además, el centro de interés Cálculo tiene la segunda cohesión más baja y la mayor cantidad de vocablos. Debido a lo anterior, es lógico que para estos dos centros de interés el índice PBM haya determinado una mayor cantidad de agrupaciones (3 y 4 respectivamente), producto principalmente de la gran cantidad de vocablos que poseen y baja cohesión. Esto último hace que las agrupaciones demasiado grandes (con una gran cantidad de lexicones), aumente la distancia intra-clúster, produciendo que el índice de validación, al compararlo con la ganancia en calidad recibida por separar las clases, valoré en mayor medida una cantidad más alta de agrupaciones para estos centros (es decir al ser menos cohesionados son mayormente separables).

Por el contrario, la mayor cohesión de los demás centros de interés y su menor cantidad de vocablos produce que las configuraciones evaluadas por el índice de validación no produzcan más de dos grupos como configuración óptima para estos casos. En otro

aspecto, el centro de interés Geometría, el más cohesionado de todos, tiene agrupaciones que difieren notoriamente en el largo promedio de sus respuestas, pudiendo inferir que, aunque la mayoría de los lexicones poseen vocablos en común (alta cohesión), la característica diferenciadora encontrada por este proceso de clustering corresponde a la cantidad de vocablos de los lexicones más que a la diferencia de vocablos entre ellos.

Para describir mejor estos resultados, el Cuadro 6 muestra cada centro de interés y una lista con las cinco palabras de más alto IDL, ordenadas de mayor a menor, para cada clúster obtenido, apreciándose de mejor forma las diferencias en las configuraciones. Por ejemplo, se puede ver para Datos y Azar que existen dos clústeres que, al menos en sus vocablos más disponibles, son casi idénticos, salvo ciertas diferencias en el orden. En adición, el tercer clúster difiere en mayor medida producto de la inclusión de vocablos que hacen referencia a los juegos de azar. Esto permite observar que el algoritmo de clúster logra agrupar a los elementos que tienen estos vocablos que, además de ser altamente disponibles, son los de mayor cantidad como ya se mostró en el Cuadro 4. En lo referente al centro de interés Cálculo, las cuatro agrupaciones son claramente diferenciables, destacando el clúster 3 que posee vocablos de las operaciones elementales, y el clúster 4, que contiene palabras que hacen referencia a las aplicaciones del cálculo como lo son el cálculo de áreas y volúmenes. En contraposición a esto, los grupos 1 y 2 de esta configuración están dominados por palabras sobre los elementos del análisis de funciones y conceptos relacionados con la derivada.

Los demás centros de interés tienen configuraciones con sólo dos grupos, ambos fácilmente diferenciables, excepto los clústeres obtenidos de Geometría, los cuales tienen sus vocablos más disponibles casi idénticos. Este efecto puede ser explicado por ser el centro de interés más cohesionado y por la diferencia en el largo de las respuestas como se mencionó anteriormente.

Por último, se debe destacar el valor obtenido por el índice de cohesión para cada clúster, donde se advierte la inversa proporcionalidad de su medida respecto a la cantidad de vocablos, lo que demuestra que este índice no es el más adecuado para la medición de la calidad de las agrupaciones obtenidas automáticamente, ya que indistintamente de las relaciones que posean los vocablos agrupados, si no hay una diferencia notoria en el largo de las respuestas, este índice considerará siempre más cohesionado a un clúster que tenga menos vocablos que a otro que posea un número mayor de ellos.

4.9. Clasificación de características

Como ya se ha mencionado, el modelo propuesto permite entrenar clasificadores que pueden ayudar a estimar o predecir características de los individuos sobre la base de los lexicones. Producto de que es muy improbable que exista una separabilidad lineal entre los lexicones asociados a las características de los sujetos, se ha decidido utilizar una máquina de soporte de vectores (SVM) (Schölkopf *et al*. 1995) para realizar las pruebas de clasificación. Este tipo de clasificador, en su versión binaria, permite a través de la utilización de kernels

Cuadro 6. Vocablos más disponibles para la mejor configuración de clústeres por centro de interés.

Centro de Interés	Clúster	Vocablos más disponibles por Clúster						
Datos y Azar	1	*probabilidad*	*moda*	*promedio*	*media*	*estadistica*	*mediana*	*muestra*
	2	*probabilidad*	*moda*	*media*	*mediana*	*estadistica*	*dato*	*azar*
	3	*probabilidad*	*estadistica*	*dado*	*juego*	*dato*	*razon*	*azar*
Cálculo	1	*derivada*	*limite*	*integral*	*funcion*	*numero*	*analisis*	*continuidad*
	2	*derivada*	*limite*	*integral*	*maximo*	*teorema*	*minimo*	*pendiente*
	3	*multiplicacion*	*suma*	*division*	*numero*	*resta*	*adicion*	*sustracción*
	4	*derivada*	*integral*	*limite*	*volumen*	*función*	*area*	*teorema*
Estructuras Alg.	1	*algebra*	*estructura*	*anillo*	*letras*	*demostracion*	*conjunto*	*matriz*
	2	*anillo*	*grupo*	*cuerpo*	*abeliano*	*conmutatividad*	*campo*	*subgrupo*
Geometría	1	*triangulo*	*angulo*	*circunferencia*	*recta*	*cuadrado*	*punto*	*rectángulo*
	2	*triangulo*	*area*	*figura*	*circunferencia*	*angulo*	*cuadrado*	*recta*
Sistemas Numéricos	1	*numero*	*reales*	*naturales*	*complejos*	*ecuacion*	*enteros*	*racionales*
	2	*numero*	*ecuacion*	*incognita*	*suma*	*letra*	*sistema*	*resta*

resolver de mejor forma la inseparabilidad lineal de algunos problemas. Para nuestro caso se ha utilizado una SVM con un kernel RBF y un valor *sigma=0.5*. Para el entrenamiento se ha utilizado una validación cruzada leave-one-out (Wong 2015), clasificando los lexicones según cuatro características: Género (Hombre o Mujer), Nivel (de 1 a 2 años, o de 3 a 4 años), Calificación (inferior o igual a 4.7, o superior a 4.7) y Universidad (UdeC o UBB).

Con el objetivo de determinar la calidad de la clasificación se han usado las métricas estándares que comúnmente son utilizadas para evaluar el desempeño de clasificadores (Marbouti *et al*. 2016): Exactitud, Precisión y Exhaustividad. El Cuadro 7 muestra las medidas de dimensión (Dim), Exactitud, Precisión y Exhaustividad para cada centro de interés y característica. Cada medida de Precisión y Exhaustividad está divida en dos clasificaciones posibles: I y II, que corresponden a cada clase de la característica que se desea clasificar. Para obtener la configuración del clasificador, se realizó una reducción dimensional mediante PCA a 2, 3, 4, 5, 10, 15, 25, 50 y 100 dimensiones, luego se entrenó y probó el clasificador para cada característica y centro de interés con tal de elegir la configuración de mayor Precisión y Exhaustividad en conjunto.

Como se puede observar en el Cuadro 7, la dimensión óptima se mantiene baja para todos los centros de interés y características. Como se podía esperar, la clase Género es la que mayor dimensión requirió, y menor Precisión y Exhaustividad obtuvo, debido a la casi ausente diferencia de género en el léxico disponible sobre los centros de interés seleccionados. Por otro lado, se puede ver que, para cada clase de característica, existe un centro de interés distinto donde se alcanzan los mejores desempeños, pudiendo destacar: Nivel en el centro de interés de Estructuras Alg; Calificación en Datos y Azar, y Cálculo; y Universidad en el centro de interés Estructuras Alg. y Geometría.

A pesar de que el detalle desglosado sobre el desempeño del clasificador es valioso para entender cuan clasificable, según el modelo propuesto, es una característica, es recomendable la visualización conjunta de todos los resultados. Para lo anterior, se puede observar la Figura 6, en la cual se muestra un gráfico con el F-score de cada clase y centro de interés. El F-score es una media harmónica entre las medidas de Precisión y Exhaustividad, siendo común su utilización en el área de recuperación de información y reconocimiento de patrones. Para nuestro caso, la versión F_1-score es una medida que le da igual importancia a las medidas de Precisión y Exhaustividad. Como se puede ver en la Figura 6, las características con las cuales el clasificador alcanza su mejor desempeño son Nivel, Calificación y Universidad. Además, se puede apreciar que estas son mejor clasificadas en los centros de interés de Cálculo, Estructuras Alg, y Geometría, obteniéndose un menor desempeño para casi todas ellas en los centros de interés Sis. Numéricos y para algunos casos en Datos y Azar.

Cuadro 7. Medidas de Dimensión, Exactitud, Precisión y Exhaustividad para cada centro de interés, los mejores resultados están resaltados en negrita

Clase	Centro de Interés	Dim.	Exactitud	Precisión I	Precisión II	Exhaustividad I	Exhaustividad II
Genero I= hombre II= mujer	Azar	3	.5820	.5714	.5962	.6557	.5082
	Cálculo	4	.5085	.5088	.5082	.4915	.5254
	Estructuras Alg.	2	.5854	.5758	.5965	.6230	.5484
	Geometría	4	.5424	.5385	.5472	.5932	.4915
	Sis. Numéricos	4	.6186	.6094	.6296	.6610	.5763
Nivel I=1 o 2 II= 3 o 4	Azar	3	.7041	.7273	.6852	.6531	.7551
	Cálculo	2	.7551	.7551	.7551	.7551	.7551
	Estructuras Alg.	3	.7959	.8222	.7736	.7551	.8367
	Geometría	2	.7449	.7609	.7308	.7143	.7755
	Sis. Numéricos	5	.5612	.5789	.5500	.4490	.6735
Calificación I= C<=4.7 II= C>4.7	Azar	2	.7349	.7111	.7632	.7805	.6905
	Cálculo	5	.6988	.7353	.6735	.6098	.7857
	Estructuras Alg.	2	.6625	.6585	.6667	.6750	.6500
	Geometría	2	.6024	.5909	.6154	.6341	.5714
	Sis. Numéricos	2	.5904	.6000	.5833	.5122	.6667
Universidad I=UBB II= UdeC	Azar	3	.6170	.6170	.6170	.6170	.6170
	Cálculo	3	.7340	.7500	.7200	.7021	.7660
	Estructuras Alg.	4	.7667	.8158	.7308	.6889	.8444
	Geometría	2	.7553	.7609	.7500	.7447	.7660
	Sis. Numéricos	2	.6064	.6042	.6087	.6170	.5957

Se debe acotar que, la naturaleza del clasificador SVM difiere con notoriedad del algoritmo de clustering k-means. Por ejemplo, la alta cohesión del centro de interés Geometría no es impedimento para clasificar, con un desempeño superior al *70%,* el Nivel y Universidad de los alumnos. Por otro lado, la clasificación es una técnica supervisada, lo que requiere integrar ejemplos al clasificador con tal de que aprenda como determinar cuándo un lexicón pertenece a una clase u otra. Sin embargo, esto sólo ocurre en la fase de entrenamiento, no siendo utilizada esta información en la etapa de prueba o validación.

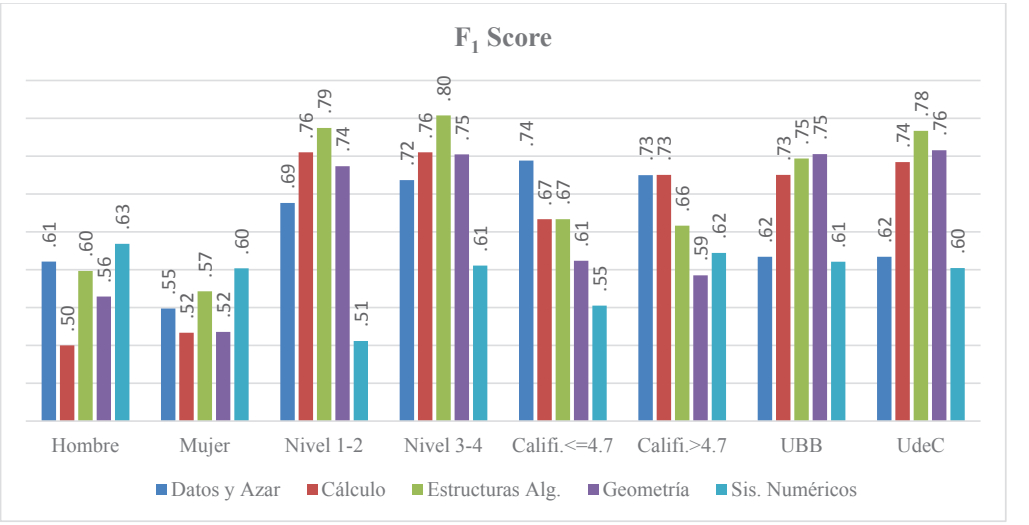

Figura 6. F_1-score para la clasificación del género, nivel, calificación y universidad según centro de interés

4. Conclusiones

Se ha desarrollado un nuevo modelo para la representación y análisis de los resultados de encuestas de disponibilidad léxica, permitiendo operar sobre ellos técnicas y métodos de análisis cuantitativos que no son posibles con el enfoque tradicional basado en estadígrafos. A partir de este modelo, se ha realizado un análisis de la disponibilidad léxica de dos grupos de estudiantes de pedagogía de dos universidades chilenas. Los resultados muestran que es factible utilizar la representación de matriz de términos pesados, permitiendo obtener medidas cuantitativas mediante la búsqueda no supervisada de grupos característicos de los centros de interés utilizando algoritmos de clustering e índices de calidad de clústeres.

En adición, el modelo desarrollado permite obtener un modelo vectorial del lexicón de los individuos sometidos a una encuesta de disponibilidad léxica. Este tipo de modelo permite la aplicación de cualquier técnica de reconocimiento de patrones que procese los datos en un espacio vectorial, no estando limitado a las técnicas presentadas en este trabajo. A esto se debe agregar que, según nuestro conocimiento, no existen investigaciones sobre la caracterización del léxico disponible mediante modelos vectoriales.

En lo referente a la eficiencia del cómputo y a la manipulación de los datos, el modelo propuesto se ha basado en los modelos vectoriales clásicos de los sistemas de recuperación de información, los cuales son bien conocidos y altamente replicables, asegurando de esta forma que desarrollar la implementación de estos modelos sea plausible en la práctica.

En otro contexto, la visualización espacial de los lexicones que se ha propuesto, al estar determinada por un modelo que considera los índices de disponibilidades de los vocablos (a través de las funciones de peso de la matriz), permite una mayor representatividad de lexicones que poseen vocablos de baja disponibilidad pero que, contrariamente, podrían ser altamente frecuentes. Esto difiere con los modelos basados en grafos, donde los resultados mostrados dependen de la secuencialidad directa de los vocablos en las listas de palabras.

Respecto a la agrupación de lexicones, se pudo mostrar que es posible a través del modelo propuesto utilizar algoritmos de clustering como k-means, desarrollando procesos de agrupación que pueden ser llevados a cabo sin intervención del analista, entregando una información más objetiva en la caracterización y agrupación de las respuestas de los individuos. En este sentido, este modelo de agrupación puede ser fácilmente utilizable como sistema de determinación automática de grupos, por ejemplo, para determinar grupos de trabajo en el ambiente educativo, basado en la disponibilidad léxica de los individuos en un centro de interés educacional, lo que podría convertirse en un sistema complementario para otros enfoques de agrupamiento como el presentado en (Pinninghoff et al. 2016), donde la determinación de grupos se realiza a través de los resultados de un conjunto de encuestas a los alumnos.

Por último, el modelo desarrollado demuestra ser factible para ser aplicado en la predicción de características asociadas al lexicón de los sujetos que, según los resultados obtenidos, pueden ser utilizados para predecir con un desempeño adecuado la calificación, el nivel y la universidad de procedencia de los alumnos.

Obras citadas

Casado, Elisa. 2002. "Prototipos de la interacción pedagógica". *Revista de Pedagogía* 23 (67): 247-279.

Chen, Kewen *et al*. 2016. "Turning from TF-IDF to TF-IGM for term weighting in text classification". *Expert Systems with Applications* 66: 245-260.

Del Valle, María *et al*. 2016. "Analyzing the Availability of Lexicon in Mathematics Education Using no Traditional Technological Resources". *Journal of Supply Chain Mana-*

gement 5 (2):144-149.
Echeverría, Max *et al*. 2008. "DispoGrafo: una nueva herramienta computacional para el análisis de relaciones semánticas en el léxico disponible". *Revista de Lingüística Teórica y Aplicada* 46 (1): 81-91.
Ferreira, Anita *et al*. 2014. "Estudio de disponibilidad léxica en el ámbito de las matemáticas". *Estudios Filológicos* 54: 69-84.
Germany, Patricia & Ninette Cartes. 2000. "Léxico disponible en inglés como segunda lengua de instrucción formalizada". *Estudios Pedagógicos* 26: 39-50.
Jain, Anil K. 2010. "Data clustering: 50 years beyond K-means". *Pattern Recognition Letters* 31 (8): 651-666.
López, Humberto. 1993. "Los estudios de disponibilidad léxica: pasado y presente". *Boletín de Filología de la Universidad de Chile* 35: 245-259.
MacQueen, James. 1967. "Some methods for classification and analysis of multivariate observations". *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1 (14): 281-297.
Marbouti, Farshid *et al*. 2016. "Models for early prediction of at-risk students in a course using standards-based grading". *Computers & Education* 103: 1-15.
Martínez, Eugenio *et al*. 2011. "Técnicas de clasificación de opiniones aplicadas a un corpus en español". *Procesamiento del Lenguaje Natural* 47: 163-170.
Michea, René. 1953. "Mots fréquents et mots disponibles, un aspect nouveau de la statistique du langage". *Langues Modernes* 47 (4): 338-344.
MINEDUC Chile. 2012. *Estándares orientadores para carreras de pedagogía en educación media*. Ministerio de Educación. Web. Disponible en: http://portales.mineduc.cl/usuarios/cpeip/File/librosestandaresvale/libromediafinal.pdf
Pakhira, Malay *et al*. 2004. "Validity index for crisp and fuzzy clusters". *Pattern Recognition* 37 (3): 487-501.
Pinninghoff, María Angélica *et al*. 2016. "Genetic algorithms as a tool for structuring collaborative groups". *Natural Computing* 16 (2): 1-9.
Osses, Sonia & Sandra Jaramillo. 2008. "Metacognición: 'Un camino para aprender a aprender'". *Estudios Pedagógicos* 34 (1): 187-197.
Salcedo, Pedro *et al*. 2013. "A bayesian model for lexical availability of chilean high school students in mathematics". En: José Ferrández *et al*. Eds. *Natural and Artificial Models in Computation and Biology – 2013.* IWINAC 2013, Part I. Berlin/Heidelberg: Springer. 245-253.
Salcedo, Pedro *et al*. 2015. "LEXMATH - A Tool for the study of available lexicon in mathematics". En: José Ferrández *et al*. Eds. *Bioinspired Computation in Artificial Systems*. IWINAC 2015, Part II. Berlin/Heidelberg: Springer. 11-19.
Salton, Gerard *et al*. 1975. "A Vector Space Model for Automatic Indexing". *Communications of the ACM* 18 (11): 613-620.
Schölkopf, Bernhard *et al*. 1995. "Extracting Support Data for a Given Task". En: Usama Fayyad & Ramasamy Uthurusamy. Eds. *Proceedings of the First International Confe-*

rence on Knowledge Discovery & Data Mining. KDD'95. 252-257.
Soumen, Chakrabarti. 2002. *Mining the web. Discovery knowledge from hypertext data*. USA: Morgan Kaufmann.
Urzúa, Paula *et al*. 2006. "Disponibilidad léxica matemática: análisis cuantitativo y cualitativo". *Revista de Lingüística Teórica y Aplicada* 44 (2): 59-76.
Valencia, Alba. 2010. "Léxico del color en Santiago de Chile". *Revista de Lingüística Teórica y Aplicada* 48 (2): 141-161.
Valencia, Alba & Max Echeverría. 1999. *Disponibilidad léxica en estudiantes chilenos*. Santiago de Chile: Ediciones Universidad de Chile–Universidad de Concepción.
Wong, Tzu-Tsung. 2015. "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation". *Pattern Recognition* 48 (9): 2839-2846.