

## Evaluación de resúmenes en español con Análisis Semántico Latente: Una implementación posible\*

### Summaries assessment in Spanish using Latent Semantic Analysis: A possible implementation

**René Venegas**  
rene.venegas@ucv.cl  
Pontificia Universidad Católica de Valparaíso  
Chile

Recibido: 19-IV-2010 / Aceptado: 19-VIII-2010

**Resumen:** Evaluar resúmenes es una tarea compleja debido a que presenta problemas relacionados con la sistematicidad en la evaluación y con la dedicación de gran cantidad de tiempo por parte de los evaluadores. Estos problemas han motivado la indagación de métodos confiables de evaluación automatizada de resúmenes. En este contexto, el objetivo de este trabajo es identificar un método eficiente de evaluación automatizada de resúmenes, basado en el Análisis Semántico Latente (LSA). Los resúmenes fueron realizados por estudiantes de educación secundaria de la ciudad de Valparaíso, Chile. Para llevar a cabo esta tarea se correlacionan los puntajes de la evaluación realizada por tres docentes a 224 resúmenes provenientes de textos predominantemente expositivos, y 129 resúmenes provenientes de textos predominantemente narrativos con los puntajes entregados por tres métodos computacionales, basados en el LSA. Los métodos utilizados son los siguientes: 1) comparación de los resúmenes con el texto fuente de donde provienen, 2) comparación de los resúmenes con un resumen consensuado por lingüistas y 3) comparación de los resúmenes con resúmenes contruidos por los tres docentes evaluadores a partir de los mismos textos fuente. Los resultados generales indican que los puntajes de los métodos 2 y 3 son similares estadísticamente a los puntajes asignados por los docentes evaluadores, cuando los resúmenes evaluados provienen de textos predominantemente narrativos. Sin embargo, esta similitud no se replica cuando los resúmenes provienen de textos predominantemente expositivos.

**Palabras Clave:** Evaluación automatizada de resúmenes, similitud semántica, Análisis Semántico Latente, textos narrativos, textos expositivos.

**Abstract:** Summary assessment is a complex task, due to the problems related to systematicity and to the time it takes. These problems have motivated the study of reliable automatized assessment methods. In this scenario, the aim of this work is to identify an efficient method for assessing summaries, based on Latent Semantic Analysis (LSA). The summaries were written by students from secondary school in the city of Valparaíso, Chile. To accomplish the mentioned aim, we correlated the scores assigned by three teachers to 244 summaries of predominantly expository texts and 129 summaries of predominantly narrative texts with the scores provided by three computational methods, based on LSA. The methods are: 1) comparison of summaries with the source text, 2) comparison of summaries with a summary developed by the consensus of a group of linguists, and 3) comparison of summaries with three summaries constructed by three language teachers. General results show that the scores assigned by method 2 and 3 are statistically similar to the scores assigned by the teachers, when the source texts are predominantly narrative. However, this similarity is not so when the source text is predominantly expository.

**Key Words:** Automatic summary assessment, text comprehension, semantic similarity, Latent Semantic Analysis, narrative texts, expository texts.

## INTRODUCCIÓN

La educación de la lecto-escritura en Latinoamérica, según diversos autores (Peronard, 1989; Peronard, Gómez, Parodi & Núñez, 1998; Parodi, 2007; PISA, 2007; Ibáñez, 2008), no ha sido lo suficientemente efectiva como para que nuestros estudiantes desarrollen las competencias necesarias para comprender y producir textos que les permitan desenvolverse adecuadamente en la sociedad actual. En general, existe consenso entre los investigadores de la comprensión del texto escrito en lo que se refiere a la necesidad de describir y evaluar los procesos psicolingüísticos relativos a la comprensión, con el fin de diagnosticar y apoyar a los estudiantes en las tareas de comprensión (Marinkovich, Peronard & Parodi, 2006; Parodi, Peronard & Ibáñez, 2010). En el caso de la evaluación se destacan múltiples opciones de técnicas evaluativas, por ejemplo, respuesta a preguntas por medio de alternativas, la técnica cloze, el desarrollo de mapas conceptuales y/o mentales, respuesta a preguntas abiertas, parafraseo y escritura de resúmenes.

En nuestro caso, nos interesa indagar en una de estas opciones, a saber, el resumen como técnica de evaluación de la comprensión. El uso de esta técnica encuentra apoyo teórico en los planteamientos de Kintsch y van Dijk (1978) y de van Dijk y Kintsch (1983). Sin embargo, desde el punto de vista evaluativo, utilizar el resumen como técnica de evaluación de la comprensión presenta algunos problemas. El primero de ellos es, sin duda, que el resumen es un producto textual, por lo que en el momento de la evaluación de la comprensión no se pueden

aislar de modo discreto los efectos de los procesos psicolingüísticos relacionados con la producción, realizados por los sujetos comprendedores. En este mismo sentido, el establecimiento de criterios de evaluación y el entrenamiento de los evaluadores es un aspecto de suyo relevante. Ahora bien, si se considera la posibilidad de establecer claramente estos criterios es probable el surgimiento de otro tipo de problemas, sobre todo, en la evaluación de grandes cantidades de resúmenes. Estos problemas estarán probablemente relacionados con variables como la sistematicidad en la aplicación de los criterios, la atención consciente o inconsciente a elementos formales (ortografía, caligrafía) de la producción escrita que pueden incidir en su evaluación, aspectos subjetivos que puedan intervenir (cansancio, estrés, malhumor, etc.) y, por sobre todo, la extensa cantidad de tiempo requerida para este tipo de evaluación.

Los problemas de sistematicidad en la evaluación así como el de los requerimientos de extensa cantidad de tiempo han motivado el estudio del resumen y su evaluación desde la perspectiva computacional. Así lo expresan, por ejemplo, Yulan, Siu y Tho (2009: 890):

“In today’s computerized world, teachers are still required to grade students’ written summaries manually. This is a very time-consuming task that reduces the amount of time teachers can devote to other duties. In order to reduce the amount of time they have to spend on grading these summaries, many teachers have chosen to reduce the number of summaries given to their students.

However, in doing so, students will have insufficient practices, thereby affecting their summary writing skills. To tackle this problem, one approach is to provide computer-assisted assessment of summary writings”.

De este modo, la evaluación automatizada de resúmenes ha sido una tarea que se ha abordado desde mediados de los años sesenta, aunque aún no se han encontrado sistemas computacionales suficientemente confiables (Mani & Maybury, 2001; Batten, 2003). No obstante, los avances realizados a partir de la lingüística computacional y el procesamiento del lenguaje natural permiten pensar que hoy se está más cerca de encontrar métodos eficientes de evaluación automatizada de resúmenes escritos (Yulan et al., 2009).

Entre los diversos métodos computacionales existentes se destacan aquellos basados en algoritmos de procesamiento del lenguaje natural (PLN), en particular el Análisis Semántico latente (LSA por su sigla en inglés), para evaluar automatizadamente diversos productos textuales, entre ellos los resúmenes escritos en lengua inglesa (Landauer, Laham, Rehder & Schreiner, 1997; Rehder, Schreiner, Wolfe, Laham, Landauer & Kintsch, 1998; Kintsch, Steinhart, Stahl, LSA Research Group, Matthews & Lamb, 2000; Landauer & Psootka, 2000; Landauer, Laham & Foltz, 2003). Estos estudios han mostrado que los métodos basados en LSA son efectivos en la evaluación automatizada de ensayos y resúmenes. Sin embargo, en el ámbito de la evaluación automatizada de resúmenes en lengua española, los estudios que utilizan el LSA aún son insuficientes (Pérez, Alfonseca, Rodríguez, Gliozzo, Strapparava & Magnini, 2005; León, Olmos, Escudero, Cañas & Salmerón, 2006; Venegas, 2005, 2006a, 2007, 2009a, 2009b; Silva, 2008; Órdenes, 2009).

Con el fin de ampliar los estudios en esta área en lengua española y aportar mayor evidencia empírica respecto de la posibilidad de evaluar automatizadamente resúmenes, nos planteamos el objetivo de identificar un método de evaluación automatizada de resúmenes, basado en el Análisis Semántico Latente, que permita evaluar eficientemente la calidad del contenido de resúmenes de textos predominantemente narrativos y expositivos escritos en español de modo similar a como lo hacen tres docentes evaluadores. Para ello, se comparan los resultados de la evaluación realizada con tres métodos basados en LSA con los resultados provenientes de la evaluación realizada por los tres expertos humanos.

En este artículo revisaremos, en primer lugar, algunos aspectos relativos al resumen como técnica de evaluación; en segundo lugar, explicaremos brevemente el funcionamiento del LSA y algunas investigaciones en las que se ha utilizado para evaluar automatizadamente resúmenes. En tercer lugar, presentaremos la investigación propuesta y los resultados. Finalmente, cerraremos este artículo con las conclusiones correspondientes.

## **I. El resumen como técnica de evaluación de la comprensión del texto escrito**

Desde el punto de vista pedagógico, el hecho de que los estudiantes produzcan un resumen es tanto una estrategia didáctica como una técnica evaluativa. En el primer caso, la estrategia estará orientada hacia el desarrollo de la capacidad de sintetizar la información que se presenta en un texto, explicando con sus propias palabras de qué se trata. En el segundo caso, estará orientada a evaluar si los estudiantes comprendieron el texto fuente. Así, el resumen se convierte en una evidencia textual de cuánto y cómo se comprenden el contenido del texto leído. Por lo mismo, el resumen se ha convertido en una de las técnicas que más se utiliza para evaluar la comprensión de un texto (Farr & Carey, 1986; Peronard, 1997; Zipitria, Elorriaga, Arruate & de Ilarraza, 2004). Ello, a pesar de que, como ya mencionamos, se presentan algunos problemas. Por ejemplo, la falta de sistematicidad en la aplicación de las pautas de evaluación y la falta del tiempo que se requiere para la evaluación de los resúmenes.

No obstante lo anterior, evaluar un resumen permite contar con una visión más precisa de lo que los estudiantes comprenden del texto, en comparación con las preguntas de selección múltiple, las pruebas de respuestas cortas (Steinhart, 2001) e, incluso, las preguntas abiertas al final de la lectura de un texto (Wade-Stein & Kintsch, 2004). Por consiguiente, escribir resúmenes parece ser la técnica que resulta más apropiada para evaluar la construcción de la macroestructura de un texto por parte de los estudiantes. En este sentido, tal como plantean Peronard (1997) y Landauer y Psootka (2000) esta técnica permite diferenciar, de modo relativamente fácil, a los estudiantes que entienden la esencia del texto de los que no lo hacen.

Como planteamos más arriba, el uso de esta técnica se sustenta en los postulados teórico-empíricos propuestos por Kintsch y van Dijk (1978) y van Dijk

y Kintsch (1983). Estos autores plantean que, en el proceso de comprensión de un texto, los lectores se representan la información global contenida en el texto por medio de la denominada macroestructura. Esta macroestructura se deriva o se infiere a partir de los conjuntos de representaciones proposicionales, que hace el sujeto, de la interrelación entre las frases, las cláusulas y las oraciones del texto, y que permiten organizar el significado local del texto en la mente del sujeto. A estas representaciones proposicionales locales se les conoce con el nombre de microestructuras. Así, el lector para construir la macroestructura, hace uso de las macrorreglas y las macroestrategias, convirtiendo las secuencias proposicionales (microestructuras) en un conjunto más pequeño de proposiciones generales. Para ello, el lector eliminará aquellas que son menos importantes para el significado global que se construye del texto, generalizará las proposiciones para convertirlas en conjuntos más abstractos y construirá nuevos significados, que representan a los significados del texto fuente.

El funcionamiento de las macrorreglas y macroestrategias es recursivo, esto quiere decir que las macroestructuras que se construyen pueden someterse a otra secuencia de macrorreglas y macroestrategias, lo que generaliza aun más el contenido del texto. Por consiguiente, la macroestructura es una representación semántica del contenido semántico global del texto. En este sentido, la macroestructura es relevante no solo para manejar información compleja en la producción del texto escrito sino que también en la comprensión de él y en la recuperación de información.

Tal como plantea van Dijk (1980), la noción de macroestructura ha sido muy útil para dar cuenta de las propiedades del discurso así como del uso de la lengua. Por ejemplo, diversos experimentos han mostrado que estos significados globales son normalmente mejor recordados por los usuarios de la lengua. En consecuencia, la noción de macroestructura permite explicar el cómo y el porqué los usuarios de una lengua son capaces de sintetizar una conversación o un texto escrito en un nuevo texto, denominado resumen. En otras palabras, un resumen es la textualización de un significado que representa el significado global de un contenido textual, ya sea escrito u oral (van Dijk, 1980).

## 2. Análisis Semántico Latente (LSA)

El LSA es una técnica de procesamiento del lenguaje natural que permite representar matemáticamente las relaciones de significado entre palabras y párrafos de uno o varios textos. Esto es posible debido a que un determinado tipo de procesamiento matemático-estadístico posibilita identificar las relaciones de significados subyacentes existentes en una gran cantidad de textos, lo que permite asignar valores de similitud semántica a la relación entre palabras y entre párrafos (Landauer, McNamara, Dennis & Kintsch, 2007). El procedimiento que permite construir esta representación matemática se sustenta en la noción de significado relacional (Kintsch, 2007), por lo que no se identifica el significado conceptual o denotativo de una palabra sino que se calcula el grado de relación contextual existente entre dos palabras o dos unidades textuales (párrafos, apartados, etc.).

La identificación de las relaciones semánticas subyacentes se realiza a través de dos grandes etapas. En la primera de ellas el corpus textual es organizado en una matriz de palabras por párrafos, de modo tal que en cada celda de la matriz se cuantifica la frecuencia de aparición de cada palabra en cada uno de los párrafos. Los datos incluidos en esta matriz inicial son normalizados a través del cálculo del peso estadístico local y global, así como de la aplicación de una fórmula de entropía (*log entropy*). Este primer procedimiento permite establecer una suerte de equilibrio entre las palabras muy frecuentes, bajándole el peso estadístico, pues aportan poca información; y las menos frecuentes, aumentándole el peso estadístico, ya que aportan mayor cantidad de información. Las más frecuentes son normalmente aquellas conocidas como palabras funcionales y las menos frecuentes aquellas conocidas como palabras de contenido. En una segunda etapa, a la matriz sometida a normalización se le aplica un algoritmo denominado Descomposición en Valores Singulares (SVD, por su sigla en inglés). Este algoritmo permite construir una matriz de menor tamaño que la original, considerando en ella solo aquellas dimensiones o vectores más representativos de las frecuencias de las palabras de la matriz original. Esta reducción dimensional relevante, pues es por medio de este proceso es que se pueden potenciar las relaciones de significado de las palabras en los textos, que de otro modo no son posibles de identificar en

el corpus (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). Cabe mencionar que tal reducción dimensional se materializa en una matriz que representa la información contenida en corpus de textos, la que en el LSA se denomina espacio semántico (Martin & Berry, 2007).

El espacio semántico construido utilizando el LSA permite calcular las similitudes semánticas existentes entre palabras y párrafos de textos, estableciendo mediciones de su representación vectorial, a través del cálculo de coseno de sus ángulos en un espacio multivectorial (Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998; Martin & Berry, 2007). Los valores de coseno corresponden a 1 para vectores con la misma dirección (esto significa que lo medido es igual) y a 0 para aquellos vectores ortogonales (perpendiculares en el espacio multivectorial, es decir, que lo medido es completamente distinto), en tanto que todos los valores intermedios corresponden a diferentes grados de similitud entre los vectores, que representan a las palabras o párrafos (Deerwester et al., 1990; Landauer et al., 1998; Manning & Schütze, 2003).

Como puede inferirse de lo descrito anteriormente, la representación matemática que se obtiene con el LSA es parcial y limitada, pues solo utiliza como dato inicial la frecuencia de las palabras en cada uno de los párrafos de los textos del corpus, no considerando otro tipo de información lingüística tal como la categorización morfológica o las relaciones sintácticas presentes en los textos.

Para más detalles sobre LSA y su funcionamiento, se recomienda leer Landauer y Dumais (1997), Landauer et al. (1998), Landauer (2000), Venegas (2003), Landauer et al. (2007), entre otros.

Ahora bien, tal como mencionamos en la introducción diversos investigadores han hecho uso del LSA para proponer métodos de evaluación automatizada de resúmenes. Ello, fundamentalmente, porque el uso del LSA permite soslayar la dificultad de los métodos basado exclusivamente en correspondencia o co-ocurrencia de palabras entre el resumen y el texto fuente (Lin & Hovy, 2003; Lin, 2004), debido a que cuantifica la similitud entre las palabras y no solo la presencia o ausencia de ellas. De este modo, diversos estudios reportan una alta correlación entre la evaluación automatizada de resúmenes y los juicios de evaluadores humanos (Landauer et al., 1998; Kintsch et al., 2000; Landauer et al., 2003). Esto implica que, en general, las

puntuaciones obtenidas con el LSA son comparables con las evaluaciones que realizan los evaluadores (profesores con experiencia) a estos resúmenes. Tales resultados han permitido, incluso el desarrollo de programa que utilizan LSA en la evaluación y retroalimentación automatizada de ensayos. Uno de los más conocidos es el *Intelligent Essay Asesor* (IEA por su sigla en inglés), desarrollado por Landauer et al. (2003). En este programa, a diferencia de otras perspectivas estadísticas, los autores plantean que el LSA, al representar matemáticamente el contenido semántico de los textos más que aspectos mecánicos de la producción escrita como la gramática, la ortografía literal y puntual, permite obtener una evaluación más realista de la comprensión del texto escrito por parte de los sujetos. Lo anterior no significa que el IEA no proporcione retroalimentación sobre los aspectos formales en la evaluación del ensayo, de hecho el programa evalúa también aspectos como el estilo, la ortografía y la coherencia sintáctica.

Los resultados obtenidos a partir del uso del LSA en el IEA concuerdan con los hallazgos de Kintsch et al. (2000) y permiten pensar en la conveniencia de utilizar el utilizar el LSA en tareas de evaluación automatizada de resúmenes en otras lenguas, dado los buenos resultados obtenidos utilizando textos en lengua inglesa. De hecho, algunas investigaciones ya han realizado avances en lengua española. Por ejemplo, Pérez et al. (2005) proponen para la evaluación automatizada de textos el sistema de evaluación de respuestas en texto libre llamado Atenea (Alfonseca & Pérez, 2004), en el cual combinan una versión modificada del algoritmo *BiLingual Evaluation Understudy algorithm* (Papineni, Roukos, Ward & Zhu, 2001), utilizando tanto técnicas basadas en correspondencias de palabras como el LSA. Los resultados de los experimentos muestran que el uso del LSA aumenta la similitud entre los puntajes dados por Atenea y las notas dadas por los profesores para el mismo conjunto de preguntas, mejorando evaluaciones realizadas anteriormente en las que solo se consideraban el uso de n-gramas y otras técnicas de PNL. Cabe señalar que para el uso del LSA en Atenea se utilizó un corpus en español traducido automáticamente al inglés, utilizando *AltaVista Babelfish*, por lo que las pruebas se realizaron únicamente en lengua inglesa (Pérez et al., 2005). Esto introduce la problemática relacionada con la lengua, pues probablemente los resultados podrían variar si se considerara la versión original y no la traducida.

Otro ejemplo, del uso del LSA en la evaluación de resúmenes en español es la investigación realizada por León et al. (2006). En ella los investigadores evalúan resúmenes muy concisos (50 palabras de largo) provenientes de textos con predominancia narrativa y expositiva, mediante el LSA, y comparan estos resultados con las evaluaciones de cuatro evaluadores expertos humanos. Los resultados obtenidos por estos investigadores muestran que los métodos holísticos en base a LSA se correlacionan mejor con los expertos, en particular cuando se trata de resúmenes breves provenientes de textos narrativos.

Otras indagaciones sobre evaluación de resúmenes utilizando LSA en español han sido las llevadas a cabo por Venegas (2006b, 2009a, 2009b). En estas investigaciones se realizaron diversas comparaciones entre el puntaje calculado por LSA y la evaluación realizada por tres docentes de lengua española a resúmenes escritos en español. Los resultados obtenidos mostraron, por una parte, que existía cierta similitud entre los puntajes asignados por el LSA y los evaluadores (Venegas, 2006b). En Venegas (2009a, 2009b), los resultados generales permitieron establecer la posibilidad de realizar la evaluación automatizada de resúmenes utilizando el texto fuente (Venegas, 2009a), en particular cuando los textos fuente son predominantemente narrativos y cuando los resúmenes han sido escritos por estudiantes considerados con un alto nivel de logro por los docentes (2009b). Estos resultados aportan evidencia a favor del uso del LSA en la evaluación de resúmenes provenientes de textos predominantemente narrativos. En tanto que en el caso de los resúmenes provenientes de textos predominantemente expositivos, en estas investigaciones como en todos los estudios revisados, se presentan resultados más bajos en lo concerniente a los puntajes otorgados por docentes y por el método LSA.

### **3. La investigación**

Tal como se ha mencionado en los apartados anteriores, esta investigación se enmarca en la

evaluación automatizada de resúmenes de textos escritos en español utilizando el LSA. En particular en esta investigación nos proponemos como objetivo identificar un método de evaluación automatizada de resúmenes, basado en el análisis semántico latente, que permita evaluar eficientemente la calidad del contenido de resúmenes de textos predominantemente narrativos y expositivos escritos en español de modo similar a como lo hacen tres docentes evaluadores. Para cumplir este objetivo, se realizan diversos análisis correlacionales, que permiten comparar el desempeño en la evaluación de los resúmenes entre los métodos computacionales, basados en LSA, y los resultados obtenidos a partir de la evaluación realizada por los docentes evaluadores expertos.

#### **3.1. El corpus: Los resúmenes**

En esta investigación, se utilizaron 373 resúmenes, realizados por estudiantes de entre 14 y 16 años, pertenecientes a liceos de educación secundaria técnico-profesional (comerciales, marítimos e industriales) de la ciudad de Valparaíso, Chile. Estos resúmenes fueron realizados por los estudiantes en el contexto de la investigación realizada en el proyecto Fondecyt 1020786 (Parodi, 2007). Tal como en Venegas (2006b, 2009a), los resúmenes fueron realizados a partir de textos fuente predominantemente expositivos y predominantemente narrativos (en adelante textos expositivos y narrativos). El tópico de cada texto expositivo se seleccionó de acuerdo al área de especialización de cada liceo, en tanto que el tópico del texto narrativo fue común para todos los estudiantes (para mayor descripción de los textos ver Parodi, 2007). Durante el proceso de digitalización de los resúmenes, los errores ortográficos fueron corregidos con el fin de que el programa LSA pudiera calcular automáticamente la similitud léxico-semántica entre cada resumen y los textos fuente. En la Tabla 1, se muestran los valores descriptivos, según el número de palabras de los resúmenes digitalizados.

Tabla I. Valores descriptivos según cantidad y extensión de los resúmenes.

	Resúmenes provenientes de textos	
	Expositivos	Narrativos
Cantidad de resúmenes	224	129
Extensión mínima del texto en palabras	8	16
Extensión máxima del texto en palabras	141	124
Extensión promedio del texto en palabras	60,0	72,2
Desviación estándar	25,7	22,6

Estos resúmenes se encuentran disponibles en [www.elgrial.cl](http://www.elgrial.cl) y corresponden al denominado corpus DETP-2004 (Discurso Escrito Técnico-Profesional).

### 3.2. El espacio semántico y cálculo de similitud léxico-semántica

En la construcción del espacio semántico se consideró un corpus de textos del español diversificado denominado COTEGE (Corpus Textual del Español General) constituido por 10.242.384 palabras. Este corpus se conformó a partir de cinco corpus multirregistro:

**1) Corpus PUCV-2003:** Corpus recolectado por el equipo de investigación FONDECYT 1020786. La conformación general del Corpus PUCV-2003 se desglosa en 90 textos que equivalen a 1.466.744 palabras. Este corpus general está dividido, a su vez, en tres subcorpus, a saber, Corpus Técnico-Científico (CTC), Corpus de Literatura Latinoamericana Escrita (CLL) y Corpus de Entrevistas Orales (CEO). Este corpus está disponible en [www.elgrial.cl](http://www.elgrial.cl)

**2) Corpus Oral del Castellano:** Corpus de textos recolectados y transcritos por la Universidad Autónoma de Madrid. Consta de 1.099.400 palabras e incluye 12 géneros orales. Este corpus está disponible en [www.llf.uam.es/corpus/corpus.html](http://www.llf.uam.es/corpus/corpus.html)

**3) Corpus de Referencia del Español Contemporáneo:** Corpus elaborado por la Universidad Autónoma de Madrid y está compuesto por el Corpus Lingüístico de Referencia de la Lengua Española en Chile y el Corpus de Referencia de la Lengua Española en la Argentina. En este corpus se recogen un total de 3.156.491 palabras de 10 tipos de géneros textuales diferentes. Este corpus está disponible en <ftp://ftp.uba.ar/pub/misc/corpus/>

**4) Corpus de Narraciones Escritas:** Corpus construido con el fin de realizar las primeras pruebas de los sistemas computacionales en LSA en español.

Este corpus cuenta con 863.981 palabras e incluye tres obras narrativas escritas (la Biblia, Alicia en el País de las Maravillas e Historia de dos Ciudades). Está disponible en [www.elgrial.cl](http://www.elgrial.cl)

**5) Corpus ARTICO:** Corpus de Artículos de Investigación Científica. Está constituido por 678 artículos publicados en revistas indexadas ScELO entre el año 2000 y 2003. Tiene un total de 3.655.768 palabras (disponible en [www.elgrial.cl](http://www.elgrial.cl)).

Cabe mencionar que esta conformación del corpus, para su posterior transformación en espacio semántico, se justifica en relación a los hallazgos de Olmos, León, Escudero y Botana (2009), en los que se establece que la dependencia de dominio temático afecta la eficiencia de la evaluación automatizada utilizando LSA. Por lo que un corpus multiregistro, en los que se consideran textos de diversa índole temática y organizacional (expositivos, narrativos, descriptivos, etc.), tendría mejores efectos en la evaluación de los resúmenes.

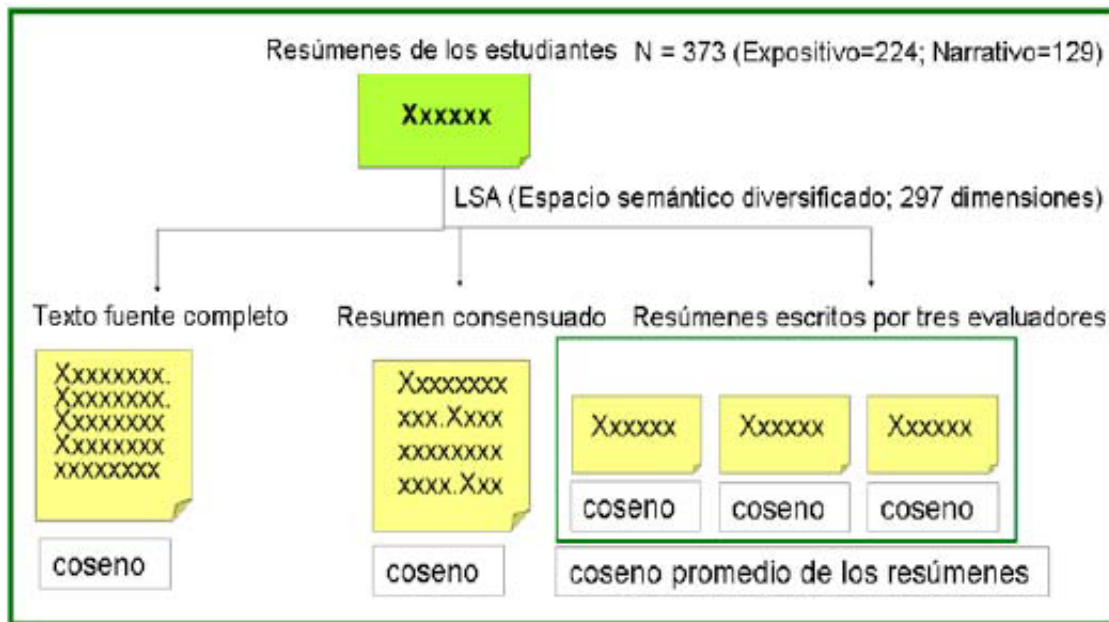
El procesamiento de COTEGE por medio del algoritmo de LSA dio como resultado el Espacio Semántico del Corpus Textual del Español General (ES-COTEGE), el que quedó constituido por 297 dimensiones y 99.966 palabras únicas. Este espacio está disponible para ser utilizado en [www.elgrial.cl](http://www.elgrial.cl) (sección comparación de textos o directamente en: <http://158.251.61.111/compareFiles/>). Cabe señalar que esta es la primera herramienta disponible gratuitamente para la comunidad científica interesada en realizar comparaciones de textos a partir del cálculo de similitud léxico-semántica con LSA de textos en español.

Utilizando ES-COTEGE se calculó el valor de coseno para cada uno de los 373 resúmenes comparándolos a partir de tres métodos, a saber: en el primer método, se calculó la similitud léxico-semántica entre cada resumen y el correspondiente texto fuente completo; en el segundo método, el

cálculo se realizó entre el resumen de cada alumno con un resumen consensuado por un grupo de investigadores en lingüística, realizado a partir de cada texto fuente. En el tercer método, se calculó la similitud léxico-semántica entre cada resumen con tres resúmenes a partir de los textos fuente, escritos por cada uno de los que serían los tres evaluadores. La Figura 1 esquematiza los métodos para el cálculo de similitud léxico-semántica utilizados en este estudio.

nueva muestra de resúmenes y se calibraron las asignaciones de puntajes. Todo esto se realizó con el fin de evitar el grado de impacto de ciertas variables, como por ejemplo, la tendencia inherente de los docentes a evaluar ciertos aspectos en desmedro de otros o evaluar aspectos no considerados en la pauta.

Para efectos de la evaluación de los resúmenes, se construyó una pauta de evaluación de los resúmenes, según cada uno de los textos fuente.



**Figura 1.** Esquema del proceso de evaluación aplicado a los resúmenes de los estudiantes utilizando los tres métodos basados en LSA.

### 3.3. Los evaluadores y la pauta de evaluación

Para la evaluación de los resúmenes, se seleccionaron y capacitaron a tres docentes de Lengua Castellana, estudiantes del programa de Magíster en Lingüística Aplicada de la Pontificia Universidad Católica de Valparaíso, Chile. Esta capacitación se llevó a cabo durante los meses de diciembre 2008 y enero de 2009, con un total de 60 horas. En esta capacitación se revisaron los aspectos teóricos relacionados con la comprensión de textos escritos (van Dijk & Kintsch, 1983; Kintsch, 1988, 1998) y su impacto en la evaluación de la comprensión. Además, se presentaron y explicaron los criterios de evaluación, se aplicó la pauta a una muestra de resúmenes, se establecieron consensos en la aplicación de los criterios y la asignación de puntaje, se reevaluó una

Para la construcción de las pautas de evaluación se consideraron los siguientes criterios de tipo semántico-cognitivo: Presencia de las Ideas Principales, Integración de las Ideas, Generalización de las Ideas, Secuencia Lógica, Macrosemantización. La pauta utilizada fue construida en base a una escala con un intervalo de seis puntos (McCarthy & McNamara, 2008). Los puntajes fueron interpretados del siguiente modo: 1: Sí, cumple absolutamente con el criterio; 2: Sí, cumple medianamente con el criterio; 3: Sí, cumple suficientemente con el criterio, 4: No cumple suficientemente con el criterio, 5: No cumple mínimamente con el criterio y 6: No cumple en nada con el criterio. La pauta fue validada (85% de acuerdo) y mejorada de acuerdo a la opinión de tres jueces expertos.



## 4. Resultados

### 4.1. Correlación entre los docentes evaluadores

En la Tabla 2 se presenta un resumen de los resultados obtenidos entre los docentes evaluadores en términos de la correlación de los puntajes asignados por ellos. Además, se presentan en la Tabla 4 tres tipos de medidas comparativas entre los docentes evaluadores (correlación promedio, correlaciones de similitud y la  $\alpha$  de Cronbach) según los resúmenes provenientes de textos predominantemente expositivos o narrativos. Esto último con el fin de poder comparar la consistencia con la que los tres docentes evaluadores realizaron las evaluaciones.

Los valores observados en las Tablas 2 y 3 muestran consistencia en la evaluación de los resúmenes provenientes de textos narrativos por parte de los docentes evaluadores, siendo esta correlación mayor que la mostrada en la evaluación de los resúmenes de textos expositivos, según el test Z de Fisher (sim  $r= 0,875$  y sim  $r=0,807$ ;  $Z=-2,14$ ,  $p= 0,0324$ ). Esto último, confirma los resultados obtenidos por León et al. (2006), en tanto que existe mayor correlación entre los docentes evaluadores al evaluar resúmenes de textos narrativos que al evaluar resúmenes de textos expositivos. Este primer resultados nos lleva a pensar que, la predominancia expositiva en un texto es una variable que no solo afecta el logro en los niveles de comprensión (Otero, León

**Tabla 2.** Correlación entre docentes evaluadores.

\*\* Los valores de las correlaciones son significativos al nivel 0,01. (n resúmenes de textos expositivos=244; n resúmenes de textos narrativos=129)

	Correlación entre docentes evaluadores según resúmenes provenientes de textos Expositivos		Correlación entre docentes evaluadores según resúmenes provenientes de textos Narrativos	
	Docente Evaluador 2	Docente Evaluador 3	Docente Evaluador 2	Docente Evaluador 3
Docente Evaluador 1	0,670**	0,515**	0,710**	0,743**
Docente Evaluador 2	---	0,564**	---	0,651**

**Tabla 3.** Medidas de comparación entre los tres evaluadores.

	Medidas de comparación entre docentes evaluadores según resúmenes provenientes de textos	
	Expositivos	Narrativos
Cantidad de resúmenes	244	129
Correlación promedio	0,583	0,701
Correlación de similitud	0,807	0,875
$\alpha$ de Cronbach	0,807	0,875

**Tabla 4.** Correlación entre los tres métodos LSA.

\*\* La correlación es significativa al nivel 0,01 bilateral

(n resúmenes de textos expositivos=244; n resúmenes de textos narrativos=129)

	Correlación entre métodos según resúmenes provenientes de textos Expositivos		Correlación entre métodos según resúmenes provenientes de textos Narrativos	
	Método 2	Método 3	Método 2	Método 3
	<b>Método 1</b>	0,667**	0,797**	0,601**
<b>Método 2</b>		0,853**		0,734**

& Graesser, 2002; Parodi, 2005), sino que además afecta la consistencia en la evaluación.

Otro aspecto a destacar es que el valor promedio de correlación entre los docentes evaluadores corresponde a  $r = 0,642$ , para todos los resúmenes. Si bien este valor puede ser considerado alto para este tipo de investigaciones, esto puede explicarse por el tiempo dedicado al entrenamiento en el uso de la pauta, pues permitió sistematizar el procedimiento de evaluación; y diferenciar con mayor objetividad las evaluaciones entre de los resúmenes provenientes de textos narrativos y de textos expositivos.

#### 4.2. Correlación entre los tres métodos

Con el fin de establecer si entre los métodos existe algún grado de relación, se realizó un análisis correlacional de los puntajes otorgados a los resúmenes por los tres métodos basados en LSA. Los resultados se muestran en la Tabla 4.

Los resultados de las correlaciones indican que entre los tres métodos existe una correlación significativa positiva alta en todos los casos. Recordemos que los métodos utilizados son:

- Método 1: comparación de los resúmenes con el texto fuente,
- Método 2: comparación de los resúmenes con un resumen consensado por lingüistas y
- Método 3: comparación de los resúmenes con resúmenes construidos por los tres docentes evaluadores a partir de los mismos textos fuente.

Como puede observarse, en primer lugar, dada las

altas correlaciones entre los métodos se podría plantear que los resultados obtenidos al evaluar la similitud léxico-semántica entre los resúmenes de los alumnos, utilizando cualquiera de los tres métodos, son relativamente similares.

En segundo lugar, acorde con los resultados podemos plantear que entre las correlaciones de los Métodos 1 y 3 y los Métodos 2 y 3 no existe diferencia estadísticamente significativa cuando los resúmenes provienen de textos expositivos, en tanto que entre todos los otros casos sí existe (Test Z de Fischer:  $M1-M2 < M1-M3$ ,  $Z = -3,13$ ,  $p = 0,0017$ ;  $M1-M2 < M2-M3$ ,  $Z = -5,07$ ,  $p = 0,000$ ;  $M1-M3 = M2-M3$ ,  $Z = -1.94$ ,  $p = 0,0524$ ).

En tercer lugar, observamos que no existe diferencia estadísticamente significativa entre las correlaciones entre los métodos cuando la evaluación se realiza a resúmenes provenientes de textos narrativos (Test Z de Fischer:  $M1-M2 = M1-M3$ ,  $Z = -1.87$ ,  $p = 0,0615$ ;  $M1-M2 = M2-M3$ ,  $Z = -1,93$ ,  $p = 0,0536$ ;  $M1-M3 = M2-M3$ ,  $Z = -0.05$ ,  $p = 0,9601$ ). Lo anterior implica que si se realizara una evaluación con cualquiera de estos métodos los resultados serían similares entre sí.

Si bien, los datos anteriores nos permiten tener una primera aproximación a la relación existente entre los valores asignados por cada método a los resúmenes, ellos no permiten tener una visión global del rendimiento obtenido por ellos. Por lo tanto, se hace necesario verificar si efectivamente los métodos basados en LSA presentan diferencias entre sí en la evaluación de los resúmenes, a partir de valores de correlación, para ellos se utilizan las medidas de evaluación sintetizadas en la Tabla 5.

**Tabla 5.** Medidas de comparación entre los tres métodos.

	Medidas de comparación entre métodos según resúmenes provenientes de textos	
	Expositivos	Narrativos
<b>Cantidad de resúmenes</b>	<b>244</b>	<b>129</b>
<b>Correlación promedio</b>	<b>0,772</b>	<b>0,689</b>
<b>Correlación de similitud</b>	<b>0,910</b>	<b>0,746</b>
<b><math>\alpha</math> de Cronbach</b>	<b>0,907</b>	<b>0,271</b>

Como observamos en la Tabla 5, existe una mayor correlación entre los métodos cuando la evaluación se realiza a partir de los resúmenes provenientes de textos expositivos que narrativos (Test Z de Fisher:  $\text{sim } r=0,910$  y  $\text{sim } r=0,746$ ,  $Z=5,13$ ,  $p=0,000$ ). Esta situación es inversa a la presentada por los docentes evaluadores, ya que los métodos basados en LSA se comportan de manera más homogénea al evaluar los resúmenes de texto expositivos que narrativos.

Estos resultados demuestran que los dos métodos que presentan mayor correlación, tanto para evaluar resúmenes de textos expositivos como narrativos, consideran resúmenes escritos por humanos, lo cual implica que en el caso de construirse un sistema computacional se debe considerar este factor. Las correlaciones que involucran al Método I son más bajas, aunque como ya sabemos sin diferencia estadísticamente significativa en el caso de los

resúmenes de textos narrativos, lo cual indicaría que a pesar de ser más fácil de implementar computacionalmente (solo se necesita el texto fuente), los resultados serían distintos al uso de los otros métodos.

Junto con la correlación entre los docentes evaluadores, por una parte, y entre los métodos, por otra, también se realizó una correlación con la extensión de los resúmenes en cuanto a la cantidad de palabras. Lo anterior, con el fin de determinar si los docentes evaluadores y/o los métodos se ven afectados y en qué medida por la extensión del resumen al asignar el puntaje de evaluación. La Tabla 6 muestra una comparación entre las correlaciones promedio, tanto de los evaluadores como para los métodos con la extensión de los resúmenes, así como la aplicación del test Z de Fisher para saber si existen diferencias.

**Tabla 6.** Comparación de la correlación entre evaluadores y métodos.

Extensión del resumen	Resúmenes provenientes de textos:	
	Expositivos	Narrativos
<b>Correlación promedio</b>		
<b>Evaluadores</b>	-0,535	-0,515
<b>Métodos</b>	-0,457	-0,504
<b>Z de Fisher</b>	1,14	0,12
<b>Probabilidad de error</b>	0,254	0,905

Tal como se observa en la tabla anterior, existe una correlación negativa media tanto entre los evaluadores como entre los métodos en relación a la extensión del texto en la evaluación de los resúmenes. Esto quiere decir que a mayor cantidad de palabras en el resumen, menor será el puntaje de evaluación otorgado (y por lo mismo mejor logro) y viceversa. Recordemos que la escala de evaluación usada en esta investigación considera que 1 es el mejor puntaje al que se puede optar y 6 al menor puntaje (ver escala de evaluación en 3.3.). Cabe mencionar que los puntajes de LSA fueron adaptados a la lógica de la escala para efectos de comparación. En síntesis, tanto los docentes evaluadores como los métodos se ven afectados por la extensión del texto en una relación cercana al 50% de los casos, no existiendo diferencia estadísticamente significativa entre ellos.

Hasta aquí podemos plantear que, por una parte, los docentes evaluadores y los métodos se diferenciarían en que los primeros evalúan más consistentemente los resúmenes provenientes de textos narrativos y los métodos los provenientes de textos expositivos y, por otra, que el tamaño del texto afecta en la misma medida tanto los evaluadores como a los métodos en la evaluación de los resúmenes.

#### **4.3. Correlación entre evaluadores y métodos en la evaluación de los resúmenes**

A partir de los datos obtenidos en la evaluación realizada por los docentes y de la asignación de puntaje otorgado a cada resumen, se ha llevado a cabo un análisis correlacional entre los evaluadores y entre el promedio otorgado por los evaluadores y cada método. Este procedimiento evidencia el grado de similitud entre los resúmenes escritos por los estudiantes y los puntajes otorgados por evaluadores y los tres métodos basados en LSA.

En la Tabla 7, se muestran los resultados correlacionales obtenidos entre los evaluadores y cada uno de los métodos, de acuerdo a si los resúmenes provienen de textos predominantemente expositivos o de textos predominantemente narrativos.

Como se observa en la tabla anterior, todas las correlaciones son significativas y positivas en términos estadísticos. Lo que en primer lugar indica que existiría mayor correlación entre los docentes evaluadores y los métodos cuando lo que se evalúa son los resúmenes provenientes de textos narrativos. Lo anterior se evidencia según el test de Fisher solo para la relación con los Métodos 2 y 3 (M1-Ev,  $Z=-1,54$ ,  $p=0,1230$ ; M2-Ev,  $Z=-2,68$ ,  $p=0,0074$ ; M3-Ev,  $Z=-2,58$ ,  $p=0,0079$ ). Las correlaciones entre Método 1 y los docentes evaluadores, por su parte, no difieren en términos estadísticos. Si bien el puntaje de correlación más alto se presenta entre los evaluadores y el Método 3 (Resumen-Resúmenes de los evaluadores), favoreciendo a los resúmenes provenientes de los textos narrativos ( $r=0,604$ ), no existe una diferencia estadísticamente significativa al compararse con los otros dos métodos (Test Z de Fisher: M1-Ev=M2-Ev,  $Z=-1,27$ ,  $p=0,2041$ ; M1-Ev=M3-Ev,  $Z=-1,5$ ,  $p=0,1336$ ; M2-Ev=M3-Ev,  $Z=-0,23$ ,  $p=0,8155$ ).

En el caso de la evaluación de resúmenes provenientes de textos expositivos, los métodos presentan magnitudes menores y, al igual que en el caso anterior, no se observa ninguna diferencia entre las correlaciones entre los métodos y los evaluadores, según el test Z de Fisher (M1-Ev=M2-Ev,  $Z=-0,39$ ,  $p=0,6965$ ; M1-Ev=M3-Ev,  $Z=-0,82$ ,  $p=0,4122$ ; M2-Ev=M3-Ev,  $Z=-0,43$ ,  $p=0,6672$ ).

Estos resultados permiten aseverar, en primer lugar, que existe una diferencia 0 estadísticamente significativa a favor de la relación entre los Métodos 2 y 3 y los evaluadores, cuando los resúmenes provienen de textos narrativos; en segundo lugar, que es posible establecer que los tres métodos posibilitan una evaluación similar de los resúmenes provenientes de textos narrativos y expositivos, escritos por los estudiantes. Sin embargo, dado que los valores correlacionales promedio entre los docentes evaluadores son mayores (ver Tabla 3), queda pendiente establecer si existe una diferencia significativa entre dichos valores y los obtenidos mediante el uso de los métodos LSA y los evaluadores (es decir, simulando la asignación de puntaje de cada método como si fuera la de un docente evaluador más). En la Tabla 8, se muestran, según cada método y tipo de texto, los valores de probabilidad obtenidos a partir de la comparación de los coeficientes de correlación.

**Tabla 7.** Comparación entre métodos LSA y los tres evaluadores.

\*\* La correlación es significativa al nivel de 0,01

(n resúmenes de textos expositivos=244; n resúmenes de textos narrativos=129)

Promedio correlacional	Resúmenes provenientes de texto	
	Expositivo	Narrativo
Método 1 – Docentes Evaluadores	0,328**	0,470**
Método 2 – Docentes Evaluadores	0,359**	0,585**
Método 3 – Docentes Evaluadores	0,393**	0,604**

**Tabla 8.** Comparación de los coeficientes de correlación obtenidos entre los docentes evaluadores y los obtenidos entre los docentes evaluadores y cada método.

\*\* La diferencia es significativa a nivel de 0,01.

Valores de P en Test Z de Fisher	Docentes Evaluadores y Método 1+Doc.Ev	Docentes Evaluadores y Método 2+Doc.Ev	Docentes Evaluadores y Método 3+Doc.Ev
Resúmenes de textos expositivos (n=244)	0,0003**	0,0013**	0,0057**
Resúmenes de textos narrativos (n=129)	0,0043**	0,1137	0,1776

Los resultados de la comparación de los coeficientes de correlación permiten establecer que el Método 1 (resumen–texto completo) se diferencia significativamente de la correlación obtenida por los evaluadores en la evaluación de cada uno de los resúmenes, ya sean provenientes de textos expositivos o narrativos. El Método 2 y el Método 3 se diferencian de las correlaciones humanas cuando los resúmenes provienen de textos expositivos. Estas diferencias permiten establecer que los métodos basados en LSA no obtienen puntajes similares a los obtenidos por los humanos en la tarea de evaluación de resúmenes de textos expositivos en español.

Por otra parte, el Método 2 (resumen–resumen consensado) y el Método 3 (resumen–resúmenes de los evaluadores) no se diferencian de la correlación obtenida por los docentes entre sí cuando se evalúan resúmenes provenientes de textos narrativos. Lo anterior confirma que los resultados obtenidos con los métodos LSA se asemejan a los obtenidos por los docentes evaluadores en la asignación de puntaje a los resúmenes provenientes de textos en español predominantemente narrativos.

## CONCLUSIÓN

Como hemos planteado en la introducción de este artículo, enmarcado en el ámbito de la evaluación automatizada de resúmenes en lengua española y motivada por los avances obtenidos en otras investigaciones con el uso del LSA, nos hemos planteado como objetivo identificar un método de evaluación automatizada de resúmenes, basado en el análisis semántico latente, que permita evaluar eficientemente la calidad del contenido de resúmenes de textos predominantemente narrativos y expositivos escritos en español de modo similar a como lo hacen tres docentes evaluadores. Para cumplir con esto, llevamos diversos análisis correlacionales que nos permitieran comparar el desempeño de los tres métodos, basados en LSA, en la evaluación de 373 resúmenes con los puntajes asignados por tres docentes evaluadores, entrenados en el uso de una pauta especialmente diseñada para evaluar la calidad del contenido de los resúmenes.

Los hallazgos empíricos nos permiten, en primer lugar, establecer que al menos dos de los tres

métodos, basados en LSA, evalúan de modo similar los resúmenes provenientes de textos narrativos escritos en lengua española. De este modo, los métodos más eficientes en la evaluación de los resúmenes corresponden al Método 2 (comparación de los resúmenes con un resumen consensuado por lingüistas) y al Método 3 (comparación de los resúmenes con resúmenes contruidos por los tres docentes evaluadores a partir de los mismos textos fuente).

En consecuencia con lo anterior, destacamos que ambos modelos, basados en LSA, incluyen el requerimiento de resúmenes, que sirvan de modelo para la evaluación, ya sean escritos consensuadamente entre especialistas a partir de los textos fuente, ya sean escritos individualmente por parte de los docentes evaluadores. Esto, a su vez, implica que la comparación entre los resúmenes de los estudiantes con el texto fuente es un método que si bien se asemeja a los otros métodos en la comparación correlacional, no lo hace respecto de los docentes evaluadores, por lo que no resultaría ser un método eficiente de evaluación de resúmenes.

Por otra parte, en relación a la tarea de evaluación de los resúmenes provenientes de textos expositivos, si bien los métodos 2 y 3 obtienen mayor correlación en relación con los docentes evaluadores que el Método 1, ninguno de los tres es estadísticamente similar a la evaluación realizada por los tres docentes evaluadores. Esto sin duda, es un resultado algo desalentador, pues si bien se sabía que estudios similares han reportado menor correlación entre los métodos y los evaluadores, en este estudio se esperaba que tal correlación fuera de una magnitud mayor entre los evaluadores, debido al entrenamiento y el uso de la pauta, y, por lo mismo, no existiera diferencia con los métodos.

Ahora bien, este resultado podría explicarse debido a que la construcción de la macroestructura que representa el contenido semántico de un texto expositivo requiere de mucho mayor conocimiento previo que la construcción de la macroestructura de un texto narrativo. Lo que afectaría no solo la textualización del resumen por parte de los estudiantes sino que también la homogeneidad con la que los tres docentes evalúan estos resúmenes. Por otra parte, para el caso de los métodos basados en LSA, este hallazgo también puede explicarse debido a la forma de funcionamiento del LSA, ya que en este solo se atiende a la información

léxica disponible en el texto, lo que implica que en los resúmenes los ítems léxicos utilizados son más estables en relación con los textos fuente expositivos (por ejemplo, uso de terminología que no varía entre el texto fuente y los resúmenes). Por el contrario, la posibilidad de uso de mayor variación léxica (o parafraseo) respecto de los textos fuente narrativos por parte de los estudiantes y los docentes, hace que entre los métodos basados en LSA se observe mayor heterogeneidad. Aspecto que en el caso de los docentes evaluadores no se da, probablemente, pues, como ya dijimos, ellos nos solo atienden a los aspectos léxicos sino que también a los estructurales y al conocimiento previo en relación a lo que significa leer textos narrativos.

Sin duda, el aporte aquí, más que dar respuestas precisas, es abrir un camino para futuras investigaciones en relación con la evaluación automatizada de resúmenes, en particular de aquellos provenientes de textos expositivos, así como de otros textos con predominancia organizacional diferente (argumentativos y descriptivos, por ejemplo). De este modo, se podría aportar más evidencia respecto de cómo la información semántica es procesada en términos de conceptualización y abstracción de los contenidos en el proceso de comprensión y, también en el de producción, siendo de hecho tanto o más relevantes que la incidencia del procesamiento de estructuras sintácticas en este tipo de procesos cognitivos.

Como se ha mostrado, el uso del LSA en la evaluación automatizada de resúmenes permite simular eficientemente la evaluación de la calidad del contenido de los resúmenes provenientes de textos narrativos, evidenciándose diferencia con respecto a los resúmenes provenientes de texto expositivos. Este resultado, que aporta predictibilidad al estudio, es congruente con lo presentado por León et al. (2006) para los textos en español, a pesar de que tanto la conformación del espacio semántico como de su dimensionalidad son diferentes. En el caso de León et al. (2006) fue de 337 dimensiones y en el nuestro de 297 dimensiones. Esto implica, que el algoritmo funciona de modo similar, sin necesariamente tener un espacio semántico constituido por los textos del mismo dominio temático (ni dimensionalidad), lo que confirma, entre otras cosas, la decisión de utilizar un espacio semántico de dominio general. Sin desestimar lo anterior, sería de suyo interesante considerar experimentar con cambios tanto en la constitución del espacio semántico como en

la dimensionalidad del espacio, con el fin de estar seguro de cuál es la cantidad óptima de dimensiones para esta tarea y evaluar si se mantiene la diferencia según la predominancia del texto.

Finalmente, podemos establecer que lo presentado aquí es un avance, a nuestro parecer importante, para pensar en la posibilidad de desarrollar un sistema computacional confiable que contemple en su construcción otras técnicas de procesamiento natural del lenguaje. Para ello se debe tener en cuenta que los métodos con mejor desempeño requieren de la inclusión de resúmenes escritos por

humanos. Lo anterior, viene a dificultar la tarea de la implementación en algún sistema de entrenamiento en la escritura de resúmenes, aunque en un grado bastante menor, puesto que para cada actividad en la que los usuarios deban producir un resumen a partir de un texto, deberá existir en el sistema una base de resúmenes que sirvan de modelo de contraste, realizados por sujetos expertos en ello. Esto si bien no hace que el sistema sea 100% automatizado le daría, por otra parte, una mayor confiabilidad a los resultados obtenidos.

## REFERENCIAS BIBLIOGRÁFICAS

- Alfonseca, E. & Pérez, D. (2004). *Automatic assessment of short questions with a Bleu-inspired algorithm and shallow NLP*. Ponencia presentada en the 4th International Conference, EsTAL 2004, Alicante, España.
- Batten, E. (2003). Project Essay Grade: PEG. En M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 39-50). Mahwah, NJ.: Erlbaum.
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.
- Farr, R. & Carey, R. (1986). *Reading. What can be measured*. Newark, Delaware: IRA.
- Graesser, A., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371-395.
- Ibáñez, R. (2008). Comprensión de textos académicos en inglés: Relación entre nivel de logro y variables involucradas. *Revista Signos. Estudios de Lingüística*, 41(67), 203-229.
- Kintsch, W. & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-94.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension construction-integration model. *Psychological Review*, 95, 163-182.
- Kintsch, W. (1998). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163-182
- Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, Matthews, C. & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, 8(2), 87-109.
- Kintsch, W. (2007). Meaning in context. En T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 89-105). Mahwah, NJ.: Erlbaum.
- Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.
- Landauer, T. K., Laham, D., Rehder, B. & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. En M. G. Shafto & P. Langley (Eds.), *Actas de The 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ.: Erlbaum.
- Landauer, T. K., Foltz, P. W. & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.
- Landauer, T. K. & Psofka, J. (2000). Simulating text understanding for educational applications with latent semantic analysis: Introduction to LSA. *Interactive Learning Environments*, 8(2), 73-86.
- Landauer, T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *The psychology of learning and motivation*, 41, 1-63.
- Landauer, T. K., Laham, D. & Foltz, P.W. (2003). Automated essay scoring: A cross disciplinary perspective. En M. Shermis & J. Burstein (Eds.), *Automated essay scoring and annotation of essays with the Intelligent Essay Assessor* (pp.87-112). Mahwah, NJ: Erlbaum.
- Landauer, T. K., McNamara, D., Dennis, S. & Kintsch, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ.: Erlbaum.



- León, J., Olmos, R., Escudero, I., Cañas, J. & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods*, 38(4), 616-627.
- Lin, Ch. & Hovy, E. (2003). *Automatic evaluation of summaries using n-gram co-occurrence statistics*. En Actas del 2003 Language Technology Conference, Edmonton, Canadá.
- Lin, Ch. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*. En actas del Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, España.
- Mani, I. & Maybury, M. (2001). *Advances in automatic text summarization*. Massachusetts: Massachusetts Institute of Technology.
- Manning, C. & Schütze, H. (2003). *Foundations of statistical natural language processing*. Massachusetts: MIT Press.
- McCarthy, Ph. & McNamara, D. (2008). *The user-language paraphrase challenge* [en línea] Disponible en: <https://umdrive.memphis.edu/pmmccrth/public/Phil%27s%20papers.htm?uniq=-xq6brv>
- Marinkovich, J., Peronard, M. & Parodi, G. (2006). *Programa de optimización de la competencia estratégica para comprender y producir textos escritos (LECTES)* [en línea]. Disponible en: [www.lectes.cl](http://www.lectes.cl)
- Martin, D. & Berry, M. (2007). Mathematical foundations behind latent semantic analysis. En T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35-56). Mahwah, NJ.: Erlbaum.
- Olmos, R., León, J., Escudero, I. & Botana, J. (2009). Efectos sobre el tamaño y especificidad de los corpus en la evaluación de resúmenes mediante el LSA y jueces expertos. *Revista Signos. Estudios de Lingüística*, 41 (69), 71-81.
- Órdenes, A. (2009). *El resumen como instrumento de evaluación de la comprensión del discurso escrito: Correlación entre evaluadores expertos*. Tesis de Licenciatura. Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.
- Otero, J., León, J. & Graesser, A. (2002). *The psychology of science text comprehension*. Mahwah, NJ.: Erlbaum.
- Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2001). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311-318). Philadelphia: Association for Computational Linguistics.
- Parodi, G. (Ed.) (2005). *Discurso especializado e instituciones formadoras*. Valparaíso: EUVSA
- Parodi, G. (2007). Comprensión y aprendizaje a partir del discurso especializado escrito: Teoría y empiria. En G. Parodi (Ed.). *Lingüística de corpus y discursos especializados: Puntos de mira* (pp. 223-258). Valparaíso: EUVSA.
- Parodi, G., Peronard, M. & Ibáñez, R. (2010). *Saber leer*. Madrid: Aguilar.
- Pérez, D., Alfonseca, E., Rodríguez, P., Gliozzo, A., Strapparava, C. & Magnini, B. (2005). About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. *Revista Signos. Estudios de Lingüística*, 38(59), 325-343.
- Peronard, M. (1989). Estrato social y estrategias de comprensión de lectura. *Lenguas Modernas*, 16, 19-32.
- Peronard, M. (1997). ¿Qué significa comprender un texto escrito? En M. Peronard, L. Gómez, G. Parodi & P. Núñez (Comps.), *Comprensión de textos escritos: De la teoría a la sala de clases* (pp. 55-78). Santiago: Andrés Bello.

- Peronard, M., Gómez, L., Parodi, G. & Núñez, P. (1998). *Comprensión de textos escritos: De la teoría a la sala de clases*. Santiago: Andrés Bello.
- PISA (2007). *PISA 2006. Science competencies for tomorrow's world*. Paris: OECD Publishing.
- Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K. & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25, 337-354.
- Silva, N. (2008). *Hacia una evaluación automatizada de resúmenes en español usando LSA*. Tesis de Licenciatura. Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.
- Steinhart, D. (2001). *Summary Street: An intelligent tutoring system for improving student writing through the use of latent semantic analysis*. Tesis doctoral, University of Colorado, Boulder, USA.
- Van Dijk, T. (1980). *Macrostructures*. Hillsdale, NJ.: Erlbaum.
- Van Dijk, T. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
- Venegas, R. (2003). Análisis Semántico Latente: Una panorámica de su desarrollo. *Revista Signos. Estudios de Lingüística*, 36(53), 121-138.
- Venegas, R. (2005). *Las relaciones léxico-semánticas en artículos de investigación científica: Una aproximación desde el Análisis Semántico Latente*. Tesis doctoral, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.
- Venegas, R. (2006a). La similitud léxico-semántica en artículos de investigación científica en español: Una aproximación desde el Análisis Semántico Latente. *Revista Signos. Estudios de Lingüística*, 39(60), 75-106.
- Venegas, R. (2006b). *Comparación de la evaluación realizada por docentes y por el Análisis Semántico Latente*. Informe de Investigación DI 184719/2006. Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.
- Venegas, R. (2007). Using Latent Semantic Analysis in a Spanish research article corpus. En G. Parodi (Ed.), *Working with Spanish corpora* (pp. 195- 216). London: Continuum.
- Venegas, R. (2009a). Towards a method for assessing summaries in Spanish using LSA. En H. Lane & H. Guesgen (Eds.), *Actas de the Twenty-Second International Florida Artificial Intelligence Research Society Conference* (pp. 113-115). Washington, DC: AAAI.
- Venegas, R. (2009b). *Evaluación de resúmenes en español: Correspondencia entre profesores y el Análisis Semántico Latente*. Informe de Investigación Final Fondecyt N° 11070225, Comisión Nacional de Ciencia y Tecnología, Chile.
- Wade-Stein, D. & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction*, 22, 333-362.
- Yulan, H., Siu, Ch. & Tho, Th. (2009). Automatic summary assessment for intelligent tutoring systems. *Computers & Education*, 53, 890-899.
- Zipitria, I., Elorriaga, J., Arruate, A. & de Ilarraza, A. (2004). *From human to automatic summary evaluation*. En Seventh International Conference on Intelligent Tutoring System, Maceió, Brasil.

\* Investigación financiada parcialmente por el Proyecto Fondecyt 11070225.