Formación Universitaria
E-ISSN: 0718-5006
citrevistas@gmail.com
Centro de Información Tecnológica
Chile

Rojas, Darío F.; del C. Zambrano, Carolina; Salcedo, Pedro A.
Metodología de Análisis de Disponibilidad Léxica en Alumnos de Pedagogía a través de la Comparación Jerárquica de Lexicones
Formación Universitaria, vol. 10, núm. 4, 2017, pp. 3-13
Centro de Información Tecnológica
La Serena, Chile

Disponible en: http://www.redalyc.org/articulo.oa?id=373552294002

Formación Universitaria
Vol. 10(4), 3-14 (2017)
doi: 10.4067/S0718-50062017000400002

Metodología de Análisis de Disponibilidad Léxica en Alumnos de Pedagogía a través de la Comparación Jerárquica de Lexicones

Darío F. Rojas, Carolina del C. Zambrano y Pedro A. Salcedo
Depto. Metodología de Investigación e Informática Educacional, Facultad de Educación, Universidad de Concepción, Victoria 280, Concepción, Chile. (e-mail: dariorojas@udec.cl; carozambrano@udec.cl; psalcedo@udec.cl)

Recibido Dic. 30, 2016; Aceptado Mar. 8, 2017; Versión final May. 21, 2017, Publicado Ago. 2017

Resumen

El presente trabajo propone una metodología de análisis de disponibilidad léxica, basada en la comparación directa de lexicones obtenidos de estudiantes de pedagogía. Esto se hace en respuesta a centros de interés sobre temáticas indicadas como estándar por el Ministerio de Educación de Chile, para la enseñanza de las matemáticas en educación media. La metodología define una nueva medida de disimilitud entre lexicones, basada en la distancia de Levenshtein, la que permite utilizar un enfoque de análisis cuantitativo mediante técnicas estadísticas y de agrupación que no son posibles en los enfoques tradicionales. Los resultados muestran la factibilidad del enfoque propuesto, permitiendo utilizar el léxico disponible como herramienta de comparación de grupos e individuos.

Palabras clave: disponibilidad léxica; clustering jerárquico; distancia de Levenshtein; interacción pedagógica

Methodology of Lexical Availability Analysis in Students of Pedagogy through Hierarchical Comparison of Lexicons

Abstract

The present work proposes a methodology of analysis of lexical availability based on the direct comparison of lexicons obtained from students of pedagogy. This is done in response to centers of interest on topics considered as standard by the Chilean Ministry of Education, for the teaching of mathematics in secondary education. The methodology defines a new measure of dissimilarity between lexicons based on Levenshtein's distance, which allows using a quantitative analysis approach including statistics and clustering techniques that are not possible in traditional approaches. The results show the feasibility of the proposed approach, allowing the use of the available lexicon as a tool for comparison of groups and individuals.

Keywords: lexical availability; hierarchical clustering; Levenshtein's distance; pedagogical interaction

INTRODUCCIÓN

La interacción pedagógica es parte fundamental del aprendizaje (Casado, 2002), y hace posible que nuevos elementos del lenguaje, tanto propios como colectivos, puedan ser adquiridos (Bernal, 2007). En esta interacción suele aparecer un caudal de palabras muy usadas, que están estrechamente relacionadas con los conceptos que contextualizan la comunicación (Michea, 1953). La caracterización de estas palabras, que corresponden al vocabulario que un grupo de individuos tiene más disponible para utilizar en un contexto específico, se denomina "léxico disponible" (López, 1993), mientras que al contexto que las producen se conoce como "centro de interés".

Es de sentido común que la interacción de comunicación alumno-profesor sea importante en los procesos de enseñanza-aprendizaje. Desde el punto de vista de la didáctica, una interacción deficiente puede obstaculizar la transposición didáctica al dificultar la transmisión del saber sabio como saber enseñado (Chevallard, 1991; Mendoza, 2005), debido a que por ejemplo, el vocabulario del profesor sobre el saber sabio de una temática específica, se puede diferenciar demasiado al vocabulario del saber enseñado necesario para los alumnos, el cual debiera estar en sintonía y coincidir para poder resultar en un aprendizaje significativo. De acuerdo con Piaget y sus postulados sobre el constructivismo (Piaget, 2001), los profesores deben tener en cuenta los conocimientos previos de sus alumnos (Marines et al., 2014), y esto, también implica conocer cuáles son los elementos comunicativos que posee el alumno para interactuar adecuadamente en función del aprendizaje.

Por otro lado, la interacción de comunicación también se puede dar entre alumnos, principalmente en actividades donde se requiera un nivel de interacción considerable para lograr el aprendizaje, como por ejemplo en el aprendizaje colaborativo (Martín et al., 2009; Laal et al., 2013). Esta interacción es muy importante y puede ser un componente esencial para la adquisición de conocimientos, habilidades y actitudes en estos contextos (Ariza y Oliva, 2000; Carrió, 2007). Debido a que el léxico disponible es contextual y es permeable a la situación y experiencia, este puede modificarse no solo por la interacción pedagógica con el profesor, sino que también por la interacción entre alumnos. En este sentido, los estudios de disponibilidad léxica, permiten caracterizar los vocablos como elementos de interacción, y obtener indicios sobre el grado de compatibilidad con el proceso de comunicación.

Los estudios del léxico disponible comenzaron en la década de los cincuenta por la UNESCO, con tal de buscar una forma de representar un vocabulario común básico de la lengua francesa, y facilitar de esta forma, el aprendizaje de dicha lengua a los habitantes no nativos de Francia (Lopéz, 1993). Para obtener el léxico disponible de un grupo de personas, se realiza un tipo de prueba de fluencia semántica denominada "prueba de disponibilidad léxica", la que consiste en indicar a los individuos escribir en orden las primeras palabras que se le vengan a la mente (lexicón) ante un estímulo denominado "centro de interés", el cual es generalmente dado como una palabra u oración simple. Estos lexicones se pueden obtener y analizar utilizando métodos tradicionales como escribir en papel las palabras en un determinado tiempo, o mediante servicios especializados como los presentados por el proyecto DispoLex (Bartol y Hernández, 2004; Mangado-Cruz, 2008) o el software LexMath (Del Valle, 2016).

La primera propuesta de análisis que utilizaba no sólo la frecuencia de los vocablos, sino que también el orden de estos, corresponde a la formulación del Índice de Disponibilidad de Vocablos (IDLV), realizada en (Loran y López Morales, 1983). Este enfoque utiliza la frecuencia de aparición de cada palabra, ponderada por un factor compuesto por una constante definida para el estudio y la posición del vocablo al interior de cada lexicón. De esta forma, una palabra alcanza un IDLV mayor (más cercano al valor 1), cuando tiene una alta frecuencia y se encuentra en las primeras posiciones de cada lexicón. Por el contrario, cuando la palabra es poco frecuente o se encuentra generalmente en las últimas posiciones de los lexicones, obtiene un valor de IDLV bajo (más cercano al valor 0).

Años después en (López y Strassburger, 1987) se presenta el Índice de Disponibilidad Léxica Ajustado (IDLA), el que mejora la formulación del IDLV incorporando en su cálculo la cantidad de individuos y un factor de ajuste, que está basado en una función exponencial y el coeficiente entre la posición del vocablo y la máxima posición alcanzada. Esta nueva fórmula estima mejor los índices de disponibilidad léxica, disminuyendo los problemas de medición presentes en la fórmula original referentes a la subestimación de los valores de disponibilidad para vocablos alejados de la primera posición.

Por otro lado, en (Ávila y Sánchez, 2010; Ávila y Sánchez, 2011), se propone un enfoque de representación para la disponibilidad léxica basada en la teoría de prototipos, considerando los centros de interés cómo núcleo prototípico y el grado de accesibilidad de las palabras como una medida de disponibilidad de los vocablos. De este modo, haciendo uso de la teoría de conjuntos difusos y sus operaciones, define el Grado de Compatibilidad de un Vocablo (GCV) como la fórmula que determina el grado de pertenencia de las palabras al conjunto que representa el centro de interés. Este índice consiste principalmente en la

multiplicatoria de factores compuestos por el cociente entre una constante experimental y la posición de los vocablos, utilizando la frecuencia absoluta de las palabras como exponente. En este sentido, el índice GCV se comporta de manera análoga a IDLV e IDLA, en lo que se refiere a las posiciones de las palabras y su frecuencia. Sin embargo, el marco teórico que lo sustenta es más robusto. Adicionalmente, el índice CRV (Coeficiente de Relación de Vocablos), definido en (Madrigal-Melchor et al., 2010), corresponde a una fórmula que calcula la correlación entre las frecuencias de posición de dos vocablos. De esta forma, si dos vocablos tienen frecuencias parecidas en las posiciones dentro de un grupo, obtendrán un valor alto de correlación, lo que se puede interpretar como un comportamiento, desde el punto de vista de la disponibilidad.

Respecto a la representación de los lexicones, se pueden encontrar formulaciones basadas en el cálculo de los índices de disponibilidad IDLV y GCV vistos anteriormente. Es así como, el Índice de Disponibilidad Léxica individual (IDLI) propuesto en (López y Strassburger, 1991) está formado por la suma ponderada del índice IDLV de cada vocablo perteneciente al lexicón. Similarmente, la propuesta de (Ávila y Sánchez, 2010), define el Índice de Descentralización Léxica del Informante (IDLINF), en base a la multiplicatoria del GCV de cada vocablo.

En otros enfoques de análisis, basados en estructuras matemáticas no convencionales, se puede considerar a (Echeverría et al., 2008), donde se propone un enfoque de análisis de los centros de interés a través de grafos, mostrando la utilidad de estas estructuras matemáticas para poder representar gráficamente la relación entre los vocablos de un grupo en base a su posición y relación con los lexicones. En ese mismo sentido, en (Salcedo et al., 2013), se propone una extensión de dicho modelo mediante un enfoque de análisis bayesiano, aplicado al estudio del léxico disponible de alumnos de educación media en centros de interés relacionados al algebra, permitiendo la cuantificación de las relaciones mediante grafos con pesos probabilísticos en las aristas.

Más recientemente, en (Callealta y Gallego, 2016), se propone una generalización y estandarización de los índices de disponibilidad basados en posición y frecuencia de las palabras. En esta investigación los índices IDLV, GCV, IDLI y IDLINF son reformulados en base a la definición de parámetros que permiten principalmente controlar los límites superiores e inferiores de los índices, permitiendo una correcta comparabilidad entre los distintos estudios de disponibilidad léxica. Además, este nuevo planteamiento permite controlar los sesgos producidos por los distintos tamaños de muestra, largo de las respuestas, casos extremos y constantes experimentales definidas en cada índice.

En el contexto educativo, varios son los trabajos que han realizado estudios de disponibilidad léxica. En (Germany y Cartes, 2000) se lleva a cabo un estudio de disponibilidad léxica del inglés en distintos tipos de establecimientos educacionales, encontrando diferencias en índices de composición grupal y conjunto de vocablos, según el tipo de establecimiento del cual provienen los alumnos. En (Ferreira et al., 2014) y (Salcedo y Del Valle, 2013) se analiza el léxico disponible de alumnos de enseñanza media en torno a contenidos relacionados con la matemática. Los resultados mostrados prueban que el léxico de los alumnos aumenta con los años de escolaridad, principalmente determinados por la cantidad de palabras y largo promedio de respuestas de los alumnos. Así mismo, en (Castañeda et al., 2016) se hace un análisis de disponibilidad léxica en centros de interés relacionados con conceptos matemáticos, para alumnos pertenecientes a carreras de ciencias e ingeniería. Los resultados de este estudio muestran que hay diferencias en la disponibilidad léxica entre alumnos que tienen un año de diferencia de permanencia en la universidad, aunque los resultados no son generalizables en todos los centros de interés y grupos, obteniéndose índices grupales con resultados dispares en cada caso.

El presente trabajo muestra una aproximación algorítmica para el análisis de disponibilidad léxica entre lexicones. Este enfoque no se basa en la frecuencia de palabras o en características del grupo, y puede complementar las metodologías de análisis existentes. Además, el enfoque propuesto permite determinar relaciones de disponibilidad léxica a nivel de los individuos directamente, permitiendo realizar comparaciones con independencia del tamaño de la muestra, e incluso entre lexicones pertenecientes a distintos centros de interés. A continuación, se presenta la metodología de análisis empleada, para luego ser aplicada en un contexto educativo, mostrando los resultados, discusión y conclusiones.

PROPUESTA METODOLOGÍA

Se propone una nueva metodología de análisis del léxico disponible basada en la relación entre lexicones. La hipótesis es que existen agrupaciones jerárquicas de lexicones en centros de interés relacionados con las matemáticas, que pueden ser determinados a través de una medida de distancia basada en el ordenamiento de los vocablos de cada lexicón, sin necesidad de utilizar la frecuencia de palabras o características grupales.

El lexicón mental en cada centro de interés es obtenido a través de una prueba de disponibilidad léxica a estudiantes de pedagogía. Luego, para cada centro de interés se obtiene la distancia entre cada par de lexicones posibles de comparar (todos con todos). Posteriormente se aplica el algoritmo de clúster jerárquico, que permite agrupar los lexicones similares (con la menor distancia) y formar una jerarquía de agrupación como las presentadas en las Figura 1 y 2. Esta jerarquía de lexicones permite agrupar en distintos niveles jerárquicos a los estudiantes, pudiendo formar grupos grandes o pequeños según la relación de cada miembro en cada grupo. Por último, se calculan los estadígrafos más comunes para cada centro de interés, aunque cabe destacar, que estos no son necesarios para el cálculo de la distancia y la agrupación jerárquica, siendo presentados cómo información complementaria.

Muestra

En nuestro estudio se realizó la prueba de disponibilidad léxica para los alumnos de la carrera de Pedagogía en Matemática y Computación de la Universidad de Concepción, considerando alumnos de cuatro niveles (de primero a cuarto año). El instrumento de recolección de datos corresponde al test de disponibilidad léxica empleado en (Valencia y Echeverría, 1999) y el tiempo límite de respuesta es de dos minutos por cada centro de interés. Se obtuvieron 394 lexicones de cinco centros de interés definidos a partir de la agrupación temática de los 21 estándares disciplinares que determina el Ministerio de Educación de Chile, para la enseñanza de las matemáticas en educación media (MINEDUC, 2012). Estos son: Sistemas Numéricos y Algebra (NUME), Cálculo (CALC), Estructuras Algebraicas (ESTR), Geometría (GEOM), y Datos y Azar (AZAR). Cada participante realizó la prueba de disponibilidad léxica sobre cada uno de los cinco centros de interés. De esta manera, cada centro de interés está compuesto por las respuestas de los mismos 79 estudiantes, aunque una de las encuestas a un centro interés se entregó vacía, lo que explica el lexicón faltante en el total. La Tabla 1 muestra la cantidad de individuos participantes por género (F=Femenino, M=Masculino), asociado al promedio y desviación estándar de la edad en cada nivel de alumno (N1, N2, N3, N4).

Tabla 1: Cantidad y Edad de los Participantes

Nivel	*N1*		*N2*		*N3*		*N4*	
Género	F	M	F	M	F	M	F	M
Cantidad	12	14	5	6	9	11	13	9
Edad	18.8±1	19.9±2.2	19.6±1.3	21.2±1.7	20.7±1.1	23.2±4	23.2±4.9	23±0.7

Estadígrafos de disponibilidad léxica

En la mayoría de los estudios de disponibilidad léxica, cinco son los índices más utilizados en investigaciones donde se requiere determinar la riqueza léxica de los sujetos (Urzúa et al., 2006). Estos son: i) *Cantidad de Lexicones (N):* equivalente a la cantidad de individuos considerados en un grupo (se obtiene un lexicón por individuo); ii) *Largo promedio de respuestas (XR):* promedio de la cantidad de vocablos indicados por los individuos de un grupo ante un estímulo o centro de interés específico; iii) *Cantidad de Palabras Distintas (NPD):* cantidad de palabras distintas en el grupo, considerando cada palabra una sola vez; iv) *Índice de Cohesión (IC):* indicador del grado de coincidencia en la respuesta de los individuos y corresponde a la razón XR/NPD; v) *Índice de Disponibilidad Léxica (IDLV):* grado de disponibilidad de cada vocablo en el grupo. La forma más común de calcular IDLV, es a través de la Ecuación (1), donde f_p es la frecuencia de aparición del vocablo 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 posición t del lexicón. Por último, la expresión λ^{p-1} es la tasa de decaimiento o factor de ponderación a la posición, y su valor va decreciendo a medida que la posición es mayor.

(1)

Se realizó para cada nivel de la carrera un solo proceso de encuesta para los cinco centros de interés. Los datos fueron transcritos, tabulados y procesados para obtener los índices grupales y el índice de disponibilidad léxica de cada palabra, según Ecuación 1, con λ=0.9.

Distancia entre lexicones

Una forma de poder determinar la relación entre dos lexicones es definiendo una medida de distancia para cuantificar cuan relacionados están. Dentro de este contexto, la distancia de Levenshtein, utilizada

comúnmente en la corrección ortográfica, permite determinar cuan disimiles pueden ser un par de secuencias de símbolos (Yujian y Bo, 2007). Debido a que los lexicones están compuestos por una secuencia ordenada de palabras, se propone considerar el lexicón de los individuos como una secuencia de símbolos, donde los símbolos corresponden a los vocablos pertenecientes al lexicón.

La distancia de Levenshtein tiene una formulación algorítmica definida para dos listados de símbolos. Esta, no requiere la utilización de características grupales, y puede usarse para la comparación directa entre dos pares de secuencias independientes de símbolos, sin tener en consideración a que agrupación pertenecen o cual es la frecuencia de cada símbolo en el conjunto. La distancia de Levenshtein se basa en la aplicación de tres operaciones elementales: inserción, eliminación y substitución; las que son aplicadas sobre los símbolos que componen una de las dos secuencias a comparar, con tal de convertir esa secuencia en la otra. Por consecuencia, la distancia de Levenshtein entrega como resultado la cantidad mínima de símbolos que deben ser eliminados, insertados o sustituidos, con tal de transformar una secuencia en otra. Por ejemplo, dado dos lexicones de alumnos como conjuntos ordenados de símbolos l_1 = {'dato', 'azar', 'modelo', 'estadigrafo'} y l_2 = {'dato', 'azar', 'estadística'}, la distancia de Levenshtein para este par de secuencias es 2, debido a que se necesitan como mínimo 2 operaciones para transformar l_1 en l_2 (sustituyendo 'modelo' por 'estadística' y eliminando 'estadigrafo'). Así, dos secuencias de vocablos independiente de su longitud y procedencia (por ejemplo, otra muestra o centro de interés) pueden ser comparados directamente y cuantificar su relación (distancia) numéricamente. Por consecuencia, esta distancia entre lexicones puede interpretarse como la cantidad mínima de vocablos que son necesarios operar, para convertir el lexicón de un individuo en el lexicón de otro.

Análisis clustering entre lexicones

Una de las ventajas de tener una medida de distancia para comparar los lexicones, es que hace posible la aplicación de técnicas de análisis de clúster que permiten determinar sub-agrupaciones derivadas de sus similitudes. El algoritmo utilizado en el enfoque propuesto corresponde al algoritmo de conglomerado jerárquico WPGMA (Weighted Pair Group Method with Arithmetic Mean) (Xu y Wunsch, 2008). Este algoritmo de tipo aglomerativo permite obtener (mediante distancia entre pares de elementos) agrupaciones de elementos por similitud, formando grupos jerárquicos de los elementos que pueden ser visualizados en forma de árbol o dendrograma. WPGMA es un algoritmo comúnmente utilizado en la filogenética, permitiendo analizar las relaciones evolutivas de las especies a través de la conformación de árboles filogenéticos de secuencias de RNA (Hori et al., 1985). En nuestro enfoque, el clúster jerárquico permitirá establecer sub-agrupaciones de lexicones por pares de individuos, permitiendo visualizar y cuantificar una jerarquía de lexicones en forma de árbol. En una primera instancia, el algoritmo considera a cada lexicón en forma individual y sin agrupar (hojas del árbol). Luego, estos se van agrupando según la distancia de Levenshtein hasta determinar un último nivel, en el cual todos los lexicones pertenecen a un solo grupo (raíz del árbol). Cabe destacar que el nivel de agrupación en la jerarquía depende de las necesidades de análisis, y aunque existen diversos criterios para determinar automáticamente la cantidad óptima de grupos, nuestro enfoque hace uso de la determinación visual de las agrupaciones a través del dendrograma.

RESULTADOS

A continuación, se muestran los resultados del análisis de disponibilidad léxica mediante la metodología propuesta. Cabe notar que cada vocablo o palabra que aparece como parte de un lexicón, no contiene tildes ni espacios en blanco, ya que estos han sido eliminados en el pre-procesamiento de la información, para disminuir diferencias debido a errores ortográficos en las respuestas. Además, algunas palabras y lexicones han sido truncadas en ciertas ocasiones para hacerlas corresponder con el espacio disponible, y poder desplegarlas al interior de gráficos y tablas sin perder legibilidad. Cuando una palabra se encuentre truncada al final de esta aparecerá el símbolo "~", indicando el lugar de corte.

Índices grupales

En la Tabla 2 se muestran los valores de los índices NPD, XR e IC para cada centro de interés y los cuatro niveles de alumnos. Como se puede observar, NPD, XR e IC son mayores a medida que aumenta el nivel para casi todos los centros de interés. En los niveles intermedios como N2 se puede apreciar el más alto de los IC. Sin embargo, esto se debe al bajo NPD producto de los pocos alumnos existentes en ese nivel (ver Tabla 1). Así mismo, el nivel N3 presenta los más bajos índices IC debido a la gran dispersión provocada principalmente por un alto NPD. Esto indica que a medida que los alumnos aumentan su nivel (permanecen más años en la universidad), su léxico disponible va aumentando (NPD, XR) y homogeneizando de manera irregular (IC). De acuerdo a lo anterior, como objetivo de análisis los demás resultados están centrados en los niveles extremos N1 y N4 con tal de resaltar las diferencias que existen entre estos.

Tabla 2: Índices grupales

Índice	*Nivel*	*AZAR*	*CALC*	*ESTR*	*GEOM*	*NUME*
NPD	N1	152	148	110	204	122
	N2	86	100	70	118	86
	N3	153	174	140	164	170
	N4	172	189	129	187	135
XR	N1	11.92	11.31	8.68	19.88	9.42
	N2	14.55	15.73	11.27	25.82	13.00
	N3	14.45	15.25	11.70	20.35	13.50
	N4	17.60	17.00	14.70	26.10	14.70
IC	N1	0.078	0.076	0.079	0.098	0.077
	N2	0.169	0.157	0.161	0.219	0.151
	N3	0.094	0.088	0.084	0.124	0.079
	N4	0.103	0.090	0.114	0.140	0.109

La Tabla 3 presenta los 15 vocablos más disponibles de cada centro de interés entre los niveles extremos N1 y N4. Los vocablos están ordenados de acuerdo a su disponibilidad (IDLV) de mayor a menor (de arriba hacia abajo). De esta se puede destacar los siguientes:

Centro de Interés AZAR: En N1 existen vocablos genéricos más asociados a conceptos comunes del azar como 'dado', 'juego' y 'acertijo'. Por el contrario, en N4 aparecen altamente disponibles vocablos como 'moda', 'muestra' y 'frecuencia', siendo estos conceptos más propios del área de Datos y Azar.

Centro de Interés CALC: En N1, los vocablos genéricos como 'numero', 'adicion' y 'sustraccion' están presentes como altamente disponibles. Por otro lado, en N4 esos vocablos bajan su disponibilidad o desaparecen, dando lugar a la valoración de vocablos más propios del área como 'diferencial', 'infinito' y 'teorema' entre otros.

Centro de Interés ESTR: En N1 hay vocablos menos relacionados con el área como 'algebra', 'conjunto', 'matriz', los que disminuyen su disponibilidad o desaparecen en favor de vocablos más específicos como 'grupo',' anillo' y 'cuerpo'.

Centro de Interés GEOM: No se observa mucha diferencia de vocablos entre los niveles, salvo en el orden de disponibilidad de estos. En ambos predominan las figuras geométricas y sus elementos. Sin embargo, en N4 aparecen algunos vocablos distinguibles como 'congruencia' y 'euclides'.

Centro de Interés NUME: Tampoco se observa mucha diferencia en los niveles, salvo el orden de los vocablos. En ambos predominan vocablos referentes a distintos conjuntos de sistemas numéricos y operaciones aritméticas.

Se pueden observar diferencias entre la mayoría de los niveles N1 y N4, tendientes a representar un vocablo más técnico y especializado en N4 que en N1, como se podría esperar. Sin embargo, esto es más evidente en AZAR, CALC y ESTR, y no tanto para GEOM y NUME.

Tabla 3: Vocablos más disponibles por Centro de Interés y Nivel

AZAR		CALC		ESTR		GEOM		NUME	
N1	N4	N1	N4	N1	N4	N1	N4	N1	N4
probabili~	probabili~	Numero	Integral	algebra	grupo	area	Triangulo	Numero	numero
estadística	Dato	Integral	Derivada	conjunto	anillo	angulo	Angulo	Algebra	ecuacion
Dado	estadística	Derivada	Limite	matriz	cuerpo	circunferen~	circunferen~	Orden	incognita
Dato	Azar	Adicion	Función	estructu~	conjunto	perímetro	Recta	Adicion	sistema
Moda	Moda	Limite	diferencial	polinomio	abeliano	triangulo	Cuadrado	ecuación	suma
Azar	muestra	sustraccion	Teorema	axioma	operacion	parábola	Punto	multiplicación	variable
Mediana	mediana	Calculo	Volumen	numeros	conmutativ~	hiperbola	Rectángulo	sustracción	resta
Media	Media	multiplicaci~	Infinito	Anillo	relación	recta	Área	conjunto	simbolo
porcentaje	frecuencia	Función	Área	binomio	demostraci~	elipse	Euclides	naturales	multiplicaci~
Razón	varianza	Variable	continuidad	cuerpo	isomorfismo	figura	Perímetro	matriz	polinomio
informaci~	Dado	División	lhopital	orden	campo	cuadrado	lado	reales	inecuacion
juego	poblacion	Raíz	variable	trinomio	asociativid~	vertice	congruencia	decimal	division
frecuencia	porcentaje	Fracción	antiderivad	ecuacion	subgrupo	lado	semejanza	matemática	operacion
acertijo	grafico	incognita	lateral	operacion	axioma	geometria	volumen	división	igualdad
grafico	promedio	Resolver	Serie	real	teorema	punto	paralela	símbolo	grado

Diferencias entre niveles de los lexicones

Los listados como los de la Tabla 3, pueden ser considerados como la respuesta grupal a un centro de interés de un nivel específico. Debido a que el método propuesto permite calcular la distancia entre dos lexicones, es posible considerar la respuesta grupal como el lexicón grupal. De esta forma, es posible obtener la distancia del lexicón de cada individuo a la respuesta global del grupo, lo que puede otorgar una medida comparativa de cuanto se distancian las respuestas de los individuos por cada nivel.

Con tal de buscar diferencias significativas entre las diferencias de cada nivel, para cada centro de interés se realizó un análisis de varianza ANOVA de un factor. Definiendo a los niveles como factores y la distancia de cada lexicón individual al lexicón del grupo como medida, se pudo constatar que el efecto del nivel cursado sobre la distancia de los lexicones al grupo, fue significativa para AZAR ($F(3, 75) = 4.56$, $p < 0.05$, $\eta^2=0.154$) y CALC ($F(3, 75) = 3.432$, $p < 0.05$, $\eta^2=0.121$). Sin embargo, para los centros de interés ESTR ($F(3, 74) = 4.309$, $p = 0.098$), GEOM ($F(3, 75) = 2.427$, $p =0.072$) y NUME ($F(3, 75) = 0.463$, $p =0.709$) no se encontró evidencia de que fueron afectados significativamente por los niveles que cursaban los alumnos.

Las pruebas post-hoc, que corresponden a la comparación por pares entre todos los niveles de cada centro de interés, revelaron para el centro de interés AZAR, que sólo había diferencias significativas entre los niveles N2 y N3 ($p<0.05$), mientras que para el centro de interés CALC las diferencias significativas sólo fueron evidentes entre los niveles N1 y N4 ($p<0.05$).

Estructura Jerárquica de Grupos

Con tal de comprobar la factibilidad de realizar la agrupación por lexicones, se ha realizado un proceso de clustering jerárquico sobre los centros de interés que presentaron diferencias significativas en sus niveles. Los resultados de este proceso, se pueden apreciar en las Figuras 1 y 2, los que hacen referencia a la estructura de los grupos correspondiente a los niveles AZAR-N3 y CALC-N4. Las figuras generadas por el algoritmo WPGMA, obtienen un dendrograma con la asociación de pares de lexicones agrupados en forma jerárquica. De este modo, por un lado, el eje horizontal “Distancia” representa una medida de disimilitud entre los lexicones o subgrupos que se muestran unidos por una línea vertical. Por otro lado, el eje vertical “Lexicones” presenta a cada uno de los lexicones informados por los alumnos.

Para efectos de una mejor utilización de los espacios, cada lexicón es descrito como una lista de hasta seis palabras (lexicones más largos que este número son truncados), en el mismo orden que han sido producidas por los informantes, siendo las de la izquierda las palabras más disponibles (las primeras informadas como respuesta). Además, para efectos de normalización, y presentar medidas acotadas gráficamente, se han considerado para la confección del gráfico sólo los vocablos más disponibles, es decir con IDLV mayor a la media del grupo.

Para interpretar los dendrograma presentados, se debe considerar la distancia entre dos elementos como el lugar del eje “Distancia” donde estos elementos se unen. Por ejemplo, en la Figura 1, el primer lexicón (superior), el cual empieza por los vocablos ‘suerte’ y ‘casino’, tiene una distancia superior a 10 con el resto de todos los lexicones, debido a que existe una línea vertical a la izquierda del 10, que une el primer lexicón con el subgrupo formado por todos los otros lexicones.

Además, con tal de ayudar en la visualización de los resultados, a cada dendrograma se le han incorporado líneas horizontales que dividen las subgrupaciones, y etiquetas compuestas por una letra mayúsculas entre paréntesis, que permiten identificar a las subagrupaciones más relevantes. Para determinar el nivel de agrupación, se ha determinado manualmente un umbral de distancia mostrado a través de una línea punteada que cruza el eje horizontal. Como se puede apreciar en las Figuras 1 y 2, se ha definido el valor de los umbrales en 9 y 8 respectivamente.

La Figura 1 del grupo AZAR-N3 presenta cinco agrupaciones destacables. El subgrupo (B) tiene una alta presencia de los vocablos ‘moda’, ‘media’ y ‘mediana’. El subgrupo (D) está principalmente cohesionado mediante los vocablos ‘dado’, ‘juego’ y ‘numero’. El subgrupo (E) tiene una alta presencia de los vocablos ‘estadística’ y ‘moneda’. El subgrupo (C) tiene una agrupación de sólo dos lexicones, ambos con los mismos vocablos iniciales ‘dato’ y ‘azar’. Por último, existe un lexicón aislado (A), producto de que el orden de sus vocablos dista considerablemente de los presentes en el grupo. En adición, se puede observar que los subgrupos (C), (D) y (E) son los más similares entre ellos. Al contrario, el subgrupo (B) es el más distinto a todos, no encontrando relación con otro subgrupo directamente, sino que relacionándose con el conjunto formado por los otros subgrupos (C), (D) y (E).

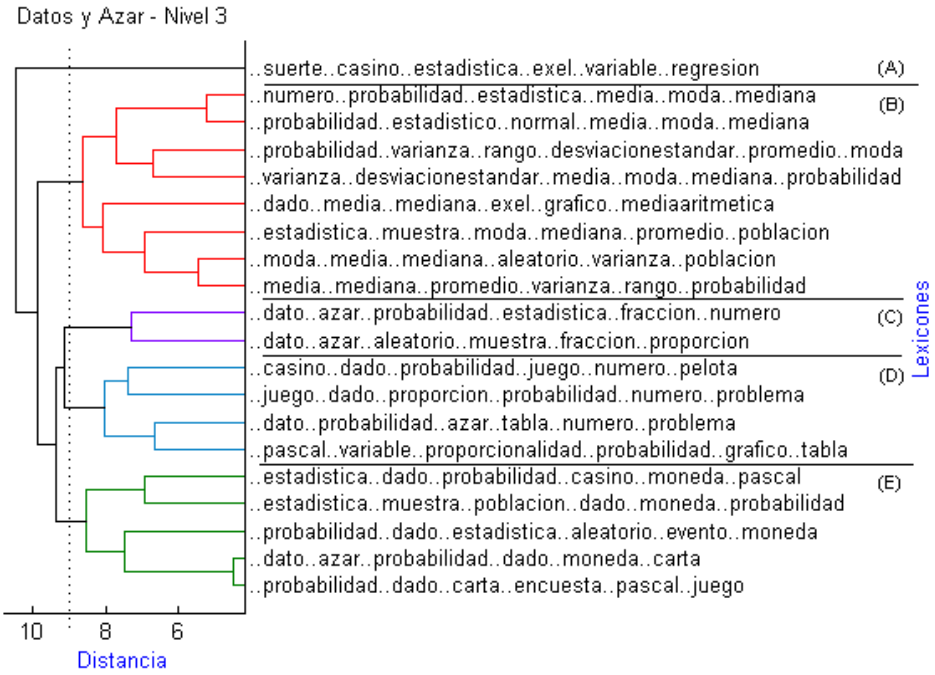

Fig 1: Jerarquías de lexicones para AZAR-N3

En la Figura 2, se detalla la estructura del grupo CALC–N4, donde se pueden visualizar cinco posibles subgrupos. Por un lado, el subgrupo (B) está dominado por los vocablos 'integral' y 'derivada', mientras que el subgrupo (C), que es similar al (B), incorpora el vocablo 'limite' dentro de las primeras posiciones. El subgrupo (D), está principalmente conformado por lexicones que comienzan por los vocablos 'limite' y 'derivada', incluyendo el vocablo 'infinito' en otras posiciones. Por otro lado, el subgrupo (E) está principalmente conformado por los vocablos 'derivada', 'limite' e 'integral', agregando vocablos más específicos como 'lhopital' y 'teorema'. Contrariamente a esto, existe un lexicón aislado que no se asemeja en el orden a otros lexicones, quedando fuera de los otros subgrupos. Además, las jerarquías permiten distinguir que los subgrupos (B) y (D) se asemejan más a los subgrupos (C) y (E) respectivamente.

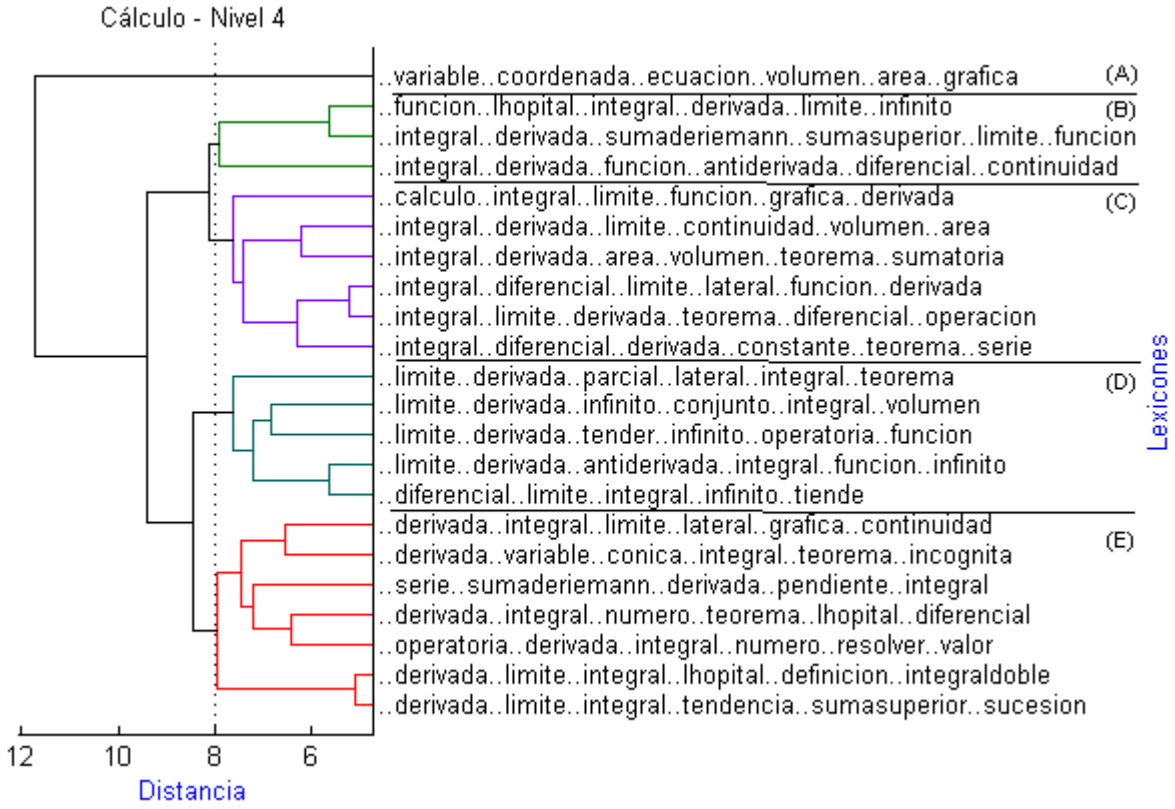

Fig 2: Jerarquías de lexicones para CALC-N4

DISCUSIÓN

Los resultados obtenidos concuerdan con otros estudios (Ferreira et al., 2014; Salcedo y Del Valle, 2013), respecto al hecho de que el léxico disponible, se incrementa a medida que aumenta el tiempo de permanencia de los alumnos en los ambientes educativos. Particularmente, a través del método propuesto, el modelo ANOVA, determinó una diferencia significativa en el léxico disponible de los alumnos, particularmente en los centros de interés AZAR y CALC. De esta forma, los contrastes existentes entre el léxico disponible de alumnos con solo un año de diferencia, concuerda con los resultados mostrados en (Castañeda et al., 2016), donde también se pudieron observar diferencias en este lapso de tiempo.

Se debe destacar el hecho, de que esta diferencia significativa, corresponde a la cantidad de transformaciones que se requieren, para convertir el léxico disponible de un nivel a otro. De esta forma, los resultados encontrados en el análisis ANOVA, indicarían que existe una diferencia significativa en la cantidad de modificaciones necesarias, para transformar el léxico disponible de AZAR-N2 en AZAR-N3, y de CALC-N1 a CALC-N4. Por el contrario, para el resto de niveles y centros de interés no se encontró evidencia para asegurar que se requiera una cantidad significativa de transformaciones, con tal de convertir los lexicones de un nivel a otro. Esto último, puede ser producto de la excesiva heterogeneidad de los lexicones al interior de cada nivel, o porque los niveles son muy parecidos entre ellos.

Como muestran los resultados de la Tabla 3, la diferencia del orden de los vocablos en los extremos de cada nivel, muestra una clara tendencia sobre la transformación del léxico disponible hacia una disponibilidad de vocablos más técnicos y propios de cada área, a medida que permanecen más tiempo los alumnos en la universidad. Por otro lado, los resultados de la Tabla 2 muestran una gran diferencia del índice IC entre los niveles N2 y N3. Sin embargo, es necesario tener en consideración que este índice no toma en cuenta el orden de las palabras, por lo que una mayor cohesión suele reflejar por construcción, una menor cantidad de palabras distintas en relación al largo promedio de los lexicones, lo que generalmente puede deberse a las diferencias de los tamaños de muestra, y no sólo a una diferencia entre los niveles de cohesión. En el mismo contexto, los resultados presentados en las estructuras jerárquicas permiten observar de forma gráfica la aglomeración de lexicones en un grupo. Así, la cohesión de la estructura se puede considerar como la cantidad de subgrupos y la estructura interna de estos.

Respecto a las consideraciones sobre la distancia de Levenshtein definida, esta no es comparable a los índices de disponibilidad léxica existentes, debido principalmente a que no es una fórmula, sino que un algoritmo. Además, esta no se determina el orden de importancia o disponibilidad de los vocablos, por lo que la propuesta no corresponde a un índice de disponibilidad léxica como IDLV, IDLA, GCV, o de comparación de vocablos como CRV. Igualmente, nuestra propuesta tampoco es un índice de disponibilidad léxica a nivel del individuo como IDLI o IDLINF, debido a que no hay una forma de representación del lexicón. En este sentido, como se indicó anteriormente, la distancia definida se puede considerar como la cuantificación de la cantidad de transformaciones necesarias para convertir un lexicón en otro, en base a operadores de secuencia de símbolos, que no asumen estructura o interpretación sobre los vocablos ni el lexicón mental. Debido a esto último, la distancia propuesta tampoco podría ser comparada con el enfoque presentado en (Ávila y Sánchez, 2010).

Lo anterior, lejos de ser una desventaja, permite considerar la relación entre lexicones como una medida de distancia que es independiente de la conformación o procedencia que estos tengan. Esto es importante, debido a que muchas de las técnicas estadísticas y algoritmos de reconocimiento de patrones, como el algoritmo de clustering WPGMA, requieren la definición de funciones de distancia con tal de asegurar un cálculo e interpretación correcta de las operaciones realizadas sobre los elementos. Además, esta distancia no requiere el cálculo o determinación de frecuencias en ningún nivel de agrupación, como si lo requieren los índices presentados anteriormente, lo que la hace invariante a los tamaños muestrales.

Para finalizar, el algoritmo de clustering utilizado intenta maximizar la homogeneidad interna de los subgrupos conformados. Esto se puede considerar, por analogía, como el intento de conformar subgrupos lo más cohesionados posibles. Por consiguiente, los resultados del clustering jerárquico permitirían complementar la interpretación del índice IC, describiendo la estructura jerárquica de la cohesión existente al interior de un grupo, teniendo en cuenta no sólo la coincidencia de los vocablos, sino que también la ordena de estos.

CONCLUSIONES

Según el método propuesto, lo mostrado en los resultados y la discusión, se puede concluir que: 1) el método de análisis de disponibilidad léxica propuesto, hace factible la comparación y análisis de lexicones entre individuos, sin la necesidad de información grupal; 2) es posible representar jerárquicamente los

lexicones de los estudiantes y mostrar de forma gráfica la conformación de las relaciones entre los alumnos, según la estructura de cohesión de los centros de interés; 3) se agrega la posibilidad de disponer un método de comparación cuantitativo entre lexicones, que permite el uso de técnicas de estadística tradicionales y algoritmos de reconocimiento de patrones que requieren una medida de distancia; 4) el método propuesto permite complementar los análisis de disponibilidad léxica tradicionales, enfocando su utilidad en la descripción de la relación entre individuos y grupos en contextos educativos; 5) se puede interpretar la distancia de lexicones definida como la cantidad de vocablos que se deben agregar, eliminar o modificar, para transformar un lexicón determinado en otro.

Adicionalmente, sobre la utilidad de la metodología empleada, esta podría potencialmente ser útil para determinar en forma automática diferencias y similitudes del léxico de los alumnos, lo que permitiría, por ejemplo, realizar conformación de grupos donde se requiera que la interacción se pueda desarrollar eficientemente, preocupación que debería ser considerada prioritaria en el aprendizaje colaborativo (Calzadilla, 2002). Por otro lado, esta misma estructura jerárquica puede ser utilizada por el profesor para determinar subagrupaciones del léxico, que le permitan desarrollar mejor el discurso, teniendo en cuenta las relaciones de disponibilidad léxica entre los alumnos, y complementando los índices de disponibilidad léxica que basan su formulación en la frecuencia de palabras e información grupal.

AGRADECIMIENTOS

Los autores agradecen a la Comisión Nacional de Investigación Científica y Tecnológica de Chile (CONICYT), en virtud de la adjudicación del Proyecto de Investigación Fondecyt 1140457.

REFERENCIAS

Ariza, A. y Oliva, S., Las nuevas tecnologías de la información y la comunicación y una propuesta para el trabajo colaborativo, V Congreso Iberoamericano de Informática Educativa, (2000)

Ávila, A., y Sánchez, J., La disponibilidad léxica. Antecedentes y fundamentos, Variación social del léxico disponible en la ciudad de Málaga. Diccionario y análisis, 37-81, (2010)

Ávila, A., y Sánchez, J., La posición de los vocablos en el cálculo del índice de disponibilidad léxica: procesos de reentrada en las listas del léxico disponible de la ciudad de Málaga, ELUA: Estudios de Lingüística. Universidad de Alicante, 25, 45-74, (2011)

Bartol, J. y Hernández, N., Dispolex: banco de datos de la disponibilidad léxica, Panel de investigación presentado en el VI Congreso de Lingüística General, Santiago de Compostela, 3-7 de mayo (2004)

Bernal, R. F., Representaciones de género de profesores y profesoras de matemática, y su incidencia en los resultados académicos de alumnos y alumnas, 43, 103-118, (2007)

Callealta, F. y Gallego, D., Medidas de disponibilidad léxica: comparabilidad y normalización, doi: 10.4067/S0718-93032016000100002, Boletín de filología (en línea), 51(1), 39-92, (2016)

Calzadilla, M.E., Aprendizaje colaborativo y tecnologías de la información y la comunicación, Revista Iberoamericana de Educación, ISSN: 1681-5653 (en línea), 1(10), 1-10, 2002. http://rieoei.org/tec_edu7.htm, Acceso: 15 de mayo (2017)

Carrió, M.L., Ventajas del uso de la Tecnología en el Aprendizaje Colaborativo, Revista Iberoamericana de Educación, ISSN: 1681-5653 (en línea), 41(4), 1-10, 2007, http://rieoei.org/1640.htm, Acceso: 15 mayo (2017)

Casado, E., Prototipos de la interacción pedagógica, Revista de Pedagogía, 23, 247-279, (2002)

Castañeda, S., Pacheco, A. y Palomares, S., Disponibilidad Léxica Matemática en estudiantes de Ingeniería y Ciencias. UNIÓN, Revista Iberoamericana de Educación Matemática, (47), 44-61, (2016)

Chevallard, Y., La transposition didactique. Du savoir savant au savoir enseigné, 2nd Ed., La Pensée Sauvage Editions, Grenoble perspectives, Encyclopedia of mathematics education, Springer, Dordrecht, (1991)

Del Valle, M., Salcedo, P. y Ferreira, A., Analyzing the Availability of Lexicon in Mathematics Education Using no Traditional Technological Resources, Journal of Supply Chain Management, 5(2), 144-149, (2016)

Echeverría, M., Vargas, R., Urzua, P. y Ferreira, R., DispoGrafo: una nueva herramienta computacional para el análisis de relaciones semánticas en el léxico disponible RLA, doi: 10.4067/S0718-48832008000100005, Revista de Lingüística Teórica y Aplicada (en línea), 46, 81-91, (2008)

Ferreira, A., Salcedo, P. y del Valle, M., Estudio de disponibilidad léxica en el ámbito de las matemáticas, doi: 10.4067/S0071-17132014000200004, Estudios Filológicos (en línea), 54, 69-84, (2014)

Germany, P. y Cartes, N., Léxico disponible en inglés como segunda lengua de instrucción formalizada, Estudios Pedagógicos, 26, 39-50, (2000)

Hori, H., Lim, B. y Osawa, S., Evolution of green plants as deduced from 5S rRNA sequences, Proceedings of the national academy of sciences, 82(3), 820-823, (1985)

Laal, M., Naseri, A.S., Laal, M., y Khattami-Kermanshahi, Z., What do we Achieve from Learning in Collaboration?, Procedia-Social and Behavioral Sciences, 93, 1427-1432, (2013)

López, H., Los estudios de disponibilidad léxica: pasado y presente, Boletín de Filología, ISSN: 0718-9303 (en línea), 35, 245-259, 1993, https://goo.gl/Ui4DXC, Acceso: 5 de mayo, (2017)

López, H. y Strassburger, C., Otro cálculo del índice de disponibilidad léxica. En Actas del IV Simposio de la Asociación Mexicana de Lingüística Aplicada, Presente y perspectiva de la lingüística computacional en México. México: Universidad Nacional Autónoma de México, (1987)

López, H. y Strassburger, C., En Actas del II Seminario sobre "Aportes de la lingüística a la enseñanza de la lengua materna", Universidad de Puerto Rico, 91-112, (1991)

Lorán, R. y López, H., Nouveau calcul de l'indice de disponibilité, Universidad de Puerto Rico, (1983)

Madrigal-Melchor, J., Enciso-Muñoz, A., Contreras-Solorio, D. A., Rivera-Juárez, J. M., y López-Chávez, J., Propuesta de Enseñanza con Base en la agrupación de Términos Marcados por el IDL y del Coeficiente de Relación entre Vocablos. Latin-American J. Physics Education, ISSN: 1870-9095 (en línea), 4, 1033, (2010)

Mangado-Cruz, M. y Areta-Lara, M., Procesamiento informático de datos para la elaboración de diccionarios de disponibilidad léxica, Actas del XXXVII Simposio Internacional de la Sociedad Española de Lingüística, 479-493, (2008)

Marines, M.D.S., Heredia, N. G., Solís, L. E., y Mena, D. A., Taller Multidisciplinario para el Desarrollo de Competencias de Comunicación Lingüística de la Investigación, doi: 10.4067/S0718-50062014000200006, Formación Universitaria (en línea), 7(2), 41-50, (2014)

Martín, J.V., Tadeo, F., Álvarez, T., y Peláez, J., Equipo Didáctico para Aprendizaje Colaborativo en Automatización e Informática Industrial, doi: 10.4067/S0718-50062009000500005, Formación Universitaria (en línea), 2(5), 31-40, (2009)

Mendoza, M. A. G., La transposición didáctica: historia de un concepto, Revista Latinoamericana de Estudios Educativos, ISSN: 1900-9895 (en línea), 1, 83-115, 2005, https://goo.gl/Fp7H22, Acceso: 15 de mayo, (2017)

Michea, R., Mots fréquents et mots disponibles, un aspect nouveau de la statistique du langage, Langues Modernes, 47, 338-344, (1953)

MINEDUC Chile 2012, Estándares orientadores para carreras de pedagogía en educación media, Ministerio de Educación, https://goo.gl/kFounE, Último acceso: 21 de nov. (2016)

Piaget, J., Psicología y pedagogía, Crítica, Barcelona, España, (2001)

Salcedo, P. y del Valle, M., Disponibilidad Léxica Matemática en Estudiantes de Enseñanza Media de Concepción, Chile, Atenas, ISSN: 1682-2749 (en línea), 1(21), 1-16, 2013, https://goo.gl/1hT57w, Acceso: 15 de mayo (2017)

Salcedo, P., Ferreira, A. y Barrientos, F., A bayesian model for lexical availability of chilean high school students in mathematics, En 5th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2013, Mallorca, Spain, June 10-14, 245-253, (2013)

Urzúa, P., Sáez, K., y Echeverría, M.S., Disponibilidad léxica matemática: análisis cuantitativo y cualitativo, doi: 10.4067/S0718-48832006000200005, RLA., Revista de lingüística teórica y aplicada (en línea), 44(2), 59-76, (2006)

Valencia, A. y Echeverría, M., Disponibilidad léxica en estudiantes chilenos, Santiago de Chile, Ediciones Universidad de Chile–Universidad de Concepción, (1999)

Xu, R., y Wunsch, D., Clustering (Vol. 10), John Wiley & Sons, (2008)

Yujian, L., y Bo, L., A normalized Levenshtein distance metric, IEEE transactions on pattern analysis and machine intelligence, 29(6), 1091-1095, (2007)