Hacia la automatización de la evaluación de resúmenes desde la experiencia cognitiva

I. Zipitria, J. A. Elorriaga y A. Arruarte

Universidad del País Vasco - Euskal Herriko Unibertsitatea (UPV/EHU)
649 P.K. , E-20080 Donostia
{iraide.zipitria, a.arruarte, jon.elorriaga}@ehu.es

Resumen: Este artículo presenta el proceso de desarrollo de un sistema para la evaluación de resúmenes basado en la experiencia de profesores expertos. El proceso de evaluación del resumen y los criterios utilizados por los expertos se han obtenido a partir de datos experimentales y se han integrado en un modelo de toma de decisiones. Por otro lado, el análisis y la comprensión del texto resumido se realiza por medio de técnicas de Procesamiento del Lenguaje Natural (PLN) y un modelo de comprensión del lenguaje natural: *Latent Semantic Analysis* (LSA).
Palabras clave: Evaluación de resúmenes, Toma de decisiones de Evaluación.

Abstract: This paper presents the process of generating a summary evaluation grading system based on teaching expert knowledge. It is founded on human practice to emulate the evaluation underneath. Expert criteria are represented by a decision making model. On the other hand, text analysis and comprehension are modelled by means of Natural Language Processing (NLP) techniques and a natural language comprehension model: Latent Semantic Analysis (LSA).

Key words: Summary Evaluation, Evaluation decision making.

1. Introducción

Una tarea importante en el campo de los entornos de enseñanza/aprendizaje inteligentes es producir un diagnóstico adecuado de la actividad del aprendiz. Entre los diversos procedimientos existentes algunos se ocupan de inferir la comprensión del aprendiz analizando las respuestas que produce en lenguaje natural. La mayor ventaja que confiere este método en comparación con los métodos cerrados es que concede al aprendiz una mayor libertad de respuesta, aportando, como consecuencia, una mayor riqueza de información a la hora de diagnosticar el nivel de comprensión adquirido. Al no tener ningún tipo de pista o límites en torno a una respuesta determinada, el aprendiz tiene una mayor responsabilidad sobre la construcción de la respuesta.

En este contexto, nuestra investigación se centra en la evaluación de respuestas recogidas en forma de resumen. Esta tarea se utiliza específicamente para diagnosticar el grado de comprensión de textos o de temas específicos (Barlett, 1932; Brown & Day, 1983; Kintsch et al., 1999) debido a que aporta una buena aproximación de la información retenida en memoria.

En los entornos de enseñanza/aprendizaje inteligentes es necesario incorporar mecanismos que permitan comprender el resumen de forma automatizada. La evaluación automática de texto libre es una tarea compleja y está condicionada a los métodos de comprensión de textos y Procesamiento de Lenguaje Natural (PLN) utilizados. Algunas soluciones de PLN se crean de forma específica para un idioma, y para hacer uso de las mismas en otros idiomas es necesario adaptarlas a las características particulares de esa lengua. De forma similar, el uso de sistemas de comprensión lingüística tales como Latent Semantic Análisis (LSA) (Landauer & Dumais, 1997), puede requerir ciertos ajustes. En el trabajo aquí presentado se aplica el modelo LSA y las técnicas de PLN para

lograr la evaluación automática de resúmenes sobre textos escritos en euskara realizados por aprendices.

Este artículo describe la metodología llevada a cabo para automatizar la evaluación de resúmenes, desde la observación humana hasta la configuración del sistema propuesto. Tras la introducción en la Sección 1, el artículo analiza el proceso de crear un resumen en la Sección 2. La Sección 3 detalla los resultados de un estudio llevado a cabo para observar la tarea que realizan los expertos para evaluar resúmenes. La Sección 4 presenta el sistema LEA, sistema capaz de realizar el proceso de evaluación de resúmenes de manera automática. Finalmente, la Sección 5 engloba conclusiones finales y posibles líneas de futuro.

2. ¿Qué implica resumir?

En la práctica educativa resumir es una estrategia de aprendizaje que habitualmente se utiliza para diagnosticar la comprensión lectora ya que implica múltiples habilidades: comprensión, abstracción, organización y reproducción de información. Un resumen se puede crear tanto de forma oral como escrita y la madurez del mismo varía en función de la edad, la habilidad de abstracción y las habilidades lingüísticas y de aprendizaje de cada persona (Manning & Schutze, 1999).

Se ha demostrado que resumir es una técnica de aprendizaje que ayuda a comprender lo escrito; la información incluida en un resumen nos da una idea general de hasta qué punto se ha comprendido el texto (Garner, 1982). Debido a ello, resumir es uno de los modos más utilizados a la hora de diagnosticar la comprensión de contenidos.

Se ha hablado mucho sobre lo que debería incorporar un buen resumen. Pero, ¿se producen resúmenes ideales? Los resúmenes no siempre contienen lo que se espera de ellos. Este tema ha sido estudiado por muchos autores que han subrayado las diferencias y similitudes existentes entre resúmenes maduros y resúmenes inmaduros. La capacidad de resumir está incluida dentro de las habilidades de escritura y su desarrollo está directamente relacionado tanto con el desarrollo cognitivo como discursivo del aprendiz. Por tanto, comparte algunas características con otros tipos de expresión escrita como informes, portafolios, etc.

Hasta la fecha numerosas investigaciones se han centrado en el estudio de la estructura del texto, la habilidad de resumir y el modelado de la comprensión de textos. (Barlett, 1932) fue uno de los primeros psicólogos que analizó el recuerdo de la información contenida en los textos y los procesos subyacentes a la calidad del recuerdo. Más tarde, (Rumelhart, 1975) produjo uno de los primeros modelos que explicaba el procedimiento implicado en la narrativa de los resúmenes. De forma paralela, (Thorndike, 1977) exploró el recuerdo de textos basándose en la complejidad del tema de los mismos. Pero fueron finalmente (Kintsch & van Dijk, 1978) quienes propusieron un modelo holístico sobre la representación y estructura de textos de resúmenes que describía la representación mental y la dificultades principales que surgen durante el proceso de desarrollo de resúmenes. Finalmente, (Lehnert, 1981) produjo un modelo capaz de explicar la narrativa de resúmenes tomando en consideración variables tanto afectivas como emocionales.

Además, hay innumerables estudios empíricos que avalan y amplían lo propuesto por estos autores. Así, (Garner, 1982, 1987) estudió las diferencias existentes entre los resúmenes realizados por estudiantes eficientes y estudiantes poco eficientes. De acuerdo con su trabajo, los estudiantes que resumían de forma eficiente también guardaban información en la memoria de manera más eficiente. De este modo demostró que hay una relación directa entre resumir de forma eficiente y recordar de forma eficiente. Por otro lado, (Manelis & Yekovich, 1984) analizan textos expositivos y su relación con procesos de aprendizaje y comprensión. (Bransford et al., 1990) hacen una reflexión sobre lo que los aprendices comprenden del texto y sobre cómo se les guía durante el proceso de aprendizaje.

Pero la cuestión clave que ha protagonizado la mayoría de las investigaciones en este contexto ha sido la búsqueda de la diferencia existente entre un resumen pobre y un buen resumen. Muchas investigaciones confirman la existencia de una clara diferenciación entre resúmenes maduros e inmaduros Pero, ¿cómo podemos identificarlos? La labor investigadora realizada hasta ahora se ha centrado principalmente en identificar el tema y las frases e ideas principales, ya que el uso adecuado de estos patrones es considerado síntoma de madurez al

resumir (Manelis & Yekovich, 1984). De forma similar, (Garner, 1982) argumentaba que aquellos que redactan resúmenes de forma eficiente también reconocen la información semántica asociada al texto, aunque esta no aparezca expresada explícitamente. Por el contrario, las dificultades observadas en la redacción de resúmenes inmaduros se relacionaban con la falta comprensión y recuerdo. Otra tendencia observada en resúmenes inmaduros es la estrategia de "*expresar todo aquello que se conoce o recuerda sobre el tema*", lo cual implica adjuntar gran cantidad de información irrelevante (Brown & Day, 1983). En conclusión, las características principales que nos ayudan a identificar resúmenes inmaduros son la inclusión de información irrelevante, la copia literal de partes del texto original y las dificultades de comprensión.

Otro factor a considerar es el nivel de conocimiento del lenguaje o idioma en el que se ha realizado el resumen. Hasta tal punto esto es así, que el conocimiento lingüístico por sí mismo puede marcar la diferencia entre un resumen maduro y uno inmaduro. Los resúmenes producidos en una segunda lengua (L2) o durante su periodo de aprendizaje llevan al aprendiz a cometer errores que no necesariamente se producirían en su lengua materna. Al resumir en L2 el aprendiz se enfrenta a dificultades léxicas y de comprensión que no se darían en su lengua propia. Con ello, aunque el resumen final es similar al producido por un estudiante inmaduro en su lengua materna, las razones subyacentes pueden ser muy diferentes (Long & Harding-Esch, 1978). Por tanto, resumir en L2 requiere adiestramiento y una evaluación específica.

Además de todo lo citado, no podemos obviar la relevancia del conocimiento previo. Para procesar la información adquirida de forma adecuada y realizar un buen resumen es necesario asociarla correctamente a la ya existente. Esto está directamente relacionado con la selección adecuada de categorías que faciliten la extracción de información y reduzcan la demanda de memoria de trabajo (Symons & Pressley, 1993). Así, una familiaridad previa con el tema puede determinar la comprensión del texto y la identificación de las ideas principales, lo cual es aplicable tanto al conocimiento de contenidos como de lenguaje.

Entonces, ¿qué implica evaluar un resumen? Sintetizando todo lo analizado anteriormente, parece claro que hay varios criterios a tener en cuenta a la hora de evaluar un resumen: nivel de madurez del resumen, nivel de conocimiento de la lengua, conocimiento previo del tema, etc. Pero aún quedan cuestiones específicas del proceso mismo de evaluación y de la toma de decisiones que no han sido resueltas en estas investigaciones. Con el objetivo de identificar estas cuestiones hemos realizado un estudio con expertos, descrito a continuación.

3. Evaluación de resúmenes: entrevistas y experimentos para observar el procedimiento de los expertos.

Con el objetivo de estructurar la tarea de evaluación de resúmenes se desarrolló un estudio que incluía entrevistas a expertos (subsección 3.1) y un experimento para observar su modo de evaluación (subsección 3.2). Se quería observar la evaluación de resúmenes para grupos críticos de aprendizaje en la redacción de resúmenes: grupos inmaduros y grupos maduros. Así, se entrevistó a profesionales de enseñanza primaria, secundaria y L2 como expertos en evaluar resúmenes inmaduros y a profesores de universidad como expertos en evaluar a aprendices que ya producían resúmenes de cierta madurez.

3.1. Entrevistas con expertos

Tras entrevistar a los expertos de los diversos contextos educativos, y en lo que se refiere a la habilidad de resumir, se concluye que en la *educación primaria y secundaria,* se realiza un adiestramiento con metodología de andamiaje (*scaffolding*); así, la habilidad de resumir se va adquiriendo paso a paso por medio de aproximaciones sucesivas. Esto coincide con lo propuesto por otros autores sobre el adiestramiento en la capacidad de escribir (Cassany, 1993; Fitzgerald, 1987; Inoue, 2005): el objetivo principal es instruir en la abstracción de ideas así como en la adquisición de las habilidades lingüísticas adecuadas.

Coincidiendo con (Manelis & Yekovich, 1984), también se señala que los estudiantes de estos niveles

educativos utilizan herramientas de ayuda al desarrollo de resúmenes como mapas conceptuales y esquemas. Además, no podemos pasar por alto la utilización de material teórico sobre: conectores, reiteración, eliminación de estrategias de copia y pega (Brown & Day, 1983), coherencia, cohesión, adecuación, gramática, etc. La enseñanza de la técnica del resumen se produce principalmente en modo instruccional, aunque también se utilizan estrategias de aprendizaje colaborativo, evaluación entre iguales y la auto-evaluación como parte del proceso de andamiaje.

Otro aspecto o consideración relevante a tener en cuenta en el nivel de educación primaria y secundaria es el tipo de texto que se utiliza, ya que se ha de intentar ajustarlo a las necesidades específicas de los aprendices. Parece ser que los textos narrativos son percibidos por los aprendices como más sencillos que los textos explicativos. Esto se asocia fundamentalmente a dos factores: por una parte, en los textos narrativos se incluye un orden cronológico o secuencial del que carecen los explicativos y, por otra, los aprendices están más habituados a trabajar con este tipo de textos.

Tal y como se ha comentado en la Sección 2, el grupo de aprendices de L2 presenta un perfil específico dentro del grupo de los aprendices que realizan resúmenes inmaduros. Es este, según los expertos, un grupo heterogéneo; incluso dentro de un mismo curso presentan grados de estudios distintos, diferencias en cuanto a la habilidad lectora, etc. No obstante, a la hora de resumir, y siempre según los expertos, los aprendices de este grupo identifican en primer lugar las ideas principales y pasan entonces a corregir aspectos del lenguaje. La relevancia y peso de cada uno de estos elementos varía en función del nivel de conocimiento de la lengua. Entre las herramientas de ayuda utilizadas, si es que el aprendiz lo estima oportuno, se encuentran los diccionarios y los mapas conceptuales. Además, cuanto mayor es el nivel de conocimiento de los aprendices de L2, más parecidas son sus necesidades a las de los aprendices que realizan resúmenes maduros.

Finalmente, los encargados de evaluar *resúmenes de estudiantes universitarios* aseguraban que en la universidad no se recibe formación explícita de cómo resumir. El adiestramiento, en caso de existir, pasa a ser una opción de auto-aprendizaje o una iniciativa individual. El uso de herramientas de apoyo queda relegado también a la iniciativa individual.

El diagrama de la Figura 1 recoge los factores o variables que hemos identificado y que convergen a la hora de determinar la calidad de un resumen. La nota o puntuación global estará sujeta a la probabilidad estimada para la ocurrencia de cada una de estas variables, sometidas, a su vez, a las decisiones de la evaluación de los expertos. Se han identificado variables referentes a la *evaluación del discurso* (se muestran con forma ovalada en la Figura 1) que recoge la coherencia, cohesión, adecuación, uso del lenguaje y comprensión; la *evaluación de conocimientos previos* (forma de pentágono en la Figura 1) se refiere al nivel del aprendiz, sus conocimientos previos sobre el tema y el contexto educativo; la *evaluación relativa al texto* (forma romboide en la Figura 1) se refiere a la ausencia/presencia, tipo, longitud y tema del texto *y* finalmente, la *oferta de herramientas de ayuda* (forma de cuadrado con borde redondeado en la Figura 1).

3.2. ¿Qué es común al evaluar? Análisis y modelado de decisiones

Tras analizar las entrevistas con los expertos y considerar el trabajo realizado hasta ahora en el área de evaluación de resúmenes (Sección 2) se han identificado las variables que inciden en la calidad de un resumen, pero, ¿cuál es la relevancia de cada variable al calcular la puntuación global del resumen? En este apartado se describe el experimento llevado a cabo para identificar las decisiones subyacentes a la estimación de la nota global del resumen.

Participantes. La mayor parte de las investigaciones (ver Sección 2) coinciden en destacar la diferenciación entre dos grupos críticos atendiendo a la madurez del resumen (Brown & Day, 1983; Garner, 1982, 1987; Taylor, 1982). Entre los resúmenes inmaduros cabe distinguir, además, aquellos resúmenes producidos durante el proceso de aprendizaje de una segunda lengua (Kozminsky & Graetz, 1986; Long & Harding-Esch, 1978). Atendiendo a todas estas premisas, se seleccionaron expertos profesionales de los tres ámbitos educativos citados (primaria y secundaria, L2 y universidad) y

con una larga experiencia en su tarea. Un total de 15 participantes: 5 de educación primaria o secundaria, 5 de enseñanza de L2 y 5 del ámbito universitario.

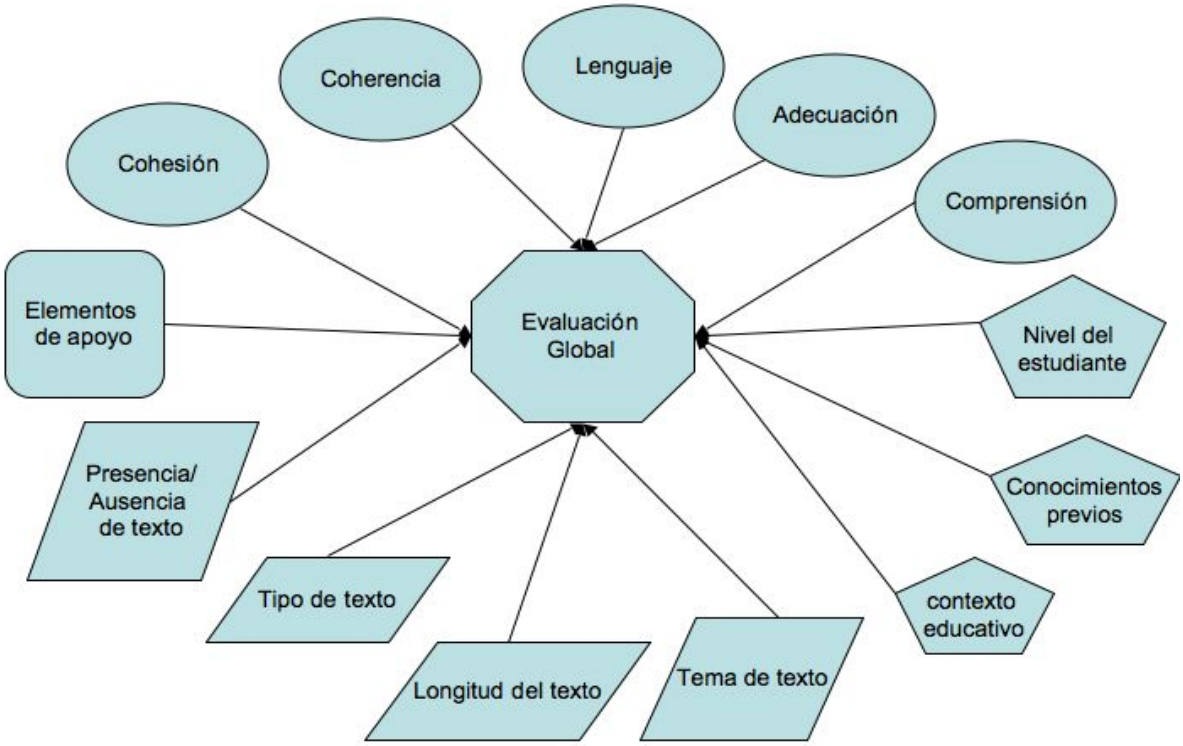

Figura 1. Modelo de evaluación de resúmenes

Método. La misma diversidad que se aplicó en la selección de participantes fue considerada al escoger la variedad de resúmenes para la prueba. Cada participante evaluó 5 resúmenes incluidos en un libreto que contenía 2 resúmenes de L2, 2 resúmenes de enseñanza primaria y secundaria y 1 resumen de universidad. Los resúmenes a evaluar eran iguales para todos los participantes y correspondían a resúmenes escritos en Euskara sobre un sólo texto original cuyo tema era *el ciclismo y el dopaje*. En el libreto también se adjuntaban las instrucciones experimentales junto con el texto en que se basaban los resúmenes. Las variables a evaluar eran: coherencia, cohesión, adecuación, y corrección lingüística y de contenido. Se proporcionaron definiciones de las variables a evaluar para cerciorarse de que fuesen comprendidas de manera uniforme. Además, los evaluadores debían incluir una valoración global del resumen. Todas estas puntuaciones fueron recogidas en una escala de 0-10.

Resultados. Por un lado, se trató de descubrir si existía algún criterio común entre los distintos evaluadores, y se realizaron estudios predictivos (mediante modelos de regresión lineal múltiple y modelos de redes Bayesianas) para analizar hasta qué punto se podían representar y predecir las respuestas de las puntuaciones. Además, se mostró mediante análisis correlativo *r* que las respuestas de los evaluadores, al margen de su origen, tenían un alto índice de congruencia. Expertos de L2 mostraron un nivel medio de correlación entre r=0.75 y r=0,96; expertos de Universidad r=0,51 y r=0,9 y el nivel medio de correlación en el contexto de la educación primaria y secundaria vario entre r=0,47 y r=0,84. Todas las correlaciones fueron significativas a nivel $p < 0,01$.

Por medio del análisis del *modelo de regresión lineal múltiple* por pasos se pudo explicar el 89% de la varianza observada, con una $F(1,71)=199,9$ y un nivel

de significación p<0,01. Tres de las variables estudiadas fueron identificadas como predictoras: coherencia, comprensión y lenguaje. Los valores de Beta mostraron el nivel de relevancia que se predijo para cada variable: β=0,47 para coherencia, β=0,38 para comprensión y β=0,16 para lenguaje (Zipitria et al., 2004).

Las observaciones obtenidas por medio del análisis regresivo fueron confirmadas y ampliadas por los *modelos de redes Bayesianas*. Los modelos Bayesianos nos proporcionan las probabilidades de las puntuaciones de cada una de las variables estudiadas con independencia de su relevancia global. Este modelado mostró que las variables eran independientes entre sí, siendo por tanto posible estudiar cada una de ellas por separado (Zipitria et al., en edición). La figura 2 muestra el ejemplo del modelo analizado en las circunstancias descritas para este experimento utilizando el entorno gráfico ELVIRA (Elvira Consortium, 2002) y considerando además para el análisis el tipo y origen del resumen y el evaluador *(rater)*.

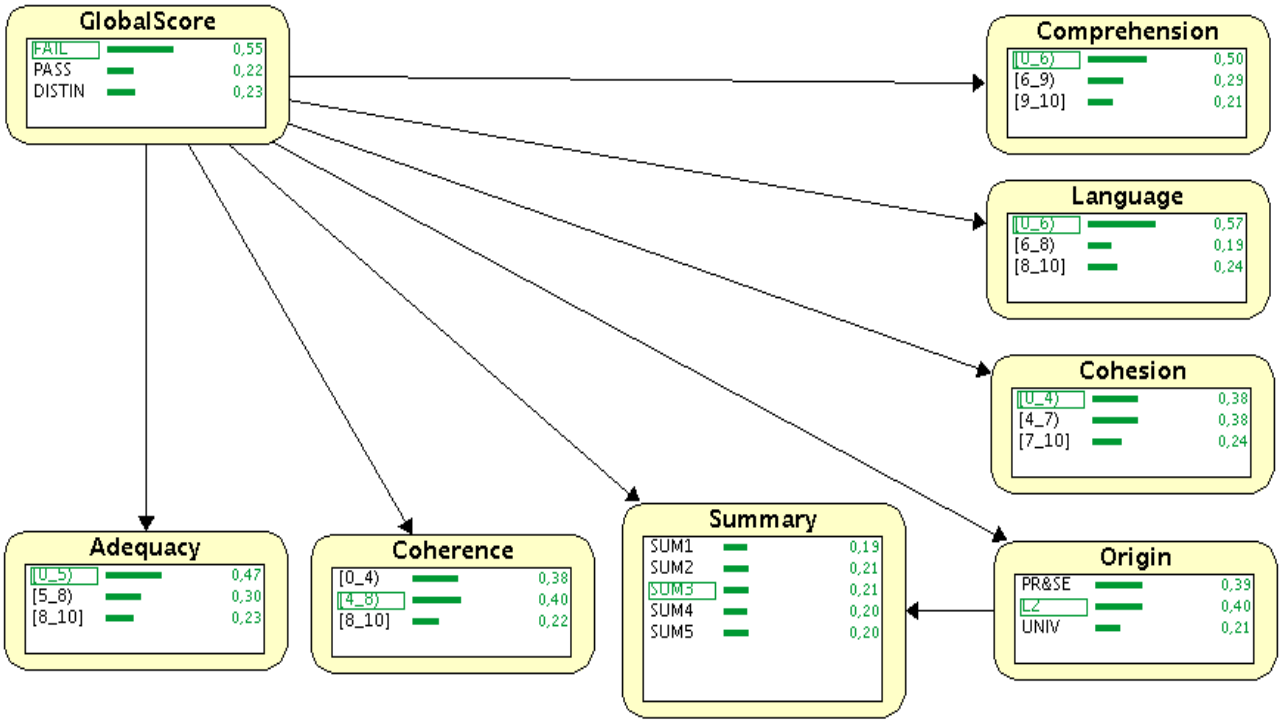

Figura 2. Ejemplo de modelado de decisión Bayesiano con la distribución a priori para cada variable

Tras discretizar los datos atendiendo a categorías de puntuación afines a los utilizados en nuestro sistema educativo, se obtuvo un modelo que predecía el 86,67% de la incertidumbre analizada. Con ello concluimos que la variable que representa al evaluador era irrelevante, ya que no aparecía en la representación gráfica del modelo. Por otro lado, se puso en evidencia la independencia de las variables seleccionadas, con lo cual se confirmaba la entidad individual de las mismas. Esto permitió emular los criterios de evaluación por medio de la selección de una variable o combinación de ellas. Además, permitió identificar la relevancia de cada una de las decisiones de evaluación en la puntuación final.

Discusión. De esta forma, se observó la tendencia a usar un cierto criterio común al decidir las puntuaciones de los resúmenes. Esto nos llevó a explorar y tratar de descubrir aquello que es común al decidir la puntuación de un resumen (Zipitria et al., en edición).

3.3. Entorno Web para la recogida de las puntuaciones de los expertos

Tras observar patrones comunes de evaluación, se creó una aplicación Web para recoger un mayor número de medidas sobre la evaluación de los

expertos. Estos nuevos datos pasan a ampliar y perfeccionar el modelo de toma de decisiones de evaluación descrito anteriormente. Después de unas instrucciones iniciales de cómo realizar la tarea, la aplicación proporciona el texto que da origen a los resúmenes y las plantillas necesarias para evaluar las variables en cada uno de los resúmenes. Finalmente, los datos obtenidos sobre la calidad de los resúmenes junto con información específica sobre cada resumen se insertan en la red Bayesiana.

La figura 3 muestra la interfaz del entorno Web que corresponde a la plantilla de evaluación mediante la que se recoge el valor de distintas variables.

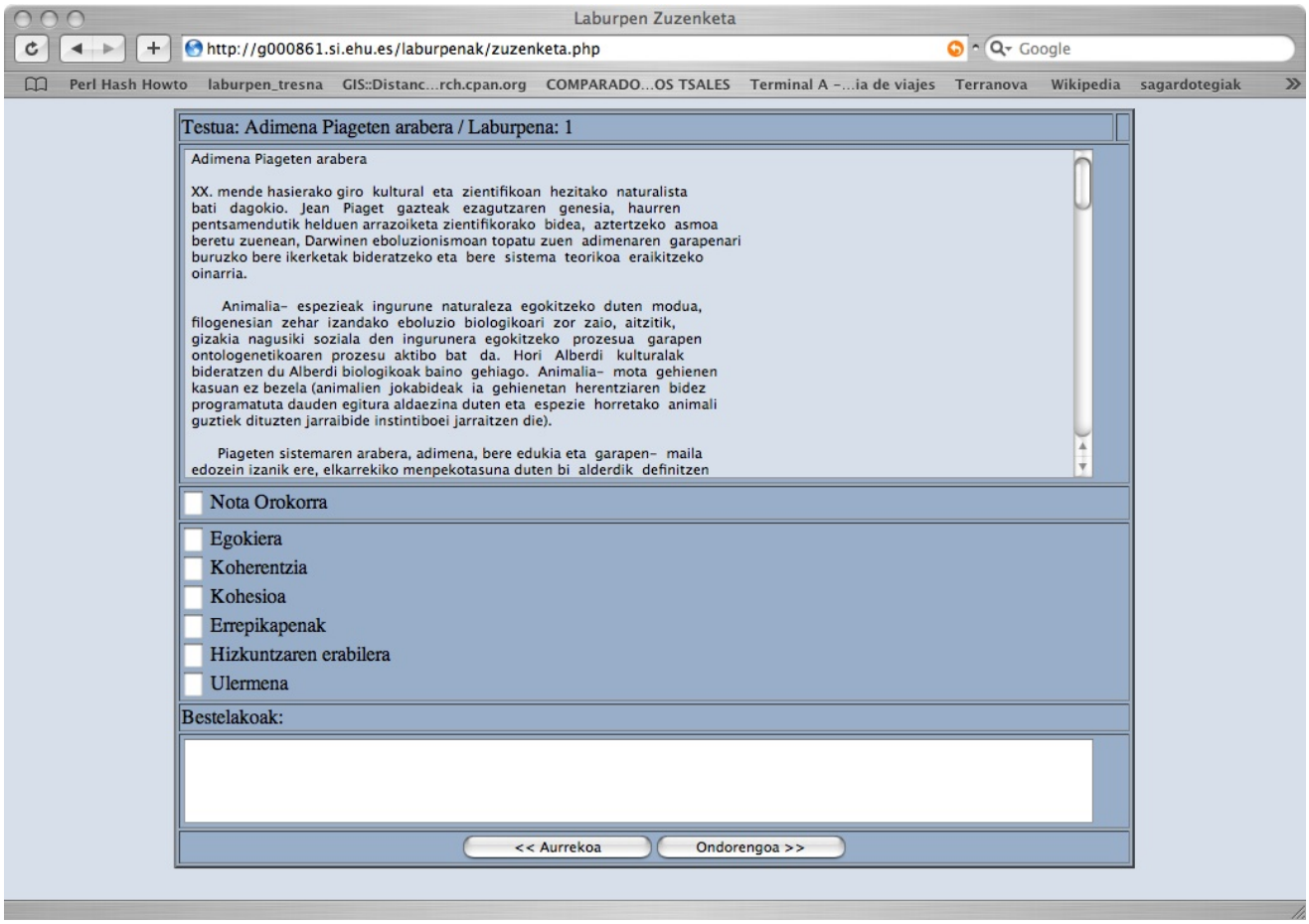

Figura 3. Ejemplo entorno Web para la puntuación de resúmenes.

4. Hacia la automatización de la puntuación global: el sistema LEA

Una vez observados los factores que intervienen en el proceso de evaluación de resúmenes y las decisiones para la estimación de la puntuación se trabaja en el desarrollo de un sistema que bajo estas premisas, sea capaz de evaluar resúmenes de aprendices de forma automática: LEA (Laburpenak Ebaluatzeko Aplikazkioa; *aplicación para la evaluación de resúmenes).*

LEA ha sido desarrollada pensando en los requisitos de la mayoría de los usuarios potenciales dentro del contexto educativo: aprendices inmaduros (L2 y escuela primaria y secundaria) y maduros (universidad). Más aún, se implementa como un instrumento funcional que podría ser utilizado en cualquier lugar provisto de ordenador e Internet. La aplicación toma en consideración tres tipos de estrategias de evaluación: evaluación instructivista (donde un instructor configura el modo de evaluación) y constructivista (el aprendiz escoge el

modo de evaluación). En ambos casos se pueden configurar los parámetros o variables que se considerarán en la evaluación o utilizar los que el sistema ofrece por defecto.

La evaluación global que aporta el sistema por cada resumen se basa en medidas de puntuaciones relativas al discurso, a las valoraciones sobre conocimientos previos, a la habituación del aprendiz al tipo de texto y al acceso del aprendiz a instrumentos de ayuda. Todo ello bajo las decisiones de evaluación calibradas según lo establecido en la Sección 3.2.

Las funcionalidades del sistema LEA se pueden resumir en: gestión de estudiantes o aprendices y grupos de estudiantes, gestión de textos a resumir, asignación de textos a estudiantes individuales o grupos de estudiantes, edición de resúmenes, gestión de las herramientas de ayuda que se pueden utilizar durante la realización del resumen, gestión de las variables a evaluar y evaluación automática de esas variables. La figura 4 muestra la arquitectura de LEA. En ella se han identificado un gestor de ejercicios y estudiantes, el visualizador del resumen para el profesor y el entorno de evaluación y realización del resumen para el estudiante. Además se incluye el módulo de herramientas de ayuda para la evaluación y el módulo de herramientas básicas de evaluación. Estas últimas se refieren a herramientas y métodos de PLN utilizados para evaluar las distintas variables.

LEA es capaz de asignar una puntuación global a un resumen teniendo en cuenta puntuaciones relativas al discurso (subsección 4.1), el nivel cognitivo y contextual del aprendiz (subsección 4.2), características relativas al texto objeto de resumen (subsección 4.3) y la utilización de herramientas de ayuda (subsección 4.4). A continuación se describe en detalle los aspectos que se consideran en la evaluación de cada uno de estos cuatro apartados.

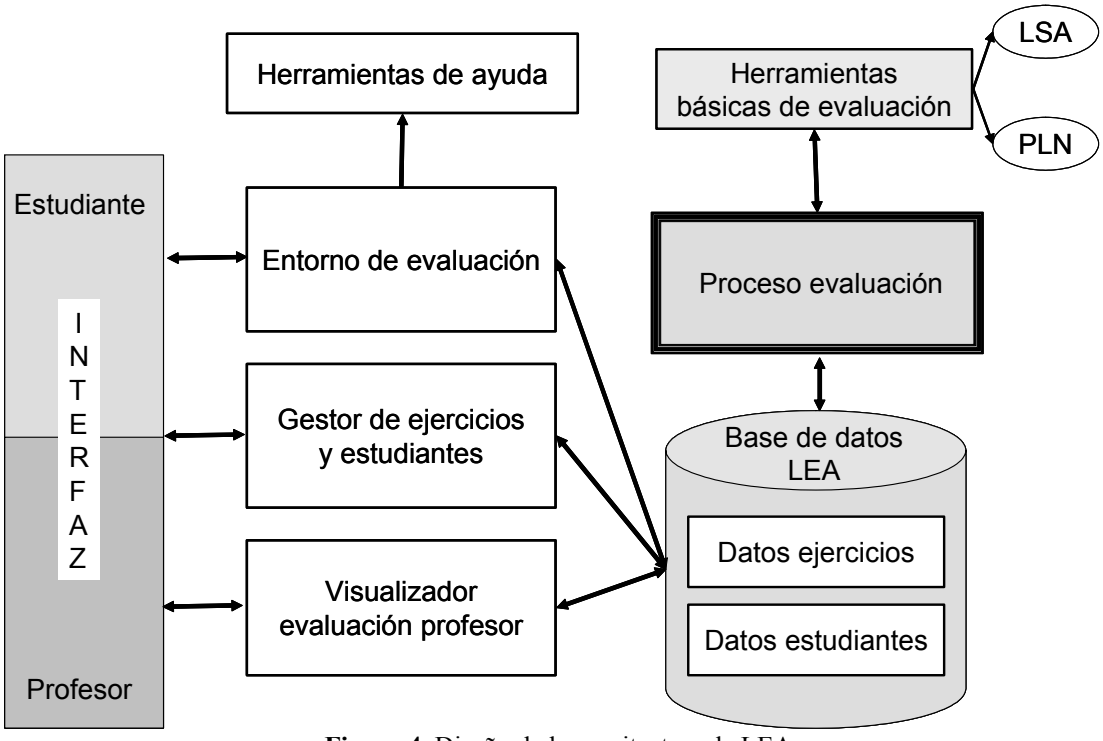

Figura 4. Diseño de la arquitectura de LEA.

4.1. Puntuaciones relativas al discurso

Las puntuaciones de evaluación relativas al discurso requieren la comprensión de lo escrito de forma automatizada.

Se ha utilizado *Latent Semantic Analysis (LSA)* para comprender el lenguaje natural. LSA es una técnica de comprensión de lenguaje natural que basa su conocimiento en la utilización de corpora adecuados. Aunque fue creado por (Deerwester *et al.,* 1990), los autores (Landauer & Dumais, 1997; Landauer et al.,

1998) comprobaron posteriormente la validez de LSA para manejar la similitud semántica de forma análoga a la humana. Además LSA ha sido utilizado ampliamente por diversos autores para modelar la semántica en varias aplicaciones (Foltz et al., 1998; Graesser et al., 2000; Kintsch et al., 2000; Wiemer-Hasting & Graesser, 2000; Wolfe et al., 1998).

Corpora. Como hemos citado anteriormente, el conocimiento semántico para la comprensión automatizada de lo escrito usando LSA se adquiere por medio de un corpus. Para obtener resultados óptimos, es condición indispensable que el corpus contenga suficiente contenido específico sobre el tema evaluado y suficiente cantidad de terminología; sin ello será imposible modelar el conocimiento del lenguaje necesario para evaluar el texto (Olde et al., 2002). Más aún, el uso de una lengua aglutinativa como el Euskara, hace que la lematización del texto produzca respuestas más precisas en aquellos casos en que los se realicen comparaciones entre frases (Zipitria et al., 2006b). Debido a esto, se ha trabajado en la selección, adecuación y filtrado de corpora para utilizarlo en esta tarea específica.

Discurso. Dentro del discurso se analiza la información semántica. Las puntuaciones en torno a la *comprensión* y *coherencia* obtenidas por medio de LSA permiten medir hasta qué punto se ha comprendido el texto y hasta qué punto las oraciones del texto están relacionadas entre sí (Foltz et al., 2000; Foltz et al., 1998; Zipitria et al., 2006a). LSA sirve también como medio para analizar la *adecuación* del vocabulario utilizado con respecto al lenguaje específico del texto. Por otro lado, para el análisis de la información del texto se utilizan diversos recursos de PLN tales como correctores, para analizar la *corrección lingüística* de lo escrito (Aduriz et al., 1997; Aldezabal et al., 2003). La *cohesión* del texto se analiza basándose en texto etiquetado (Aduriz et al., 2004). Una vez que se ha analizado el texto, se calculan las puntuaciones de cada una de las medidas de evaluación que pasarán a formar parte de la red de toma de decisiones.

4.2. Nivel cognitivo y contextual

Tanto las observaciones de los expertos entrevistados como investigaciones previas, confirman la relevancia de tener en cuenta el *nivel del aprendiz* (Brown & Day, 1983; Garner, 1982, 1987; Symons & Pressley, 1993). Debido a ello la aplicación permite escoger la categoría del ejercicio resumen a realizar.

El *conocimiento previo* sobre un tema determinado es otro elemento que ha sido identificado como crítico tanto para producir un resumen (Symons & Pressley, 1993) como para seleccionar una tarea adecuada a un nivel determinado (ver Sección 3.1). En nuestra aplicación la información contextual sobre el aprendiz es especificada por el educador.

4.3. Texto

El entorno de evaluación permite obtener los datos relativos al texto objeto de resumen: ausencia/presencia de texto, tipo, longitud y tema del texto. Estos datos serán tenidos en cuenta a la hora de determinar la puntuación global del resumen.

El sistema permite controlar la *presencia/ausencia* de texto posibilitando o negando al aprendiz revisar el texto de lectura durante el proceso de producción del resumen.

También permite controlar el *tipo de texto* atendiendo a sus características de morfología y contenidos, categorizándolo de acuerdo a las características de los aprendices destinatarios.

La *longitud del texto* es otro factor que puede resultar determinante tanto a la hora de procesar la información leída como a la hora crear un resumen adecuado. Considerar este factor permite observar el impacto que la longitud del texto tiene en las puntuaciones producidas en cada nivel.
No sólo la longitud del texto a leer sino también la longitud del texto resumido son factores determinantes a la hora de producir un buen resumen. Un resumen ha de tener una longitud mínima para poder incluir toda la información imprescindible, y no superar una longitud máxima donde ya sólo cabrían reiteraciones e información superflua. El sistema es capaz de alertar sobre los límites entre los que ha de crearse un resumen adecuado.

Finalmente, el *tema del texto* es otra condición que puede determinar la facilidad o dificultad que implica

crear un resumen. El sistema permite categorizar los textos en temas. La figura 5 corresponde a la pantalla de gestión de textos en la que se recogen las características *título, tema, descripción, género, nivel* y *número de palabras* del texto seleccionado para resumir *"Bilboko Abra lehen eta orain"*. A través de esta ventana se puede, además, asignar el texto como texto a resumir bien a estudiantes individuales o a grupos de estudiantes.

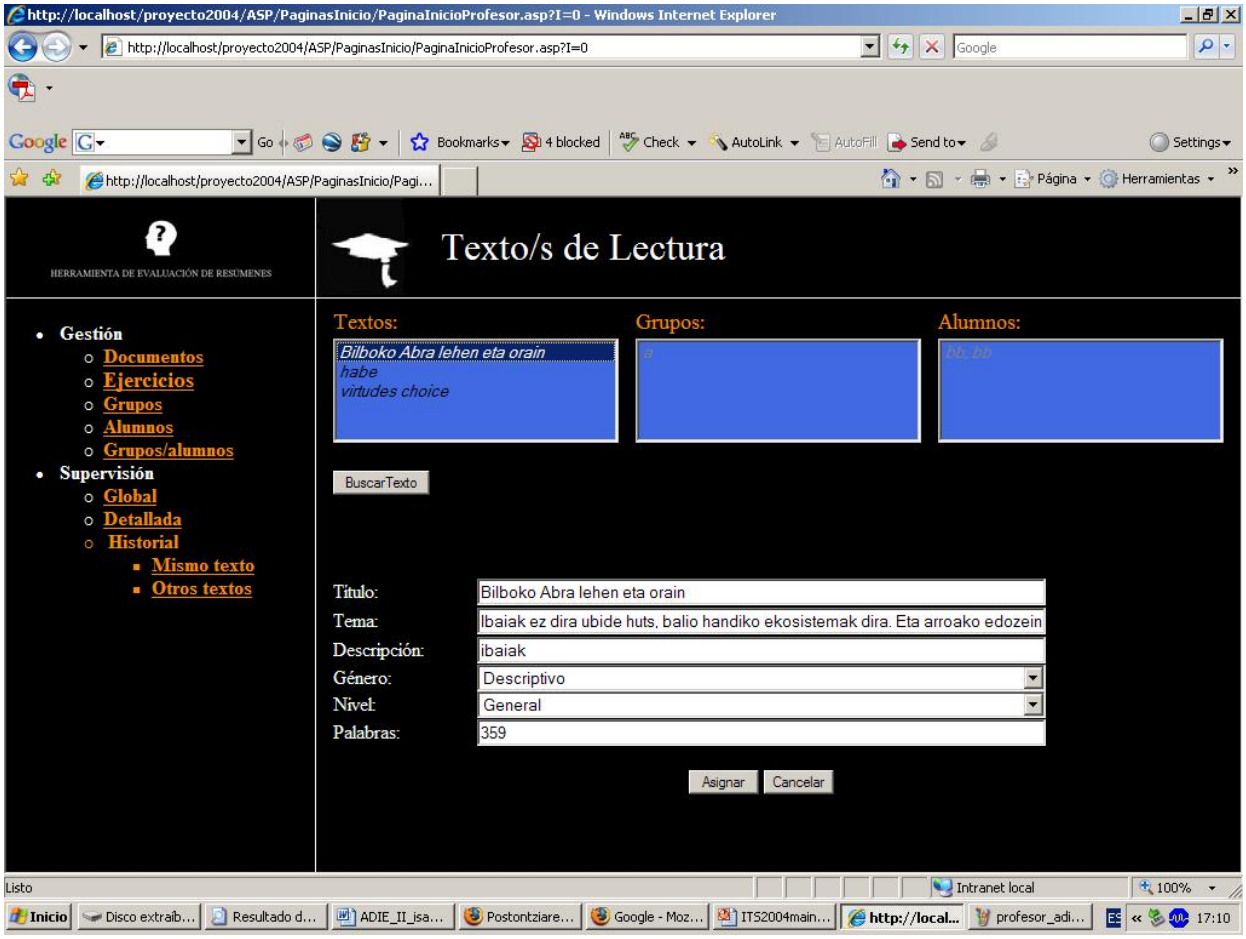

Figura 5. Interfaz de educador para la gestión de textos.

4.4. Herramientas de ayuda

En lo que se refiere a la técnica del resumen, la relevancia de la utilización de herramientas de ayuda está avalada por el uso que la enseñanza tradicional ha hecho de ellas. En al área de los resúmenes, las *herramientas de ayuda* han sido consideradas una necesidad común a los diferentes contextos educativos analizados. Debido a ello, en caso de el aprendiz necesite usarlas, el sistema cuenta con un abanico de *herramientas de ayuda* que responde a las necesidades individuales observadas en las Secciones 2 y 3. No obstante, el sistema permite incorporar nuevas herramientas en función de las necesidades controlando su acceso. El instructor puede configurarlas para permitir o negar el acceso a las herramientas e incluso permitir un acceso parcial. Su utilización o no, e incluso su utilización parcial influirán en la nota final del resumen.

4.5. Toma de decisiones para la puntuación global

La toma de decisiones ha sido identificada como un paso necesario a la hora de evaluar (Genesee & Upshur, 1996). El análisis de modelos predictivos descrito en la Sección 3.2, ha aportado un marco de toma de decisiones de evaluación sobre el que se

asientan las medidas de evaluaciones parciales identificadas en el modelo de evaluación de resúmenes (ver Figura 1). Se ha observado que durante el proceso de instrucción se toman decisiones que luego son consideradas al producir una evaluación (Zipitrial et al. - *en edición*). Así el modelo de evaluación final recoge componentes importantes del proceso mismo de resumir.
El modelo de evaluación de resúmenes generado (sección 3.2) posibilita cumplir un doble objetivo: por un lado permite ponderar la diagnosis realizada de los diferentes factores a tomar en cuenta en la evaluación de resúmenes bajo una medida global y, por otro, permite tomar esta decisión de evaluación con ponderaciones similares a las humanas basándonos en el comportamiento mostrado por los evaluadores expertos.

Así, la nota final obtenida no es una simple suma de los componentes de evaluación identificados, sino el modelo más probable de lo que se esperaría de la decisión de evaluadores expertos.

5. Conclusiones y planes de futuro

El trabajo presentado en este artículo parte de la observación de cómo los expertos realizan la evaluación de resúmenes para luego proceder a su automatización. El análisis realizado sobre la forma en la que evalúan los expertos, junto con lo aportado por la revisión bibliografía, ha proporcionado información sobre los criterios usados en la toma de decisiones a la hora de evaluar un resumen. La toma de decisiones ha sido modelada tanto en modo regresivo como por medio de redes Bayesianas.

Todo esto ha hecho posible establecer un criterio de evaluación de resúmenes de manera similar a como lo harían los expertos. La puntuación global se obtiene a mediante ponderaciones de puntuaciones de las variables parciales, según la probabilidad indicada por el modelo. Más allá de hacer una media de puntuaciones, la nota final se obtiene tomando en consideración el criterio o relevancia de cada variable. Se han identificados variables referentes a la *evaluación discursiva*, la *evaluación de conocimientos previos,* la *evaluación relativa al texto* y el *uso de herramientas de ayuda*.

Además, se ha creado la aplicación informática LEA capaz de generar de forma automática la nota final asociada a un resumen. Para automatizar las puntuaciones parciales se han utilizado técnicas informáticas de Procesamiento Lenguaje Natural y *Latent Semantic Analysis*.

En la actualidad se están afinando las puntuaciones parciales generadas mediante la combinación de diferentes herramientas de comprensión de lenguaje natural. No podemos además dejar de lado el constante enriquecimiento con nuevos datos del modelo creado en base a conocimiento experto. Esto permitirá ajustar y mejorar las decisiones de evaluación generadas.

Agradecimientos

Este trabajo está co-financiado por la Universidad del País Vasco/*Euskal Herriko Unibertsitatea* (UE06/19), la Diputación Foral de Guipúzcoa/*Gipuzkoako Foru Aldundia* bajo un programa asociado a la Unión Europea y por el Ministerio de Ciencia y Tecnología a través del programa CICYT (TIN2006-14968-C02-01). Queremos agradecer especialmente a todos los profesores y alumnos que participaron en la experiencia y a los revisores de la revista ADIE, por sus sugerencias en la mejora del artículo.

Referencias

Aduriz, I., Alegria, I., Artola, X., Ezeiza, N., Sarasola, K., & Urkia., M. (1997). A Spelling Corrector for Basque Based on Morphology. *Literary and Linguistic Computation* **12**(1).

Aduriz, I., Aranzabe, M., Arriola, J., Ilarraza, A. D. d., Gojenola, K., M.Oronoz, & Uria, L. (2004). *A Cascaded Syntactic Analyser for Basque. Computational Linguistics and Intelligent Text Processing*. 2945 LNCS Series.

Aldezabal, I., Aranzabe, M., Arrieta, B., Maritxalar, M., & Oronoz, M. (2003). Toward a Punctuation Checker for Basque. *Proceedings of ATALA Workshop of punctuation*.

Barlett, F. C. (1932). *Remembering; a Studty in Experimental and Social Psychology*: Cambridge University Press.

Bransford, J. D., Kinzer, C., & Risko, V. (1990). *Dimensions of Thinking and Cognitive Instruction.*

Brown, A. L., & Day, J. D. (1983). Macrorules for Summarizing Text: The Development of Expertise. *Journal of Verbal Learning and Verbal Behavior* **22**(1): 1-14.

Cassany, D. (1993). *Didáctica de la corrección de lo escrito* (Vol. 108). Spain: Editorial Graó, de IRIF SL.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*.

Elvira Consortium. (2002). Elvira: An Environment for Creating and Using Probabilistic Graphical Models. *Electronic Proceedings of the First European Workshop on Probabilistic Graphical Models*.

Fitzgerald, J. (1987). Research on Revision in Writing. *Review of Educational Research* **57**(4): 481-506.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments* **8**(2).

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes* **25**: 285-307.

Garner, R. (1982). Efficient Text Summarization. Costs and Benefits. *Journal of Educational Research* **75**(5): 275-279.

Garner, R. (1987). Strategies for Reading and Studying Expository Text. *Educational Psychologist* **22**(3-4): 299-312.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*: Canbridge University Press.

Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group. (2000). Using Latent Semantic Analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments* **8**: 129-148.

Inoue, A. B. (2005). Community-based assessment pedagogy. *Assessing Writing* **9**(3): 208-238.

Kintsch, E., Steinhart, D., Stahl, G., & the LSA Research Group. (2000). Developing Summarization Skills through the Use of LSA-Based Feedback. *Interactive learning environments* **8**(2): 87-109.

Kintsch, W., Patel, V. L., & Ericsson, K. A. (1999). The role of long-term working memory in text comprehension. *Psychologia* **42**: 186-198.

Kintsch, W., & van Dijk, T. A. (1978). Toward a Model of Text Comprehension and Production. *Psychological Review* **85**(5): 263-393.

Kozminsky, E., & Graetz, N. (1986). First vs second language comprehension: some evidence from text summarizing. *Journal of Research in Reading* **9**(1): 3-21.

Landauer, T., & 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., & Foltz, P. W. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes* **25**: 259-284.

Lehnert, W. G. (1981). Plots Units and Narrative Summarization. *Cognitive Science* **4**: 293-331.

Long, J., & Harding-Esch, E. (1978). *Language Interpretation and Communication.*

Manelis, L., & Yekovich, F. R. (1984). Analysis of Expository Prose and Its Relation to Learning. *Journal of Structural Leraning* **8**(1): 29-44.

Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*: The MIT Press.

Olde, B. A., Franceschetti, D. R., Karnavat, A., Graesser, A. C., & TRG. (2002). *The right stuff: Do you need to sanitize your corpus when using latent semantic analysis?* 24rd Annual Conference of the Cognitive Science Society.

Rumelhart, D. E. (1975). *Representation and Understanding: Studies in Cognitive Science.* New York.

Symons, S., & Pressley, M. (1993). Prior knowledge affects text search success and extraction of information. *Reading Research Quarterly* **28**(3): 251-259.

Taylor, B. M. (1982). Text Structured and Children's comprehension and memory for expository material. *Journal of Educational Psychology* **74**: 323-340.

Thorndike, P. W. (1977). Cognitive Structures in Comprehension and Memory in Narrative Discourse. *Cognitive Psychology* **9**: 77-110.

Wiemer-Hastings, P., & Graesser, A. (2000). Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments* **8**(2): 149–169.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text:Matching readers and texts by Latent Semantic Analysis. *Discourse Processes* **25**(309-336).

Zipitria, I., Arruarte, A., & Elorriaga, J. A. (2006a). Observing lemmatization effect in LSA coherence and comprehension grading of learner summaries. In K. D. Ashley, M. Ikeda & T.-W. Chan (Eds.), *Lecture Notes in Computer Science* 4056 (pp. 595-603): Springer.

Zipitria, I., Elorriaga, J. A., & Arruarte, A. (2006b). *LSA learner sentence comprehension in agglutinative and non-agglutinative languages.* Proceedings of the ITS 2006 Workshop on Teaching with Robots, Agents and NLP, Jhongli, Taiwan.

Zipitria, I., Elorriaga, J. A., Arruarte, A., & Díaz de Ilarraza, A. (2004). From Human to Automatic Summary Evaluation. In J. C. Lester, R. M. Vicari & F. Paraguaçu (Eds.), *Lecture Notes in Computer Science 3220* (pp. 432-442): Springer.

Zipitria, I., Larrañaga, P., Armañanzas, R., Arruarte, A., & Elorriaga, J. A. (en edición). What is behind a summary evaluation decision? *Behavior Research Methods*