Universidad de Concepción
Dirección de Postgrado
Facultad de Humanidades y Arte - Programa
de Doctorado en Lingüística

CONICYT
Ministerio de Educación
Gobierno de Chile
Fondecyt
Fondo Nacional de Desarrollo
Científico y Tecnológico

Evaluación automática de la estructura semántica de pirámide invertida en noticias escritas

Tesis para optar al grado de Doctor en Lingüística

SERGIO ANDRÉS HERNÁNDEZ OSUNA
CONCEPCIÓN - CHILE
2016

Profesora guía: Dra. Anita Ferreira Cabrera
Dpto. de Español – Fac. de Humanidades y Arte
Universidad de Concepción.

A Karina Andrea, Sergio Tomás y Diego Andrés,
nombrados en el estricto orden en que llegaron a mi vida
y la cambiaron para siempre.
:-)

A mis padres, por todo.

A la memoria de Hugo Enrique Olea Morales,
periodista, maestro y amigo.

Agradecimientos

Luego de casi nueve años en los programas de postgrado en Lingüística y una tesis de magíster sin agradecimientos ni dedicatorias (pues ya era alumno del doctorado al defender), llega el momento de cerrar esta etapa.

He de confesar que en mis primeras clases en el magíster, en marzo de 2007, sentí pánico pues no sabía ni siquiera lo que era un fonema y las diez personas que compartían el curso de Fonología conmigo sí manejaban el concepto... Vaya lío. Quise retirarme. Simultáneamente, tomé el curso de Introducción a la Lingüística Aplicada, en que todos mis compañeros eran profesores de inglés, por lo que el enfoque de la asignatura estaba bastante claro: enseñanza de segundas lenguas. El resultado de esto es que decidí abandonar y llegué a convencerme de haber tomado una decisión errada al postular al magíster.

Cuando iban tres semanas de clases y yo tenía -casi- asumida mi decisión de retirarme del programa, una de las profesoras del curso de lingüística aplicada se acercó a mí al final de la clase. Me dijo que le daba la impresión de que yo andaba un poco perdido -todavía agradezco la consideración de no decirme absolutamente perdido-. Yo le dije que era verdad y que quería cambiarme. Y en el limbo de los malentendidos quedaron mis palabras, yo quería cambiarme a un programa de magíster que se dictaba en la Facultad de Cs. Sociales de la UdeC, pero la profesora lo interpretó como que quería dejar la asignatura. Se quedó en silencio unos segundos. Me miró fijamente y me propuso darme lecturas diferenciadas a las del resto de mis compañeros, enfocadas en temas de lingüística computacional, con el fin de que que yo no me retirara -del curso-. Luego de plantearme el acuerdo me sonrió y yo dije que sí, que aceptaba y que no me cambiaría -pero del programa, aunque no lo verbalicé-. Luego ocurrió tal cual dijo Joan Manuel Serrat en 1987, “bienaventurados los que están en el fondo del pozo porque de ahí en adelante sólo cabe ir mejorando” y, para mí, aquello fue cierto desde ese día.

Por todo lo anterior, espero que estas palabras alcancen a reflejar aunque sea en parte toda la gratitud que le debo a la profesora que se me acercó al final de esa clase y que me guió en todo momento en mi estancia en el magíster y en el doctorado. Llegar a esta etapa, a esta última curva del camino, hubiera sido

imposible sin el apoyo y la guía de la Dra. Anita Ferreira Cabrera. Ella merece el crédito por todo lo bueno que he logrado en los programas de lingüística; los errores son por mi cuenta.

A ella también le agradezco la posibilidad de ser parte, como tesista de doctorado, del Proyecto Conicyt, Fondecyt regular, 1110812: "Un Sistema Tutorial Inteligente para la Focalización en la Forma en la Enseñanza del Español como Lengua Extranjera" y del Proyecto Conicyt, Fondecyt regular, 1140651: "El *feedback* correctivo escrito directo e indirecto en la adquisición y aprendizaje del español como lengua extranjera". Ambos proyectos, de los cuales la Dra. Ferreira fue y es directora, me apoyaron para poder concretar esta tesis y me dieron la posibilidad de formarme y aprender en una situación real de trabajo científico y académico.

En estas palabras tampoco quiero olvidar a todos los profesores cuyos cursos tomé en estos casi nueve años, pues son parte importante de mi formación aunque mi línea de investigación sea de otra área. Cada uno de ellos me dejó algo que posibilitó que llegara hasta el final. ¡Gracias sinceras y totales! Particularmente, quiero nombrar al Dr. Gastón Salamanca, porque fue él quien me recibió en mi primera clase de magíster y detectó que mis conocimientos de Fonología eran lo suficientemente malos como para justificar mi cara de espanto cada vez que mis compañeros respondían a sus preguntas sobre la materia y yo no tenía idea de nada. Sí, nada de nada. Pero los astros encontraron la línea adecuada, conocí a la Dra. Ferreira, y entendí el desafío que me planteaba el curso de Fonología. Hoy no sólo puedo decir que sé lo que es un fonema, sino que disfruté dicha asignatura y me siento orgulloso del rendimiento que alcancé a final de semestre viniendo de la oscuridad absoluta. Gracias, profesor Salamanca: supo motivarme, supo desafiarme, supo encantarme y, gracias a Dios, yo supe responderle. Esos momentos bonitos no los voy a olvidar.

Estos nueve años también me permitieron conocer a muchas personas con quienes compartimos las aulas (y algunas otras cosas). Un abrazo para ellos, especialmente para mis compañeros de cohorte Kerwin Livingstone, Daniel Pereira, Jorge Lillo, Nahum Lafleur y Emerita Bañados.

Como esta tesis es sólo una etapa de un proyecto que pretendo continuar, vaya toda mi gratitud para los doctores -en estricto orden alfabético, según sus

apellidos- Giovanni Parodi, Pedro Salcedo y Carlos González, integrantes de la comisión evaluadora, que con sus generosas observaciones al manuscrito, me permitieron mejorar mi trabajo de cara a lo que vendrá.

También quiero agradecer a la Universidad de Concepción por su apoyo económico y exención de parte de mi carga académica en estos últimos dos años, con el fin de poder finalizar mi doctorado de la mejor forma posible. A la Facultad de Ciencias Sociales y al Departamento de Comunicación Social, por haberme respaldado en las peticiones señaladas ante la UdeC. Y, muy especialmente, quiero nombrar aquí a los decanos de mi Facultad, Dr. Jorge Rojas y Dr. Bernardo Castro, y a los directores del Dpto. Prof. Hugo Olea, Prof. Héctor Alarcón, Prof. Carlos Oliva y Dr. Claudio Jofré por su apoyo como autoridades para lograr esto.

Por último, agradezco a mis colegas Isabel Plaza y Tito Matamala, junto con los 20 estudiantes del curso de Producción de Textos Periodísticos, por haber sido parte de esta investigación que cierra mi estancia en el Doctorado en Lingüística.

¡Muchas gracias a todos!, muy especialmente a quienes no nombré, pese a que me motivaron a terminar. Mis disculpas para ellos. No fue nada personal, sólo la memoria jugando a las escondidas y engañándome con la bendita trampa del olvido. La próxima vez, prometo recordar. Que así sea.

Resumen

La presente tesis se sitúa en el contexto disciplinar de la Lingüística Aplicada, específicamente, en el ámbito de la interdisciplina denominada lingüística informática o computacional.

Su objetivo es desarrollar un componente, a nivel de prototipo, que evalúe el proceso de escritura de noticias, enfocándose en la evaluación de la coherencia textual y en el aspecto semántico estructural de éstas (estructura de pirámide invertida). Esta estructura consiste en que entre el titular y el *lead* (primer párrafo) del texto, debe entregarse la información más relevante de la noticia al lector.

Dicho componente está construido en base a *scripts* programados en Python 3 y algunas aplicaciones ya disponibles y de código abierto. Su funcionamiento se basa en el Análisis Semántico Latente (LSA) y las noticias con que trabaja, que por exigencia del LSA deben pertenecer a un dominio temático restringido, son noticias políticas extraídas del diario chileno La Tercera.

En la construcción del componente se consideró que éste pueda adaptarse fácilmente a una estructura informática mayor, con el fin de poder concretar la idea futura de desarrollar un sistema tutorial inteligente destinado a apoyar la enseñanza de la escritura de noticias en estudiantes de periodismo.

Los resultados de la investigación arrojaron que la evaluación que realiza la máquina es similar a la efectuada por un evaluador humano experto, lo que permite considerar efectivo el funcionamiento de la herramienta diseñada.

Uno de los aportes más interesantes del presente trabajo es que aplica el método de la evaluación automática de la coherencia textual mediante LSA, con un propósito diferente del fin para el que fue originalmente concebido, logrando evaluar acertadamente la estructura semántica-jerárquica de una noticia denominada pirámide invertida.

Índice de Contenidos

Introducción.....1
Contexto.....1
Definición del problema.....2
Estructura de la tesis.....6
Primera parte: Marco teórico.....9
Capítulo 1: Conceptos fundamentales.....9
1.1. Lingüística aplicada.....9
1.1.1. Lingüística computacional.....10
1.2. Procesamiento del lenguaje natural.....12
1.3. Sistemas tutoriales inteligentes.....16
1.3.1. ELE-Tutora: un sistema tutorial inteligente enfocado en la enseñanza del español como lengua extranjera.....20
1.4. Análisis Semántico Latente.....23
1.4.1. El Análisis Semántico Latente: teoría de la adquisición, inducción y representación del conocimiento.....24
1.4.2. El Menón.....25
1.4.3. El problema de Platón: la pobreza del estímulo.....27
1.4.4. El Análisis Semántico Latente y el problema de Platón.....28
1.4.5. ¿Constituye el Análisis Semántico Latente una explicación al problema de Platón?.....32
1.5. El Análisis Semántico Latente como herramienta para la medición de coherencia textual.....35
1.5.1. Coherencia y cohesión.....35
1.5.2. La medición de coherencia textual mediante Análisis Semántico Latente.....42
Capítulo 2: La noticia: su estructura y su producción escrita.....46
2.1. Discurso especializado, discurso profesional y discurso académico.....46
2.2. La noticia.....48
2.2.1. El titular.....52
2.2.2. El *lead*.....52
2.2.3. El cuerpo de la información.....54
2.3. Estructura de la noticia como discurso.....54
2.3.1. El titular y el encabezamiento.....57

2.3.2. El episodio: los acontecimientos principales en el contexto y sus antecedentes....58
2.3.3. Consecuencias....58
2.3.4. Reacciones verbales....58
2.3.5. Comentarios....59
2.4. La noticia: ¿un tipo particular de resumen?....60
Segunda parte: Marco metodológico....65
Capítulo 3: Metodología....65
3.1. Diseño de investigación....65
3.2. Objetivos....66
3.2.1. Objetivo general....66
3.2.2. Objetivos específicos....66
3.3. Hipótesis....66
Capítulo 4: Diseño e implementación del módulo (a nivel de prototipo) de análisis semántico, enfocado en la predicción de la coherencia textual....68
4.1. Recopilación automática del corpus....68
4.2. Preparación automática del corpus....72
4.3. Evaluador automático de coherencia textual....78
4.4. Resultados de las pruebas realizadas al módulo 1....83
4.5. Consideraciones sobre las pruebas realizadas al módulo 1....89
Capítulo 5: Diseño de un módulo corrector ortográfico....91
5.1. Errores ortográficos en el texto de entrada....91
5.2. Arquitectura y funcionamiento del corrector ortográfico....92
5.2.1. Procesamiento palabra a palabra....93
5.2.2. Procesamiento como una cadena en busca de patrones de error complejos....99
5.2.3. La necesidad de un corrector ortográfico dinámico....101
5.2.4. Entrega de resultados....104
5.2.5. Sobre el corrector ortográfico diseñado....107
Capítulo 6: Construcción de un módulo (a nivel de prototipo) para evaluar la jerarquización en la producción escrita de noticias....109
6.1. Ampliación del corpus....109
6.2. Construcción del módulo de evaluación de la pirámide invertida....111
6.2.1. Modelo planteado para la evaluación de la estructura de pirámide invertida en una noticia....111

6.2.2. Estructura del tercer módulo.....115
6.3. Diseño y aplicación de la tarea de escritura de una noticia.....120
Capítulo 7: Presentación y análisis de resultados.....126
7.1. Pilotaje del componente.....126
7.2. Presentación de los resultados de la actividad práctica con estudiantes 129
7.2.1. Comparación 1: el titular de la noticia procesada con un titular tipo 130
7.2.2. Comparación 2: el *lead* de la noticia procesada con las preguntas fundamentales.....133
7.2.3. Comparación 3: el titular y el *lead* (como conjunto) de la noticia procesada con las preguntas fundamentales.....135
7.2.4. Comparación 4: párrafos siguientes al *lead* con los datos adicionales del punteo construido.....137
7.3. Análisis de los resultados de la actividad práctica con estudiantes.....139
7.3.1. Análisis de los resultados del componente.....139
7.3.2. Análisis de la comparación realizada entre los resultados de la máquina y los humanos.....142
7.4. Valoración de los resultados de la actividad práctica con estudiantes....151
Capítulo 8: Conclusiones.....154
8.1. Limitaciones y proyecciones.....159
Referencias.....165
Anexo 1: Práctico aplicados a los estudiantes de periodismo.....173

Índice de Gráficos

Gráfico 1: Puntajes promedio de todas las pruebas.....89
Gráfico 2: Comparación 1, titular de la noticia procesada con un titular tipo.....132
Gráfico 3: Comparación 2, *lead* de la noticia procesada con las preguntas fundamentales.....134
Gráfico 4: Comparación 3, titular y *lead* como conjunto de la noticia procesada con las preguntas fundamentales.....136
Gráfico 5: Comparación 4, párrafos siguientes al *lead* con los datos adicionales. 138
Gráfico 6: Gráficos de las cuatro comparaciones realizadas.....145

Índice de Figuras

Figura 1: *Continuum* de textos en ámbitos académicos y profesionales (Parodi, 2007).....48
Figura 2: Pirámide invertida (Fundeu, 2015).....50
Figura 3: La pirámide invertida y la pirámide tradicional.....51
Figura 4: Estructura hipotética de un esquema informativo (Van Dijk).....57
Figura 5: Esquema de funcionamiento del módulo 1.....83
Figura 6: Arquitectura del corrector ortográfico.....93
Figura 7: División de la noticia en dos partes.....113
Figura 8: Estructura del componente y sus tres módulos.....116
Figura 9: Esquema preliminar del componente dentro del módulo del tutor.....162

Índice de Tablas

Tabla I: Resultados de la comparación de noticias completas pertenecientes a distintos medios....88
Tabla II: Detalle de los corpus utilizados....110
Tabla III: Resultados del pilotaje....127
Tabla IV: Interpretación de las correlaciones....130
Tabla V: Resultados de la comparación 1....131
Tabla VI: Resultados de la comparación 2....133
Tabla VII: Resultados de la comparación 3....135
Tabla VIII: Resultados de la comparación 4....137
Tabla IX: Correlaciones entre los espacios semánticos (comparación 1)....140
Tabla X: Correlaciones entre los espacios semánticos (comparación 2)....140
Tabla XI: Correlaciones entre los espacios semánticos (comparación 3)....141
Tabla XII: Correlaciones entre los espacios semánticos (comparación 4)....141
Tabla XIII: Resumen de las correlaciones (humanos-máquina)....143
Tabla XIV: Correlaciones humanos-máquina (comparación 1)....148
Tabla XV: Correlaciones humanos-máquina (comparación 2)....148
Tabla XVI: Correlaciones humanos-máquina (comparación 3)....149
Tabla XVII: Correlaciones humanos-máquina (comparación 4)....149

Introducción

En 1950, cuando Alan Turing publicó “Computing machinery and intelligence”, metafóricamente, hizo que el mundo girara. En dicho artículo planteó el *Test de Turing*, que consiste -explicado someramente- en probar la capacidad de una máquina para exhibir un *comportamiento inteligente* similar, o indistinguible, del de un ser humano. Nombrar aquí todos los desarrollos surgidos tras la propuesta de Turing es imposible, pero basta mirar en nuestra vida cotidiana para tener una idea de cómo ha evolucionado la tecnología en 66 años.

Este avance recién descrito, ha permeado todos los ámbitos de nuestra vida. Incluso en una de las actividades más importantes para el ser humano, cuyo surgimiento data de alrededor del año 3.300 a. de J. C. y es considerado como el límite entre la Prehistoria y la Historia. Esta actividad es la escritura.

Contexto

La escritura es una actividad compleja. El profesor Daniel Cassany de la *Universitat Pompeu Fabra* de Barcelona ha dedicado gran parte de su investigación a la comunicación escrita. En la introducción a su obra “Construir la escritura”, Cassany (1999) cuenta que la cita de Bereiter y Scardamalia que abre el texto la había querido emplear en “Describir el escribir” (uno de sus libros anteriores), pero el editor le sugirió suprimirla porque “resultaba inadecuado y desmoralizador iniciar una obra sobre el aprendizaje de la escritura resaltando su notable dificultad”. El pasaje en cuestión es el que sigue: “Escribir un ensayo extenso es probablemente la tarea constructiva más compleja que se espera que realice la mayoría de los seres humanos en alguna ocasión. Aunque existan muchas otras actividades de dificultad parecida o superior -como diseñar y construir edificios, investigar de modo experimental, coreografiar una secuencia de baile, presentar una demanda judicial o reestructurar una empresa- la mayoría de las personas no deben realizarlas. Estas tareas están reservadas a personas con talento y formación especiales. Pero la escuela y la sociedad parecen creer que casi todo el mundo debe ser capaz de producir un ensayo coherente de cuatro mil palabras sobre un tema como, por ejemplo, la participación de los discapacitados en el deporte. No

obstante, la complejidad de esta tarea puede rivalizar con la de las otras, que sólo asignamos a los elegidos" (Bereiter y Scardamalia, 1983; Cassany, 1999). En el pasaje anterior queda clara la dificultad que involucra escribir. Si bien sólo se enfoca en los ensayos es indudable que hay otros tipos de textos tanto o más complejos de construir, como por ejemplo, cualquier tipo de discurso especializado, como es el caso del texto -noticia- con que se trabaja en esta tesis.

En el escenario recién expuesto, en que se comprende con claridad la dificultad que implica la producción escrita, es que surge la necesidad de investigar en torno a una herramienta computacional que pueda aportar en un aspecto específico de dicha tarea.

Definición del problema

Como se señaló en el apartado anterior, el presente estudio se enfoca en un aspecto particular relacionado con la escritura, esto es, la evaluación de la producción escrita de la noticia.

La motivación de la que nace esta investigación es de la percepción de las debilidades que muestran los estudiantes de periodismo al producir este tipo de texto, que se puede considerar como la presentación resumida de una situación real. Una de las características principales de la noticia, que podría considerarse como una particular clase de resumen, es la estructura semántica de la pirámide invertida. Ésta es una estructura en la cual el clímax de la narración se sitúa al comienzo, lo que implica que entre el titular y el primer párrafo debe entregarse en forma sucinta la información más relevante al lector. Aprender a producir esta pirámide invertida es una de las mayores dificultades para quien se está formando como periodista (Warren, 1975; Martínez Albertos, 2004; Martín Vivaldi, 2004).

El texto con que se trabajará -la noticia- es un tipo particular de discurso especializado, que se enseña a producir en el ámbito académico (carreras de periodismo en pregrado, en el caso de Chile) para utilizarse en el campo profesional por quienes obtengan el título de periodista.

En el escenario descrito, la investigación que se presenta en estas páginas se sitúa en el contexto disciplinar de la Lingüística Aplicada (LA), específicamente en el ámbito de la interdisciplina denominada lingüística informática o computacional

(siguiendo a Pastor, 2004).

La problemática lingüística específica en la cual se enfoca el estudio es la coherencia textual y la correcta construcción de la estructura semántica llamada pirámide invertida, en el discurso especializado escrito denominado noticia. Para hacerse cargo de lo anterior, que busca resolver un problema concreto relacionado con el uso de la lengua, se utilizará el apoyo de la tecnología acorde con el ámbito de la interdisciplina en que se sitúa la investigación.

En consecuencia, el objetivo de la tesis es desarrollar un componente, a nivel de prototipo, que evalúe el proceso de escritura de noticias, enfocándose en la evaluación de la coherencia textual y en el aspecto semántico estructural de éstas (estructura de pirámide invertida).

La razón para construir este componente es para incorporarlo a futuro en un sistema tutorial inteligente (STI, por sus iniciales). Un STI es “un programa para la enseñanza-aprendizaje guiado por el computador cuya finalidad última es la facilitación de los procesos de aprendizajes personalizados. Estos sistemas son capaces de comportarse como un experto, tanto en el dominio de conocimiento que enseña como en el dominio pedagógico, donde es capaz de diagnosticar la situación en la que se encuentra el estudiante y, de acuerdo con ello, ofrecer una acción o solución que le permita progresar en el aprendizaje” (Ferreira, 2007). El componente en cuestión se construye, con el fin de poder concretar a futuro la idea de desarrollar un sistema tutorial inteligente destinado a apoyar la enseñanza de la escritura de noticias en estudiantes de periodismo. Un componente como el construido será una pieza importante sobre la que sustentar un desarrollo como el recién descrito.

En este sentido, es importante señalar que este componente lo que realiza es detectar errores en la construcción de la estructura semántica denominada pirámide invertida. Dado que se construye para insertarse en un STI, es necesario señalar que hay experiencias previas de sistemas de este tipo que tienen módulos dedicados a la detección de errores. Por ejemplo, ELE-Tutora, un sistema tutorial inteligente para la enseñanza del español como lengua extranjera que tiene la capacidad para detectar los errores de los estudiantes y apoyarlos en la reflexión necesaria para que sean capaces de corregirlos y no volver a cometerlos. ELE-

Tutora, entre sus componentes, cuenta con un analizador sintáctico que permite identificar y clasificar los errores gramaticales en español como lengua extranjera (Ferreira y Kotz, 2010).

Volviendo al componente de la presente tesis, para lograr que éste sea capaz de realizar las evaluaciones de la coherencia textual y de la pirámide invertida, desde el punto de vista tecnológico, se utilizará el Análisis Semántico Latente (LSA, por sus iniciales en inglés), que es una técnica matemático-estadística que mide la relación semántica entre las palabras de un texto y que ha probado tener un amplio número de aplicaciones en los estudios que sobre la técnica se han realizado. Por exigencia del LSA, los textos que se utilicen deben pertenecer a un dominio temático restringido, por ello las noticias que se emplean en este trabajo serán sobre política.

De una manera muy sintética, se puede adelantar que el Análisis Semántico Latente es un método estadístico automático para la representación del significado de las palabras y pasajes de texto, basado en el análisis de extensas entradas (*inputs*) de texto. A partir de estas extensas entradas, se genera un espacio semántico y en éste las palabras, oraciones y el texto completo se representan a través de vectores. La cercanía o lejanía entre dos parejas de vectores se determina calculando el coseno entre ellos y este valor indica la relación semántica entre las dos unidades textuales seleccionadas: mientras mayor sea el valor del coseno, mayor será la cercanía de ambas unidades en el espacio semántico y, por ende, mayor la relación semántica (Kintsch, Steinhart, Stahl & LSA research group, 2000).

El componente, a nivel de prototipo, lo conforman dos módulos principales: una herramienta que sea capaz de realizar una evaluación automática de la coherencia textual de los textos producidos por los estudiantes y otro módulo que, utilizando la salida del anterior, realice la evaluación de la estructura semántica de la pirámide invertida. Ambos prototipos basan su funcionamiento en la técnica de Análisis Semántico Latente. A lo anterior hay que agregar que también se construyó un módulo enfocado en la corrección ortográfica de los textos de entrada, ya que uno de los problemas que se pueden presentar al emplear Análisis Semántico Latente y, en general, cualquier método que trabaje con textos como entrada son los errores que este texto pudiera tener en su superficie. Específicamente, en el

caso del LSA, los errores que más afectan al resultado final son los errores ortográficos. No así otro tipo de errores como, por ejemplo, los de concordancia que, generalmente, perderían relevancia tras el proceso de lematización. Por lo tanto, el componente -a nivel de prototipo- desarrollado lo integran en total tres módulos, a los que se llamará de ahora en adelante como *módulo 1*, *módulo 2* y *módulo 3*. Cada uno de éstos son:

Módulo 1 (evaluador automático de coherencia textual).
Módulo 2 (analizador y corrector ortográfico).
Módulo 3 (evaluador automático de la estructura semántica de pirámide invertida).

Uno de los aportes más interesantes del presente trabajo es que al operar conjuntamente los módulos 1 y 3 en el componente prototipo, se aplica el método de la evaluación automática de la coherencia textual (Foltz, Kintsch y Landauer, 1998) con un propósito diferente del fin para el que fue originalmente concebido, logrando evaluar acertadamente la estructura semántica-jerárquica de una noticia (pirámide invertida), como se verá más adelante. Es necesario notar que no se tienen antecedentes de trabajos similares, que utilicen el Análisis Semántico Latente para ello, ni tampoco que empleen la lengua española.

Los análisis de la estructura de pirámide invertida que realice el componente serán comparados con el juicio de humanos. La idea es que las evaluaciones realizadas por el componente sean equivalentes a las efectuadas por dichos jueces, con el fin de que éste sea una herramienta efectiva. De esto último, que el componente sea capaz de realizar un análisis equivalente al de un humano, se desprende que se busca que la máquina tenga un comportamiento inteligente, situación acorde a un trabajo inserto en la interdisciplina de la lingüística computacional dentro de la Lingüística Aplicada y que busca construir un componente para un STI.

El diseño de investigación empleado es no experimental, transeccional correlacional. Es no experimental, pues no hay manipulación de una variable, y es transeccional correlacional porque se trabaja con una muestra de textos de estudiantes de periodismo, que se recopiló en un momento específico y único; y dichos textos se evalúan por la máquina y por evaluadores humanos, y se buscan

relaciones entre ambas evaluaciones a través de un estadístico que las correlaciona.

Por último, como una forma de proyectar a futuro el trabajo realizado en la presente tesis, se consideró al construir la herramienta que ésta pueda adaptarse fácilmente a una estructura informática mayor, con el fin de poder concretar la idea de desarrollar un sistema tutorial inteligente destinado a apoyar la enseñanza de la escritura de noticias en estudiantes de periodismo. Y he aquí las dos grandes razones para realizar todo el trabajo que implica esta tesis: la primera es que un prototipo como el construido será una pieza importante sobre la que sustentar un desarrollo futuro como el recién descrito. La importancia de esto -y la segunda razón para el trabajo desarrollado en este estudio- radica en que la noticia es el texto más importante en la formación de pregrado de un estudiante de periodismo y es un texto complejo de producir para los alumnos, ya que exige comprender el hecho al que se refiere y tener la capacidad de resumirlo a su médula, para transmitirlo de la forma más escueta posible, sin despojarlo de la información fundamental que involucra. Por ello, se considera que trabajar en pos de la construcción de un STI para apoyar la enseñanza del proceso de escritura de noticias, será un aporte para la comunidad académica, al enfocarse tanto en la producción de la estructura semántica de la noticia, como en su precisión lingüística.

Estructura de la tesis

En relación a su estructura, la tesis está organizada en dos partes. La primera de éstas se titula “Marco teórico” e incluye dos capítulos. La segunda se denomina “Marco metodológico” y seis capítulos la integran.

En el capítulo 1 se presentan los conceptos fundamentales del marco teórico del trabajo realizado. Entre éstos están la Lingüística Aplicada, los sistemas tutoriales inteligentes y la evaluación de coherencia textual mediante Análisis Semántico Latente.

En el capítulo 2, que cierra el marco teórico, se trata específicamente la noticia (o información periodística, como se le llama también), situándola en el ámbito del discurso especializado, el discurso académico y el discurso profesional.

La idea fundamental es su conceptualización como un tipo particular de resumen, que utiliza en su escritura una estructura semántica conocida como pirámide invertida.

En el capítulo 3, que abre el marco metodológico, se describe el diseño de la investigación, se definen los objetivos de la misma y se establece la hipótesis de trabajo.

En el capítulo 4, se aborda el diseño e implementación de un módulo de análisis semántico, enfocado en la predicción de la coherencia textual, que será la base del desarrollo del segundo módulo de este trabajo: el que evaluará la estructura de pirámide invertida. Un punto importante descrito en este capítulo es el método de recopilación automática del corpus que se utilizó.

En el capítulo 5 se describe el diseño de un módulo corrector ortográfico, que busca mejorar los textos que ingresen al segundo prototipo, con el fin de disminuir la cantidad de errores de este tipo que puedan afectar la evaluación que reciba cada noticia. Una de las características más relevantes del desarrollo logrado, es su carácter dinámico, esto es, que puede mejorarse su funcionamiento *a posteriori* sin necesidad de conocimientos de programación e informática.

En el capítulo 6 se trata sobre el diseño del módulo para evaluar la correcta construcción de la pirámide invertida en la producción escrita de noticias. Además, se detalla la estructura del prototipo, se aborda la ampliación del corpus y construcción de nuevos espacios semánticos multidimensionales y se expone el modelo planteado para la evaluación de la estructura de pirámide invertida en una noticia.

En el capítulo 7 se presentan los resultados obtenidos, para luego exponer el análisis que se realizó de dichos resultados. Dentro de este análisis, se indican los problemas detectados y se señalan las probables causas de los mismos. Además, se entrega una valoración de los resultados obtenidos a la luz de los objetivos de la investigación y de la hipótesis de trabajo utilizada.

Finalmente, en el capítulo 8 se exponen las conclusiones extraídas de la investigación realizada.

Es importante señalar que la presente tesis fue posible gracias al apoyo del Proyecto Conicyt, Fondecyt regular, 1110812: "Un Sistema Tutorial Inteligente para

la Focalización en la Forma en la Enseñanza del Español como Lengua Extranjera", del cual el autor fue tesista de doctorado desde marzo de 2011 a marzo de 2014. Lo mismo hay que señalar del Proyecto Conicyt, Fondecyt regular, 1140651: "El *feedback* correctivo escrito directo e indirecto en la adquisición y aprendizaje del español como lengua extranjera", del cual el autor fue tesista de doctorado desde marzo de 2014 a la fecha.

Primera parte: Marco teórico

Capítulo 1: Conceptos fundamentales

1.1. Lingüística aplicada

Como se dijo en la introducción, la investigación que se presenta en estas páginas se sitúa en el contexto disciplinar de la Lingüística Aplicada (LA), específicamente en el ámbito de la interdisciplina denominada lingüística informática o computacional (siguiendo a Pastor, 2004).

La lingüística aplicada puede concebirse como una dimensión particular de la lingüística. Si se entiende a la lingüística como el estudio de las ciencias del lenguaje y de las lenguas naturales, la identidad de la lingüística aplicada surge de dos rasgos que la definen: su interdisciplinariedad y su finalidad práctica. El objetivo de la lingüística aplicada es la resolución de problemas relacionados con el lenguaje o derivados del uso lingüístico que se plantean en la vida cotidiana. En todo caso, este enfoque en lo práctico no obsta para que la lingüística aplicada se fundamente en presupuestos teóricos y elabore una teorización propia de las materias de las que se ocupa (Pastor, 2004).

La lingüística aplicada, entonces, realiza según Pastor propuestas para solucionar cuestiones concretas relativas a la comunicación o la información, partiendo siempre de la realidad lingüística y no de su idealización. Por lo anterior, la disciplina en cuestión considera al lenguaje desde la perspectiva de su contexto social, político y económico, y no como un mero sistema de signos, descontextualizado o alejado del uso.

El “Diccionario de términos clave ELE” del Centro Virtual Cervantes (2016) señala que “el área en que la lingüística aplicada tal vez ha experimentado una mayor evolución es la enseñanza y aprendizaje de segundas lenguas. Pastor (2004) agrega que a esta área luego se le fueron sumando nuevos campos de desarrollo de la disciplina: la adquisición y enseñanza de las lenguas maternas; la traducción e interpretación; la lexicografía y la terminología; la política y la planificación lingüísticas; el tratamiento de las patologías del lenguaje; y la lingüística informática o computacional.

Este último ámbito, la interdisciplina denominada lingüística informática o computacional, es el ámbito en que se circunscribe la presente tesis. En el aspecto terminológico, se aclara, que se prefiere el nombre de lingüística computacional, para seguir el que se utiliza en la convocatoria del Congreso para 2016 de la Asociación Española de Lingüística Aplicada, AESLA.

1.1.1. Lingüística computacional

Pastor (2004) señala que este campo se aboca al estudio de las cada vez más complejas y diversificadas aplicaciones y aportaciones de la informática al análisis del lenguaje y viceversa. Con toda certeza ésta es el área de la lingüística aplicada que ha progresado de un modo más espectacular en los últimos años. Se trata de un campo científico interdisciplinar bastante reciente: su investigación y desarrollo data de la década de 1950. “Su objetivo no es otro que incorporar en los ordenadores las habilidades en el manejo del lenguaje natural humano y facilitar así el tratamiento informatizado de las lenguas y de su estudio. Según dónde se quiere poner el énfasis, este campo de trabajo ha recibido distintas denominaciones, además de las ya mencionadas: ingeniería lingüística, tecnologías del habla o del lenguaje, procesamiento del lenguaje natural o industrias de la lengua” (Pastor, 2004).

Vista como se señaló al final del párrafo anterior, la lingüística computacional incluiría dentro de sí algunas materias un tanto diversas. Por lo anterior, es importante precisar un poco más el contexto dentro de la interdisciplina en que se ubica el presente estudio.

La lingüística informática tiene tres líneas de investigación principales. La primera línea se denomina *lingüística informática*, en sentido estricto, o *informática aplicada a la lingüística*. Este campo se aboca a la utilización de programas informáticos para la investigación lingüística, es decir, para el estudio científico de las lenguas. Visto así, hay que decirlo, se puede referir a todas las subdisciplinas de la lingüística que emplean herramientas informáticas; sin embargo, Pastor (2004) señala que “se reserva más específicamente a aquellas áreas en las que estos instrumentos tienen más incidencia: la lingüística de corpus, la lingüística estadística, la estilometría, la lingüística histórica computacional, la informática

aplicada a la sociolingüística y la lexicografía asistida por ordenador".

La segunda línea de investigación la constituye la *lingüística computacional* por antonomasia, que tiene tres objetivos: la elaboración de modelos lingüísticos en términos formales e implementables computacionalmente; la aplicación de estos modelos a cualquier nivel de descripción lingüística; y la comprobación automatizada de la consistencia de una teoría lingüística o sus predicciones (Pastor, 2004).

Por último, la tercera línea le da una orientación más tecnológica a la lingüística informática (entendida en el sentido más amplio) y consiste en el diseño y elaboración de sistemas informáticos capaces de trabajar con enunciados orales y escritos en lenguajes naturales. Algunas de las aplicaciones lingüísticas de la informática en que se concreta esta línea son: las herramientas de ayuda a la escritura (diccionarios electrónicos o sistemas de verificación de ortografía); herramientas de ayuda a la traducción; aplicaciones de las tecnologías del habla (programas de dictado y los sistemas de conversión de texto a voz); sistemas de gestión documental (sistemas de extracción de información y los de recuperación de información textual); aplicaciones didácticas para la enseñanza de lenguas; y sistemas de diálogo (como las interfaces de consulta en lenguaje natural a bases de datos y los sistemas automáticos de diálogo por línea telefónica) (Pastor, 2004).

Desde el punto de vista de este trabajo, como se adelantó en la introducción, esta tercera línea de investigación sería la más adecuada, ya que el componente -a nivel de prototipo- que se busca construir está pensado para integrarse en un sistema tutorial inteligente para apoyar la enseñanza de la escritura de noticias en estudiantes de periodismo, es decir, en una aplicación didáctica para la enseñanza de lenguas.

Como se puede observar de lo expuesto hasta aquí, la lingüística computacional posee un carácter marcadamente interdisciplinar, tal como se adelantó desde el inicio de 1.1. Dentro de las disciplinas con las que se relaciona, es interesante para los fines de este trabajo indicar que "desde el punto de vista de su vinculación con la informática, la lingüística computacional se considera una especialidad del campo de la inteligencia artificial, que es la parte de la informática que trabaja en el diseño de máquinas inteligentes" (Pastor, 2004). Dado que esta

tesis busca desarrollar un componente que sea capaz de realizar evaluaciones automáticas (de la coherencia textual y de una estructura particular de las noticias) y que éstas sean equivalentes a las efectuadas por un humano, es indudable que esta vinculación de la lingüística aplicada con la inteligencia artificial es un lugar en que la presente tesis se sitúa con comodidad. En este punto, es necesario abordar la relación de la lingüística computacional con el procesamiento del lenguaje natural.

1.2. Procesamiento del lenguaje natural

Lavid (2005) señala que la lingüística computacional (LC, por sus iniciales) "es un área interdisciplinaria entre la Lingüística y la Informática que se ocupa de la construcción de sistemas informáticos capaces de procesar el lenguaje humano. Esta definición concuerda con la que proporciona Allen (1995) para el área denominada *Procesamiento de Lenguaje Natural"* (PLN, por sus iniciales). El citado autor, Allen (1995), señala que el objetivo del PLN es "crear modelos computacionales del lenguaje lo suficientemente detallados que permitan escribir programas informáticos que realicen las diferentes tareas donde interviene el lenguaje natural". De lo anterior, Lavid (2005) concluye que "el objetivo de la lingüística computacional y del Procesamiento de Lenguaje Natural es el mismo: diseñar programas informáticos que puedan *emular* la capacidad lingüística humana".

Es importante señalar que Lavid en su trabajo sólo afirma que hay una coincidencia en el objetivo entre la lingüística computacional y el Procesamiento de Lenguaje Natural, pero jamás afirma que sean lo mismo.

Diversa es la postura de Villandre (2010) que señala que "este término, traducción del inglés 'Natural Language Processing', alterna con el de lingüística computacional para referirse a la línea de investigación básica dentro del campo de intersección entre el lenguaje y los ordenadores. Es más, en la actualidad lingüística computacional y Procesamiento de Lenguaje Natural se tienden a identificar, por lo que ambos términos se pueden considerar sinónimos".

En el presente trabajo se sigue una postura más cercana a la de Lavid, entendiendo que hay similitudes entre ambas áreas -lingüística computacional y

Procesamiento de Lenguaje Natural- pero no se puede afirmar que sean lo mismo y, por ende, tratar a sus denominaciones como sinónimos. Se reconoce, eso sí, que en la presente investigación se utilizan elementos provenientes del Procesamiento del Lenguaje Natural, por ello se considera necesario abordarlo dentro de esta revisión teórica.

Ahora bien, con el fin de ir precisando conceptos, para poder hablar del procesamiento automático del lenguaje natural, es necesario aclarar qué se entiende por las tecnologías de la lengua. María Antonia Martí (2003) señala que "por tecnologías de la lengua o ingeniería lingüística se entiende los programas que procesan el lenguaje humano con los siguientes objetivos: mejorar la comunicación en todas sus modalidades y facilitar el acceso a la información por encima de las barreras que impone la distancia, el uso de lenguas distintas o el modo en que tiene lugar la comunicación, ya sea hablado o escrito. Estas tecnologías tratan de superar las restricciones que tiene el uso del lenguaje, que hasta hace poco se limitaba a la comunicación directa entre personas, para hacer posible interacciones, servicios y aplicaciones en entornos multilingües, que combinen la comunicación oral y escrita y posibiliten un manejo más eficaz de la información. Se trata, en último término, de aplicar los conocimientos sobre la lengua al desarrollo de sistemas informáticos, con el fin de que puedan reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas".

Ahora, en relación al Procesamiento de Lenguaje Natural (o NLP, según sus iniciales en inglés: Natural Language Processing), el Diccionario Informático Alegsa (2016) lo define como una "rama de la inteligencia artificial que analiza, entiende y genera los lenguajes que los humanos usan naturalmente, para relacionarse con la computadora. Uno de los retos inherentes al procesamiento del lenguaje natural es enseñarle a las computadoras a entender las formas humanas de aprender y usar un lenguaje". Como podemos ver el Procesamiento de Lenguaje Natural involucra por un lado el análisis y la comprensión del lenguaje natural (el procesamiento propiamente tal) y, por otro lado, también comprende la generación automática de lenguaje natural.

Pero ya que se ha hablado hasta aquí de lenguaje natural, hay que precisar qué se entiende por éste. El término *lengua natural* designa una variedad lingüística

o forma de lenguaje humano con fines comunicativos; en otras palabras, es el medio que utilizamos de manera cotidiana para establecer nuestra comunicación con las demás personas. Por ejemplo: inglés, español, alemán, francés, etc. En contraposición a éste, existe el lenguaje formal o artificial, que es aquel que ha sido creado intencionadamente por el hombre con un fin y no ha surgido de forma espontanea en alguna comunidad. Dicho lenguaje -artificial- está formado por símbolos y formulas, y tiene como objetivo fundamental formalizar la programación de computadoras o representar simbólicamente un conocimiento. Por ejemplo: lenguajes de programación (C, C++, C#, Basic, Java, Pearl y Python), lógica matemática o lenguas planificadas como el esperanto.

Ahora bien, a continuación se entregará una breve panorámica del Procesamiento de Lenguaje Natural. Para ello, se partirá con una diferenciación entre las dos áreas que lo comprenden: el procesamiento del lenguaje natural propiamente tal y la generación de lenguaje natural. La distinción se realizará en base a definir este último.

“La generación de lenguaje natural es el proceso de construcción de un texto en lenguaje natural (por medios informáticos) para la comunicación con fines específicos. Si en una actividad de procesamiento de lenguaje natural lo esencial es pasar de la forma del texto a una representación abstracta del mismo, en generación el proceso es inverso: se pasa de una representación abstracta a un texto” (Badia, 2003).

Por lo tanto, teniendo clara la diferencia entre procesamiento y generación de lenguaje natural, es posible pasar a revisar los principales puntos a los que se ha abocado el campo del procesamiento del lenguaje natural propiamente tal, esto es, la disciplina que persigue llegar de un texto a la representación abstracta de éste, *comprensible* por un computador, a través de procedimientos informáticos y lingüísticos.

La parte más desarrollada de este campo es el análisis del lenguaje, pues tras la mayoría de las operaciones que realiza un computador hay alguna actividad de análisis lingüístico. Lo anterior es muy obvio en la traducción automática, por ejemplo, pero también en actividades más sencillas como la corrección ortográfica que realiza un procesador de textos, por ejemplo. Entonces, se puede afirmar que el

análisis del lenguaje es el proceso automático de determinar, aunque sea parcialmente, la estructura lingüística de un texto o fragmento de texto. El objetivo de este proceso es siempre conseguir una representación lingüística que facilite alguna tarea humana o automática posterior. Visto así, el análisis del lenguaje tiene un alcance muy extenso, de manera que comprende la determinación de cualquier clase de estructura en las secuencias lingüísticas (Badía, 2003):

- La estructura morfológica, es decir, la determinación de la estructura interna de las palabras en forma automática. Aquí podemos encontrar aplicaciones informáticas como los *stemmers* o lematizadores.
- La estructura sintáctica, esto es, la determinación de las relaciones estructurales entre las palabras en forma automática. Aquí nos encontramos con aplicaciones como los parseadores o analizadores sintácticos.
- La estructura semántica, que se enfoca en determinar en forma automática la estructura de significado de una secuencia de palabras. Aquí podemos hallar métodos o procedimientos como el Análisis Semántico Latente, que se emplea en el presente trabajo y se explicará en 1.4.

Además, a las ya nombradas, en algunas ocasiones se agrega la estructura textual que consiste en determinar la estructura de organización de la información.

No obstante, hay que tener presente que el proceso de análisis del lenguaje no ha alcanzado el mismo grado de éxito en los resultados que arroja para cada uno de los casos expuestos. Los análisis morfológicos y sintácticos de los textos se consiguen con niveles adecuados de acierto. En relación a los resultados de los análisis semánticos y textuales, han logrado un nivel de éxito inferior a los dos primeros (Badia, 2003).

Para finalizar, es importante señalar que lo afirmado en el párrafo anterior se relaciona directamente con el presente estudio y es uno de los desafíos del mismo, ya que justamente se enfoca en uno de los tipos de análisis que menor grado de éxito ha logrado en sus resultados: el análisis semántico; para el caso de esta tesis, el análisis semántico del texto denominado noticia.

1.3. Sistemas tutoriales inteligentes

Como se dijo en la Introducción, esta tesis busca desarrollar un componente a nivel de prototipo -para un sistema tutorial inteligente (STI)- capaz de evaluar en forma automática la coherencia textual y la correcta construcción de la estructura semántica de pirámide invertida en las noticias. Por ello, si bien no está entre los objetivos del presente trabajo construir un STI, se considera necesario incluir dentro del marco teórico un apartado que dé cuenta de este tipo de aplicaciones tecnológicas orientadas al aprendizaje y enseñanza de lengua.

Un sistema tutorial inteligente (STI, por sus iniciales) “es un programa para la enseñanza-aprendizaje guiado por el computador cuya finalidad última es la facilitación de los procesos de aprendizajes personalizados. Estos sistemas son capaces de comportarse como un experto, tanto en el dominio de conocimiento que enseña como en el dominio pedagógico, donde es capaz de diagnosticar la situación en la que se encuentra el estudiante y, de acuerdo con ello, ofrecer una acción o solución que le permita progresar en el aprendizaje” (Ferreira, 2007).

Venkatesh, Naganathan y Uma Maheswari (2010) agregan que los sistemas tutoriales inteligentes ofrecen los beneficios de la educación cara a cara y personalizada, pero en forma automática. El desafío está en *traspasar* a los computadores la experiencia, habilidades y modo de acción de un tutor humano, más allá de las restricciones del espacio, tiempo y las de naturaleza socioeconómicas. Dado el potencial que presentan, los sistemas tutoriales inteligentes han sido objeto de una creciente cantidad de investigación a través de los años. El concepto de sistemas tutoriales inteligentes es amplio y abarca cualquier programa de computador que contenga algún grado de *inteligencia* y que pueda ser usado para el aprendizaje. Los STI derivan de la enseñanza asistida por computador, también conocida como modelo CAI (por su sigla en inglés, Computer-Aided Instruction). Ferreira, Salcedo, Kotz y Barrientos (2012) agregan que “mientras que en los CAI la unidad atómica de discurso era la pregunta, aquí (en los STI) la unidad básica es la etapa de razonamiento individual”.

Ferreira et al. (2012), siguiendo el esquema de Carbonell (1970) señalan que la arquitectura básica de un sistema tutorial inteligente incluye los módulos del tutor, el módulo del estudiante y el módulo del dominio, además de la interfaz. “El

núcleo de un STI es el módulo del tutor que incluye suficiente conocimiento sobre un área particular para proporcionar respuestas ideales a preguntas y corregir no solo un resultado final sino cada pequeña etapa de razonamiento intermedio. Esto permite mostrar y modelar una forma correcta de resolver un problema". En relación al modelo del estudiante, los citados autores señalan que éste debe ser dinámico. "Para dar soporte a este entrenamiento paso a paso, con frecuencia crean y actualizan un modelo del estudiante, que refleja las reglas correctas que el STI cree que el estudiante conoce –algunas de las que se encuentran en el sistema experto o modelo de estudiante ideal. Cada vez que el estudiante comete un error, el STI diagnostica el problema –posiblemente actualizando el modelo del estudiante– y a continuación intenta remediarlo con un consejo muy detallado acerca de cómo el sistema experto habría operado en esta etapa. Este proceso se repite a cada paso en la evolución hacia la solución completa de un problema" (Ferreira et al., 2012).

Por último, sobre el módulo del dominio, los autores indican que es un componente que a menudo recibe escasa mención. "Este contiene el conocimiento específico del área de enseñanza en cuestión, el que luego es contrastado con el del estudiante para poder estimar la diferencia y actualizar el modelo del alumno y sus creencias. Este modelo que requiere una buena representación formal de cada nodo de conocimiento y de sus enlaces, en diversas disciplinas y de una forma gráfica, es representado por mapas conceptuales o grafos. En los STI se ha preferido la representación de este conocimiento a través de redes semánticas complementadas con marcos, lo que permite un enlace de cada nodo y una descripción de cada uno de ellos" (Ferreira et al., 2012).

Un STI puede variar enormemente de acuerdo al nivel de *inteligencia* de sus componentes. Por ejemplo, un proyecto centrado en la inteligencia en el modelo de dominio puede generar soluciones a problemas complejos y novedosos para que los estudiantes puedan tener siempre nuevos problemas en los que practicar, pero que sólo podría disponer de métodos sencillos para la enseñanza de esos problemas, mientras que un sistema que se concentra en múltiples o nuevas maneras para enseñar un tema en particular puede encontrar suficiente una representación menos sofisticada de ese contenido (Venkatesh et al., 2010).

Graesser, VanLehn, Rosé, Jordan y Harter (2001), por su parte, agregan que los sistemas tutoriales inteligentes son claramente una de las más exitosas empresas de la inteligencia artificial. Señalan que hay una larga lista de STI que han sido testeados en humanos y que han probado facilitar el aprendizaje. Hay tutores, por ejemplo, en álgebra, geometría y lenguajes computacionales, por mencionar algunos.

Entre los sistemas tutoriales inteligentes que se han construido, uno de los que acumula más años de desarrollo es el AutoTutor. Tras de este sistema está el Grupo de Investigación en Tutoría (*Tutoring Research Group*) de la Universidad de Memphis.

El AutoTutor es un tutor computacional que tiene la capacidad de sostener conversaciones con los estudiantes utilizando lenguaje natural. El AutoTutor simula los patrones de discurso de los tutores humanos y un número de estrategias de enseñanza ideales. Presenta al estudiante una serie de preguntas o problemas que debe resolver desde un currículo preestablecido, y se compromete en una iniciativa de diálogo colaborativo con el estudiante, mientras lo ayuda a construir la respuesta. El AutoTutor posee en su interfaz un agente animado, que es el que dialoga con el estudiante a través de un motor de habla. Fue diseñado para ser un compañero de conversación que comprende, habla, puntualiza y muestra emociones. El AutoTutor ha sido desarrollado -y testeado- para asistir en el aprendizaje de temas en física newtoniana, alfabetización computacional, mostrando impresionantes resultados de aprendizaje comparado con las mediciones pretest. Por último, es importante señalar que utiliza como método de representación del conocimiento el Análisis Semántico Latente (Graesser, Penumatsa, Ventura, Cai y Hu, 2007).

Otro sistema tutorial inteligente que se ha desarrollado es el *iSTART*. Esta aplicación también nace al alero de la Universidad de Memphis y, a diferencia del AutoTutor, se enfoca en la enseñanza de lengua, específicamente en la enseñanza de estrategias de lectura. La sigla *iSTART* quiere decir en español Entrenador Interactivo de Estrategias para la Lectura y el Pensamiento Activo. En relación al funcionamiento de *iSTART*, opera como una aplicación para Internet y comienza con una introducción a la autoexplicación y las estrategias de lectura, realizada por tres

agentes autómatas: un agente profesor y dos agentes alumnos. El estudiante humano observa cómo el agente profesor interactúa con los agentes alumnos para enseñarles las estrategias de lectura. Estas estrategias incluyen monitoreo de la comprensión, paráfrasis (reformular el texto en otras palabras), predicción (es decir, predecir lo que dirá el texto a continuación), elaboración (utilizando conocimientos y experiencias previas para comprender el texto) y puenteo (*bridging*, comprender la relación entre las distintas oraciones del texto) (McNamara, 2004; Boonthum, Levinstein y McNamara, 2007).

Al final de cada una de las secciones, "el alumno responde un pequeño cuestionario para evaluar su comprensión de la estrategia. Cada cuestionario incluye cuatro preguntas de selección múltiple que cubren las definiciones básicas de las estrategias y evalúan la habilidad del alumno para elegir explicaciones que las ejemplifican. Después de la introducción, el estudiante avanza a la sección de demostración en la cual dos nuevos agentes, Merlín y Genio, hacen una demostración de las estrategias mientras se autoexplican el texto. El estudiante, de esta forma, identifica las estrategias que se están usando en los ejemplos. En la última sección, el estudiante practica la autoexplicación de textos científicos y recibe retroalimentación de Merlín. Para entregar esta retroalimentación al estudiante, el sistema *iSTART* debe evaluar las autoexplicaciones en una serie de dimensiones. Primero determina si la autoexplicación es muy corta o simplemente una repetición de la oración. Después decide si la autoexplicación es relevante al tópico del texto de la oración, comparándola con una serie de palabras asociadas. Finalmente, evalúa la calidad de la autoexplicación en términos de número de palabras y número de asociaciones (en oposición a palabras tomadas directamente de la oración). De acuerdo con esta evaluación, Merlín solicita al estudiante la acción apropiada (por ejemplo, agregar mayor información) o le brinda retroalimentación (por ejemplo "ya", "muy bien", "excelente"). Merlín también le pregunta qué estrategias usó durante la autoexplicación y, en algunos casos, le pide al estudiante que use otras estrategias si solamente ha empleado paráfrasis o monitoreo de la comprensión" (McNamara, 2004). Al igual que el AutoTutor, el *iSTART* también emplea el Análisis Semántico Latente, en este caso, para evaluar las autoexplicaciones de los estudiantes (Boonthum et al., 2007).

1.3.1. ELE-Tutora: un sistema tutorial inteligente enfocado en la enseñanza del español como lengua extranjera

Para concluir el apartado dedicado a los sistemas tutoriales inteligentes, se presentará un sistema de este tipo: la ELE-Tutora (anteriormente ELE-Tutor) -que constituye el Aula Virtual del Programa Español como Lengua Extranjera de la Universidad de Concepción-. La razón para destacar este STI es que tiene la capacidad para detectar los errores de los estudiantes y apoyarlos en la reflexión necesaria para que sean capaces de corregirlos y no volver a cometerlos. ELE-Tutora, entre sus componentes, cuenta con un analizador sintáctico que permite identificar y clasificar los errores gramaticales en español como lengua extranjera (Ferreira y Kotz, 2010). En este sentido, es importante señalar que el componente que tiene como objetivo desarrollar la presente tesis -y que se construye para insertarse en un STI- lo que realiza es detectar errores en la construcción de la estructura semántica denominada pirámide invertida. Por ello, es importante la revisión de una experiencia previa de un STI que tiene módulos dedicados a la detección de errores.

La ELE-Tutora, que originalmente se llamó ELE-Tutor, es un STI para la enseñanza del español en hablantes no nativos de nuestra lengua. Actualmente está en funcionamiento, integrado en el aula virtual del Programa de Español como Lengua Extranjera de la Universidad de Concepción. La pantalla de acceso a la plataforma se puede visitar desde http://ele.udec.cl/aulavirtual/?lang=es. El acceso al aula es restringido a quienes participen de los cursos que se ofrecen, eso sí.

Su planteamiento inicial fue el modelo de sistema tutorial inteligente propuesto por Ferreira, Moore y Mellish (2007). Uno de sus primeros componentes fue el *parser* o analizador sintáctico desarrollado por Ferreira y Kotz (2010). Posteriormente, se diseñó la arquitectura del STI que se estructura a partir de tres componentes básicos: el módulo del dominio, el módulo del estudiante y el módulo del tutor (Ferreira et al., 2012). En la construcción de la ELE-Tutora se dio especial importancia al módulo del estudiante, el cual es fundamental para que el STI logre adaptar su comportamiento a las necesidades del alumno, ya que es dicho módulo el cual recopila y procesa información sobre cada uno de los estudiantes que interactúan con el sistema (Barrientos, Ferreira y Salcedo, 2012).

Una de las características más relevantes de este STI, que lleva más de diez años de desarrollo y perfeccionamiento, es su capacidad para detectar los errores de los estudiantes y apoyarlos en la reflexión necesaria para que sean capaces de corregirlos y no volver a cometerlos. La base de este logro viene desde las primeras publicaciones sobre el STI. En 2007, Ferreira señala que "uno de los temas más investigados en el área de los sistemas tutoriales inteligentes ha sido la identificación e implementación de estrategias de *feedback* que faciliten el aprendizaje del estudiante. Gran parte de la investigación ha estado orientada al tratamiento de los sistemas de enseñanza de habilidades procedimentales en áreas tales como álgebra, física, programación computacional, etc. Sin embargo, se ha puesto poco énfasis en los estudios e investigaciones sobre este tipo de estrategias en la enseñanza de lenguas". Tomando como base la baja cantidad de investigaciones sobre el tema, la autora realiza un estudio experimental, con el fin de explorar evidencia empírica acerca de la efectividad de las estrategias de *feedback* en estudiantes que interactúan con una aplicación no presencial. Lo anterior, con el fin de proporcionar guías de orientación efectivas para los investigadores que desarrollan sistemas tutoriales inteligentes para lenguas extranjeras o bien sistemas para el aprendizaje de lenguas asistido por computadores inteligentes (del inglés, *Intelligent Computer Assisted Language Learning*, ICALL). La experiencia consistió en comparar tres tipos de estrategias de *feedback* en un ambiente *e-learning*. Los resultados revelaron que las estrategias que elicitan la respuesta son más efectivas para tratar los errores gramaticales (Ferreira, 2007; Barrientos et al., 2012).

Los trabajos de Ferreira (2006 y 2007) sirvieron como base para el generador de *feedback* correctivo de la ELE-Tutora. "Una vez que el sistema ELE-TUTORA identifica un error a través de su analizador automático, debe seleccionar una estrategia de *feedback* correctivo ad-hoc a dicho tipo de error y producirla en español por medio del generador automático (de lenguaje natural). El *feedback* correctivo generado puede ser seguido por diferentes tipos de respuestas por parte de los estudiantes:

- Una respuesta inmediata que contiene el error reparado ya sea por

autoreparación o por reparación del sistema. Esto indica que el estudiante ha notado el error, por ende, la respuesta reparada se constituye en un indicio de mejoramiento en el aprendizaje. Una expresión reformulada por parte del estudiante da indicios de que la correlación entre la forma del estudiante y la forma final se ha notado, ello implica un paso hacia la adquisición.

- Una respuesta que todavía contiene el error. Esto puede ocurrir porque el estudiante no ha notado la forma correcta provista por el sistema o bien porque el estudiante no tiene el conocimiento previo necesario para autocorregir su error. En estos casos, el tutor trata el error sin reparación con una estrategia de *feedback* alternativa (elicitación o clarificación).
- Una respuesta en la cual el estudiante repara el error original, pero su enunciado presenta un nuevo error. En este caso, el tutor selecciona una estrategia metalingüística acorde con el tipo de error gramatical presentado y el nivel de proficiencia del estudiante" (Ferreira et al., 2012).

Para finalizar, es necesario decir que uno de los últimos trabajos del equipo de desarrollo de ELE-Tutora se enfocó precisamente en los errores de los estudiantes, un tema que interesa para los fines del presente estudio. Para ello, realizaron un análisis en el corpus ELE-UdeC, que integran 418 textos de aprendientes de español como lengua extranjera. Estos textos fueron recolectados durante tres intervenciones lingüísticas en los años 2014 y 2015 con el objeto de describir la interlengua de los aprendices e identificar los errores lingüísticos más frecuentes según el nivel de proficiencia y la lengua materna. Participaron 62 sujetos y estuvieron distribuidos en el nivel A2+ con 26 (42%) y B1 con 36 (58%). La lengua materna de los sujetos corresponde al alemán 20 (32%), francés 17 (27%), inglés 17 (27%), portugués 2 (3%), sueco 2 (3%), checo 2 (3%), italiano 1 (2%) y ruso 1 (2%). Según la cohorte, cada sujeto escribió seis textos en 2014 y siete en 2015. La metodología para el procesamiento y análisis de los textos se basó en los procedimientos de la Lingüística de Corpus, específicamente en la línea de investigación de Corpus de Aprendices y los planteamientos del Análisis de Errores Asistido por Computador. Los resultados mostraron un total de 8731 errores

clasificados en cuatro categorías del criterio lingüístico: gramática (categorías gramaticales con 29%), coherencia textual (concordancia sintáctica de género y número con un 18%), léxico (con 13%) y ortografía (acentual, literal, dierética y diacrítica con un 39%). De este total se observa una mayor frecuencia en el uso erróneo de la ortografía con 3448 errores, seguida de las categorías gramaticales con 2552, la coherencia con 1569 y el léxico con 1162 errores (Ferreira, 2015).

1.4. Análisis Semántico Latente

Para Goldman, Golden y van den Broek (2007) el Análisis Semántico Latente (LSA[1], por sus iniciales en inglés, *Latent Semantic Analysis*) "es un enfoque computacional del significado de las palabras que se basa en la coocurrencia de palabras en textos impresos de las cuales se derivan los espacios semánticos que reflejan las relaciones de significado entre las palabras". Boonthum et al. (2007) lo definen como una técnica que utiliza procesos estadísticos para extraer y representar los significados de las palabras. Los significados son representados en términos de su semejanza con otras palabras en un extenso corpus de documentos.

Landauer, Foltz y Laham (1998) definen al Análisis Semántico Latente como una técnica matemático-estadística, totalmente automática, para extraer e inferir relaciones de uso contextual esperado de palabras en pasajes del discurso. No es un procesamiento del lenguaje natural tradicional o programa de inteligencia artificial; no utiliza diccionarios construidos por humanos, ni bases de conocimiento, ni redes semánticas, ni gramáticas, ni analizadores sintácticos ni morfologías o similares; y toma como única entrada el texto segmentado en palabras, definidas como cadenas de caracteres únicos y separadas en pasajes significativos o muestras como frases o párrafos.

Por su parte, McCarthy, Briner, Rus y McNamara (2007) señalan que el Análisis Semántico Latente está basado en la idea de que las palabras (o grupos de palabras) aparecen en algunos contextos pero no en otros.

Landauer (2002), para explicar el funcionamiento del LSA, señala que el significado de un pasaje de texto está contenido en sus palabras, y que todas sus palabras contribuyen a formar su significado. A esto, agrega, que dos pasajes,

1 La sigla se usará de esta forma durante todo el texto de la tesis.

aunque tengan diferentes palabras, podrían tener un significado similar. Esto se puede resumir asumiendo que el significado de un pasaje de texto es igual a la suma del significado de las palabras que lo componen. El autor lo representa con el siguiente esquema:

$$\text{meaning of word}_1 + \text{meaning of word}_2 + \ldots + \text{meaning of word}_n = \text{meaning of passage}$$

Una vez presentada, a modo de introducción, esta breve reseña sobre qué es el Análisis Semántico Latente, es momento de entrar a analizar de un modo más preciso esta técnica matemático-estadístico (o teoría de la representación del conocimiento o enfoque computacional).

1.4.1. El Análisis Semántico Latente: teoría de la adquisición, inducción y representación del conocimiento

Uno de los textos más importantes para el desarrollo del Análisis Semántico Latente fue publicado en 1997 y escrito por Thomas Landauer y Susan Dumais. En éste, los citados autores basándose en trabajos e investigaciones previas, dieron forma a lo que en la actualidad se conoce como Análisis Semántico Latente. El título de la publicación es *"A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge"* ("Una solución al problema de Platón: El Análisis Semántico Latente, teoría de la adquisición, inducción y representación del conocimiento").

Los autores postulan al Análisis Semántico Latente como una nueva teoría general de adquisición y representación del conocimiento. "Esta teoría descansa en la noción de que algunos dominios de conocimiento contienen inmensos números de interrelaciones débiles o latentes, que si son aprovechadas se pueden amplificar produciendo aprendizaje a través de procesos de inferencia. El método de inducción propuesto depende de la reconstrucción de un sistema de relaciones de similitud múltiples en un espacio multidimensional. Se supone que la coocurrencia de eventos, en particular de palabras, en contextos locales se generan y se reflejan por su similitud en algún lugar de este espacio multidimensional" (Venegas, 2003).

En su artículo, Landauer y Dumais, comienzan describiendo el problema de la inducción en la adquisición de conocimiento, esto es, el hecho de que las personas parecen saber mucho más de lo que ellos pudieron haber aprendido en la experiencia que han tenido. Siguiendo lo planteado en el *problema de Platón* (o de la *pobreza del estímulo*), sitúan el foco del asunto concretamente con respecto al aprendizaje de la lengua materna; pero se centran en el aprendizaje de vocabulario en niños en edad escolar ya que estiman que es un ambiente apropiado para realizar las pruebas que requieren.

No se entrará a discutir, los experimentos realizados ni la forma en que el Análisis Semántico Latente se empleó en éstos. La dirección que se tomará es discutir sobre si el Análisis Semántico Latente proporciona una explicación -o solución, como señalan Landauer y Dumais- al denominado *problema de Platón*, un dilema que viene desde, aproximadamente, el 386 a 382 a. de J.C., años entre los que se supone que Platón escribió su diálogo titulado *Menón*.

1.4.2. El Menón

Las fechas de nacimiento y muerte del filósofo griego Platón son levemente imprecisas: 428 o 427 a. de J. C. y 348 o 347 a. de J. C. Pero más allá de la incertidumbre entre un año u otro, lo importante del sabio es su obra. Entre ella encontramos los llamados *Diálogos Socráticos* que incluyen la *Apología de Sócrates, Critón (sobre el deber), Eutifrón (sobre la santidad), Fedón (sobre el alma), Fedro (sobre la belleza), Banquete (sobre el amor) y Menón (sobre la virtud)*.

Para el presente trabajo, el interés se enfoca en *Menón o sobre la virtud*. Dicho diálogo, en el que intervienen como principales interlocutores Menón y Sócrates (además de Anyto y un esclavo de Menón), se inicia con la siguiente pregunta de Menón a Sócrates: “¿Podrías tú decirme, Sócrates, si la virtud se adquiere por instrucción o por el ejercicio, o si, no dependiendo de la instrucción ni del ejercicio, le es dada al hombre por la naturaleza, o de cualquiera otra manera?“ (Platón, 1960). En el texto, como se desprende de la pregunta recién citada, se plantean interrogantes sobre qué es la virtud, si ésta es posible de enseñarse o si se adquiere por la práctica. Dado que no es el tema del presente trabajo el de la virtud, se adelanta que el diálogo finalmente concluye que la virtud no es factible

de enseñarse sino que, tal cual se señala en el texto, “la virtud se presenta como un don divino en aquellos que la poseen. ¿Qué hay de cierto en esto? No lo sabremos con seguridad hasta que, antes de averiguar de dónde le viene al hombre la virtud, no nos decidamos a inquirir lo que la virtud es en sí misma” (Platón, 1960).

Como ya se indicó entre los interlocutores del diálogo se encuentra un joven esclavo de Menón y es, precisamente, el pasaje en donde éste interviene, el que interesa para el foco de este trabajo. En un momento del diálogo Sócrates señala: “Así, el alma, inmortal y renaciendo muchas veces, habiendo contemplado todas las cosas, sobre la tierra y en la morada de Hades, nada hay que no haya aprendido”. De lo anterior se desprende la idea de que el conocimiento no se aprende sólo en la vida presente, sino que se adquiere a través de las sucesivas reencarnaciones del alma. Ante la idea anterior, Menón responde a Sócrates: “Pero ¿te limitarás a decir simplemente, que no aprendemos nada, y que lo que llamamos aprender es reminiscencia?”. Y Sócrates contesta: “Me pides una lección, y acabo de sostener que no se aprende nada y que no se hace sino recordar” (Platón, 1960).

De lo planteado en el párrafo anterior, hay que dejar en claro que el tema de la reencarnación del alma es un punto en el cual no se entrará ni sobre el que se emitirá juicio alguno. Sin embargo, la idea principal de la afirmación de Sócrates es factible de seguir siendo abordada. Esta idea es que el conocimiento no es algo que se pueda aprender -ni enseñar-, sino que está ya en el ser humano.

En este punto del diálogo, Menón le pide a Sócrates que le demuestre su afirmación señalándole que “si tienes algún modo de mostrarme lo que dices, no dejes de hacerlo”. A lo que Sócrates responde: “No es nada fácil, pero me esforzaré, en homenaje a nuestra amistad. Llama a alguno de los muchos servidores que te acompañan, el que quieras, y te haré ver lo que deseas”. Así entra a escena el esclavo en el diálogo entre Menón y Sócrates.

Sócrates lo que hace es demostrar que el esclavo, pese a no tener educación alguna en el tema, conoce principios de Geometría. Sócrates le señala a Menón que no le enseñará nada al joven esclavo, sino que sólo le hará preguntas que sirvan de guía en el proceso de la demostración. Así le formula una serie de interrogantes acerca del tamaño de la figura geométrica conocida como cuadrado y sobre la longitud de las líneas que lo forman. De la lectura del pasaje se puede deducir que

Sócrates utiliza algún tipo de apoyo visual para facilitar la comprensión de los problemas que plantea al esclavo con sus preguntas. Lo sorprendente es que el esclavo, pese a no tener instrucción alguna sobre el tema planteado, es capaz de responder con acierto a las interrogantes que Sócrates le propone, demostrando de algún modo *conocer* algunos de los planteamientos de Pitágoras (Platón, 1960).

En la última oración del anterior párrafo, conocer fue escrito utilizando letra cursiva, por el hecho de que, precisamente, la idea de conocer es la que constituye el punto más relevante -o el foco principal, si se quiere- de lo planteado en el *Menón* para los fines del presente trabajo. La pregunta que surge de lo expuesto es: ¿cómo el esclavo de Menón es capaz de responder a las preguntas de Sócrates sin tener educación sobre el tema acerca del cual fue interrogado (Geometría)? Si tendrá o no respuesta la pregunta, por ahora, quedará en suspenso.

1.4.3. El problema de Platón: la pobreza del estímulo

La formulación del argumento de la *pobreza del estímulo* es atribuida a Noam Chomsky. Dicho planteamiento es generalmente aceptado como la base más sólida de la posición innatista sostenida por los partidarios de la gramática generativa en relación con el lenguaje humano.

El argumento de la *pobreza del estímulo* trata de captar el contraste entre, por una parte, el tipo de datos accesibles por los niños en el proceso de adquisición de su lengua materna y, por otro, la competencia gramatical a que llegan como resultado de ese proceso. El principio básico del argumento es que la articulación de la amplitud y riqueza de los conocimientos sobre el lenguaje de cualquier niño normal contrasta fuertemente con la escasez y la imperfección de los datos de que dispone (Longa y Lorenzo, 2008).

Como se puede colegir de lo expuesto, el argumento de la *pobreza del estímulo* se basa en la idea de que dicho estímulo es insuficiente en relación a los buenos resultados obtenidos en el aprendizaje de la lengua materna por parte del niño. De ahí que al problema de la *pobreza del estímulo* también se le conozca como el *problema de Platón*, aludiendo a lo que este filósofo plantea en *Menón*, en el pasaje en que Sócrates logra demostrar que un esclavo sin instrucción sobre el tema, es capaz de efectuar razonamientos que implicarían que él conozca de

Geometría.

Venegas (2005) señala sobre el punto que la adquisición de la lengua “ha sido un problema ampliamente debatido, sin embargo, la explicación que ha predominado en la ciencia cognitiva es la que sugiere, siguiendo a Chomsky, que la exposición por parte de los niños al lenguaje de los padres no provee evidencia adecuada para explicar el hecho de que sean capaces de producir y entender construcciones sintácticas y léxicas nunca antes oídas [...] Éste es el problema de la *pobreza del estímulo* o *problema de Platón*”.

Neil Smith (2001) también aborda el tema y agrega un pensamiento de Russell sobre el asunto: “El aprendizaje del lenguaje es un ejemplo de la *pobreza del estímulo* cuando terminamos sabiendo más de lo que hemos aprendido. Bertrand Russell lo expresó de una forma que a Chomsky le gusta citar: '¿Cómo es posible que los seres humanos, cuyos contactos con el mundo son breves, personales y limitados, consigan a pesar de todo llegar a saber tanto como saben?'”.

1.4.4. El Análisis Semántico Latente y el problema de Platón

Se señaló en 1.4 que para el Análisis Semántico Latente, el significado de un pasaje de texto es igual a la suma del significado de las palabras que lo componen:

> significado de la palabra 1 + significado de la palabra 2 + ...+ significado de la palabra n = significado del pasaje

A partir de esta forma de representar el significado de las palabras, es posible aceptar que la forma en que éstas son usadas en diferentes pasajes de texto, permite inferir el significado de las palabras y de sus combinaciones. Landauer (2002) da el siguiente ejemplo:

System 1
ecks + wye + aye = foo
ecks + wye + bie = foo

Ecks y *wye* siempre coocurren en el mismo contexto; *aye* y *bie*, por su parte,

no coocurren nunca en contextos similares. Al considerar conjuntamente las dos ecuaciones se obtiene que *aye* y *bie* deben tener el mismo significado, pero no se señala nada en absoluto de la relación entre ambos términos. Así pues, la forma de utilizar los datos empíricos de asociación para aprender el significado de las palabras, es claramente insuficiente. Por lo tanto, no podemos asumir que las palabras tienen significados similares en la medida en que tienden a aparecer juntas en un mismo contexto.

Luego Landauer añade dos nuevas ecuaciones al ejemplo:

System 2
ecks + wye + aye = foo
ecks + wye + bie = foo
ecks + wye + cee = bar
ecks + wye + dee = bar

Con esto el autor señala que *cee* y *dee* son también sinónimos. Finalmente agrega otras dos ecuaciones:

System 3
aye + cee = oof
bie + dee = rab

Por último, Landauer concluye, que para ser consistente con las ecuaciones previamente expuestas, en las cuales *aye=bie* y *cee=dee*, debemos asumir que estas dos últimas ecuaciones tienen el mismo significado (*oof=rab*), a pesar de que no tienen ninguna palabra en común (Landauer, 2002; Venegas, 2003).

Como se puede desprender de las ecuaciones que plantea Landauer, el Análisis Semántico Latente *aprende* (en letra cursiva para que quede claro que una técnica matemático-estadística no tiene la capacidad humana de aprender) el significado de las palabras al analizar las coocurrencias de éstas en extensos corpus textuales.

Lo interesante del enfoque teórico del Análisis Semántico Latente es que “permite pensar en una representación del conocimiento fundada en los textos mismos, es decir, se defiende una postura empírico-inductiva de la adquisición y

representación del conocimiento. Al respecto, Landauer y Dumais (en el citado artículo *A solution to Plato's problem...*, de 1997) sostienen que la propiedad inductiva del aprendizaje por el que las personas adquieren mucho más conocimiento del que parece estar disponible en la experiencia, es un verdadero misterio. Éste, como señala Venegas (2005), es el conocido problema de "la pobreza del estímulo" o "problema de Platón".

Deerwester, Dumais, Furnas, Landauer y Harshman (1990) señalan que existe un poderoso mecanismo en la mente de los niños que puede utilizar la información finita que recibe para transformar al niño en un usuario competente de la lengua materna. En este sentido, los autores que defienden la postura inductiva del aprendizaje plantean que la manera en que la mente resuelve este problema inductivo básico es explicable a partir del mecanismo del Análisis Semántico Latente y su acomodación simultánea de un número muy grande de relaciones de coocurrencia locales en un espacio multidimensional.

El supuesto que subyace al Análisis Semántico Latente es que “las similitudes y diferencias en el significado de las palabras pueden, en gran medida, ser inducidas desde las similitudes y diferencias que ocurren en el contexto del discurso. A su vez, similitudes en el significado de una unidad lingüística mayor a la palabra pueden ser inducidas en gran medida desde la combinación (en sentido matemático) de las palabras que contiene. Este supuesto implica que el determinante del significado verbal que generalmente domina es la elección de palabras y la combinación de las palabras en la expresión; por lo tanto, para muchos propósitos el orden de las palabras en los pasajes puede ser ignorado en la estimación del significado con una mínima pérdida de exactitud. La pregunta respecto de si estos supuestos son correctos ha sido indagada y respondida afirmativamente, por medio de la evaluación de la habilidad de los modelos computacionales basados en estos supuestos para simular un amplio rango de fenómenos verbales humanos” (Venegas, 2005).

Entonces, como ya se señaló, Landauer y Dumais (1997) consideran al Análisis Semántico Latente como una “teoría de la adquisición, inducción y representación del conocimiento”. Es decir, comprendiendo bien y cuidadosamente la afirmación recién citada, para los autores se podría explicar la representación del

conocimiento en el cerebro humano a través de un método matemático-estadístico, que se basa en la coocurrencia de las palabras en grandes corpus textuales, aprovechándose de la noción de que algunos dominios de conocimiento contienen inmensos números de interrelaciones débiles o latentes, que si son aprovechadas se pueden amplificar produciendo aprendizaje a través de procesos de inferencia.

Por supuesto, que la idea del párrafo anterior es absolutamente discutible. De hecho, las críticas al planteamiento no han sido pocas. Como resultado de estas críticas, Thomas Landauer, uno de los autores del artículo *A solution to Plato's problem...* señaló en 2002 que si cualquiera de sus escritos sobre el Análisis Semántico Latente han dado motivos para creer que éste puede ser considerado como una teoría completa del lenguaje y el conocimiento humano, o incluso sobre la semántica léxica, lo lamenta profundamente. El Análisis Semántico Latente es, en efecto, una teoría sobre lenguaje y el conocimiento humano, pero no sobre todo lo relacionado con éstos, señala. El Análisis Semántico Latente es, por supuesto, incompleto como una teoría del lenguaje, o incluso como una teoría de la semántica verbal. No incluye ningún modelo de producción del lenguaje, o de los procesos dinámicos de la comprensión. Ni tampoco trata sobre las convenciones del discurso y la conversación, o con factores pragmáticos en la semántica, agrega. Sin embargo, el Análisis Semántico Latente proporciona una base para una exploración posterior. Da ejemplos de un cálculo eficaz para algunos aspectos importantes del problema y abre caminos que fueron cerrados previamente por suposiciones incorrectas, concluye (Landauer, 2002).

Como cierre a esta discusión de argumentos, se puede decir que más allá de la retractación parcial de Landauer, la posición que se asume en el presente trabajo es que el Análisis Semántico Latente en ningún caso es una teoría capaz de explicar el proceso cognitivo que se da al interior del cerebro humano. Si bien para muchas de las aplicaciones en que se ha utilizado el Análisis Semántico Latente, al comparar los resultados obtenidos por éste con los obtenidos por humanos, se ha comprobado que ambos resultados se correlacionan, esto en ningún caso permite sostener que un método matemático-estadístico como el LSA pueda ser comparable al funcionamiento del cerebro humano. Por lo mismo, es difícil aceptar al Análisis Semántico Latente como una teoría de la adquisición, inducción y representación

del conocimiento; sin embargo, dentro del presente trabajo se reconoce que es un método matemático-estadístico muy poderoso para extraer e inferir relaciones de uso contextual esperado de palabras en pasajes de discurso, es decir, para representar el significado de las palabras.

1.4.5. ¿Constituye el Análisis Semántico Latente una explicación al problema de Platón?

Para intentar responder a la interrogante que encabeza este apartado, es necesario comenzar recordando el planteamiento de Landauer y Dumais (1997) sobre que el Análisis Semántico Latente sería una teoría de la adquisición, inducción y representación del conocimiento; lo anterior, junto a la posterior retractación parcial de Landauer, lleva a tener como primera intuición ante el problema planteado una tendencia a creer que la respuesta a la pregunta sería un no. Sin embargo, es mejor por ahora no tomar posición alguna, ya que es muy diferente considerar al Análisis Semántico Latente como una teoría de la adquisición, inducción y representación del conocimiento, que verlo como una posible explicación al *problema de Platón* o de la *pobreza del estímulo*.

Si se vuelve al *problema de Platón* o de la *pobreza del estímulo,* se tiene que no hay explicación sobre cómo los seres humanos con una información limitada son capaces de aprender tanto; para mayor claridad se trae nuevamente a colación la cita que Smith (2001) realiza de Bertrand Russell: "*¿Cómo es posible que los seres humanos, cuyos contactos con el mundo son breves, personales y limitados, consigan a pesar de todo llegar a saber tanto como saben?"*. Ahora enfoquemos el asunto desde el punto de vista lingüístico en que lo plantea Chomsky (y no de un modo más general como aparece en *Menón).* Esto es, que el aprendizaje de la lengua materna es "un ejemplo de la *pobreza del estímulo* cuando terminamos sabiendo más de lo que hemos aprendido" (Smith, 2001). En este punto es preciso llevar las ideas recién expuestas al campo del Análisis Semántico Latente. Ya se explicó que este método trabaja con una lengua natural; es más, el único *input* que recibe una máquina que emplea la técnica son textos escritos y a ellos aplica los procedimientos matemático-estadísticos con que trabaja. Entonces lo que plantean Landauer y Dumais en su artículo de 1997 *A solution to Plato's Problem...*, se podría

sintetizar de la siguiente manera: el Análisis Semántico Latente sería una *explicación* (se prefiere esta palabra por sobre *solución* en el presente trabajo) al hecho de que los seres humanos con un estímulo lingüístico pobre lleguen a convertirse en hablantes competentes de una lengua en particular. ¿Cómo se explica la afirmación recién hecha? La técnica de Análisis Semántico Latente lo que hace es que, mediante un método matemático-estadístico de análisis de coocurrencias de las palabras, llega a potenciar las relaciones débiles o latentes que hay entre éstas en corpus textuales extensos. Basado en este principio de calcular la similitud semántica, la técnica es capaz de realizar diferentes tareas como la evaluación computacional de la calidad de los resúmenes y ensayos; determinar la autoría de un ensayo a partir de los elementos presentes en el texto; optimización de textos para determinados grupos de lectores, basándose en el conocimiento de los lectores y proyectando las dificultades que tendrán al leer nuevos textos (Landauer y Dumais, 1997; Venegas, 2003).

Siguiendo con lo planteado por Landauer y Dumais, ellos señalan que el cerebro humano realizaría una operación análoga a la que realiza el Análisis Semántico Latente, esto es, que algún mecanismo en el cerebro sería capaz de potenciar las relaciones débiles o latentes y, a partir de estas relaciones, sería capaz de inducir el *conocimiento* para cuyo origen no hay explicación (Landauer y Dumais, 1997).

Ya fueron expuestos los argumentos que dio Landauer en 2002 al retractarse parcialmente de sus dichos. Pero, sin entrar a emitir juicios todavía, y siguiendo lo propuesto por los autores, se asumirá -por ahora- que es cierto lo que afirman. Entonces para explicar un problema como el que plantea Platón en su diálogo *Menón*, que va más allá del ámbito lingüístico (y no tomando en cuenta la posibilidad de encontrar una explicación en las reencarnaciones del alma), es oportuno preguntarse ¿cómo llegan a adquirir los humanos conocimientos que no son únicamente relativos a convertirlos en hablantes competentes de una lengua y que no les han sido enseñados? Y relacionando el punto con el Análisis Semántico Latente podríamos preguntarnos ¿cómo puede un método matemático-estadístico que utiliza como materia prima la lengua natural, ser considerado como una solución (explicación) al *problema de Platón*? En otras palabras, ¿cómo podría el

Análisis Semántico Latente *aprehender* lo que hay en el mundo a través de un *input* puramente lingüístico?

En este punto es necesario recordar la pregunta planteada al inicio de este apartado: ¿es el Análisis Semántico Latente una forma de explicar el llamado Problema de Platón? Para los fines de este trabajo, y sin dilatar más el asunto, la respuesta será un no. No es posible aceptar que un método matemático-estadístico aplicado por una computadora pueda ser considerado, como ya se dijo, como una teoría de la adquisición, inducción y representación del conocimiento en los seres humanos. O sea, no puede ser considerado como un modelo teórico con la capacidad de explicar el procesamiento cognitivo que realiza el cerebro humano. Pretender lo contrario suena un tanto a ciencia ficción. Siguiendo con el punto, tampoco se acepta que sea una explicación (ni respuesta y menos solución) al denominado *problema de Platón* o de la *pobreza del estímulo*. No se considera que el Análisis Semántico Latente, es decir, un método matemático-estadístico que se basa en la coocurrencia de las palabras en grandes corpus textuales, pueda explicar el *misterio* que subyace al *problema de Platón*. Sin embargo, tampoco se puede descartar de plano que el cerebro humano sea capaz realizar una tarea análoga al procesamiento de coocurrencias, con el fin de potenciar interrelaciones débiles o latentes que haya entre la información que almacena, y que las utilice para producir el aprendizaje por medio de procesos de inferencia; pero así como no se puede descartar, tampoco se puede afirmar.

Pese a lo recién expuesto, para los fines de este trabajo, se acepta la utilidad del Análisis Semántico Latente como una técnica poderosa que permite realizar procesamientos lingüísticos a partir de un *input* basado en textos escritos y lograr resultados que se correlacionen positivamente con los de un humano ante tareas similares; con esto último no se quiere desmentir lo señalado en el párrafo anterior, sino que se acepta sólo como una prueba de la efectividad del método para simular algunos procesamientos que realiza el cerebro humano. Visto de esta forma y sin ningún otro afán, el Análisis Semántico Latente es una herramienta con un potencial de desarrollo enorme y, por supuesto, aún sin explotarse por completo. Pero pretender que solucione un dilema como el planteado por Platón en el *Menón* o por Chomsky en relación al aprendizaje de la lengua materna, es ir demasiado lejos con

el entusiasmo por las bondades del método. De hecho, como se señaló más arriba y para cerrar la idea, se sostiene que el Análisis Semántico Latente ni soluciona el *problema de Platón*, ni tampoco se acerca a explicarlo.

1.5. El Análisis Semántico Latente como herramienta para la medición de coherencia textual

En este último apartado del presente capítulo se abordará un tema que se relaciona de un modo directo con el objetivo general de esta tesis, es decir, la evaluación automática de la coherencia textual que es lo que se busca que realice el componente, a nivel de prototipo, que se construirá.

Para esto, es necesario clarificar primero qué se entenderá por coherencia textual en el marco del presente trabajo, sobre todo en su relación con el concepto de cohesión.

Por último, se abordará la evaluación automática de coherencia textual mediante el método del Análisis Semántico Latente que fue explicado en 1.4.

1.5.1. Coherencia y cohesión

Gerardo Álvarez (1995) señala que el *texto* es una configuración lingüística. El mismo autor agrega que este *texto* "resulta, por una parte, de operaciones enunciativas que realiza el locutor y, por otra parte, de operaciones seriales que permiten a este mismo locutor conectar las oraciones individuales para constituir secuencias cohesivas y coherentes".

De lo anterior, se desprende que un texto tiene como requisito que las oraciones que lo forman se constituyan como secuencias cohesivas y coherentes. De inmediato, también, se puede colegir que coherencia y cohesión son términos relacionados, sin embargo, no son lo mismo. La cohesión alude al "aspecto formal, gramatical de las relaciones que existen de una oración a otra en el texto" y la coherencia "designa el aspecto mental, conceptual de la relación que se postula entre los hechos denotados" (Álvarez, 1995). En la definición anterior queda claro que la coherencia es una función cognitiva. Pero esto no es tan claro para todos los autores, como lo señalan De Vega, Díaz y León (1999), quienes si bien sostienen la postura de que la coherencia es una función cognitiva, dan a entender que hay

autores que no adhieren a esta idea. “El discurso no es una colección azarosa de frases, sino que éstas tienen una unidad temática o coherencia. La coherencia es una característica que, generalmente, se ha atribuido al discurso (v. gr., un texto es coherente o incoherente). Sin embargo, los psicolingüistas consideran que la coherencia es más bien una función cognitiva. Dicho de otro modo: la coherencia está en la mente y no en el texto”. De esta forma, el lector de un texto, según los autores, entiende las relaciones de coherencia entre las diversas partes del texto gracias, en gran medida, a sus conocimientos pragmáticos del mundo.

Como ejemplo de lo anterior se presentan estos dos microtextos citados por De Vega et al. (1999):

> (a) Al terminar de cenar en el restaurante con mis padres llamé al camarero y le pedí la cuenta.
> (b) Al terminar de cenar en casa de mis padres llamé al camarero y le pedí la cuenta.

Si bien ambos textos son parecidos en términos gramaticales e incluso en las palabras que utilizan, sólo el primero resulta perfectamente coherente. La coherencia o incoherencia viene determinada, en este caso, por el conocimiento de mundo sobre *los restaurantes* y *las cenas familiares* que comparten tanto el autor como el lector.

Desde otra perspectiva, Jurafsky y Martin (2008), dos autores no provenientes del campo de la psicolingüística, sino que del ámbito de la Lingüística Aplicada, señalan que coherencia y cohesión son términos que a menudo se confunden. Para ellos la cohesión se refiere a la forma en que las unidades textuales son enlazadas. Señalan que una relación cohesiva es como una clase de pegamento que agrupa dos unidades en una sola unidad mayor. Por otro lado, para los citados autores la coherencia alude a las relaciones de significados entre dos unidades. Una relación de coherencia explica cómo el significado de diferentes unidades textuales puede combinarse para construir un significado discursivo mayor.

En la visión de Jurafsky y Martin ya no se aprecia con claridad que la coherencia sea una función cognitiva, como sí queda claro en lo expresado por De Vega et al. (1999). Por lo mismo, es interesante exponer un ejemplo de coherencia

desde la perspectiva de Jurafsky y Martin:

(c) John escondió las llaves del auto de Bill. Él estaba ebrio.
(d) John escondió las llaves del auto de Bill. A él le gustan las espinacas.

Mientras que la mayoría de las personas no encontrarían nada fuera de lo común en la primera oración, la segunda les parecería extraña. ¿Por qué ocurre esto? Si bien ambas oraciones están correctamente construidas, algo parece estar equivocado en las sentencias que se suceden en (d). Entonces, quien se enfrenta a ellas, podría perfectamente cuestionarse cuál es la relación entre esconder las llaves de un auto y que a alguien le gusten las espinacas. Al realizar esta pregunta, lo que se está cuestionado es la coherencia del pasaje.

De esta misma forma, quien lee las oraciones podría intentar construir una explicación que le dé coherencia al pasaje. Por ejemplo, que alguien le ofreció a John espinacas -que le gustan muchísimo- a cambio de esconderle las llaves del auto a Bill. De hecho, si se considera un contexto en el cual esto ya es conocido, el pasaje se percibe como correcto. ¿Por qué ocurre esto? La razón es que esta conjetura permite a quien lee el texto, identificar el gusto de John por las espinacas como la causa de que esconda las llaves del auto de Bill, lo que explica la conexión que habría entre ambas sentencias. El hecho de que los lectores traten de identificar tales conexiones, demuestra la necesidad de establecer coherencia como una parte de la comprensión del discurso.

Esta última idea la reafirman Singer y Zwaam (2003), quienes señalan que para que un texto sea comprensible, debe ser coherente: los lectores deben poder identificar las relaciones entre las ideas del texto. Por su parte, Van Dijk (citado en Álvarez, 1995) agrega: “Un texto es una secuencia de oraciones; pero no cualquier secuencia de oraciones constituye un texto. Para que una secuencia de oraciones constituya *texto*, es decir sea aceptada como un texto *coherente* en una interacción determinada, tiene que cumplir con ciertas normas de buena formación textual [...] Generalmente, las exigencias de la *buena formación textual* se engloban bajo los términos de *cohesión* y *coherencia*”. Por lo tanto, si un texto carece de esta buena formación textual, no tendrá coherencia ni cohesión y, en consecuencia, los lectores

no podrán identificar las relaciones entre las ideas de este texto y les será incomprensible.

Para Cassany (1988) la coherencia es “la propiedad del texto que selecciona la información (relevante/irrelevante) y organiza la estructura comunicativa de una manera determinada (introducción, apartados, conclusiones, etc.). Teun Van Dijk propuso la noción de macroestructura para caracterizar estos dos aspectos. La macroestructura de un texto es la representación abstracta de la estructura global de su significado. Es un tipo de esquema que contiene todas las informaciones del texto y las clasifica según su importancia y sus interrelaciones (gráficamente tiene forma de árbol con corchetes o flechas que se ramifican). Los escritores competentes dominan este tipo de estructuras y las utilizan para construir y organizar el significado del texto”. En lo relativo a la cohesión, el mismo autor señala que “las diferentes frases que componen un texto se conectan entre sí formando una densa red de relaciones. Los mecanismos que se utilizan para conectarlas se denominan formas de cohesión y pueden ser de distintos tipos: repeticiones o anáforas (la aparición recurrente de un mismo elemento en el texto, a través de la sinonimia, la pronominalización o la elipsis), relaciones semánticas entre palabras (antonimia, hiponimia), enlaces o conectores (entonación y puntuación, conjunciones), etc.”. El autor cita como ejemplo el fragmento que se expone a continuación:

> María fue a la tienda de animales y compró un ratón. En casa lo tiñó de color rosa. Por la noche se lo puso al hombro y se presentó en el último bar moderno.

Cassany agrega que en el pasaje recién expuesto se pueden encontrar las siguientes formas de cohesión: repeticiones (María, que es el sujeto de todas las frases, y ratón, que se pronominaliza en dos ocasiones), relaciones semánticas (animal y ratón) y enlaces (la conjunción 'y', signos de puntuación). “Así pues, la cohesión es la propiedad del texto que conecta las diferentes frases entre sí mediante las formas de cohesión. Estos mecanismos tienen la función de asegurar la interpretación de cada frase en relación con las demás y, en definitiva, asegurar la comprensión del significado global del texto. Sin formas de cohesión, el texto

sería una lista inconexa de frases y la comunicación tendría grandes posibilidades de fracasar, puesto que el receptor debería conectar las frases por sí solo, sin ninguna indicación del emisor y con un elevado margen de error" (Cassany, 1988).

Hasta aquí puede notarse que es un poco difícil formarse una idea clara y precisa de qué es la coherencia y de qué es la cohesión. Lo anterior es debido a que los lindes entre un concepto y otro no son claros y varían según los autores. Esta idea también la recoge Cassany (1988) quien señala que hay que tener en cuenta que los conceptos de coherencia y de cohesión varían según los estudios: "Van Dijk clasifica de coherencia muchos puntos que aquí (en su ejemplo de María) tratamos de cohesión". Señala que él sigue la distinción, según la cual la coherencia es de naturaleza principalmente semántica y trata del significado del texto, de las informaciones que contiene, mientras que la cohesión es una propiedad superficial, de carácter básicamente sintáctico que trata de cómo se relacionan las frases entre sí.

Louwerse (2004) también da cuenta de que los dos términos a menudo han sido usados de manera un tanto confusa. Señala que algunos investigadores no usan estos términos, pero distinguen entre diferentes tipos de coherencia. Otros usan cohesión para la estructura de superficie del texto y coherencia para los conceptos y relaciones que subyacen en esta superficie. Sin embargo, otros usan el término coherencia para referirse a una *interrelacionalidad* global en el texto, mientras que reservan cohesión para unidades lingüísticas menores en el texto. Finalmente, otros describen la coherencia como coherencia semántica y cohesión como la manifestación gramatical de la coherencia subyacente.

Con respecto al significado de los conceptos citados, Álvarez (1995) señala que no existe "unanimidad entre los autores respecto a la extensión de ambos conceptos. La cohesión textual designa a las relaciones visibles entre las oraciones en la superficie textual. Pero, para algunos la cohesión concierne específicamente los fenómenos de mantención de los referentes. En cambio, para otros, la noción de cohesión incluye todas las funciones que pueden ser usadas para señalar relaciones entre los elementos que aparecen en la superficie textual. Para algunos, la distinción misma entre cohesión y coherencia es poco clara e incluso inútil y usan sólo uno de los términos". Reafirmando, entonces, lo ya dicho: existe poco consenso

acerca de los significados de cohesión y coherencia. Por supuesto, en este trabajo no se pretende zanjar el problema, sino que simplemente dar cuenta de que existe.

También, cabe señalar que a menudo se suele distinguir entre la *coherencia local*, que establece el lector entre los contenidos próximos (por ejemplo, dos oraciones consecutivas); y la *coherencia global*, que establece entre contenidos muy distantes o distribuidos a lo largo del texto (De Vega et al., 1999).

Para finalizar este apartado, se presenta un último ejemplo sobre coherencia (o cohesión), esta vez original, y que se propone como aporte a este trabajo:

(e) Los primeros ministros se saludaron con un apretón de manos.
(f) Los primeros ministros se saludaron con un beso en la boca.

La secuencia en (e) no presenta al leerla ningún problema aparente. El que dos autoridades de países diferentes se saluden con un apretón de manos al encontrarse, no produce ningún problema para la comprensión de la oración: es perfectamente coherente. En el caso de la secuencia (f) sí que el hipotético lector tendrá más de algún problema cuando llegue al final de ella y se percate de que el saludo de los dignatarios fue con un beso en la boca. En una cultura como la nuestra, un saludo de este tipo sería, posiblemente, un escándalo. Por lo tanto, al enfrentarse al texto, el lector tenderá a encontrarlo mucho menos coherente que la secuencia en (e). De hecho, la más probable explicación en nuestra cultura occidental sea intentar darle coherencia al pasaje suponiendo para ambos ministros la condición de homosexuales, lo que tampoco constituye una buena explicación, pues situaciones como ésa, en nuestra cultura, pertenecen mayoritariamente al ámbito de lo privado, no a un saludo público y protocolar.

Sin embargo, en otras culturas y con otros conocimientos de mundo, la situación descrita en la secuencia (f) podría ser perfectamente normal. De hecho, el beso que se narra entre dos cancilleres es real y ocurrió en 1979 entre el primer ministro Erich Honecker, de la desaparecida República Democrática Alemana, y su homólogo Leónidas Breznev de la también extinta Unión Soviética. Sucedió durante la conmemoración del trigésimo aniversario de la República Democrática Alemana, en junio de 1979. Es más, en Rusia y en otros países, en tiempos actuales sigue

existiendo el beso en la boca como saludo entre personas del mismo sexo, sin ningún tipo de connotación adicional como podría añadírsele en nuestra cultura occidental.

De Vega (1998) al referirse a las insuficiencias de la psicología cognitiva y de la Inteligencia Artificial señala que “en psicología se ha desarrollado un buen número de modelos de memoria semántica, del léxico mental, o de los esquemas. Todos estos modelos reflejan, sobre todo, el conocimiento que tenemos las personas sobre ciertas regularidades recurrentes del medio, que almacenamos como estructuras permanentes en la memoria. En algunos casos también se intenta explicar la génesis y modificación de estas estructuras de conocimiento, a menudo como un proceso de extracción de regularidades estadísticas a partir de ejemplares o de estímulos singulares. Por ejemplo, aprendemos el concepto mesa o el esquema del restaurante extrayendo un patrón promedio a partir de multitud de experiencias con mesas y restaurantes individuales, respectivamente”.

Lo que ocurre en las secuencias presentadas anteriormente se explica por lo que señala De Vega: en nuestra mente almacenamos conceptos y esquemas sobre regularidades recurrentes del medio y las aplicamos con el fin de intentar *dar orden* a las ideas que nos encontramos al leer un texto; este *dar orden*, se debe entender como es obvio, en el sentido de *dar coherencia*, con el fin de que el texto sea coherente y, por ende, comprensible, como señalaron Singer y Zwaam (2003). Sin embargo, lo anterior no explica completamente por qué se percibe como incoherente la secuencia (f), más bien explica por qué percibimos como coherente la secuencia (e), ya que está en nuestros esquemas que dos hombres se saluden con un apretón de manos. El mismo De Vega soluciona el punto cuando señala que “si entiendo un texto en el que se habla de un restaurante es porque activo o instancio el esquema del restaurante. Sin embargo, considero este tipo de análisis profundamente erróneo, pues codificar situaciones es algo más que recuperar información de la memoria semántica. Cuando codificamos una situación representamos una configuración única de parámetros que es, además, dinámica ya que hemos de actualizarla de tiempo en tiempo, a medida que los parámetros situacionales van cambiando”. En el caso del saludo de los primeros ministros con un beso en la boca es evidente que no basta con aplicar el esquema que tenemos

almacenado en la memoria (de seres occidentales y de nacionalidad chilena, para el caso), pues es insuficiente para determinar si la secuencia es coherente o no. De hecho, entender estos parámetros como dinámicos, es decir, que pueden cambiar, es la forma de explicar por qué una oración que según los conocimientos de mundo disponibles en nuestra cultura es incoherente, si vamos más allá de ésta, deja de serlo. Probablemente, visto de esta forma, el asunto de la coherencia sea algo un tanto relativo; sin embargo, no es un objetivo de este trabajo ahondar en dicho punto.

Sin desconocer la discusión anterior y teniendo presente que la conceptualización de coherencia y cohesión no es algo en que haya acuerdo entre los distintos autores, para efectos de la presente investigación, enmarcada en el campo de la Lingüística Computacional, asumiremos la definición de coherencia textual propuesta por Daniel Jurafsky y James Martin en 2008. Como ya se dijo, para ellos la coherencia alude a las relaciones de significados entre dos unidades. Una relación de coherencia explica cómo el significado de diferentes unidades textuales puede combinarse para construir un significado discursivo mayor. La anterior definición se ajusta perfectamente a la técnica de Análisis Semántico Latente que se empleará en el presente trabajo (la citada técnica se explicará en 1.4).

1.5.2. La medición de coherencia textual mediante Análisis Semántico Latente

En 1998 Peter Foltz, Walter Kintsch y Thomas Landauer publicaron “The measurement of textual coherence with Latent Semantic Analysis” (en español, “La medición de coherencia textual con *Análisis Semántico Latente”*) (Foltz et al., 1998). En dicho artículo señalan que el método primario para usar *LSA* con el fin de realizar predicciones de coherencia, consiste en comparar una unidad de texto con otra unidad de texto vecina, esto con el objetivo de determinar en qué grado están relacionados semánticamente ambos segmentos. Estas unidades pueden ser oraciones, párrafos o, incluso, palabras aisladas o libros completos. Este análisis debe ser hecho a todos los pares de unidades de texto vecinas, con el fin de caracterizar la coherencia de todo el texto en su conjunto. Agregan que el uso de oraciones como la unidad básica de texto es un tipo de segmentación apropiada.

Esto porque representan una pequeña cantidad de información textual (en promedio de tres a siete proposiciones) y es, relativamente, consistente con la cantidad de información que el cerebro retiene en la memoria de corto plazo. Además, el utilizar la oración como unidad básica, facilita la segmentación del texto mediante métodos automáticos (Foltz et al., 1998).

Los mencionados autores señalan que el poder del Análisis Semántico Latente para determinar el grado de relación semántica proviene del análisis de un extenso número de textos de ejemplo. Entonces, para realizar predicciones de coherencia de un texto en particular, primero es necesario tener un conjunto de textos similares al texto a analizar; esto es, que contengan una gran proporción de los términos usados en dicho texto y que éstos aparezcan en distintos contextos, con el fin de tener un marco que permita su comparación con el texto específico cuya coherencia queremos predecir por medio del *LSA*.

El Análisis Semántico Latente representa cada palabra del texto procesado -así como unidades textuales mayores- en el espacio semántico a través de vectores. Las unidades textuales mayores -como una oración- son representadas en este espacio semántico multidimensional como el promedio ponderado de los vectores de los términos que contiene. La relación semántica entre dos unidades de texto puede ser comparada determinando el coseno entre los vectores de ambas unidades. Por lo tanto, para evaluar la coherencia entre la primera y la segunda oración de un texto, el coseno entre los vectores de las dos oraciones debe ser determinado. Por ejemplo, dos oraciones que utilicen exactamente los mismos términos y con la misma frecuencia, tendrán un coseno de 1; mientras que dos oraciones que usen términos que no sean semánticamente relacionados, tenderán a tener cosenos cercanos a 0, o menores aún. En los niveles intermedios, las oraciones que contienen términos de significado relacionado, incluso si no son los mismos términos o raíces, tendrán cosenos que se acercarán a 0 ó 1, según el grado de relación semántica que tengan (Foltz et al., 1998).

Con el fin de probar la efectividad del Análisis Semántico Latente como medio para evaluar la coherencia de un texto, los citados autores compararon el rendimiento del LSA con dos estudios previos sobre coherencia textual. Aquí se expondrá brevemente la experiencia con el segundo de ellos.

Este corresponde al realizado por McNamara, D. Kintsch, E., Butler Songer, N., & Kintsch, W. en 1996 y que se titula: “Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text”. Su objetivo era examinar cómo los conocimientos previos de los lectores interactúan con la coherencia de un texto. Ellos utilizaron un artículo extraído desde una enciclopedia de ciencias sobre las enfermedades al corazón y agregaron o suprimieron información con el fin de variar la coherencia local y global. Estos cambios en el texto arrojaron cuatro versiones del mismo: una versión con una máxima coherencia (CM), una versión con una alta coherencia local, pero una baja coherencia global (Cm), una versión con una baja coherencia local, pero una alta coherencia global (cM) y una versión con una baja coherencia local y global (cm). La experiencia arrojó como resultado que los lectores con un menor grado de conocimientos previos sobre el tema del artículo, lograban su más alto beneficio del texto con una máxima coherencia (CM); mientras que los lectores con un mayor grado de conocimientos previos sobre el tema, lograban altos beneficios incluso del texto con baja coherencia local y global (cm). Debido a que los lectores con un menor grado de conocimientos previos fueron los más afectados por los efectos de incrementar o disminuir la coherencia, sus resultados de comprensión fueron utilizados para compararlos con las predicciones del Análisis Semántico Latente.

McNamara et al. (1996) encontraron una interacción en las preguntas posteriores a la experiencia (*posstest questions*), entre los textos con máxima y mínima coherencia y el nivel de conocimiento de los sujetos. Los sujetos con un menor grado de conocimientos previos sobre el tema mostraron con mayor fuerza el efecto de qué texto habían leído (esto es, si leyeron un texto con mayor o menor grado de coherencia). Por lo tanto, Foltz et al. (1998), compararon los resultados posteriores a la experiencia de los sujetos con bajo conocimiento previo con las predicciones del LSA.

Esta comparación arrojó que la medición de coherencia del Análisis Semántico Latente se correlacionó altamente con el promedio de los resultados de las pruebas posteriores a la experiencia. Dada la preparación que habían efectuado McNamara y el resto de los investigadores en los textos usados al manipular su coherencia, el experimento logró una prueba rigurosa e interesante de la eficacia de

la medición de coherencia mediante la técnica del LSA, pues exigía detectar similitudes de significado subyacente, en ausencia de una repetición literal de palabras que facilitara la tarea. Y, por supuesto, los resultados concuerdan con la idea de que un texto altamente coherente debería ser más útil para la construcción de un texto base bien vinculado en los lectores con bajo conocimiento previo sobre un tema en particular (Foltz et al., 1998).

De lo hasta aquí expuesto sobre la medición de coherencia mediante Análisis Semántico Latente, se resalta que es necesaria la presencia de un corpus de entrenamiento para la construcción del espacio semántico, a la luz del cual se analizarán los textos. Este corpus debe tener dos características fundamentales: debe contener una gran cantidad de textos y estos textos deben pertenecer a un dominio temático definido. De sobra está decir que los textos a analizar deberán pertenecer al mismo dominio temático del que provienen los que integran el corpus de entrenamiento.

Capítulo 2: La noticia: su estructura y su producción escrita

En el presente capítulo se presentará el concepto de noticia, la estructura de este tipo de texto y su producción escrita. Se señala en este punto que es posible, aunque pudiera ser discutible, el considerar a la noticia como un tipo particular de resumen. Sobre esto último, es especialmente interesante tener en cuenta lo que se presentará al final de este capítulo: si se le considera como un tipo particular de resumen, es posible explicar algunas competencias que deben desarrollar lo estudiantes de periodismo para aprender a producirla: la capacidad de poder condensar una situación o hecho en sus líneas fundamentales, sin perder ningún dato que sea esencial para la correcta comprensión del acontecimiento descrito.

El primer paso es conceptualizar a la noticia no desde el punto de vista periodístico, sino desde el punto de vista lingüístico, como discurso especializado.

2.1. Discurso especializado, discurso profesional y discurso académico

Parodi (2007) señala que “el uso del término discurso especializado se encuentra, en la actualidad, ampliamente aceptado por los estudiosos del lenguaje; no obstante, se debe reconocer que su utilización no surgió sino solo hace unos pocos años”. El autor agrega que este concepto no ha sido empleado siempre con el mismo sentido ya que en “un comienzo vino, simplemente, a reemplazar a la noción de lenguas con propósitos específicos. En la actualidad, en cambio, la noción de DE (discurso especializado) se enfrenta, por un lado, de modo más específico y, por otro, también se concibe desde una mirada más amplia y globalizadora” (Parodi, 2005 y 2007).

¿Pero cómo se relaciona el discurso especializado con el discurso académico y el profesional. El mismo Parodi (2005 y 2007) propone aproximarse a la noción de discurso especializado, reconociendo la heterogeneidad de los textos que pueden incluirse en el concepto (textos -aclara- con ciertos rasgos muy prototípicos, eso sí), planteando una escala de gradación. De esta forma, el discurso especializado “debe necesariamente entenderse como un *continuum*, en el que se alinean textos a lo largo de una gradiente diversificada que va desde un alto hasta un bajo grado de

especialización. Debido a esto mismo, es que concebimos al DE (discurso especializado) como un hiperónimo de las nociones de DA (discurso académico) y DP (discurso profesional)" (Parodi, 2007).

En 2009, Parodi refinó aún más esta idea del *continuum* para comprender la relación entre los conceptos de discurso especializado, académico y profesional. "El lenguaje escrito es el medio preferente mediante el cual se crea, fija y transmite el conocimiento disciplinar; específicamente, a través de aquellos géneros prototípicos que andamian la construcción inicial de saberes especializados y que, gradualmente, van cimentando la integración a una comunidad discursiva particular. Desde este contexto, en mi opinión, los géneros académicos y profesionales se operacionalizan a través de un conjunto de textos que se organizan a través de un *continuum* en el que se van concatenando desde los textos escolares generales hacia los académicos universitarios y los profesionales. Esta progresión se concibe desde una persona en formación académica, a través de la cual se debe ir paulatinamente enfrentando escenarios y géneros diversos".

En el contexto del presente estudio, se estima muy interesante y motivadora la propuesta del Dr. Parodi, ya que plantea el discurso especializado como un continente que incluye al discurso académico y profesional en un *continuum*. En esta investigación se trabaja con el discurso periodístico, específicamente con la producción escrita de noticias, que sin duda es un tipo de discurso especializado, pero es un caso particular, ya que es especializado en su producción pero no así en su comprensión. Lo anterior se refiere a que no cualquiera puede escribir una noticia, sino que debe pasar por un proceso de formación académica para lograrlo, esto es, estudiar periodismo en una universidad; sin embargo, este texto que es especializado en su producción, no lo es en su comprensión, ya que la noticia escrita no se dirige a una comunidad discursiva específica, sino que es un texto que es leído cotidianamente por cualquier persona, con el único requisito de que sepa leer en la lengua en que está escrita la noticia. Por lo anterior, se estima, que la propuesta de Parodi (2005, 2007 y 2009) se ajusta de muy buena forma a los fines del presente estudio que trabaja con la producción escrita de noticias, un texto que debe aprender a escribirse en un contexto académico, por medio de la formación de pregrado (carreras de periodismo en nuestro país, en otros es parte del postgrado),

para una vez dominado, producirlo en un ámbito profesional y dirigido a un público general y no a una comunidad discursiva específica (como serían quienes se interesan en la lectura de artículos científicos sobre un tema en particular).

En la Figura 1 se presenta gráficamente la conceptualización planteada por Parodi. Según el citado autor, la imagen muestra cómo se van concatenando en este *continuum* "desde el discurso escolar general, hacia el académico universitario y el profesional en el medio laboral" (Parodi, 2007).

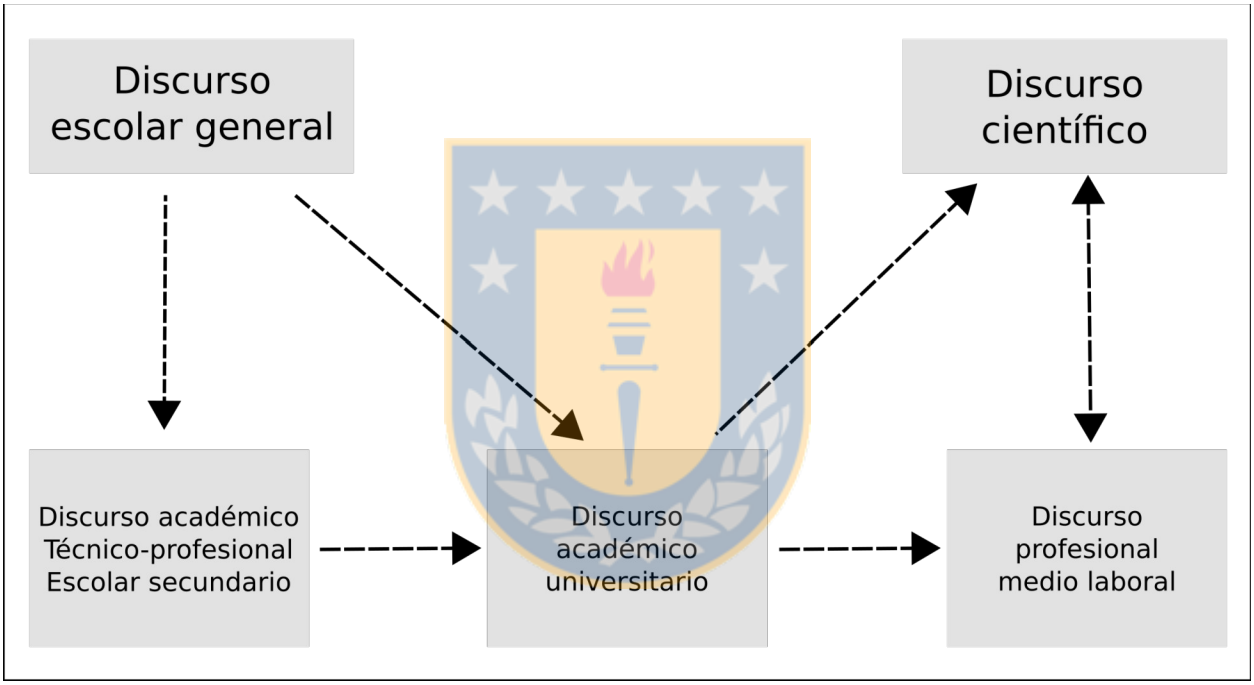

Figura 1: *Continuum* de textos en ámbitos académicos y profesionales (Parodi, 2007).

A continuación se analizará en profundidad la estructura de la noticia, este particular tipo de discurso que es especializado en su producción, ya que al ser el texto con que se trabajará en este estudio, requiere un análisis más completo y detallado.

2.2. La noticia

En periodismo, por convención, se distinguen tres géneros básicos:

Periodismo Informativo, Periodismo Interpretativo y Periodismo de Opinión. Así puede ser hallado en innumerables manuales sobre la materia, pudiendo haber pequeños matices de diferencia según los distintos autores. Desde ya hay que aclarar que los lindes entre estos géneros no son taxativos, sino que más bien tienden a ser difusos, pues cumplen una finalidad similar a la de los géneros literarios: servir de orientación, además de ser un principio de clasificación. A esto hay que agregar el fin pedagógico que también cumplen, como forma de facilitar el manejo de las estructuras del periodismo a quienes están iniciándose en dicha labor.

En el presente trabajo, nos enfocaremos en la *noticia* o *información periodística*, tipo de texto perteneciente al género llamado Periodismo Informativo.

En una primera definición, más bien genérica, Martínez Albertos (2004) concibe a la noticia como "un hecho verdadero, inédito o actual, de interés general, que se comunica a un público que puede considerarse masivo, una vez que (este hecho) ha sido recogido, interpretado y valorado por los sujetos promotores que controlan el medio utilizado para la difusión".

Otra definición de noticia señala que es "el género periodístico por excelencia que da cuenta de un modo sucinto pero completo, de un hecho actual o actualizado, digno de ser conocido y divulgado, y de innegable repercusión humana" (Martín Vivaldi, 1993).

Por su parte, Charaudeau (2003), define noticia como "un conjunto de informaciones que se remiten a un mismo espacio temático, que provienen de una determinada fuente, que tienen un carácter de novedad y que pueden ser tratadas de diversas maneras".

Finalmente, y en una visión que completa el panorama, Martínez Albertos (2004) afirma que la información es el género "más escueto, más descarnado, más fuertemente ceñido al puro esqueleto del hecho o acontecimiento que se quiere transmitir. Es, diríamos, el género periodístico más rigurosamente objetivo en su propósito teórico y desde el punto de vista de la apariencia formal del lenguaje utilizado por el periodista reportero".

Esta *noticia* o *información periodística* consta normalmente de tres partes perfectamente diferenciadas: el titular, el *lead* o párrafo inicial, y el cuerpo de la

información. Estas partes se organizan de acuerdo a la estructura de pirámide invertida. Ésta se puede ver en el esquema siguiente:

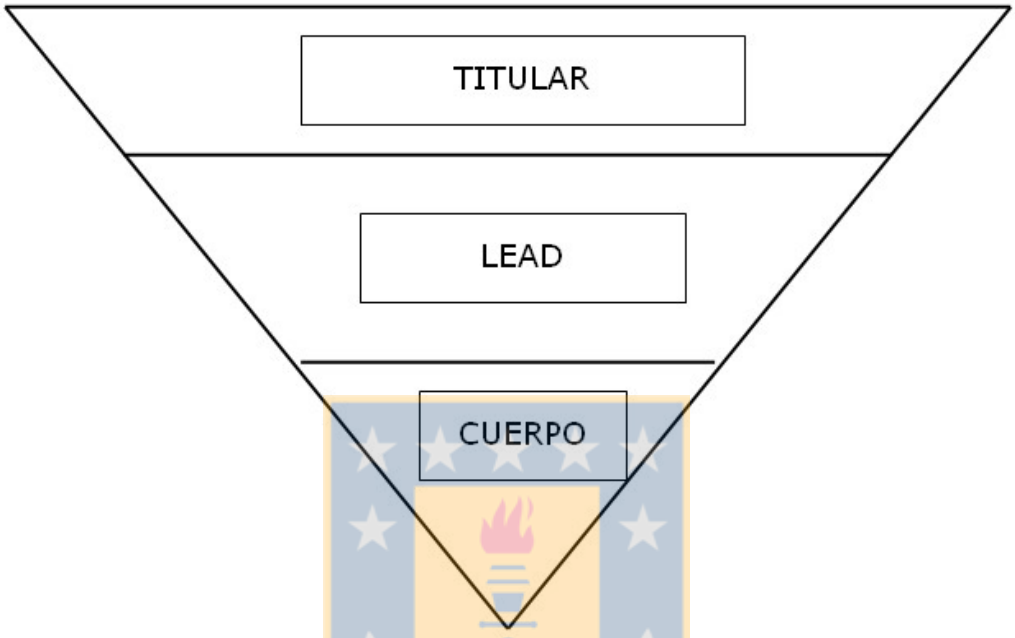

Figura 2: Pirámide invertida (Fundeu, 2015).

Esta *pirámide invertida* es una estructura en la cual el clímax de la narración se sitúa al comienzo. Desde el primer momento se dice al lector todo lo verdaderamente importante del acontecimiento, de forma que lo accesorio de la información vaya ubicándose en los párrafos posteriores, en otras palabras, esta estructura consiste en que entre el titular y el *lead* del texto (sus dos primeras partes) debe entregarse la información más relevante de la noticia al lector. Esta pirámide invertida se contrapone a la estructura de pirámide tradicional, utilizada por la ficción y el drama, en que la tensión del relato se dispone de manera creciente a medida que éste avanza (ver Figura 3).

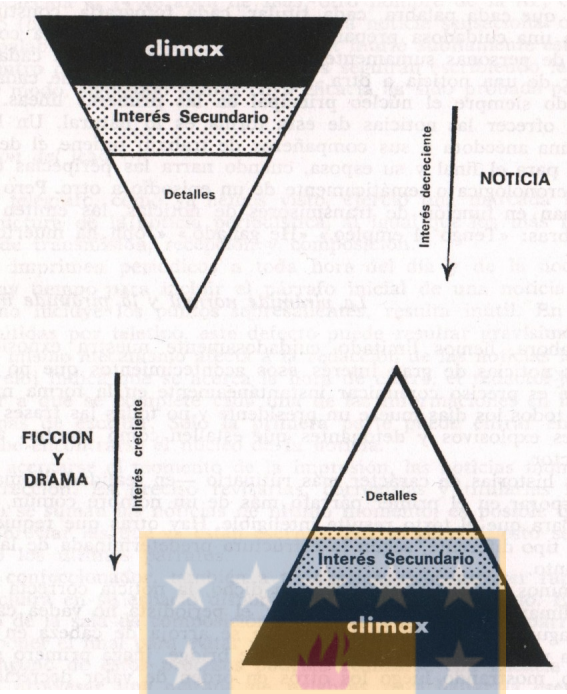

Figura 3: La pirámide invertida y la pirámide tradicional.

El origen de la *pirámide invertida* está en la práctica periodística de Estados Unidos, concretamente en la agencia *Associated Press*, como consecuencia de la Guerra de Secesión (1861-1865). “Hasta entonces los periodistas norteamericanos utilizaban para sus despachos el relato estructurado sobre el orden cronológico de los hechos, tal como ellos habían aprendido de los periodistas británicos. Pero durante la Guerra Civil no podían confiarse demasiado del telégrafo, dada la incertidumbre de los acontecimientos bélicos. Y en cuanto se hacían con uno de estos aparatos de transmisión, los reporteros de guerra empezaban haciendo un breve sumario de los acontecimientos -quién, qué, cuándo, dónde, por qué- antes de arriesgarse a telegrafiar una versión detallada. Si la conexión se mantenía en línea, el periodista podía explayarse a sus anchas. Pero en aquellas circunstancias nadie podía asegurar la continuidad del enlace telegráfico entre el frente de batalla y las redacciones de la Costa Este [...] De allí en adelante, la agencia *Associated Press*, introdujo a la *pirámide invertida* como norma obligada de estilo para todos sus reporteros y, desde este núcleo original, se expandió a todo el mundo occidental” (Martínez Albertos, 2004).

2.2.1. El titular

El titular se define como “cada uno de los títulos de una revista, de un periódico, etc., compuesto en tipos de mayor tamaño” (RAE, 2014). El titular en periodismo debe atraer la atención del lector y reflejar con claridad el contenido de la noticia. Además, debe prescindir del punto seguido y, por lo general, se recomienda que incluya un verbo (en Periodismo Informativo debe incluirlo).

Otros elementos que pueden acompañar al titular, aunque son optativos, son el epígrafe o antetítulo y la bajada.

La Real Academia Española (2014) define al epígrafe como una “cita o sentencia que suele ponerse a la cabeza de una obra científica o literaria o de cada uno de sus capítulos o divisiones de otra clase”. Si bien no contempla a la noticia en su definición, ésta le es plenamente aplicable a ella, pues el epígrafe es una cita o sentencia que se ubica antes del título, en una tipografía de menor tamaño que éste y cuyo fin es ampliar la idea del titular o contextualizarla. El Mercurio, en su sitio destinado a difundir el periodismo entre los estudiantes, señala que “el epígrafe es una palabra o frase que va sobre el título y contribuye a contextualizarlo, entregando datos que ayudan a enmarcar la información” (El Mercurio de los estudiantes, 2011).

La bajada se ubica después del título, generalmente en cursiva, y su función es destacar otras ideas o conceptos relacionados a la noticia que no aparecen expresados en el epígrafe o el titular. “La bajada es una oración que va debajo del título y que entrega una información distinta a éste. Su función también es llamar la atención del lector para incitarlo a seguir leyendo la noticia, entregando datos, ideas o conceptos relevantes o novedosos en relación a la información que se está entregando” (El Mercurio de los estudiantes, 2011).

2.2.2. El *lead*

Si bien en español debiera traducirse *lead* por arranque, entrada o comienzo, la fuerte influencia de la tradición periodística norteamericana ha incidido en que, en la mayor parte de la literatura sobre el tema, se mantenga el anglicismo. Hay que señalar eso sí, que la Real Academia Española no reconoce dicha forma, sino que recomienda utilizar en su lugar la palabra *entradilla*. Para efectos de este

trabajo, seguiremos a la mayoría de los autores y utilizaremos *lead*.

La función esencial de este *lead* es atraer la atención del lector y “busca condensar sinópticamente toda la noticia en aquellos datos esenciales para una cabal comprensión de la misma. En el *lead* informativo se destaca la esencia o los datos más sobresalientes del acontecimiento del que se quiere dar noticia. Este *lead* se conoce también con el nombre de *lead de sumario* y su técnica de realización aparece vinculada a la rutina profesional del quehacer periodístico que se designa con el nombre de la fórmula de las 5 W's” (Warren, 1975).

La mencionada fórmula de las 5 W's consiste en las cinco preguntas que hay que intentar responder para la correcta estructuración de un *lead* informativo. Su nombre deriva del grafema inicial de los siguientes pronombres y adverbios del idioma inglés:

Who?	----------	¿Quién?
What?	----------	¿Qué?
When?	----------	¿Cuándo?
Where?	----------	¿Dónde?
Why?	----------	¿Por qué?

Gonzalo Martín Vivaldi (2004) agrega un sexto elemento, que corresponde a la pregunta *¿cómo?*, aunque señala que en algunas ocasiones éste ya está contenido en el *¿qué?*, y , otras veces, la respuesta al *¿por qué?,* lleva consigo una respuesta al *¿cómo?*

Desglosando el objetivo de la respuesta a cada una de estas preguntas puede resumirse como sigue:

a) *¿Quién?* Sujeto de la información.
b) ¿Qué? El hecho, lo que ha sucedido.
c) ¿Cuándo? Factor tiempo (año, día, hora o minuto. La precisión de la fecha depende del hecho).
d) ¿Dónde? El sitio, el lugar en que se produjo el acontecimiento.
e) ¿Por qué? La causa, elemento fundamental que nos da la razón de lo que ha pasado.

f) ¿Cómo? El método, la manera de producirse el hecho.

Según el hecho en que se base una noticia en particular, podría ser más relevante dar respuesta a unas que a otras de estas preguntas. Eso dependerá de cada caso. Por ejemplo, si el texto trata sobre la muerte de un personaje conocido, tal vez las preguntas más importantes de responder sean todas: quién, qué, cuándo, dónde, porqué y cómo. Sin embargo, en el caso de que la noticia sea sobre el estreno de una película, tal vez las preguntas más relevantes sean qué, cuándo y dónde, quedando quién, por qué y cómo en un segundo plano.

2.2.3. El cuerpo de la información

Está constituido por el resto del escrito una vez dejado aparte el titular y el *lead del sumario*. El cuerpo de la información, que constituye el vértice de la pirámide invertida (que apunta hacia abajo), se debiera disponer también en orden de importancia decreciente en función de los elementos básicos que dan significación y relieve a la noticia, tal como aparece diseñada en sus líneas maestras por el *lead* (Martínez Albertos, 2004). Hay que señalar eso sí, que no siempre este orden decreciente de los datos en el cuerpo de la noticia es demasiado preciso, por lo que muchas veces no es fácil apreciar la jerarquización en esta parte del texto, con la misma claridad que puede realizarse entre sus partes fundamentales como un todo: titular, *lead* y el cuerpo como conjunto.

2.3. Estructura de la noticia como discurso

En 2.1 se presentaron las nociones de discurso especializado, discurso académico y discurso profesional. Siguiendo a Parodi (2005, 2007 y 2009) se señaló que su planteamiento del discurso especializado como un continente que incluye al discurso académico y profesional en un *continuum*, se ajusta muy bien al presente estudio. En este sentido, la noticia se puede considerar como discurso especializado en su producción, pero no en su comprensión. Lo anterior, debido a que ésta es un texto dirigido a un público general y no a una comunidad discursiva determinada (como un artículo científico); sin embargo, para su producción -como texto escrito en el caso de este trabajo- se requiere de una formación académica específica que

posibilita que un estudiante pueda aprender a producir este tipo de texto y convertirse en periodista, que es el profesional que se dedica a su escritura.

En este punto, teniendo claridad del ámbito en que se ubica la noticia desde el punto de vista lingüístico y, además, sabiendo cómo se conceptualiza ésta desde los estudiosos del periodismo, es momento de presentar un análisis de la noticia como texto (o discurso) periodístico propiamente tal, pero analizadas sus estructuras desde el ámbito lingüístico, es decir, desde el punto de vista de un lingüista. Para ello, se recurrirá, esencialmente, al trabajo "La noticia como discurso" de Teun Van Dijk. Hay que señalar desde ya que este autor realiza su análisis enfocado, principalmente, en las noticias aparecidas en medios escritos, por lo que se ajusta de excelente forma a los objetivos del presente trabajo. Así, señala el holandés sobre su texto, que "la mayoría de las veces, nos referiremos a la noticia en la prensa, es decir, el discurso o los artículos periodísticos publicados cotidianamente en los diarios" (Van Dijk, 1990).

Una de las ideas más importantes que señala Van Dijk, para los fines de esta tesis, es que "los esquemas periodísticos realmente existen, y tanto los periodistas como los lectores los utilizan al menos implícitamente en la producción y la comprensión de la noticia".

El autor holandés señala que una de las características más llamativas y típicas de la realización o elaboración temática del discurso periodístico es su carácter troceado. Esto es, cada tema se presenta en partes, no como un todo, como es el caso de otros tipos de discurso. "Esta característica estructural tiene su origen en el principio global de la organización de la relevancia en la noticia. Este principio sostiene que el discurso periodístico se organiza de manera tal que la información más importante o relevante se pone en la posición más destacada, tanto en el texto tomado como un todo, como en las oraciones. Esto significa que para cada tema, la información más importante se presenta primero. Cuando la información importante de otros temas ya se ha presentado, los temas anteriores se reintroducen con detalles de menor nivel. De esta manera, en lugar de una realización izquierda-derecha de los temas a partir de una estructura temática, tiene lugar una realización arriba-abajo, si es que esta organización arriba-abajo de lo general a lo particular también coincide con la dimensión importante-menos

importante” (Van Dijk, 1990).

En la cita anterior, aunque no la llame por su nombre, Van Dijk deja en claro que la estructura de *pirámide invertida* es una realidad efectiva en el discurso periodístico, que ordena la información expuesta en un orden decreciente de importancia.

Esta característica estructural de la noticia es también el resultado de una estrategia de producción, como señala el autor, que considera las limitaciones de la relevancia y las estrategias de lectura posibles, de modo que los lectores obtengan primero la información más importante. “La lectura parcial, en ese caso, no provocará una comprensión parcial sino sólo la pérdida de algunos detalles de nivel más bajo. Por último, la producción de la noticia tradicional tiene limitaciones en cuanto al tamaño. La organización global permite a los editores cortar los párrafos finales de un relato periodístico sin perder la información esencial” (Van Dijk, 1990). Relacionado con este último punto, Martínez Albertos (2004), señala que una de las razones de la utilidad de la *pirámide invertida* es que, si la noticia está correctamente escrita de acuerdo a ella, permite una segura manipulación posterior del texto. Cuando por exigencias de espacio hay que acortar el texto, “se pueden ir tirando tranquilamente los párrafos situados al final del relato con la certeza de que son los menos interesantes del escrito”.

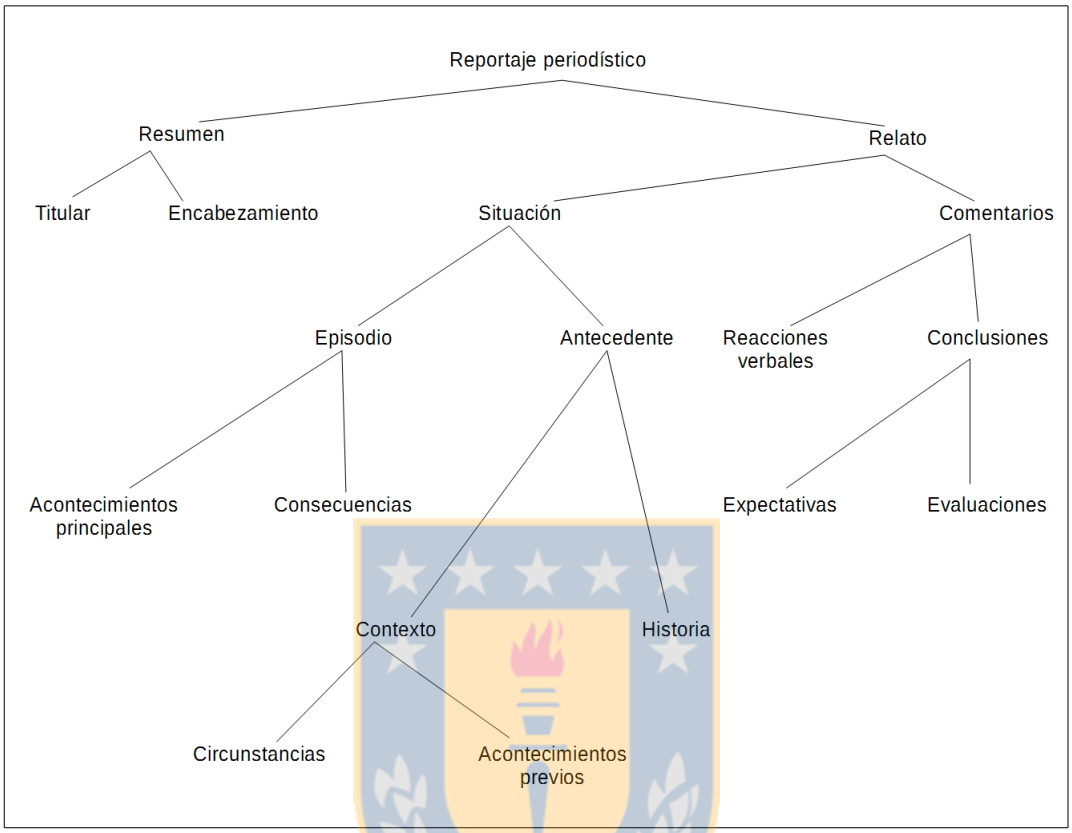

Figura 4: Estructura hipotética de un esquema informativo (Van Dijk).

En la figura anterior, se presenta un esquema con todas las categorías que Van Dijk identificó en el discurso periodístico. Aunque el mismo autor señala una limitación. “Si hablamos con propiedad, sólo el titular y los sucesos principales deben hallarse obligatoriamente en un discurso periodístico mínimamente bien construido; categorías como antecedentes, reacciones verbales y comentarios son opcionales”.

2.3.1. El titular y el encabezamiento

El *titular* desempeña la función de resumir los temas del discurso periodístico; es decir, su sola lectura ubica al lector de inmediato sobre cuál es el tema esencial del texto. “El titular precede al encabezamiento y juntos preceden al resto del ítem informativo. Su función estructural también es clara: juntos expresan los principales temas del hecho. Es decir funcionan como un resumen inicial”.

Agrega, Van Dijk, que “los encabezamientos pueden expresarse separadamente y en negrita o pueden coincidir con la primera oración temática del texto”. Es indudable, pese a que no lo expresa con claridad Van Dijk, que en los *encabezamientos* puede reconocerse lo que Carl Warren (1975) llamó *lead de sumario*.

2.3.2. El episodio: los acontecimientos principales en el contexto y sus antecedentes

Los *acontecimientos principales* deben situarse en el contexto en el cual surgen. Generalmente, este contexto viene dado por los sucesos previos, probablemente ya relatados en otro texto periodístico, que crean el marco en que se ubica la información actual. En este sentido, no hay que confundir estos sucesos previos con los antecedentes, pues estos últimos tienen una naturaleza histórica o estructural más comprehensiva (Van Dijk, 1990).

2.3.3. Consecuencias

El valor informativo de los acontecimientos sociales y políticos se halla parcialmente determinado por la seriedad de sus consecuencias. “Mediante la discusión real o posible de las consecuencias, un discurso periodístico puede otorgar coherencia causal a los acontecimientos informativos. A veces, las consecuencias son incluso más importantes que los propios acontecimientos informativos principales. En ese caso, los temas de la categoría de las consecuencias, pueden tener la misma posición jerárquica que el tema de los sucesos principales, e incluso pueden llegar a convertirse en el tema de más alto nivel y reflejarse en los titulares” (Van Dijk, 1990).

2.3.4. Reacciones verbales

Las reacciones verbales son una categoría específica del esquema periodístico que puede considerarse como un caso especial de consecuencia. Los acontecimientos informativos más importantes siguen un procedimiento estándar para conseguir los comentarios de participantes importantes o líderes políticos destacados. “La categoría de las reacciones verbales viene señalada por los

nombres y los roles de los participantes periodísticos y por citas directas o indirectas de declaraciones verbales. Por lo general, esta categoría se sitúa después de la de sucesos principales, contexto y antecedentes, hacia el final del discurso periodístico, aunque previamente pueden mencionarse importantes reacciones en el ítem, con las restricciones adicionales del ordenamiento por relevancia" (Van Dijk, 1990).

2.3.5. Comentarios

Finalmente, está la categoría de los comentarios, opiniones y evaluaciones del periodista o del propio periódico. Aun cuando muchos productores de noticias comparten la visión ideológica de que el hecho y la opinión no deben mezclarse, esta última categoría de los comentarios aparece frecuentemente en la noticia, si bien a veces de una forma indirecta. "La categoría de los comentarios consiste en dos subcategorías principales: evaluación y expectativas. La evaluación se caracteriza por las opiniones evaluativas sobre los acontecimientos informativos actuales; la categoría de las expectativas formula consecuencias políticas o de otro tipo sobre los sucesos actuales y la situación. Puede, por ejemplo, intentar predecir acontecimientos futuros" (Van Dijk, 1990).

Por último, el mismo Van Dijk señala que, si bien las categorías que ha identificado poseen una naturaleza hipotética, una amplia investigación empírica ha demostrado que por lo general el discurso periodístico adopta estas categorías. En este sentido, José Luis Martínez Albertos, en su texto "El lenguaje periodístico" (1989), reproduce el diagrama de la estructura hipotética del esquema informativo propuesto por Van Dijk y señala que "el análisis del discurso aplicado al relato periodístico ha demostrado tener especial relevancia. Este método permite, por ejemplo, examinar los modelos textuales que sirven de vehículo de comunicación y hacernos una idea de cómo los textos de los periódicos adquieren sentido para los lectores. Y, al mismo tiempo, estos modelos proporcionan útiles claves para comprender cómo los periodistas dan sentido al mundo en el texto de las noticias o cómo los lectores comprenden estos textos".

2.4. La noticia: ¿un tipo particular de resumen?

Como se señaló al principio de este capítulo, se podría considerar a la noticia como un tipo particular de resumen. Con el fin de poder conceptualizar a la noticia como tal, se presentan algunas ideas sobre el resumen en sí.

Para Gonzalo Martín Vivaldi (2004) el resumen “es la exposición que sintetiza la información esencial de un texto oral o escrito”. Agrega que “resumir no es tan fácil como pudiera parecer a primera vista. En general, se corre el peligro de escamotear lo esencial y de caer en lo accesorio”.

Cervera, Hernández, Pichardo y Sánchez (2006), por su parte señalan que “la acción de resumir implica sintetizar aquello que un texto nos dice. Todo resumen, por tanto, significa una transformación de un texto (*texto original*) en otro texto (*resumen*) que refleje de forma general y breve las ideas principales del anterior y deje al margen las secundarias. Desde una perspectiva formal, además, debe presentar coherencia con la estructura del texto original y cierto paralelismo en su exposición”.

El resumen, por sus características, también ha sido considerado como una forma o medida para evaluar la comprensión de textos. “El resumen puede ser definido como un texto en segundo grado que presenta lo esencial de un texto-fuente anterior. Realizar un resumen implica identificar la coherencia del texto-fuente y producir un nuevo texto más breve. Esto supone omitir la información ya dada, generalizar la información nueva e integrar ideas coherentemente. De este modo observamos que la producción de un resumen supone haber construido una representación concisa del texto original” (Kintsch y van Dijk, 1975; Palinscar y Brown, 1984, citados en Irrázabal et al., 2006).

De lo anterior, se colige que el “resumen no es otra cosa que un texto y por lo tanto debe mantener los mismos requisitos de coherencia y cohesión, así como derivarse de la actuación de los mismos procesos cognitivos. Aún más, en el resumen se agrega, a la compleja actividad de comprender, un esfuerzo por producir oral o de forma escrita lo comprendido” (Irrázabal et al., 2006).

Lo anterior da cuenta de que la tarea de aprender a construir correctamente un resumen no es algo sencillo, sobre todo asumiendo que implica primero la comprensión del texto en que se basa. En este sentido, Steinhart (2001) señala que

para resumir un texto, la persona debe leer y comprender el material, aislar las ideas principales y transmitir estas ideas en forma sucinta. Agrega que estas tareas claramente involucran procesamientos cognitivos más profundos que simplemente leer un texto, ya que requiere que el individuo posea habilidades críticas que le permitan enfocarse en las ideas principales del texto y separarlas de los detalles. La habilidad de resumir, concluye, permite desarrollar una comprensión profunda de textos complejos y, adicionalmente, articular esa comprensión de forma que los conocimientos adquiridos puedan ser compartidos.

Por último, Jurafsky y Martin (2008) señalan que resumir un texto es el proceso de extraer la información más importante de éste, con el fin de producir una versión abreviada del mismo, para una tarea y un usuario concreto. Los citados autores, que provienen del campo de la Lingüística Computacional, en otras palabras, aluden a que el proceso de producir un resumen consiste en producir un texto de extensión menor que el texto original y que este nuevo escrito debe contener la información más relevante contenida en el texto primitivo. Además, Jurafsky y Martin señalan dos puntos importantes, éstos son, que el resumen se realiza en el marco de una tarea específica y, por ende, con un fin específico y, además, que se dirige a un usuario determinado. Por ejemplo, de un artículo científico se podría producir un resumen para que sea el *abstract* del mismo artículo, o bien, se podría producir un resumen que explique el contenido del texto en palabras sencillas para fines de divulgación del texto hacia un público general y no para una comunidad discursiva determinada.

Como se pudo vislumbrar en las deficiniones presentadas, siempre se habla de que el resumen se realiza a partir de un texto, ya sea éste oral u escrito. En este punto es importante tener en cuenta que por *texto* se puede entender no sólo la obra fruto de la escritura sino, en general, cualquier producto de la actividad lingüística y del saber expresivo, ya sea éste de carácter textual, icónico o sonoro (Coseriu, 1981). Lo anterior no implica que, para los fines de este trabajo, al decir que la noticia puede considerarse como un tipo particular de resumen, se asuma que cualquier situación no lingüística se considere como un texto posible de resumirse, aunque pudiera parecer posible afirmar tal cosa siguiendo a Coseriu. Al proponer, con mucha cautela, considerar a la noticia como un tipo particular de

resumen se alude a que casi en la totalidad de las ocasiones el trabajo del periodista se basa en fuentes lingüísticas, como se explicará a continuación.

Teniendo claridad, entonces, lo dicho en los párrafos precedentes sobre el resumen, sobre su relación con el texto fuente y la dificultad que implica su correcta construcción, es momento de articular esto, como se adelantó, con la idea señalada de que la noticia puede considerarse como un tipo particular de resumen.

Van Dijk (1990) señala que "la producción de noticias debe analizarse principalmente en términos del procesamiento del texto [...] Con la frase *procesamiento del texto* no sólo queremos decir que un texto periodístico está siendo procesado, es decir, escrito en varias etapas o fases. La expresión también implica que la mayor parte de la información utilizada para escribir un texto periodístico ingresa en forma discursiva: los reportajes, las declaraciones, las entrevistas, las reuniones, las conferencias de prensa, otros mensajes de los medios, los comunicados de prensa, los debates parlamentarios, los juicios en los tribunales, las documentaciones policiales, etc.". De lo anterior se desprende como señala el citado autor que la mayor parte de las noticias "no se basan en la observación inmediata de los acontecimientos informativos. La mayoría de las noticias deducen su información a partir del discurso. Debemos distinguir, en este caso, entre un discurso que es por sí mismo un acontecimiento periodístico, como las declaraciones de importantes políticos o la publicación de un importante informe o libro, y el discurso que se utiliza solamente por su contenido informativo, no por el valor periodístico del acontecimiento comunicativo en el cual ha sido producido". El citado autor tiene un punto bastante claro en su afirmación: los periodistas -en la mayoría de los casos- construyen el texto de su noticia basándose en fuentes discursivas, ya que obviamente es imposible pensar que por cada noticia publicada hubo un periodista observando cómo ocurría el suceso que la origina (imaginemos, por ejemplo, un choque de automóviles, que es un acontecimiento inesperado).

El mismo van Dijk (1990) agrega que "el resumen tiene lugar en cada nivel del texto fuente y del procesamiento del texto periodístico. La explicación de una conferencia de prensa, de una entrevista, un juicio o un extenso informe, supone por lo general un resumen. El importante rol del resumen en la producción periodística llega a ser obvio cuando nos damos cuenta de que permite al reportero:

1) reducir textos extensos a textos breves; 2) comprender detalles locales de la información del texto fuente relativos a sus macroestructuras; 3) definir la información más importante o relevante de los textos fuente; 4) comparar diferentes textos fuente en relación con sus temas comunes y prioridades; 5) utilizar el resumen como una guía ya preparada y, en consecuencia, como un ejemplo de control semántico básico para escribir el texto periodístico y para deducir titulares, y 6) utilizar el resumen como un plan o diseño para un texto periodístico y para la discusión con los colegas y editores. Debido a la gran cantidad de posibles textos fuente y la complejidad de su información, el resumen es el proceso central de una producción y control periodísticos efectivos, una vez que se ha realizado la selección primaria. Es la estrategia principal para la reducción de la complejidad informativa".

Por último, sobre los elementos de titulación de la noticia -titular, por un lado, y epígrafe y bajada si es que se incluyen- Van Dijk (1990) señala que "su función estructural también es clara: juntos expresan los principales temas del hecho. Es decir funcionan como un resumen inicial". Nuevamente, la idea de resumen aparece ligada a la de noticia, para el caso, enfocándose específicamente en el titular (y el epígrafe y la bajada).

En síntesis, lo afirmado por van Dijk nos permite entender que la noticia puede considerarse como un tipo de resumen, que se basa principalmente en otros medios discursivos, por la baja probabilidad de que un periodista esté presente en el momento de producirse el hecho noticioso (por ejemplo, un choque, un asesinato, un robo, etc.). Con el objeto de no centrarse sólo en lo afirmado por el autor holandés, se analizarán a continuación parte de las definiciones de noticia ya entregadas.

La noticia según Martín Vivaldi (1993) "da cuenta de un modo sucinto pero completo, de un hecho actual o actualizado, digno de ser conocido y divulgado, y de innegable repercusión humana". Es decir, la noticia se refiere a un hecho y da cuenta de éste de un modo sucinto, es decir, breve, resumido.

Martínez Albertos (2004) añade que la noticia o información es el género "más escueto, más descarnado, más fuertemente ceñido al puro esqueleto del hecho o acontecimiento que se quiere transmitir". O sea, para este autor, la noticia

es escueta, carente de adornos y busca representar y transmitir los aspectos fundamentales o esenciales del hecho a que se refiere; no busca contar el hecho con todos sus detalles, sino que representarlo en forma resumida, prescindiendo de todo lo accesorio.

Carl Warren (1975), aludiendo al *lead* de la noticia, señala que en éste "se destaca la esencia o los datos más sobresalientes del acontecimiento del que se quiere dar noticia. Este *lead* se conoce también con el nombre de *lead de sumario*". La razón de llamarlo *lead de sumario* es, precisamente, porque busca presentar en un párrafo todos los antecedentes esenciales de la noticia, siguiendo la técnica de responder a las preguntas fundamentales del periodismo (5 o 6 W's), según se explicó anteriormente. Este *lead,* en otras palabras, busca presentar los antecedentes esenciales del hecho en la forma más resumida posible, o sea, realizando un sumario de éstos.

Según lo precedente, entonces, la noticia lleva en sí la idea de resumen, ya que -valga la redundancia- se le puede considerar como un tipo de resumen, lo que coincide a plenitud con lo planteado por van Dijk. Por ello, para los fines de este trabajo se entenderá a la noticia como un tipo particular de resumen, en la línea de lo explicado por el autor holandés, con toda la complejidad que implica por ende su construcción; dificultad que reafirma el objetivo de la presente tesis de construir un componente que sea capaz de evaluar la coherencia textual de una noticia y la correcta construcción de la estructura semántica denominada pirámide invertida.

Finalmente, siguiendo la definición de resumen Cervera et al. (2006) y adaptándola a la noticia, se puede decir -y proponer con la cautela ya señalada al inicio- que la acción de escribir una noticia implica sintetizar aquello que un hecho nos presenta. Toda noticia, por tanto, significa una transformación de un hecho (que puede entenderse como el conjunto de textos que dan cuenta de éste) en un texto (noticia) que refleje de forma general y breve los elementos principales del hecho y deje al margen lo accesorio.

Segunda parte: Marco metodológico

Capítulo 3: Metodología

3.1. Diseño de investigación

El alcance inicial del estudio puede definirse como correlacional, ya que se busca crear un componente automático cuya evaluación de la estructura semántica de la pirámide invertida sea similar a la realizada por un humano. En otras palabras, hay dos variables en juego: la evaluación realizada por la máquina de la citada estructura semántica y la efectuada por un humano. Lo que se busca es medir ambas variables y conocer la relación o grado de asociación que exista entre éstas en una muestra o contexto en particular (Fernández Collado, Hernández Sampieri y Baptista, 2014).

En relación al diseño de la investigación realizado en la presente tesis, se optó por un diseño no experimental, ya que las dos variables implicadas (evaluación de la máquina y evaluación de los humanos) no inciden de manera alguna sobre la otra, por lo tanto no hay manipulación de una de ellas. Dentro de los diseños no experimentales, se optó por un diseño transeccional, pues los datos se recolectaron en un solo momento, en un tiempo único. La anterior decisión se tomó debido a que en la investigación se trabaja con una muestra de textos de estudiantes de periodismo, que se recopiló en un momento específico. Por último, se optó por un diseño transeccional correlacional, ya que el estudio describe la relación entre dos variables. La elección del diseño transeccional correlacional es porque la presente investigación persigue medir el grado de correlación entre las dos variables mencionadas (evaluación de la máquina y evaluación de los humanos), con el fin de establecer similitudes entre las mediciones realizadas por la máquina y por los humanos. Finalmente, dado que no hay relación de causalidad alguna entre las dos variables, la hipótesis formulada -en consecuencia- es de tipo correlacional (Fernández Collado et al., 2014).

3.2. Objetivos

3.2.1. Objetivo general

Desarrollar un componente, a nivel de prototipo, que evalúe el proceso de escritura de noticias, enfocándose en la evaluación de la coherencia textual y en el aspecto semántico estructural de éstas (estructura de pirámide invertida), capaz de obtener un rendimiento equivalente o superior a la realizada por un evaluador humano experto.

3.2.2. Objetivos específicos

- Desarrollar un módulo, a nivel de prototipo, capaz de evaluar la coherencia textual en noticias.

- Desarrollar un módulo, a nivel de prototipo, capaz de evaluar en forma automática el aspecto semántico estructural de una noticia, esto es, relevancia y jerarquización de la información presentada, en textos pertenecientes a un dominio temático específico.

- Comparar la evaluación efectuada por el componente -de la relevancia y jerarquización de la información presentada en una noticia- con la realizada por evaluadores humanos.

3.3. Hipótesis

H_1: La evaluación automática de la estructura semántica (pirámide invertida) de las noticias escritas, pertenecientes a un dominio temático específico, tiene un rendimiento equivalente o superior a la realizada por un evaluador humano.
H_0: La evaluación automática de la estructura semántica (pirámide invertida) de las noticias escritas, pertenecientes a un dominio temático específico, no

tiene un rendimiento equivalente o superior a la realizada por un evaluador humano.

En la hipótesis 1, el término *equivalente* no debe entenderse en el sentido de que la evaluación entregada por la herramienta prototipo, sea igual a la entregada por los humanos. Es decir, no se pretende probar que si la máquina asigna una evaluación X a un texto determinado, los evaluadores humanos también asignen X. El sentido en que debe entenderse *equivalente* es que ambas evaluaciones -máquina y humanos- sigan tendencias similares y se correlacionen, esto indicaría que la evaluación automática realizada por el componente prototipo es correcta y que puede utilizarse como información útil para perfeccionar el texto producido.

Capítulo 4: Diseño e implementación del módulo (a nivel de prototipo) de análisis semántico, enfocado en la predicción de la coherencia textual

La metodología de trabajo empleada para el desarrollo del módulo 1, de análisis semántico, incluyó tres etapas de desarrollo, más una serie de pruebas. En el presente capítulo se presentan las etapas aludidas: recopilación automática del corpus, preparación automática del mismo y la construcción del evaluador automático de coherencia textual. Por último, se dará cuenta de las pruebas realizadas para testear el correcto funcionamiento de este módulo, el primero de los tres módulos que integran el componente que se busca construir en el presente trabajo.

4.1. Recopilación automática del corpus

En esta etapa se trabajó en un computador con sistema operativo Linux, específicamente con la distribución Ubuntu 14.04. Dicha distribución es muy sencilla de utilizar e instalar e, incluso, se puede correr en una máquina virtual dentro de un computador con Windows o Mac OS, utilizando una aplicación como Oracle VirtualBox (http://www.virtualbox.org), sin necesidad de formatear el equipo, ni abandonar el sistema operativo que se utilice normalmente.

Linux, en sus distintas versiones, cuenta con una consola de comandos o Terminal, cuyo intérprete de comandos por defecto -en la mayor parte de sus distribuciones- es Bash. Dentro de las aplicaciones que se pueden invocar mediante comandos de Bash se encuentra GNU Wget. Según se define en el sitio del Proyecto GNU (2014), GNU Wget es un paquete de software libre para la recuperación de archivos a través de HTTP, HTTPS y FTP, que son los protocolos de Internet más utilizados. Es una herramienta de línea de comandos no interactiva, por lo que puede ser fácilmente invocada desde *scripts.* En otras palabras, mediante el uso de esta aplicación se pueden descargar archivos que se encuentren alojados en Internet y esto incluye a los documentos escritos en HTML (o similares) que almacenan gran parte del código que los navegadores interpretan para presentar las páginas de un sitio web. Es decir, mediante el ingreso del comando *wget* en la Terminal de Linux es posible descargar, por ejemplo, una o varias de las noticias

ubicadas en una determinada dirección de Internet.

El tipo de texto que se utiliza para esta investigación, como se señaló al inicio, lo constituye el género discursivo noticias políticas. Ello se debe al gran volumen de material que se produce diariamente sobre el tema, lo que se vio reflejado finalmente en el tamaño del corpus y el mayor esmero que hay en la redacción de las mismas; si bien esto último puede considerarse algo subjetivo, en los medios de prensa se escoge con mayor cuidado a los periodistas y editores que se aboquen a este tema, por la importancia mediática del mismo.

Las noticias sobre política fueron extraídas del sitio web del diario La Tercera, que es un medio de circulación nacional en Chile. La razón para elegir éste es que en su sitio en Internet cuenta con un canal dedicado a las noticias sobre política, las que se encuentran ubicadas dentro del directorio localizado en http://www.latercera.com/noticia/politica. Si se ingresa esta dirección en el navegador, arrojará un error de página no encontrada o de acceso restringido; sin embargo, todas las noticias sobre política que publica el citado medio, se encuentran alojadas en este directorio y accesibles sólo por algunos días. Para recopilarlas, se utilizó la aplicación *wget*. Ésta permite explorar este directorio y descargar todo el contenido que haya en él al momento de la revisión. A continuación se presenta una línea de comandos similar a la utilizada en este trabajo, pero agregando algunas opciones más, con el fin de poder explicarlas para demostrar de una forma más completa las posibilidades de GNU Wget:

```
wget --include noticia/politica/ --wait=20 --limit-rate=20K -r -p -N -R jpg,jpeg,gif,png,js,swf
--accept shtml -U Mozilla http://www.latercera.com/
```

Lo primero es invocar el comando, simplemente, escribiendo su nombre, *wget*. A continuación, con *--include* se indica la ruta al directorio a examinar (noticia/politica/). Luego, con *--wait=20* --limit-rate=20K se ordena a la aplicación que espere 20 segundos entre el fin de la descarga de un archivo y el inicio de la descarga del siguiente; además, limita la velocidad de descarga a 20 kb/s. Esto último con el fin ético de no saturar el servidor del medio y, a la vez, protegerse de un posible bloqueo a la IP del equipo utilizado, en caso de no emplear respetuosamente el ancho de banda y los recursos del servidor al que se accede.

A continuación, el comando incluye la opción *-r* que implica que la descarga de archivos será recursiva, o sea, buscará archivos dentro de los subdirectorios que haya. La opción *-p* le indica al programa que descargue todos los archivos que son necesarios para la correcta visualización del documento HTML. La opción *-N* permite que si en el computador ya existe un archivo igual al que se va a descargar, se solicite al servidor la fecha de la última modificación del fichero y, únicamente, si es más reciente del que ya se tiene, lo descargará. La opción *-R* permite excluir ciertos patrones (o extensiones de archivo, para el caso) que no se desea que se descarguen (en el ejemplo: jpg, jpeg, gif, png, js, swf). La opción *--accept* le indica a la aplicación el tipo de archivos que se quiere descargar, para el presente trabajo, los con extensión *.shtml.* La opción *-U Mozilla* permite que el GNU Wget sea detectado como un navegador en particular, Mozilla Firefox, en este caso. Por último, se incluye la dirección raíz del sitio de La Tercera.

En el trabajo realizado, no se utilizaron las opciones *-p* ni *-R*, ya que sólo interesaban los archivos con extensión *.shtml*. Hay que señalar que el comando escrito más arriba igual serviría para el objetivo de obtener sólo los archivos citados, pero como se señaló, podrían eliminarse las opciones mencionadas, con el fin de evitar código inútil.

Hasta aquí se presentó una forma de descargar los archivos necesarios para conformar el corpus, mediante una línea de comandos de Bash. El siguiente paso es automatizar la tarea de recopilación de texto.

Dicha automatización, en un primer momento se realizó con el comando *cron* de Linux, que sirve para automatizar la ejecución de procesos. Sin embargo, también se utilizó para automatizar la tarea el símil de *cron* que trae Windows: el Programador de Tareas. Se explicará sólo esta segunda opción, en el entendido de que la mayoría de las personas utilizan el sistema operativo de Microsoft. En caso de querer emplear el comando *cron* en ambiente Linux, en Internet hay muchos sitios sobre cómo crear un archivo *crontab* para automatizar la tarea de ejecutar la línea de comandos presentada más arriba.

Previo a la automatización, hay que poder usar *wget* en Windows. Para esto, primero hay que instalar la aplicación Cygwin que es una colección de herramientas GNU y Open Source, que proporcionan una funcionalidad similar a una distribución

de Linux en ambiente Windows. La aplicación se puede descargar desde http://www.cygwin.com/, sitio en el cual también hay instrucciones para configurarla correctamente e instalar aplicaciones GNU en Cygwin, para el presente caso, Wget. Es fundamental asegurarse de que se agregue la ruta a Cygwin en la variable de entorno PATH de Windows (más información sobre cómo hacer esto se puede revisar en http://www.computerhope.com/issues/ch000549.htm).

Una vez definida la línea de comandos e instalado Cygwin, queda el paso final que es escribir un *script* en Windows para poder ejecutar desde el Programador de Tareas. Este paso, aunque parezca complejo, es bastante sencillo. Simplemente, hay que abrir un editor de texto plano, por ejemplo el Bloc de Notas de Windows, y copiar las siguientes líneas, asumiendo que Cygwin se instaló en su ruta por defecto (C:\cygwin\bin):

```
@echo off

C:
chdir C:\cygwin\bin
bash --login -i -c 'wget --include noticia/politica/ --wait=20 --limit-rate=20K -r -p -N -R
jpg,jpeg,gif,png,js,swf --accept shtml -U Mozilla http://www.latercera.com/'
```

Nótese que entre comillas simples, al final, va la línea de comandos para operar Wget (se copió la versión más extensa que se explicó más arriba). Luego, se debe guardar este archivo asignándole un nombre a elección y asegurarse de agregarle manualmente la extensión *.bat* (no *.txt*). El archivo resultante es un archivo Batch, que es una pequeña aplicación, que al ejecutarse lleva a cabo las instrucciones que contiene. En el presente caso, conectarse al sitio de La Tercera para descargar las noticias que necesitamos para el corpus.

Lo último que resta es utilizar el Programador de Tareas de Windows para configurar que el archivo Batch se ejecute automáticamente. Las instrucciones para realizar esto se pueden encontrar en la propia ayuda del sistema operativo[2]. En el caso que se expone en el presente trabajo, se eligió ejecutar el archivo todos los días a las 8, a las 12, a las 16, a las 20 y a las 0 horas. Mediante este método se recopiló el corpus original de 7.165 textos (cómo se verá más adelante, éste no fue

2 También se puede visitar la siguiente dirección web: http://windows.microsoft.com/es-xl/windows/schedule-task#1TC=windows-7

el único corpus utilizado).

Con el método descrito, cada vez que el Programador de Tareas de Windows ejecute el *script* para recopilar textos, se abrirá la consola de comandos del sistema y permanecerá funcionando hasta que se complete la tarea. Obviamente, el tiempo para esto será variable: dependerá, entre otros posibles factores, de la cantidad de textos que encuentre y de la velocidad de conexión. Una complicación de lo recién expuesto, es que la aparición de la consola puede ser un elemento distractor en caso de estar trabajando en el computador o, simplemente, se corre el riesgo de que uno mismo o alguien más cierre la consola e impida la ejecución completa de la tarea. Una buena forma de prevenir lo anterior, es creando otro *script*, que se ejecute en segundo plano e invoque al que anteriormente se guardó con la extensión *.bat*. De esta forma, la consola no aparecerá y la ejecución de la tarea será invisible para el usuario del equipo. El código para construir este nuevo *script* es el que se presenta a continuación:

```
set objshell = createobject("wscript.shell")
objshell.run "Ruta-al-script\script.bat", vbhide
```

Al igual que el código anterior, éste debe copiarse en un editor de texto plano y guardarlo con el nombre que se desee, pero con la extensión *.vbs*. En la segunda línea, luego de *objshell.run* hay que incluir la ruta completa al *script .bat*. Luego de lo anterior, hay que guardarlo y configurar el Programador de Tareas para que ejecute el *script .vbs* y no el *.bat*. Con esto, la recopilación de textos será en segundo plano y no interferirá en absoluto con el usuario del equipo en los momentos de ejecución de la tarea.

4.2. Preparación automática del corpus

El funcionamiento de esta etapa del proceso se basa en un *script* programado en Python 3 (https://www.python.org/). El *script* construido se hace cargo de los textos desde que se reciben a través de la recopilación automática que se realiza de éstos desde el sitio web del diario La Tercera, como se explicó en el punto anterior, hasta que ya están preparados para incorporarlos al prototipo propiamente tal que

permite realizar consultas sobre relación semántica y coherencia textual. Dicho *script* funciona sin problemas sobre los sistemas operativos Linux (probado en Ubuntu 14.04 y Linux Mint 16 y 17) y Windows (probado en Windows 7, 8 y 8.1).

El primer paso dentro del desarrollo fue que el *script* tomara los textos obtenidos desde La Tercera en formato *.shtml* y los transformara en un archivo de texto plano, eliminando las marcas, los caracteres especiales y toda aquella información que, si bien no corresponde a un lenguaje de marcas para la web, si es considerada inútil desde el punto de vista del analizador semántico. Algunos ejemplos de esto último, son el nombre del autor de la noticia, los titulares que enlazan a otras informaciones del medio, a las noticias más leídas o a las redes sociales desde las cuales se puede seguir a La Tercera, entre otros contenidos que no aportan a la meta buscada con el analizador.

Cada archivo *.shtml* con una noticia, por ende, tiene una gran cantidad de información que se debe eliminar, para dejar sólo el texto de ésta con sus elementos de titulación. Para realizar dicha tarea, se realizaron múltiples pruebas, sobre todo con parsers de HTML, algunos de los cuales estaban implementados como módulos dentro de Python. Los resultados de las pruebas no fueron enteramente satisfactorios, ya que si bien estas pequeñas aplicaciones realizaban su tarea con algún grado de éxito, no siempre lograban limpiar por completo las marcas que contenían los archivos. A lo anterior, hay que agregar el hecho de que dichas aplicaciones están orientadas a eliminar las marcas de un lenguaje, pero no ayudan en absoluto en la criba de otros elementos textuales innecesarios como el nombre del autor o los enlaces a otras noticias del sitio; más aún, al eliminar las marcas del lenguaje era todavía más complicado diferenciar el texto de la noticia de estos agregados.

Por lo recién explicado, se optó por prescindir de la utilización de herramientas prediseñadas como los *parsers* de HTML y enfocarse en la construcción desde cero de un *script* a la medida de los archivos recopilados, que eliminara todo lo innecesario de éstos y los dejara listos para su utilización. En este escenario, el lenguaje de marcas de los archivos fue algo que -paradojalmente- permitió la realización de la tarea con éxito. Lo anterior, debido a que en la construcción del *script* se aprovechó la segmentación del documento que realizan

estas marcas, para así poder determinar qué porciones debían eliminarse y cuáles debían conservarse.

El *script* primero trabaja con todos los archivos *.shtml* que haya en un directorio determinado y realiza sobre ellos operaciones recursivas con el fin de ir eliminado el contenido considerado innecesario. Un ejemplo que clarifica mucho esta etapa, es la utilización de las marcas para eliminar porciones inútiles de código y texto. Cada archivo recopilado comienza con:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
```

Después de esta etiqueta, viene una gran cantidad de información y marcas inútiles para el objetivo del trabajo, hasta llegar a la etiqueta:

```
<h1 class="titularArticulo">
```

Entre una y otra, en promedio hay 47.260 caracteres de información que no sirve. Por lo tanto, en este caso, el *script* lo que realiza es simplemente borrar las etiquetas y todo lo que está incluido entre éstas, con lo que se limpia gran parte del archivo. Para ello, se utilizó la siguiente línea de código en Python:

```
1 import re                    #Esta línea importa el módulo de expresiones regulares de
Python.
2
3 archivo_entrada = open(’entrada.txt’, 'r')
4 archivo_salida = open(’salida.txt’, 'w')
5 texto_archivo = archivo_entrada.read()
6 eliminar = re.compile('<!DOCTYPE.*?titularArticulo">', re.I | re.S)
7 texto_archivo2 = eliminar.sub("", texto_archivo)
8 archivo_salida.write(texto_archivo2)
9 archivo_entrada.close()
10 archivo_salida.close()
```

Los números que están antes de cada línea no pertenecen al código, sino que indican el número de la línea, para facilitar la explicación posterior. Hay que advertir que el código presentado es completamente funcional, pero se ha modificado del original, con el fin de facilitar su lectura. Lo que hace el código es abrir y leer el

archivo de entrada (líneas 3 y 5) y abrir el de salida para su escritura (línea 4). Luego, empleando el módulo de expresiones regulares de Python que se importó en la primera línea, busca el patrón que comienza con '*<!DOCTYPE' y termina con 'titularArticulo*"*>'* y lo elimina al escribir el archivo de salida (líneas 6 y 7); los signos *'.*?'* que se emplean para unir ambas porciones del patrón indican que se incluirá en el texto a borrar, lo que sea que haya entre las dos porciones del patrón, en este caso, como ya se indicó, aproximadamente 47.260 caracteres de información innecesaria. De la forma ejemplificada, se puede borrar gran cantidad de información inútil del archivo .*shtml*. Por último, se escribe el texto modificado en el archivo de salida (línea 8) y se cierran los archivos abiertos (líneas 9 y 10).

Los nombres asignados a los archivos son arbitrarios y se pueden cambiar al escribir el *script*. De hecho, en el *script* construido no se lee y se escribe un archivo cada vez que se desea ejecutar una instrucción, sino que ésta tarea se realiza sólo al comienzo y al final del *script*, almacenando las salidas intermedias en la memoria del equipo. Además, para abrir los archivos se empleó la sentencia *with* que no obliga a declarar que éstos se cierren (por ejemplo, "*with open ('entrada', 'r') as archivo_entrada:").* Pero, como se dijo, aquí se modificó el código empleado con el fin de que sea más fácil de entender para alguien que no tenga conocimientos de Python.

Otros reemplazos que realiza este *script* son borrar las líneas en blanco que quedan después de las operaciones aplicadas, cambiar la codificación de caracteres a UTF-8 si es que estuviera, por ejemplo, en ANSI y, también, cambiar los caracteres especiales por los normales del español (por ejemplo, reemplaza *á* por *á*).

Por ejemplo, para realizar esto último, se utiliza nuevamente el módulo de expresiones regulares de Python, con el fin de buscar estos patrones y reemplazarlos por los deseados. El código empleado para esto es el siguiente:

```
1 import re
2
3 reemplazos = {'&aacute;':'á', '&eacute;':'é', '&iacute;':'í', '&oacute;':'ó',
'&uacute;':'ú'} #(se omitió aquí el resto de los reemplazos, por razones de espacio).
4 archivo_entrada = open('entrada.txt', 'r')
5 archivo_salida = open('salida.txt', 'w', encoding='utf8')
6 for line in archivo_entrada:
```

```
7     for src, target in reemplazos.items():
8         line = line.replace(src, target)
9     archivo_salida.write(line)
10 archivo_entrada.close()
11 archivo_salida.close()
```

El *script* también se hace cargo de un problema particular que tienen los textos: los elementos de titulación. Normalmente en periodismo se aceptan de tres tipos, uno imprescindible como el titular y dos optativos como el epígrafe o antetítulo (que va sobre el titular) y la bajada o subtítulo (que como su nombre indica, va debajo del titular). Si bien en los textos recopilados, al igual que en la mayoría de las noticias para Internet, se prescinde del epígrafe, el problema hallado es que no había un uso uniforme de la bajada. En otras palabras, algunas noticias tenían sólo titular y otras incluían titular y bajada. Por ello, este *script* se hace cargo de reconocer cuándo una noticia tiene bajada e incluye una marca antes y después de ésta, con el fin de destacarla para su uso posterior, en que podría decidirse incluir la bajada en los análisis a realizar o prescindir de ella. De la misma manera, incluye una marca para el titular. La marca de inicio del titular es *zzzinictitularzzz* y la de término es *zzztitularfinzzz*; para el caso de la bajada, las marcas son *zzzinicbajadazzz* y *zzzbajadafinzzz*. Como se puede ver, la elección de las marcas es absolutamente arbitraria.

La última operación de relevancia que realiza el *script* es agregar a cada archivo procesado las marcas exigidas por la aplicación de Análisis Semántico Latente, Infomap-NLP, que permite la construcción del espacio semántico en base al corpus textual. Estas marcas son <DOC> y <TEXT> (en ese orden) antes del inicio de cada texto, y </TEXT> y </DOC> al final de cada una de las noticias. El código para lo anterior es como sigue:

```
1 archivo_entrada = open('entrada.txt', 'r')
2 archivo_salida = open('salida.txt', 'w')
3 inicio = ('<DOC>\n<TEXT>\n') + archivo_entrada.read()
4 final = inicio + ('\n</TEXT>\n</DOC>\n')
5 archivo_salida.write(final)
6 archivo_entrada.close()
7 archivo_salida.close()
```

Como una muestra de la potencia y versatilidad de Python, se expone a

continuación una sola línea de código que realiza las mismas tareas que las líneas 1 a 5 del código recién presentado, aunque es un poco más complicada de leer:

```
open('salida.txt', 'w').write('<DOC>\n<TEXT>\n' + open('entrada.txt').read()
+'\n</TEXT>\n</DOC>\n')
```

Finalmente, el *script* une todos los archivos resultantes en un solo fichero de texto plano. Esto se realiza gracias a los módulos *glob, shutil* y *os* de Python. En el ejemplo que sigue, se asume que los archivos a unir están ubicados en el directorio llamado *textos*, que se encuentra junto al script:

```
1 import glob,shutil,os
2
3 ruta = os.getcwd() + '/textos/'
4 with open('salida_marcas.txt', 'wb') as archivo_salida:
5     for filename in glob.glob(os.path.join(ruta, '*.txt')):
6         with open(filename, 'rb') as archivo_entrada:
7             shutil.copyfileobj(archivo_entrada, archivo_salida)
```

Lo hasta aquí presentado es la forma de procesamiento de datos y metodología de funcionamiento del *script.* Obviamente, se pueden realizar muchas tareas adicionales escribiendo las líneas de código para ello, con el fin de adaptarlo a las necesidades puntuales de cada situación. Por ejemplo, en el caso específico del *script* construido se definió una función que permite elegir si el archivo de salida lo queremos con o sin las marcas (*zzzinictitularzzz, zzztitularfinzzz*, *zzzinicbajadazzz, zzzbajadafinzzz*) y con o sin la bajada del texto. Dicha función nos da a elegir, mediante una pregunta que se imprime en la consola de comandos, si queremos el archivo completo y con marcas (opción 1), sin marcas y con bajada (opción 2), o sin marcas y sin bajada (opción 3). Lo anterior se logra gracias al uso del condicional *if* en el código, como se muestra a continuación, en que se presenta la función escrita (llamada *elegir*), que trabaja sobre el archivo *salida_marcas.txt* generado en los pasos descritos más arriba:

```
def elegir(pregunta, reclamo='Se equivocó de tecla'):
    while True:
        ok = input(pregunta)
        if ok in ('1'):
```

```
            return True
        if ok in ('2'):
            with open('salida_marcas.txt') as archivo_entrada:
                paso1 = filer.read()
                archivo_salida = open('salida_conbajada.txt', 'w')
                paso2 = re.sub('zzzinictitularzzz','', paso1)
                paso3 = re.sub('zzztitularfinzzz','', paso2)
                paso4 = re.sub('zzzinicbajadazzz','', paso3)
                paso5 = re.sub('zzzbajadafinzzz','', paso4)
                archivo_salida.write(paso5)
                archivo_salida.close()
            os.remove('salida_marcas.txt')
            return True
        if ok in ('3'):
            with open('salida_marcas.txt') as archivo_entrada:
                paso6 = filer.read()
                archivo_salida = open('salida_sinbajada.txt', 'w')
                patter = re.compile('zzzinicbajadazzz.*?zzzbajadafinzzz\n', re.I | re.S)
                paso7 = patter.sub("", paso6)
                paso8 = re.sub('zzzinictitularzzz','', paso7)
                paso9 = re.sub('zzztitularfinzzz','', paso8)
                archivo_salida.write(paso9)
                archivo_salida.close()
            os.remove('salida_marcas.txt')
            return True
        print(reclamo)

elegir('¿Qué archivo quiere: completo con marcas(1), completo sin marcas(2), sin marcas
y sin bajada(3)? =')
```

Por último, es importante señalar que la ejecución del *script*, si es que se procesan muchos archivos de texto a la vez (los 7.165, por ejemplo) demanda una gran cantidad de recursos del computador, ya que el procesamiento de cantidades ingentes de cadenas de texto es una tarea pesada para cualquier equipo. Lo anterior, junto con las especificaciones de la máquina en que se ejecute el *script*, influirá en el tiempo que tarde en completarse la tarea.

4.3. Evaluador automático de coherencia textual

Para el desarrollo del evaluador automático de coherencia textual se utilizó Python y NLTK (Natural Language Toolkit, disponible en http://www.nltk.org/). Dicho prototipo se construyó en base a dos *scripts*: el primero se encarga únicamente de tomar el archivo con los textos entregados por la etapa de preparación de los

mismos y, en base a este fichero, previa lematización de los textos que lo conforman, construye el espacio semántico multidimensional con el que se trabajará. La tarea de construir el espacio semántico en Infomap-NLP[3] es bastante sencilla y se puede realizar rápidamente por un humano, por ello no se explicará mayormente este primer *script*. Sin embargo, la parte de invocar al lematizador que incluye, es exactamente igual a cómo opera en el segundo *script*, por lo tanto se detallará únicamente para este último. Dicho *script*, a diferencia del primero es algo más complejo, y se encarga, expresándolo en términos simples, de las diferentes tareas que se necesitan para realizar la comparación de las unidades textuales de cada noticia procesada en el prototipo.

A diferencia de los *scripts* de la etapa anterior, los dos recién mencionados necesitan trabajar en un ambiente Linux, ya que la aplicación Infomap-NLP[4] funciona nativamente en un sistema operativo de este tipo. Sin embargo, debería ser factible operar los *scripts* en un sistema con Windows, si es que se instalara Infomap-NLP utilizando la aplicación Cygwin, ya que según las notas de lanzamiento de la última versión de Infomap, lo anterior es posible. Sin embargo, para efectos del presente trabajo, esto no ha sido testeado. En el caso del presente trabajo, para facilitar la tarea, se creó una máquina virtual con la aplicación Oracle VirtualBox (https://www.virtualbox.org/), en la cual se utilizó el sistema operativo Ubuntu 14.04, sobre el que se instaló Infomap. VirtualBox se instaló sobre un computador con Windows 8.1, por lo que no hubo necesidad de formatear el PC ni ninguna otra complicación adicional. De hecho, sólo se utilizaron 30GB de disco duro en el equipo con Windows y la instalación completa no tomó más de 45 minutos.

Como se mencionó antes, ambos *scripts* realizan un proceso de lematización. Para ello, se empleó el Snowball Stemmer, incluido en NLTK, ya que trabaja con

3 Previamente a la construcción del espacio, hay que considerar lo que señalan Manning, Raghavan y Schütze (2008), en el sentido de que hay palabras sumamente comunes que son de escaso valor en la tarea de seleccionar documentos que concuerden con la petición de un usuario. Estas palabras son excluidas del vocabulario de la aplicación que recupera información y se les llama *stop words* (traducido como palabras de parada, palabras de relleno o palabras de paso). En el Diccionario Tecnológico (2015) del sitio web chileno Usando.info, que trata sobre Arquitectura de la Información, Usabilidad y Accesibilidad de sitios web, se entrega otra definición de *stop words* y señala que “son aquellas palabras que por ser comunes, los buscadores ignoran para asegurar la calidad de los resultados de lo que se busca. Normalmente entran en esta categorías las proposiciones y conjunciones”. En el caso del presente trabajo, se modificó el código de Infomap-NLP para incluir en éste el mismo listado de *stop words* del español que se empleó en Hernández (2010) y Hernández y Ferreira (2010).

4 Obviamente, las instrucciones sobre cómo instalar Infomap-NLP no se incluyen, pero pueden revisarse en http://infomap-nlp.sourceforge.net/doc/user_manual.html

lengua española y, además, con un grado de precisión bastante aceptable. En ambos casos, para poder utilizar Snowball, los textos deben ser preparados previamente, eliminando los caracteres que no son admitidos (tildes, comas, signos de interrogación, signos de exclamación, etc.). Para realizar esto, se utilizó el módulo de expresiones regulares de Python, tal cual se explicó en el apartado anterior. En el caso de la lematización, el código que se empleó es el siguiente:

```
from nltk.stem.snowball import SnowballStemmer          #Esta línea importa el
lematizador desde NLTK.

archivo_entrada = open('entrada.txt', 'r')
archivo_salida = open('salida.txt', 'w')
for linea in archivo_entrada:
  linea1 = re.sub('\n', ' <zzz>\n', linea)
  singles = []
  stemmer = SnowballStemmer("spanish")
  for plural in linea1.split():
    singles.append(stemmer.stem(plural))
    linea2 = (' '.join(singles)).replace('<zzz>','\n').lstrip()
  archivo_salida.write(linea2)
archivo_entrada.close()
archivo_salida.close()
```

Para la construcción del espacio semántico, el *script* se basó en las instrucciones para realizar esta tarea en forma manual, entregadas en el sitio de Infomap-NLP[5] y se utilizó la construcción del modelo en un directorio de trabajo y no en un directorio permanente, porque de esta forma se tiene mayor control sobre el mismo. La única gran repercusión de esto, es la forma de realizar las posteriores consultas, para lo cual sólo hay que indicar la ruta al directorio del modelo y utilizar el comando *associate* de Infomap. La forma de realizar las consultas al modelo en forma manual, también se indican en la URL de la nota al pie de esta página, pero es básicamente la estructura que sigue:

```
associate -t -c espacio_modelo presidente
```

En el ejemplo anterior, lo que estamos realizando es consultar cuáles son las palabras que tienen una mayor relación semántica con *presidente*, en el espacio

5 http://infomap-nlp.sourceforge.net/doc/user_manual.html

semántico denominado como *espacio_modelo* (el nombre asignado es arbitrario).

En el caso del *script* construido, no se busca averiguar las palabras que mayor relación semántica tienen con una palabra en particular, sino que el objetivo es comparar todas las unidades textuales adyacentes -párrafos para el caso-, de una noticia en particular. El código que se presenta más abajo es el utilizado para realizar esta tarea, el que debería agregarse a continuación del código para lematizar ya presentado (se reproduce la línea a partir de la cual debería añadirse):

```
import subprocess

modulo_salida = []
for linea in entrada:
#—Código omitido—#
      linea2 = (' '.join(singles)).replace('<zzz>','\n').lstrip()
      linea_texto = line2.split()
            comandos_infomap = ['associate', '-q', '-m', '/home/nombre/espacio2', '-c',
'espacio_modelo']
      procesar = comandos_infomap + linea_texto
      vectores = subprocess.check_output(procesar)
      modulo_salida .append(vectores)
  modulo_salida = [x for x in salidas if x != b' ']
```

El prototipo, en esta etapa, toma cada uno de los textos de entrada por separado, los segmenta en párrafos y calcula el vector que representa a cada uno de estos párrafos. Luego, entrega los valores de cada vector, con el fin de poder pasar a la última tarea que realiza este *script* del prototipo: la comparación de todas las unidades textuales adyacentes, a través del cálculo del coseno del ángulo que forman los vectores que representan a cada unidad. Sin entrar a definir conceptos matemáticos en forma profunda, situación que no persigue este trabajo, hay que tener claro que para cada par de vectores adyacentes (A,B) se calcula el coseno entre ellos de la siguiente forma:

$$\cos(A,B)=\frac{A*B}{\|A\|*\|B\|}$$

Lo anterior significa que el coseno del ángulo formado por dos vectores es igual al producto escalar entre los vectores dividido por el producto de sus módulos.

También se puede representar de la siguiente manera:

$$\cos(A, B) = \frac{\sum_{i=1}^{n} a_i b_i}{\|A\| * \|B\|}$$

Las operaciones para realizar dicho cálculo se realizan con la ayuda del módulo Numpy de Python (http://www.numpy.org/). El código utilizado en el *script* es el que sigue y también debería ser agregado luego del código que calcula los vectores:

```
import numpy as np
from numpy import linalg as LA

modulo_salida  = []
for linea in entrada:
#—Código omitido—#
    modulo_salida  = [x for x in salidas if x != b' ']
    valores =[]
    for par1, par2 in zip(modulo_salida[0::1], modulo_salida[1::1]):
        a = np.genfromtxt([par1])
        b = np.genfromtxt([par2])
        coherencia = (np.dot(a, b))/(LA.norm(a)*LA.norm(b))
        valores.append(coherencia)
promedio = np.mean(valores)
```

El código anterior toma el valor de los vectores que representan a cada una de las unidades textuales definidas (párrafos) y compara el primer valor con el segundo, el segundo con el tercero y así sucesivamente, hasta llegar a la última pareja. Los puntajes de coherencia obtenidos al comparar dos unidades textuales se promedian y este resultado es el puntaje de coherencia del texto en su conjunto.

Una vez realizadas todas estas operaciones, el prototipo arroja como salida los resultados en un archivo de texto plano, que permite revisarlos y someterlos a algún tipo de procesamiento posterior, como se requerirá en las etapas siguientes del trabajo. A continuación, se muestra un ejemplo de la salida final, en que se copia el titular de la noticia, los puntajes de coherencia para cada pareja de vectores y, por último, el puntaje del texto como unidad.

Walker: "Esto es un tema político más que disciplinario"

0.66224260856399708
0.73623747060453659
0.79895212424025353
0.67184566234764576
0.52964567682386254
0.605330035033259

_____
0.667375596269

El funcionamiento del prototipo desarrollado (módulo 1), puede resumirse gráficamente en la Figura 5:

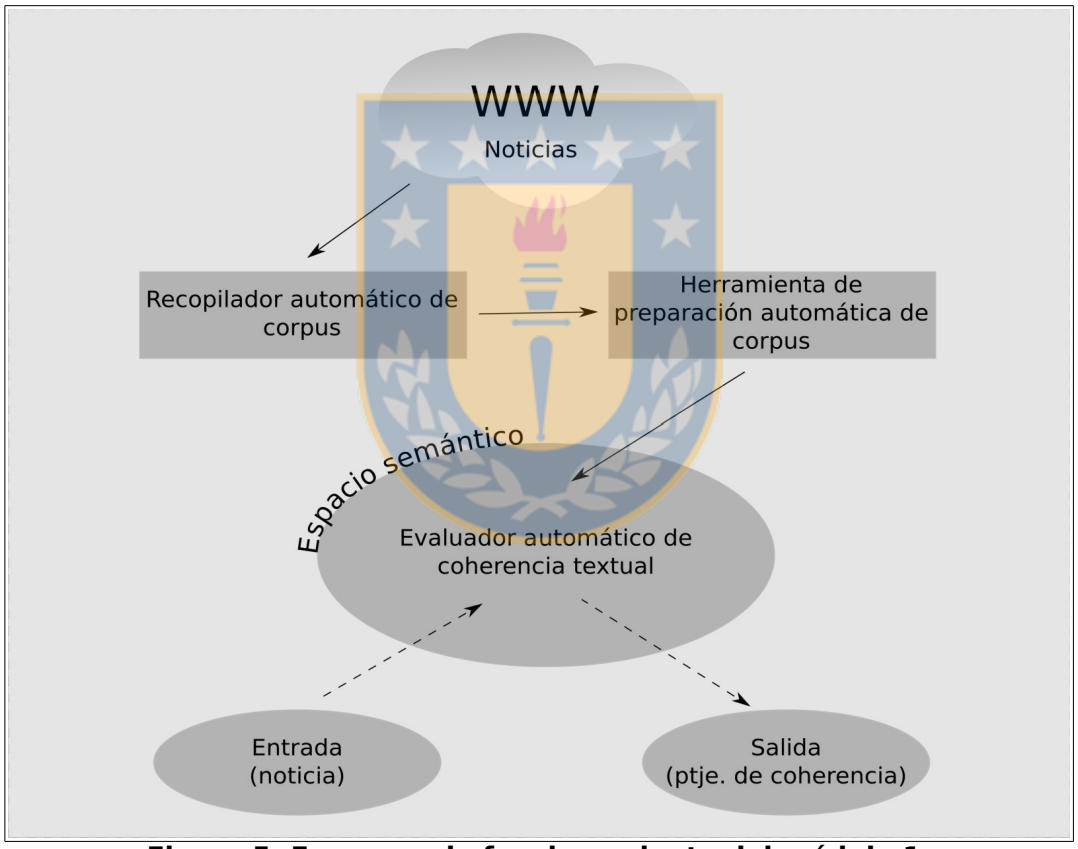

Figura 5: Esquema de funcionamiento del módulo 1.

4.4. Resultados de las pruebas realizadas al módulo 1

La forma de comprobar si la máquina evalúa acertadamente la coherencia textual de una noticia es comparando los resultados que entrega con el juicio de evaluadores humanos. En Hernández y Ferreira (2010), se utilizó y validó un método de análisis similar al propuesto en el presente trabajo, por lo que no fue necesario

replicar toda la tarea de comparación de los resultados.

El módulo, eso sí, se sometió a una serie de pruebas en que se alteraba intencionadamente la coherencia textual con el fin de ver si era sensible a estos cambios: manipular la coherencia de textos, crear textos sin sentido uniendo párrafos extraídos desde diferentes noticias (cuya única relación era pertenecer al dominio en que se enfocó el trabajo) y se compararon textos que presentaban la misma noticia, pero extraída de medios de prensa diferentes (para ello se trató a cada noticia como una única unidad textual).

Entre las pruebas realizadas se compararon oraciones construidas, alusivas a temas pertenecientes al dominio. Por ejemplo, al cotejar la secuencia "*el diputado votó en contra de la ley*" con "*el senador rechazó el proyecto*" el puntaje de coherencia que arroja el módulo es de 0,81; dicha cifra refleja la similitud semántica de ambas oraciones y también da cuenta de los matices que la diferencian. Si se cambia la segunda oración por "*el ministro fue a ver la presidenta*" el puntaje de la comparación desciende a 0,13, dando cuenta de la poca relación entre ambas secuencias.

Otra prueba consistió en tomar noticias con un puntaje de coherencia relativamente alto (superior a 0,60) y manipular el texto cambiando palabras, con el fin de alterar la coherencia textual. Por ejemplo, a continuación se presenta el titular y el primer párrafo de una noticia en particular que tiene un puntaje promedio de 0,69 (considerando el texto completo):

> **Titular de comisión de RR.EE. del Senado y dichos de Sabag: "No representan el espíritu de Chile"**
> El presidente de la comisión de RR.EE. del Senado, Francisco Chahuán (RN), desestimó la polémica que generaron en Bolivia los dichos de su par de la Cámara de Diputados, Jorge Sabag (DC), quien sostuvo en una entrevista con el diario El Sur que "a Chile le ha ido mejor con armas que con diplomacia", en relación al inicio de los alegatos del juicio entre ambos países en la Corte Internacional de Justicia de La Haya.

El puntaje que arroja la comparación entre el titular y el primer párrafo es de 0,71. Sin embargo, al cambiar algunas palabras en el primer párrafo el puntaje desciende a 0,45, por lo que el módulo demuestra ser sensible a los cambios que

alteran la coherencia. La misma tendencia se observó en las otras nueve pruebas realizadas. A continuación se presenta el párrafo modificado, destacando en cursivas las palabras cambiadas.

> El *intendente* de la comisión de *vivienda* del *ministerio*, Francisco Chahuán (RN), desestimó la polémica que generaron en Bolivia los dichos de su par de la *Moneda*, *Fuad Chaín* (DC), quien sostuvo en una entrevista con el diario El Sur que "a *Concepción* le ha ido mejor con armas que con diplomacia", en relación al inicio de los alegatos del juicio entre ambos países en la Corte Internacional de Justicia de La Haya.

Otra prueba realizada, fue tomar cinco textos que tuvieran un puntaje superior a 0,60 y construir un único texto combinando el titular de un texto, el primer párrafo de otro, el segundo párrafo del siguiente y así sucesivamente. Para realizar este constructo, se decidió que los cinco textos originales debían pertenecer al mismo medio. A continuación se presenta el texto resultante de la combinación de cinco noticias de Bío Bío Chile.

> **Diputado que pidió minuto de silencio por Pinochet: “Le debemos un eterno agradecimiento”**
>
> Alcaldes de la UDI acordaron levantar la candidatura del jefe comunal de Las Condes, Francisco de La Maza, a la presidencia de la colectividad. Mario Olavarría, afirmó que ni el senador Hernán Larraín, ni el diputado Javier Macaya son los que el gremialismo necesita.
>
> “Yo espero que algún día tengamos un gobierno realmente de derecha en nuestro país y ahí podamos reescribir la historia realmente, como corresponde, y no como se ha escrito todos estos años“, agregó.
>
> En la sesión de hoy, concurrió nuevamente el ex jefe de la División de Administración y Finanzas, Gabriel Aldana, dado que los parlamentarios querían aclarar sus afirmaciones a la luz de otras versiones ante la Comisión, emitidas con posterioridad a su primer testimonio.
>
> La mantención de la cautelar fue valorada por el abogado de Carlos Lavín y Carlos Délano, Julián López. “El hecho de estar sometido a una medida cautelar de menor intensidad facilita el trabajo de la defensa, que nos va a permitir demostrar que la participación que se les atribuye a mis representados no es la que la fiscalía y los acusados han sostenido hasta ahora”, sostuvo.

Como era esperable, para el texto recién presentado el puntaje fue de 0,26. Lo mismo sucedió en los otros nueve textos construidos con puntajes que fluctuaron

entre 0,21 a 0,34. Una última prueba realizada, es la comparación de la misma noticia publicada en diferentes medios. Sin duda, ésta es una de las experiencias más interesantes porque compara textos de diferentes fuentes, pero referidos a un mismo hecho. Para ello, se utilizaron textos de La Tercera (www.latercera.com), Emol (www.emol.com), Cooperativa (www.cooperativa.cl) y Bío Bío Chile (www.biobiochile.cl).

A continuación se detallan las cinco pruebas realizadas para el caso de una de esas noticias en particular, para finalmente presentar una tabla y un gráfico que resumen todos los resultados que arrojaron las diez noticias a que se aplicaron estas pruebas.

El texto íntegro de la noticia en que se enfocó el ejemplo, titulada “Mariana Aylwin entrega descargos a tribunal supremo DC y dice que solicitud de expulsión es injusta" (La Tercera), se comparó con el texto de “Mariana Aylwin y petición de expulsión: Vengo a defenderme de una acusación infundada e injusta” (Emol). Los resultados de dicha comparación, entre dos textos diferentes pero enfocados exactamente en el mismo tema, arrojaron un puntaje de 0,94, que indica una altísima relación semántica entre ambos escritos, lo que demuestra la efectividad de la medición realizada.

En una segunda prueba, se mantuvo el texto publicado por La Tercera, pero ahora se le comparó con otra noticia relacionada que publica Emol, sobre la opinión del presidente del partido político aludido en las noticias anteriores, titulada “Presidente de la DC: Más que un tema disciplinario es político”. El resultado, obviamente, disminuye a 0,78. Dicho puntaje también expresa una alta relación semántica entre los textos, ya que tratan sobre el mismo tema, pero la baja en el puntaje se explica porque no son la misma noticia, sino que sólo son textos relacionados.

En una tercera prueba, se comparó la misma noticia de La Tercera con un texto sobre política de Emol, pero que no tenía relación con el fondo de la noticia original. Este último se titula “Rossi abandona indignado comisión que analiza indicaciones de la reforma educacional”. En esta comparación, el puntaje disminuyó aún más, al arrojar sólo 0,54, lo que da a entender que se trata de noticias diferentes, aunque pertenecientes a un mismo dominio temático (política).

En una cuarta prueba, se comparó el mismo texto utilizado en los testeos anteriores, con una noticia ajena al dominio temático elegido, titulada “Quintanilla repite su tercer lugar y avanza a la cuarta posición de la General del Dakar en motos”. En esta ocasión, el puntaje entregado por el módulo fue de 0,36, que es bastante bajo, como era esperable.

Por último, en una quinta prueba se comparó el texto original con una noticia de un dominio temático diverso, pero además externa a Chile, país en el que se produjeron todas las noticias anteriores. La nota se titula “Actor de Star Wars y James Bond muere durante ensayo de obra de teatro” y en la comparación se obtuvo el más bajo puntaje de todas las pruebas con sólo 0,32.

En lo referente a la comparación de textos completos, se realizaron las cinco pruebas descritas anteriormente a diez noticias en total, como ya se señaló, y en todos los casos, las bajas de puntaje presentaron tendencias similares. En seis de las pruebas se utilizó como texto base uno de La Tercera, que es el medio del que se extrajeron las noticias con que se construyó el corpus empleado para crear el espacio semántico, y se las comparó con cinco noticias extraídas de otros medios: Emol, Cooperativa y Bío Bío Chile. En dos pruebas se utilizó una noticia de Emol como base y se las comparó, en cada caso, con cinco noticias sacadas de La Tercera. Por último, se realizaron dos pruebas utilizando como base una noticia de Cooperativa y otra de Bío Bío Chile, con el fin de realizar la experiencia en textos publicados en sitios informativos que no tienen su raigambre en la prensa escrita como Emol y La Tercera, sino que en radios. La idea de la experiencia era utilizar en su mayoría textos que no fueran sacados de La Tercera, por ello en total -incluyendo textos base y de comparación- se emplearon 16 extraídos del citado medio y 44 de otros: 12 de Emol, 16 de Cooperativa y 16 de Bío Bío Chile. A lo anterior, hay que añadir que la selección de los textos base se realizó al azar de entre noticias recopiladas desde el sitio de La Tercera a lo largo de seis meses y que no forman parte del corpus integrado en el módulo. El universo total para esta selección fue de 2510 textos. Se eligieron diez, y de ésos, seis se utilizaron como texto base (los de La Tercera); para los cuatro casos restantes, se tomó el tema de la noticia de La Tercera y se buscó en los otros medios ya señalados la noticia equivalente. En la Tabla I se presentan los resultados de toda la experiencia recién descrita.

Tabla I: Resultados de la comparación de noticias completas pertenecientes a distintos medios.

Título noticia base	Medio texto base	Prueba 1	Prueba 2	Prueba 3	Prueba 4	Prueba 5	Medio textos comparados
Mariana Aylwin entrega descargos a tribunal supremo DC y dice que solicitud de expulsión es injusta	La Tercera	0,94	0,78	0,54	0,37	0,32	Emol
Fulvio Rossi por dichos de rector Sánchez sobre aborto: "Aquí nadie está por sobre la ley, y él tampoco"	La Tercera	0,86	0,75	0,55	0,41	0,45	Emol
De la Maza renuncia a directiva UDI y se agudiza crisis interna	La Tercera	0,89	0,84	0,6	0,42	0,32	Bío Bío
Diputado que pidió minuto de silencio por Pinochet: "Cómo se le ocurre que voy a estar arrepentido"	La Tercera	0,92	0,78	0,56	0,37	0,3	Bío Bío
Velasco reaparece junto a Amplitud y profundiza distancias con Nueva Mayoría	La Tercera	0,9	0,79	0,51	0,4	0,28	Cooperativa
Pizarro y Chahín fijan plazo de 48 horas para explorar lista unitaria	La Tercera	0,87	0,76	0,59	0,31	0,23	Cooperativa
Bachelet cita a comité político extraordinario para dar instrucciones sobre agenda anticorrupción	Emol	0,83	0,65	0,51	0,34	0,3	La Tercera
SQM: Contadora admitió entrega irregular de dinero a Martelli para campaña presidencial de Frei	Emol	0,94	0,78	0,52	0,4	0,35	La Tercera
Corte mantiene en arresto domiciliario a ex controladores de Penta	Bío Bío	0,88	0,87	0,54	0,45	0,3	Cooperativa
Ministra Ximena Rincón alertó que su correo electrónico fue hackeado	Cooperativa	0,91	0,78	0,56	0,44	0,43	Bío Bío

Como se puede ver en la Tabla I, la tendencia a la baja de los puntajes es similar en todos los casos. La única vez en que esto no se cumple es en la prueba 4 de la segunda noticia, ya que el módulo arrojó que la relación semántica entre la noticia base y una noticia nacional ajena al dominio fue de 0,41, en circunstancias que la comparación con la noticia internacional ajena al dominio fue de 0,45. Se considera como no relevante el caso anterior, sobre todo por la tendencia a no haber una diferenciación significativa entre las comparaciones de la pruebas 4 y 5, en donde los puntajes tienen una separación cuantitativa mucho menor que el resto de las pruebas. Por último, en el Gráfico 1 se muestra el promedio de los puntajes obtenidos en las cinco pruebas realizadas a las diez noticias, con el fin de ver cómo varía la medición de una forma general, a medida que se reduce la relación semántica entre los textos comparados.

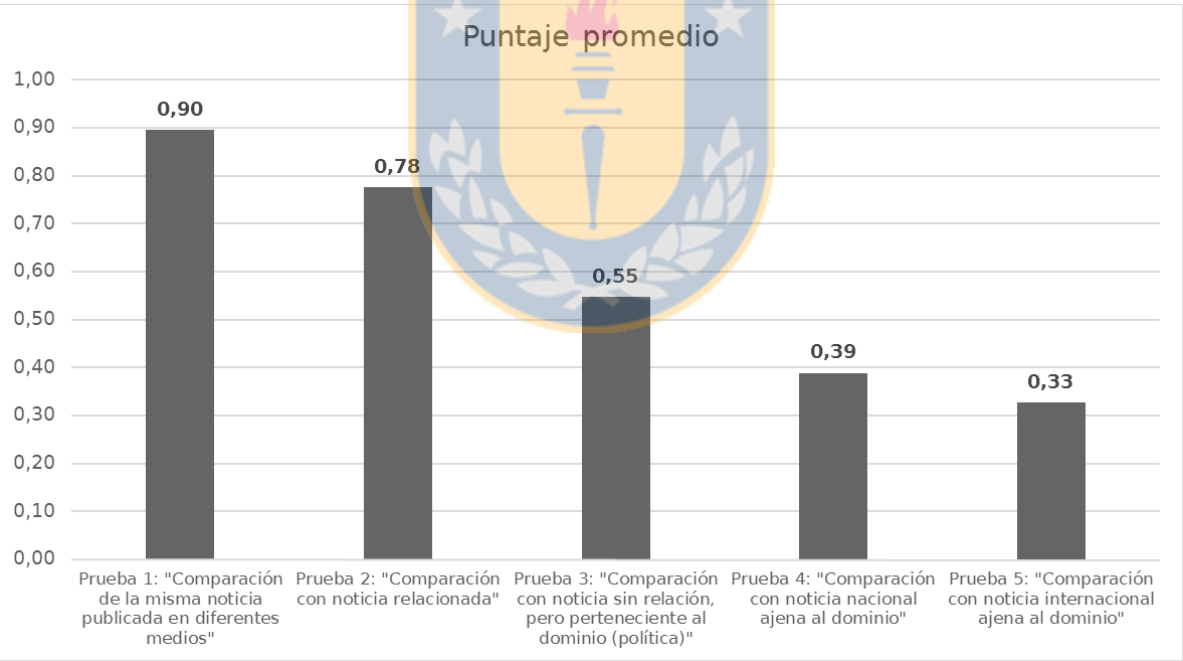

Gráfico 1: Puntajes promedio de todas las pruebas.

4.5. Consideraciones sobre las pruebas realizadas al módulo 1

De las pruebas efectuadas en el testeo del módulo se puede colegir que no afecta a su correcto funcionamiento, el que las noticias analizadas provengan de un medio de prensa distinto al utilizado para recopilar las noticias que conforman el

corpus, en base al cual se construyó el espacio semántico multidimensional -La Tercera en este caso-. Si bien esto era algo esperable, ya que en un medio de prensa no es un solo periodista quien escribe todos los textos, sino que un equipo de redacción, era importante despejar la duda mediante pruebas empíricas. Lo anterior, es algo fundamental para cualquier utilización futura del módulo, ya que permite procesar textos escritos por cualquier autor y, en este caso en particular, estudiantes de periodismo.

Es importante, además, señalar que en el testeo realizado hubo que corregir manualmente la ortografía de los textos utilizados, ya que los errores de este tipo alteraban los puntajes arrojados por el módulo 1. Por lo mismo, si bien no está entre los objetivos de este trabajo, se determinó la necesidad de construir un módulo corrector ortográfico para el componente, que pueda trabajar conjuntamente con el módulo 1 (y también con el módulo 3, como se verá más adelante); más aún, considerando que los textos a analizar serán de estudiantes de pregrado.

Para finalizar, es necesario señalar que la herramienta diseñada, al basarse en *scripts* de texto, permite su adaptación a múltiples proyectos en que se requiera realizar análisis semántico. De hecho, bastaría con cambiar el corpus a partir del cual se construye el espacio semántico y adaptar algunas líneas de código para tener una herramienta personalizada para trabajar en otro ámbito que se desee, diferente a las noticias sobre política.

Capítulo 5: Diseño de un módulo corrector ortográfico

En el presente capítulo se describirá la construcción de un módulo, a nivel de prototipo, de un analizador y corrector ortográfico, programado en Python 3, que pueda trabajar en forma conjunta con el módulo 1 -analizador semántico- (y con el módulo 3), con el fin de apoyar su funcionamiento ante este tipo específico de error. Éste es el segundo módulo de los tres que integran el componente que tiene como objetivo desarrollar el presente trabajo.

Hay que agregar que en la actualidad este módulo opera en el STI ELE-Tutora (descrito en 1.3.1), como parte de la vinculación del presente trabajo con el Proyecto Conicyt, Fondecyt regular, 1110812: "Un Sistema Tutorial Inteligente para la Focalización en la Forma en la Enseñanza del Español como Lengua Extranjera".

5.1. Errores ortográficos en el texto de entrada

Uno de los problemas que se pueden presentar al emplear Análisis Semántico Latente (LSA) y, en general, cualquier método que trabaje con textos como entrada (*input*) son los errores que este texto pudiera tener en su superficie. Específicamente, en el caso del LSA, los errores que más afectan al resultado final son los errores ortográficos. No así otro tipo de errores como, por ejemplo, los de concordancia que, generalmente, perderían relevancia tras el proceso de lematización.

Entendiendo que el LSA es un enfoque computacional del significado de las palabras que se basa en la coocurrencia de éstas en textos, a partir de los cuales se derivan los espacios semánticos que reflejan las relaciones de significado entre las palabras, es fácil imaginarse que si una palabra está mal escrita va a tener un impacto en el resultado final. Este error puede producirse en dos etapas: en los textos a partir de los que se construye el espacio semántico multidimensional o en los textos que se procesen a la luz de este espacio semántico para su análisis.

En el caso de los textos a partir de los cuales se deriva el espacio semántico, si tuvieran errores el impacto va a ser ínfimo, ya que se trata de grandes cantidades de texto. En el caso del módulo 1, descrito en el capítulo anterior, se trata de 7.165 textos, con un total de 3.042.957 palabras; por ello, si una palabra aparece mal

escrita, la probabilidad indica que saldrá escrita correctamente en bastantes más ocasiones; por otra parte, si aquella fuera la única vez que aparece, o sólo apareciera unas pocas más, se trataría de una palabra sin relevancia para el dominio temático del corpus.

Diferente es el caso de un texto en particular que se está procesando a la luz del espacio semántico. Aquí un error ortográfico sí va afectar la medición. Por ejemplo, en uno de los textos presentados en el capítulo anterior, una noticia de La Tercera titulada *Walker:* "*Esto es un tema político más que disciplinario*", cuyo puntaje de coherencia es de 0,67, se introdujeron dos errores ortográficos en el primer párrafo: se cambió *Walker* por *Wallker* y *política* por *pokítica,* lo que produjo un descenso del puntaje a 0,63. Si bien puede aparecer que la diferencia de puntaje no es demasiado grande al considerar el promedio del texto (0,04 puntos), si nos centramos sólo en las líneas implicadas tenemos que el puntaje original entre el titular y el primer párrafo desciende de 0,66 a 0,52 (0,14 puntos); y entre el primer y segundo párrafo baja de 0,74 a 0,65 (0,9 puntos), lo que da una idea clara del impacto no menor que pueden tener los errores ortográficos en el desempeño del módulo 1 al analizar un texto en particular.

Por lo anterior, se consideró necesaria la etapa de desarrollar un módulo que revisara la ortografía de los textos de entrada, con el fin de que opere conjuntamente con el módulo 1 y apoye su funcionamiento.

5.2. Arquitectura y funcionamiento del corrector ortográfico

A continuación se explica cómo se abordó la construcción de un corrector ortográfico para el español, programado en Python 3. El trabajo se realizó en un equipo con Linux, pero el código final opera sin problemas en Windows. El presente apartado se centrará en la arquitectura propuesta para éste (Figura 6). Los números en el esquema se utilizan para apoyar la explicación de cada etapa que se dará en los puntos siguientes.

La entrada que utiliza el corrector ortográfico, empleando la terminología de Python, es una cadena de texto. Dicha cadena se procesa de dos formas, según se ve en la Figura 6: palabra a palabra (punto 1) y como una cadena en busca de patrones de error complejos (punto 8). Primero se analizará el procesamiento

palabra a palabra.

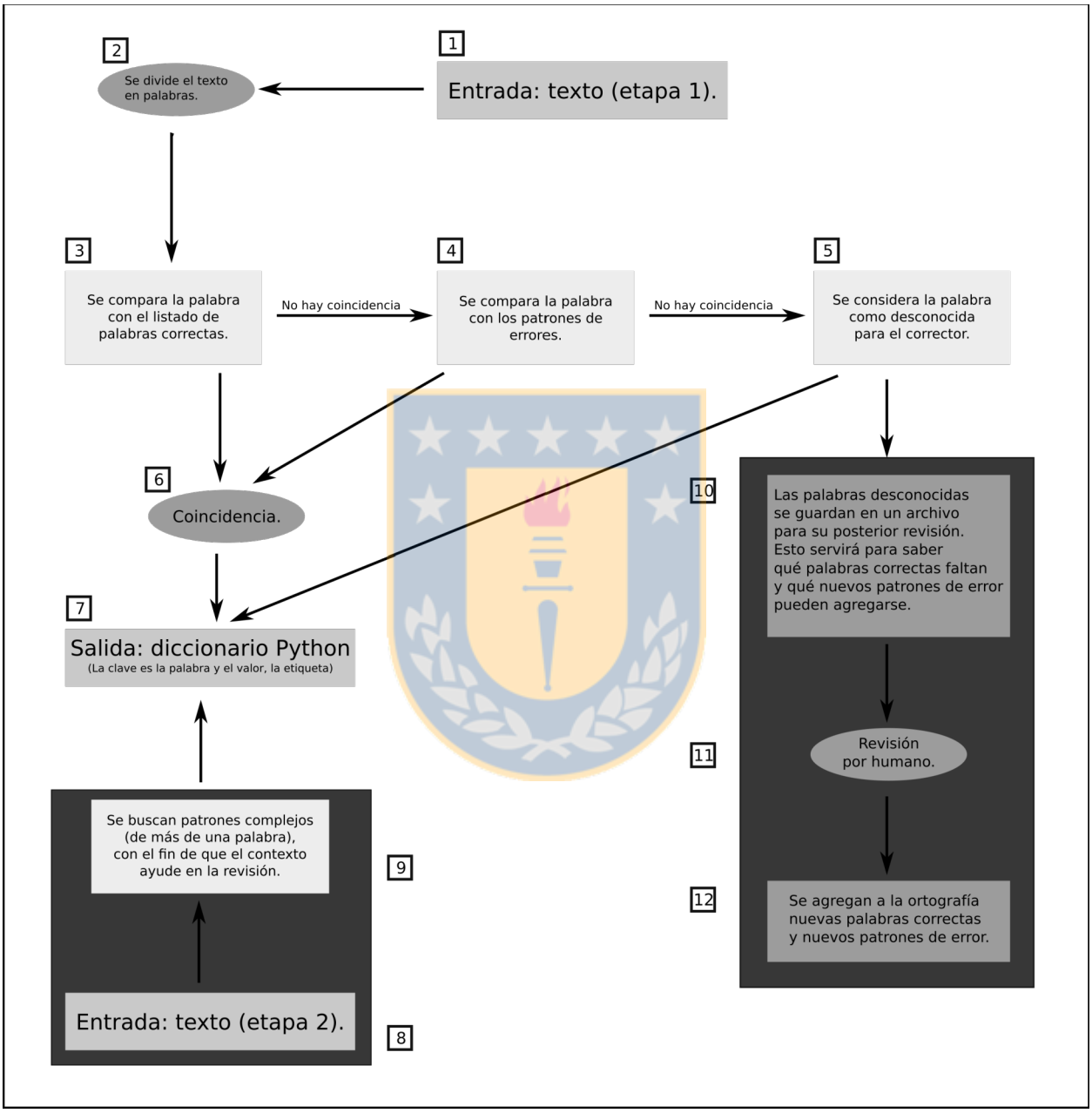

Figura 6: Arquitectura del corrector ortográfico.

5.2.1. Procesamiento palabra a palabra

En esta etapa, como se indica en el punto 2, la cadena de texto se divide en palabras independientes para su procesamiento. Esto se hace gracias a la función

tokenize incluida en NLTK (Natural Language Toolkit, disponible en http://www.nltk.org/), la que permite separar las palabras de una cadena de los signos de puntuación de la misma. Luego, se seleccionan sólo los elementos que contienen grafemas del español, con sus correspondientes diacríticos si es que fuera el caso. En otras palabras, se seleccionan sólo los elementos que contengan las letras de nuestro alfabeto (incluida la eñe), más las vocales tildadas y la vocal ü (con diéresis). Lo anterior deja fuera del análisis a los signos de puntuación, los números y cualquier otro signo ortográfico ajeno al español, como por ejemplo, los acentos breve (˘) y circunflejo (^). La idea de excluir los signos de puntuación, obviamente importantísimos en un corrector ortográfico, es porque esta primera etapa del procesamiento se centra sólo en la escritura de cada palabra como unidad. De esta forma, por ejemplo, se busca detectar errores como escribir *aviación* de la siguiente manera: *aviasion*.

La revisión que realiza el corrector para cada palabra, comienza cotejando si ésta se encuentra en un listado de palabras correctas del español (punto 3). Siguiendo con el ejemplo anterior, si el término de entrada fuera *aviación*, el corrector lo encontraría en el listado aludido y lo reconocería como bien escrito. En este caso, al hallarse la coincidencia que se señala en el punto 6, la palabra se almacenaría como salida en un diccionario de Python llamado *resultados simples*, se etiquetaría en este diccionario como sin error (con la etiqueta 'SIN_ERROR') y se obviarían los pasos de los puntos 4 y 5 (más adelante se explicarán el funcionamiento del diccionario y las etiquetas en el corrector).

En una primera versión, este listado de palabras correctas se construyó en base a expresiones regulares (de un modo análogo a la búsqueda de patrones de error, como se verá más adelante). Sin embargo, el procesamiento de un texto de 10 palabras tomaba algo más de 7 segundos. Por ello, en una segunda versión, se prefirió no emplear las expresiones regulares y se optó por confeccionar el listado como un simple archivo de texto plano, en el cual las palabras están separadas por un salto de línea y tienen un espacio en blanco antes de cada una de ellas. El corrector, entonces, realiza una búsqueda para cotejar si la palabra se encuentra o no en el listado (el archivo en que se almacenan se llama *listado_palabras_original*): si está en el listado la considera escrita correctamente, en caso contrario pasa a la

etapa de búsqueda de errores, como se verá más adelante. El espacio en blanco y el salto de línea son una forma arbitraria de marcar el inicio y el final de cada palabra (así como en las expresiones regulares se realiza esto con “^” y “$”), con el fin de evitar que haya falsas coincidencias; si no existieran estos límites y, por ejemplo, se procesara la forma verbal mal escrita *beo* (*veo*), el corrector arrojaría una coincidencia, entre otras, con las palabras *beodo* o *beorí* y consideraría a *beo* como escrita correctamente. Con la forma de cotejo recién explicada, a diferencia del uso de expresiones regulares, los tiempos de procesamiento para el mismo texto de 10 palabras disminuyeron a menos de un segundo. Lo anterior se basa en la siguiente función escrita en Python 3:

```
1 def palabras_correctas(patron):
2   with open(os.path.join(os.getcwd() + '/listado_palabras/listado_palabras_original'), 'r',
encoding='utf8') as correctas:
3       for line in correctas:
4           if (' '+patron+'\n') in line:
5               return('encontrada')
```

En este punto es importante detenerse a explicar cómo se llega a obtener un listado de los lemas del español y su variabilidad léxica.

Lamentablemente, a día de hoy, no existe en Internet un lexicón de nuestra lengua que sea *open source* y que permita su descarga para utilizarse en proyectos. Sólo se puede conseguir un listado de los lemas incluidos en el Diccionario de la Lengua Española de la Real Academia, lo que obviamente es insuficiente a la hora de pensar en el corrector ortográfico, ya que no se incluyen las formas derivadas (por ejemplo, aparece *perro*, pero no *perrito* ni *perritos*; aparece el verbo *cantar*, pero no sus conjugaciones en los diferentes modos y tiempos). Para solucionar este problema, se desarrolló un *script* en Python, anexo al corrector ortográfico, que permite crear un listado de palabras del español y sus formas derivadas a partir de un texto cualquiera (obviamente que esté en español y se encuentre en un formato apropiado). El funcionamiento del *script* requiere ubicar los textos que se desee utilizar como entrada en la carpeta llamada *textos* (todos los archivos de texto plano que se quiera) y ejecutar el *script*. Éste devolverá una lista ordenada de las palabras únicas utilizadas en todos los textos de entrada. Así, por ejemplo, si en

éstos aparece mil veces la palabra *perritos*, el *script* sólo la incluirá una vez en el listado de salida. Un punto importante a tener en cuenta es que los textos de entrada estén correctamente escritos, ya que si alguno de ellos incluye errores ortográficos, éstos se traspasarán al listado de palabras correctas, lo que iría en desmedro del funcionamiento posterior del corrector. Para finalizar el punto, como textos de entrada para construir el listado de palabras correctas del español se utilizaron los lemas de nuestra lengua contenidos en el diccionario de la RAE y algunas obras en español que se pueden descargar desde el sitio web del Proyecto Gutenberg (http://www.gutenberg.org/browse/languages/es). En este último caso, hubo que revisar que los textos no estuvieran escritos en variantes antiguas de nuestra lengua. Con esto se logró un listado bastante más completo de las palabras correctas de nuestra lengua, que incluye los lemas del Diccionario de la Lengua Española de la RAE y, además, muchas formas derivadas. Como es de suponer, no se puede afirmar que el listado de palabras obtenido sea completo e incluya todas las palabras de nuestra lengua y sus posibles derivaciones. Por ello, como se verá más adelante, el corrector ortográfico diseñado se planteó como un corrector dinámico, que permita agregar nuevas palabras al listado original.

Retomando el procesamiento de un texto dado por el corrector, en el caso de que no haya coincidencia en el cotejo del punto 3 (listado de palabras correctas), se pasa al punto 4. En éste se compara la palabra con distintos patrones de error, construidos en base a expresiones regulares[6], que se almacenan en un archivo de texto plano. Dichas expresiones se leen del archivo y se almacenan como tuplas[7] de Python, en un listado llamado *errores_originales_simples*. La lectura de las expresiones regulares desde el archivo se realiza con el siguiente código:

6 Las expresiones regulares (llamadas RE -por sus iniciales en inglés- o patrones de expresiones regulares) son, esencialmente, un sencillo y altamente especializado lenguaje de programación incrustado dentro de Python y puesto a disposición a través del módulo re. Al usar este sencillo lenguaje, se deben especificar las reglas para el conjunto de posibles cadenas que se desean buscar; este conjunto puede contener frases en inglés (o en otra lengua), o direcciones de correo electrónico, o lo que se quiera. También se pueden utilizar las expresiones regulares para modificar una cadena de texto o dividirla de diferentes maneras (documentación de Python Software Foundation, disponible en https://docs.python.org/3/howto/regex.html?highlight=regular%20expression).

7 Las tuplas son secuencias inmutables, normalmente utilizadas para almacenar colecciones de datos heterogéneos y se escriben entre paréntesis normales: (). Las tuplas también se utilizan para casos donde se requiere una secuencia inmutable de datos homogéneos (documentación de Python Software Foundation, disponible en https://docs.python.org/3/library/stdtypes.html?highlight=tuple#tuple).

```
errores_originales_simples = []
for linea in codecs.open(os.path.join(os.getcwd() +
'/errores_originales/errores_simples'), 'r', encoding='utf8'):
    tupla = ast.literal_eval(linea)
    errores_originales_simples.append(tupla)
```

La función[8] en que se basa el cotejo de las palabras de entrada con los patrones de error es la siguiente:

```
def corregir_ortografia_simples(sentencia):
    outext = ''
    error = ''
    s = sentencia
    for couple in buscar_reemplazar_simples:
        correcta = re.sub(couple[0], couple[1], s)
        incorrecta = re.search(couple[0], s)
        if incorrecta:
            return[incorrecta.group(0), correcta, couple[2]]
        s=correcta
    return ["", "", ""]
```

Un ejemplo abreviado de una expresión regular utilizada se presenta a continuación:

```
(r'''(^[Aa]bdica|^[Zz]onifica)+(cion$)''', r'''\1ción''', "ORTO_OMI_AGUDA"),
```

En el ejemplo anterior, se buscan las palabras *abdicación* o *zonificación* escritas sin tilde (con error). En caso de encontrar una coincidencia con alguna de ellas, se obtiene como salida una lista[9] de Python con el patrón encontrado (la palabra con error), la forma correcta y la etiqueta que identifica al error. Las etiquetas de error que se utilizaron son las propuestas por Ferreira et al. (2014), considerando únicamente las enfocadas en los errores ortográficos. Esta lista, al igual que las palabras correctas encontradas (como ya se explicó), se almacena en el diccionario de Python que arroja como salida final el corrector ortográfico. Un

8 Esta función se basa en una escrita por el Prof. Daniel Campos, a quien se agradece el facilitarla con el fin de incluirla en este trabajo.

9 Las listas de Python son un tipo de datos compuestos -el más versátil de los que incluye este lenguaje-, que se utilizan para agrupar otros valores. Se escriben como una lista de valores separados por comas (ítems) encerrados entre corchetes: []. Las listas pueden contener elementos de diferentes tipos pero, por lo general, se utilizan para agrupar elementos de un mismo tipo. Por ejemplo, un listado de números de teléfonos móviles: [98563487, 96846389, 99140711] (documentación de Python Software Foundation, disponible en https://docs.python.org/3.5/tutorial/introduction.html#lists).

ejemplo de esta lista, para el caso de abdicación es:

```
['abdicacion', 'abdicación', 'OMI_ORTO_AGUDA']
```

En caso de no haber una coincidencia con algún patrón de error, y asumiendo que en el paso anterior tampoco la hubo con alguna palabra correcta, el sistema continúa a la siguiente etapa (punto 5). Aquí, los casos de palabras que no arrojan coincidencia en ninguno de los pasos anteriores, se etiquetan como palabras desconocidas para el sistema (la etiqueta empleada es 'PALABRA_DESCONOCIDA'). Éstas también se agregan al diccionario de Python, que contiene las salidas de las dos etapas anteriores.

El citado diccionario -junto con las listas y las tuplas- es una forma de almacenar datos en Python y se compone de dos elementos: clave y valor. Cada una de estas parejas debe tener siempre estos dos componentes y un diccionario puede incluir tantas parejas como se quiera, separadas unas de otras por una coma. A continuación se presenta un ejemplo de diccionario en Python:

```
diccionario = {"clave":"valor", "clave":"valor",...}
```

Los datos del diccionario[10] pueden ser recuperados posteriormente, con el fin de realizar las operaciones que se deseen, como por ejemplo, corregir el texto o entregar *feedback* a la persona que escribe sobre qué errores cometió. Con el fin de mostrar los resultados del corrector ortográfico, se ingresó en éste el siguiente pasaje:

> A la salida del estadio de Concepcion, los hinchas lanzaron piedras contra el bus de Wanderers.

10 Para ampliar un poco más la explicación dada, se puede decir que la mejor forma de concebir un diccionario de Python, es pensar en éste como un conjunto desordenado de claves y valores, con el requisito de que las claves sean únicas (dentro de un diccionario). Un par de llaves crea un diccionario vacío: {}. Dentro de éstos, se colocan separados por *dos puntos* las claves y sus correspondientes valores y cada una de estas parejas se separa de otras mediante *comas*. Las principales operaciones en un diccionario son almacenar un valor con alguna clave y extraer el valor utilizando la clave. Los diccionarios pueden incluir listas, tuplas o, incluso, otros diccionarios. Por ejemplo, imaginemos utilizar un diccionario de Python para almacenar números de teléfonos, asociados a los nombres de sus propietarios; así cada nombre propio será una clave y el número de teléfono su valor: {'pedro': 98563487, 'juan': 96846389, 'diego': 99140711} (documentación de Python Software Foundation, disponible en https://docs.python.org/3.5/tutorial/datastructures.html#dictionaries).

Con el fin de tener una muestra de todas las posibilidades de salida en el diccionario resultante, se introdujo a propósito un error ortográfico al no tildar Concepción. Además, el nombre Wanderers es desconocido para el corrector pues no ha sido incorporado a éste aún. La salida que se obtiene es la que sigue:

{'Concepcion': ['Concepcion', 'Concepción', 'OMI_ORTO_AGUDA'], 'de': 'SIN_ERROR', 'del': 'SIN_ERROR', 'hinchas': 'SIN_ERROR', 'la': 'SIN_ERROR', 'lanzaron': 'SIN_ERROR', 'bus': 'SIN_ERROR', 'contra': 'SIN_ERROR', 'salida': 'SIN_ERROR', 'los': 'SIN_ERROR', 'A': 'SIN_ERROR', 'Wanderers': 'PALABRA_DESCONOCIDA', 'el': 'SIN_ERROR', 'piedras': 'SIN_ERROR', 'estadio': 'SIN_ERROR'}

Como se puede ver en todos los casos, la clave es la palabra ingresada. El valor, por otro lado, en la mayoría de los casos es la etiqueta SIN_ERROR. En dos palabras, sin embargo, esto difiere. Como ya se adelantó, el nombre *Wanderers* es desconocido para el corrector, por lo tanto, como valor se obtiene la etiqueta PALABRA_DESCONOCIDA. La otra salida distinta es justamente para la palabra en que se omitió tildar, esto es, *Concepcion*. En este caso el valor es una lista de Python que contiene la palabra con el error, su forma correcta y la etiqueta que identifica el error, es decir, la omisión de la tilde en una palabra aguda: OMI_ORTO_AGUDA.

Todo lo explicado hasta aquí es para realizar el análisis de las palabras del texto aisladas, es decir, consideradas cada una de ellas como unidad.

5.2.2. Procesamiento como una cadena en busca de patrones de error complejos

En el punto 8 de la Figura 6, se toma nuevamente el mismo texto de entrada, pero esta vez no se analiza palabra a palabra, sino que se revisa toda la cadena en busca de patrones de error complejos (punto 9) que necesitan de contexto para poder identificarlos. Por ejemplo, consideremos la siguiente oración como texto de entrada:

No se que pasa en México.

En el pasaje anterior se omitieron intencionadamente las tildes de *se* y *que*. Si esta oración se sometiera al análisis palabra a palabra, no arrojaría error alguno, ya que las formas *se* y *que* sin tilde existen en el español. Sin embargo, al leerlas en contexto podemos saber que *se* no es un pronombre personal, sino que la primera persona del presente de indicativo del verbo saber; y *que* es una forma interrogativa indirecta y no el pronombre relativo. Por lo tanto, si se busca *no se que* como expresión regular, cuyo patrón complejo de error ha sido incorporado previamente al corrector, la aplicación puede reconocer que hay dos errores ortográficos en la oración y arrojar como salida la forma correcta *no sé qué*. En el siguiente ejemplo se presenta la salida que arroja el corrector tras procesar el pasaje *no se que pasa en México* (con los tildes omitidos en *sé* y *qué*). Como se puede ver, en primer lugar, la salida presenta *no se que* con error y como una construcción compleja y después cada una de las palabras por separado son catalogadas como *sin error*, incluso a las que se omitió el tilde. Esto demuestra lo que permiten las reglas contextuales:

```
{'No se que': ['No se que', 'No sé qué', 'ORTO_OMI_TILDES'], 'pasa': 'SIN_ERROR', 'se': 'SIN_ERROR', 'en': 'SIN_ERROR', 'No': 'SIN_ERROR', 'que': 'SIN_ERROR', 'México': 'SIN_ERROR'}
```

Una de las limitaciones más obvias de estas reglas contextuales es la infinidad de combinaciones que ofrece el español, por lo que sería imposible preverlas todas; sin embargo, son una forma eficiente de solucionar casos como el recién expuesto y lograr que el corrector ortográfico funcione de forma más precisa. (Al igual que en el caso de las palabras aisladas, también pueden añadirse nuevos patrones complejos de error gracias a la arquitectura dinámica del corrector, que se explicará en el apartado siguiente).

En este caso, los patrones de error complejos también se almacenan en un archivo de texto y se leen con el siguiente código:

```
1 errores_originales_complejos = []
2 for linea in codecs.open(os.path.join(os.getcwd() +
'/errores_originales/errores_complejos'), 'r', encoding='utf8'):
```

```
3    tupla = ast.literal_eval(linea)
4    errores_originales_complejos.append(tupla)
```

5.2.3. La necesidad de un corrector ortográfico dinámico

De lo expuesto en 5.2.1 y 5.2.2 se puede prever que el corrector adolecerá de dos problemas con total certeza: (1) no se puede asegurar que la lista de palabras correctas constituirá un listado total de toda la variabilidad léxica de los lemas del español, ni tampoco que contendrá los nombres propios que podría utilizar un hablante de nuestra lengua; (2) menos aún se puede afirmar que en los patrones simples de error (palabras) y complejos (contextuales) se contemplarán todas las posibilidades de error que podrían cometerse en nuestro idioma. Lo anterior se ve agravado porque el corrector ortográfico se basa en un *script* programado en Python, lenguaje que no todo el mundo maneja, lo que complica la tarea de ingresar nuevas palabras y nuevos patrones de error. Una dificultad adicional es que este corrector ortográfico está pensado para ser utilizado -junto con los otros módulos del componente- en una aplicación montada en un servidor (por ejemplo, un sistema tutorial inteligente, aunque perfectamente podría adaptarse a otras aplicaciones por la versatilidad de Python), por lo que para modificarlo debería tenerse acceso directo al servidor. En caso de poder acceder, hay que tener en cuenta que muchos servidores están montados en Linux y muchas veces sin interfaz gráfica, por lo que cualquier operación se realiza a través de comandos de textos. Otra alternativa sería conectarse remotamente al servidor y editar el código del corrector, lo que tampoco es tarea sencilla.

Todo lo descrito en el párrafo anterior deviene en que el corrector, una vez terminado, se transformaría en una aplicación rígida, que sería muy difícil de modificar por quienes administren el sistema en que éste se utilice. Por esto, se propone que el corrector tenga una estructura dinámica y que pueda ser modificado por una persona sin conocimientos de Python a través de una interfaz gráfica preparada para ello. Para lograrlo es imperativo contemplar esta posibilidad en la arquitectura del corrector y desarrollar el código necesario para este funcionamiento dinámico, y conectarlo posteriormente con la interfaz gráfica del

área de administración que utilice la aplicación en que se inserte (a modo de ejemplo, podría utilizarse en un gestor de contenidos como Moodle o alguna aplicación hecha a medida).

Los puntos 10, 11 y 12 de la arquitectura propuesta se enfocan en esta parte del corrector ortográfico. A partir del punto 5, cuando una palabra se califica como desconocida, puede ser por dos razones: la palabra está correctamente escrita, pero no se incluye en el listado de lemas y derivaciones que tiene el corrector; o la palabra contiene uno o más errores ortográficos, pero el patrón de error no está en el corrector. También puede darse el caso de que el sistema no arroje la etiqueta palabra desconocida, ya que para reconocer el error se necesita de un patrón complejo que no existe entre los incluidos (por ejemplo, que no hubiera una regla para reconocer *no sé qué* cuando se omiten las tildes). Todos estos problemas pueden ser detectados por las personas que administren el sistema; sin embargo, solucionarlos no es algo sencillo, como ya se explicó.

El primer paso para enfrentar el punto es que el *script* construido guarda las palabras etiquetadas como desconocidas en un archivo de texto plano, como un listado, separadas por saltos de línea (también podría ser separadas por comas, si se quisiera; nos parece que el listado facilita su lectura por un humano). De esta forma, este archivo puede ser fácilmente leído desde la interfaz gráfica antes propuesta, para su revisión por los administradores del sistema web en que operará el corrector. A lo anterior, hay que agregar que el código que realiza la lectura de este archivo, de una forma análoga a la explicada para la creación del listado de palabras correctas, eliminará las palabras repetidas, para facilitar el trabajo del revisor. Desde esta misma interfaz -aclaramos, interfaz hipotética, pues en este trabajo sólo se construyó el *script*-, quien administre podría seleccionar las palabras de ese listado que estén correctamente escritas y agregarlas a través de un formulario al archivo adicional de palabras correctas que posee el corrector ortográfico y, posteriormente, éste las leerá para realizar el cotejo de las entradas de texto que reciba. La idea de crear un archivo adicional fue con el fin de mantener separado el listado original (denominado listado_palabras_original) del

listado de palabras correctas agregadas posteriormente (denominado listado_palabras_agregadas), con el fin de mantener un mayor control sobre las modificaciones que se realicen (por ejemplo, que el revisor cometiera un error al agregar una palabra). Para lograr esto último, sólo hubo que modificar la función palabras_correctas que se mostró anteriormente, para que quede de la siguiente forma:

```
1 def palabras_correctas(patron):
2     with open(os.path.join(os.getcwd() + '/listado_palabras/listado_palabras_original'), 'r',
encoding='utf8') as correctas1, open(os.path.join(os.getcwd() +
'/listado_palabras/listado_palabras_agregadas'), 'r', encoding='utf8') as correctas2:
3         for line in correctas1, correctas2:
4             if (' '+patron+'\n') in line:
5                 return('encontrada')
```

Para el caso de añadir patrones simples de error, es decir a nivel de palabra, la forma elegida para realizarlo es agregando reglas por cada etiqueta de error, las que están previamente definidas con el fin de mantener el orden del sistema (como se señaló, se utilizaron las etiquetas propuestas por Ferreira et al., 2014). Estos patrones (expresiones regulares) agregados por los administradores o revisores del hipotético sistema también deben almacenarse en un archivo de texto plano; por lo mismo, se consideró un archivo por cada etiqueta (para mantener el orden, como ya se dijo). A continuación se presenta el código para poder leer las expresiones regulares desde un archivo de texto, para la omisión de las tildes en palabras agudas, graves y esdrújulas:

```
1 errores_agregados_simples = []
2 for linea in codecs.open(os.path.join(os.getcwd() +
'/errores_nuevos/patrones_nuevos_ORTO_OMI_AGUDA'), 'r', encoding='utf8'):
3     tupla = ast.literal_eval(linea)
4     errores_agregados_simples.append(tupla)
5 for linea in codecs.open(os.path.join(os.getcwd() +
'/errores_nuevos/patrones_nuevos_OMI_ORTO_LLANA'), 'r', encoding='utf8'):
6     tupla = ast.literal_eval(linea)
7     errores_agregados_simples.append(tupla)
8 for linea in codecs.open(os.path.join(os.getcwd() +
'/errores_nuevos/patrones_nuevos_ORTO_OMI_ESDRUJULA'), 'r', encoding='utf8'):
9     tupla = ast.literal_eval(linea)
10    errores_agregados_simples.append(tupla)
```

El citado código lee los patrones de error almacenados en los siguientes archivos:

- patrones_nuevos_ORTO_OMI_AGUDA
- patrones_nuevos_ORTO_OMI_LLANA
- patrones_nuevos_ORTO_OMI_ESDRUJULA

Para el caso de los patrones complejos de error que se añadan se optó por almacenarlos en un único archivo de texto plano llamado patrones_nuevos_COMPLEJOS. El código utilizado para leerlo es:

```
1 errores_agregados_complejos = []
2 for linea in codecs.open(os.path.join(os.getcwd() +
'/errores_nuevos/patrones_nuevos_COMPLEJOS'), 'r', encoding='utf8'):
3     tupla = ast.literal_eval(linea)
4     errores_agregados_complejos.append(tupla)
```

Hay que dejar en claro que para poder agregar patrones de error, la persona que realice la tarea debe tener conocimiento de expresiones regulares, lo que de todas formas complica la labor. Por ejemplo, para agregar un error que contemplara escribir la palabra *área* sin tilde, la regla a añadir sería la que sigue:

```
(r'''\barea\b''', r'''área''', "ORTO_OMI_ESDRUJULA")
```

En todo caso, lo anterior se podría solucionar creando un asistente que ayude a ingresar la regla, lo que permitiría facilitar la tarea. Dicho asistente podría requerir que la persona sólo deba escribir la palabra en forma incorrecta en un formulario de ingreso de texto, luego la palabra correcta en otro campo de texto y, finalmente, seleccionar la etiqueta que describa el error de un listado. Para lograr esto, debiera programarse un *script* con el fin de que se encargue de automatizar la escritura del resto del código que utilizan las expresiones regulares.

5.2.4. Entrega de resultados

Como resumen de lo hasta aquí presentado, el corrector ortográfico en primer lugar coteja las palabras de entrada con un listado de palabras correctas.

Luego busca los patrones de error almacenados en las siguientes variables creadas en Python:

- errores_originales_simples = []
- errores_originales_complejos = []
- errores_agregados_simples = []
- errores_agregados_complejos = []

Sin embargo, y para respetar el orden de procesamiento, previo al trabajo recién descrito, el corrector combina, por una parte, los patrones de error simples (originales y agregados) en la variable *buscar_reemplazar_simples*; y, por otra, los patrones de error complejos (originales y agregados), en la variable *buscar_reemplazar_complejos*. Lo anterior, con el fin de facilitar el procesamiento posterior. El código que efectúa la tarea descrita es el que sigue:

```
1 buscar_reemplazar_simples = errores_originales_simples+errores_agregados_simples
2 buscar_reemplazar_complejos =
errores_originales_complejos+errores_agregados_complejos
```

Por último, en base a estas dos variables, el corrector realiza las tareas ya descritas a lo largo de los apartados precedentes. Para ello, primero ejecuta el procesamiento palabra a palabra que busca si cada término de la entrada está presente en los listados de palabras correctas (originales y agregadas). Luego pasa a revisar si están presentes los patrones simples de error (originales y agregados). Finalmente, en caso de que no haya resultados en las dos etapas previas, asume que la palabra es desconocida para el sistema. Los resultados de todo lo anterior, los guarda en un diccionario llamado *resultados_simples*. El código que se presenta a continuación realiza las tareas descritas:

```
1 resultados_simples = {}
2 with open('entrada', 'r') as filer:
3     tokenizar = nltk.word_tokenize(filer.read())
4     regex = re.compile('(^[a-zñA-ZÑáéíóúüÁÉÍÓÚÜ]+$)')
5     limpiar = [m.group(1) for l in tokenizar for m in [regex.search(l)] if m]
6     texto_limpio = "\n".join(limpiar)
7     for palabra in texto_limpio.split():
8         if palabras_correctas(palabra) == 'encontrada':
```

```
9           resultados_simples.update({palabra: 'SIN_ERROR'})
10      elif (corregir_ortografia_simples(palabra))[0] == palabra:
11          resultados_simples.update({palabra: (corregir_ortografia_simples(palabra))})
12      else:
13          resultados_simples.update({palabra: 'PALABRA_DESCONOCIDA'})
14          with open(os.path.join(os.getcwd() + '/listado_palabras/desconocidas'), 'a') as
descon:
15              descon.write(palabra+'\n')
```

En el caso de los patrones complejos de error, el siguiente código ejecuta la tarea de buscarlos y almacena sus resultados en el diccionario *resultados_complejos*:

```
1 resultados_complejos = {}
2 with open('entrada', 'r') as filer:
3    tokenizar = nltk.word_tokenize(filer.read())
4    regex = re.compile('(^[a-zñA-ZÑáéíóúüÁÉÍÓÚÜ]+$)')
5    limpiar = [m.group(1) for l in tokenizar for m in [regex.search(l)] if m]
6    texto_limpio_cadena = " ".join(limpiar)
7    for couple in buscar_reemplazar_complejos:
8        correcta = re.sub(couple[0], couple[1], texto_limpio_cadena)
9        incorrecta = re.search(couple[0], texto_limpio_cadena)
10       if incorrecta:
11           salida = [incorrecta.group(0), correcta, couple[2]]
12           resultados_complejos.update({incorrecta.group(0): salida})
```

Finalmente, se combinan los resultados de los dos diccionarios que se obtienen de los pasos anteriores, en un único diccionario llamado simplemente *resultados*. Éste es la salida final de todo el procesamiento realizado. El código para este último paso es:

```
resultados = dict(resultados_simples, **resultados_complejos)
```

Como salida arroja una similar a la ya presentada en el ejemplo expuesto en 5.2.1 que se reproduce nuevamente:

Entrada:
A la salida del estadio de Concepcion, los hinchas lanzaron piedras contra el bus de Wanderers.

Salida:
{'Concepcion': ['Concepcion', 'Concepción', 'OMI_ORTO_AGUDA'], 'de': 'SIN_ERROR', 'del': 'SIN_ERROR', 'hinchas': 'SIN_ERROR', 'la': 'SIN_ERROR', 'lanzaron': 'SIN_ERROR', 'bus': 'SIN_ERROR', 'contra': 'SIN_ERROR', 'salida': 'SIN_ERROR', 'los': 'SIN_ERROR', 'A': 'SIN_ERROR', 'Wanderers': 'PALABRA_DESCONOCIDA', 'el': 'SIN_ERROR', 'piedras': 'SIN_ERROR', 'estadio': 'SIN_ERROR'}

Finalmente, a este diccionario de Python que constituye la salida se aplican los procedimientos descritos en 5.2.3 que hacen que este corrector ortográfico sea dinámico, y pueda mejorar su funcionamiento con el tiempo y la ayuda de un revisor humano.

5.2.5. Sobre el corrector ortográfico diseñado

En relación al planteamiento del corrector ortográfico realizado en los puntos precedentes y la construcción del mismo que se llevó a cabo, es necesario precisar, fundamentalmente, que se tiene plena conciencia de que su funcionamiento no es perfecto; en otras palabras, no es un corrector ortográfico capaz de reconocer todos y cada uno de los errores ortográficos posibles de cometer en la lengua española. De hecho, es probable que tal corrector ortográfico no haya sido construido aún, ya que debería adaptarse a las infinitas alternativas de error que puede cometer una persona al escribir un texto en nuestra lengua. Por ello, para los fines del presente trabajo y de lo que se espera realizar en el futuro, se deben reconocer las limitaciones del corrector construido.

Lo anterior, no quiere decir que se estime que la herramienta diseñada para revisar la ortografía de los textos sea pobre; por el contrario, se reconoce su potencia, sobre todo, por el hecho de ser un corrector dinámico, que puede ser adaptado a los escenarios que vayan surgiendo en la aplicación concreta que se haga de la herramienta (en nuestro caso, corregir los textos de entrada del componente destinado a evaluar la pirámide invertida de las noticias, en el sistema tutorial inteligente que se pretende construir a futuro).

Para las mejoras de la herramienta que se pretenden realizar a futuro, una de las ideas que se maneja, es incorporar en su funcionamiento un analizador sintáctico o *parser*, con el fin de poder determinar la categoría gramatical de las diferentes palabras, para desambiguar de una manera más precisa y funcional

cadenas confusas como la citada *no sé qué* (omitiendo uno o los dos tildes por error). Lo recién descrito sería un paso importante hacia potenciar aún más el funcionamiento de la herramienta, con el fin de volverla más precisa y ampliar su espectro de reconocimiento de errores. Pero, como se dijo, es algo en lo que aún se debe trabajar en su planificación y, posteriormente, en su desarrollo e implementación. La idea es dejar constancia de que el corrector construido se considera una herramienta potente, pero perfectible y es una tarea que queda abierta y planteada.

Capítulo 6: Construcción de un módulo (a nivel de prototipo) para evaluar la jerarquización en la producción escrita de noticias

En el presente capítulo se presentará el proceso seguido para construir y testear el funcionamiento del tercer y último módulo que incluye el componente, esto es, el módulo para evaluar la estructura semántica de pirámide invertida en noticias escritas, enfocado principalmente en el titular y en el *lead*. Este módulo programado en Python 3 asigna un puntaje de -1 a 1 a algunos aspectos específicos de las noticias -como se explicará más adelante-, con el fin de calificar si la jerarquización de los datos que se presentan, en respuesta a las preguntas fundamentales, es adecuada o no. Para testear su correcto funcionamiento, se compararon los resultados de la evaluación del prototipo con la efectuada por humanos.

Dado que el componente en que se integra este tercer módulo (junto a los módulos 1 y 2) busca a futuro insertarse en un sistema tutorial inteligente, destinado a apoyar la enseñanza de la escritura de noticias en estudiantes de periodismo, la experiencia se realizó con textos producidos por estudiantes de Primer Año de la Carrera de Periodismo de la Universidad de Concepción, del curso de Producción de Textos Periodísticos, en el que precisamente aprenden a escribir el tipo de textos en que se enfoca este trabajo: la *noticia* o *información periodística*. Más adelante se explicará cómo se recopilaron estos textos.

6.1. Ampliación del corpus

Landauer y Dumais (1997) señalan que el Análisis Semántico Latente es un método matemático-estadístico que se basa en la coocurrencia de las palabras en grandes corpus textuales. De lo anterior se desprende que una de las condiciones más importantes para su correcto funcionamiento es que cuando se construya el espacio semántico multidimensional, se haga a partir de un corpus conformado por una gran cantidad de textos. Como se indicó en el capítulo 4 de este trabajo, el módulo de análisis semántico enfocado en la predicción de la coherencia textual que se construyó, utiliza para su funcionamiento un espacio semántico creado a

partir de un corpus de 7.165 noticias sobre política, con un total de 3.042.957 palabras.

Obviamente, un corpus de dicho tamaño es considerable. Por ejemplo, en Hernández y Ferreira (2010) se utilizó un corpus de 1.505 noticias policiales, con un total de 386.537 palabras. Por otra parte, Landauer, Foltz y Laham (1998) utilizaron un corpus de 4,5 millones de palabras, que es casi un tercio más grande que el del presente trabajo. Por ello, atendiendo a que durante el desarrollo de la tesis el *script* que recopila el corpus siguió funcionando día a día, se logró reunir más textos y con éstos se conformó un nuevo corpus.

La idea con lo anterior no fue abandonar el corpus original, sino que aprovechar los dos corpus para construir espacios semánticos diferentes con los que contrastar el funcionamiento de la técnica, que se espera sea similar. Hay que señalar que los textos de ambos corpus corresponden a noticias publicadas en el sitio en Internet de La Tercera y ningún texto se repite entre un corpus y otro. Además, se aprovechó la instancia para crear un tercer corpus surgido de la combinación de los dos primeros, con lo que finalmente se contó con tres corpus que permitieron crear tres espacios semánticos distintos en vista a las mediciones que se realizarán. El detalle de lo anterior se presenta en la Tabla II:

Tabla II: Detalle de los corpus utilizados.

Etiqueta del corpus	Cantidad de textos	Cantidad de palabras
DK1 (original)	7.165	3.042.957
DK2 (nuevo)	6.794	3.247.463
DK1y2 (combinado)	13.959	6.290.420

La idea de esto es que se pueden realizar mediciones para comparar la evaluación de los textos entregada por cada uno de los espacios semánticos (creados a partir de cada corpus), lo que sirve para observar mejor el funcionamiento del Análisis Semántico Latente y corroborar lo que se sospecha: que sus evaluaciones debieran ser similares ya que el dominio temático de los corpus es el mismo.

6.2. Construcción del módulo de evaluación de la pirámide invertida

El módulo 1, cuya construcción se trató en el capítulo 4, se centra en el análisis semántico de las noticias, enfocándose específicamente en la predicción de la coherencia textual. Lo anterior fue un paso necesario en el desarrollo de este trabajo de tesis por dos razones: la primera es que como se dijo desde la introducción, la tarea que se describe en estas páginas es parte del objetivo futuro de diseñar un sistema tutorial inteligente (STI) que apoye la enseñanza de escritura de noticias en los estudiantes de periodismo. En dicho sistema, se requerirá de un módulo que evalúe la coherencia textual de los escritos producidos, con el fin de retroalimentar a sus usuarios sobre este aspecto. La segunda razón es que dicho módulo de evaluación de coherencia textual es necesario para el funcionamiento del módulo 3, cuya construcción se describirá a continuación, ya que deben operar en forma conjunta con el fin de que este último aproveche las tareas que realiza el módulo 1; ambos módulos, más el módulo 2 (corrector ortográfico), permiten realizar la tarea final del componente. La idea de este nuevo módulo es evaluar -aprovechando el trabajo de los dos primeros módulos- la correcta jerarquización de los datos de una noticia: estructura de pirámide invertida, esto es, presentación de la información desde lo más importante a lo menos importante, enfocándose especialmente en el titular y el *lead*.

La idea de coordinar el funcionamiento del módulo 1, que evalúa la coherencia textual (capítulo 4) con el nuevo módulo que se describe en este capítulo es uno de los aportes más interesantes del presente trabajo, ya que busca aplicar la técnica automática de evaluación de coherencia textual con un fin distinto del original, este nuevo fin es evaluar si la información de una noticia está correctamente presentada según la estructura de la pirámide invertida. Para poder entender a cabalidad lo anterior, es necesario conocer primero la idea en que se basa su funcionamiento.

6.2.1. Modelo planteado para la evaluación de la estructura de pirámide invertida en una noticia

Como se señaló en 2.2.3 y 2.3, en la literatura dedicada al periodismo se reconoce que la noticia tiene una estructura conocida como pirámide invertida. Van

Dijk (1990), a través del análisis de noticias desde un enfoque lingüístico, concluye que este esquema particular del periodismo de presentar la información desde lo más relevante a lo menos relevante existe y es efectivo. De hecho, el holandés señala que “sólo el titular y los sucesos principales deben hallarse obligatoriamente en un discurso periodístico mínimamente bien construido; categorías como antecedentes, reacciones verbales y comentarios son opcionales” (Van Dijk, 1990). Martínez Albertos (2004) indica que esta estructura permite que, si la noticia está correctamente escrita de acuerdo a ella, cuando por exigencias de espacio hay que acortar el texto, “se pueden ir tirando tranquilamente los párrafos situados al final del relato con la certeza de que son los menos interesantes del escrito”. Y el mismo Van Dijk (1990) corrobora lo anterior, pues señala que “la organización global permite a los editores cortar los párrafos finales de un relato periodístico sin perder la información esencial”.

En resumen, la pirámide invertida es una estructura válida para la producción escrita de noticias -desde un enfoque periodístico y uno lingüístico-. Lo más importante en dicha estructura de significado es el titular y el *lead* que resumen los sucesos principales, que deben responder a las cinco -o seis, según se explicó en 2.2.2- preguntas fundamentales: qué, quién, dónde, cuándo, por qué y cómo. Del mismo modo, también se señaló que no siempre será necesario responder en el titular y el *lead* a todas las preguntas fundamentales; recuérdese el ejemplo presentado en 2.2.2 de que la noticia sea sobre el estreno de una película, en que las preguntas más relevantes serían qué, cuándo y dónde, quedando quién, por qué y cómo en un segundo plano.

De lo anterior, -sin desconocer la división de la noticia en titular, *lead* y cuerpo- es posible afirmar que la noticia se puede dividir en dos partes principales: el titular y el *lead*, por un lado; y los párrafos posteriores, por otro (ver Figura 7). En la primera parte se agrupan los dos estadios superiores de la pirámide invertida y en la segunda el vértice inferior.

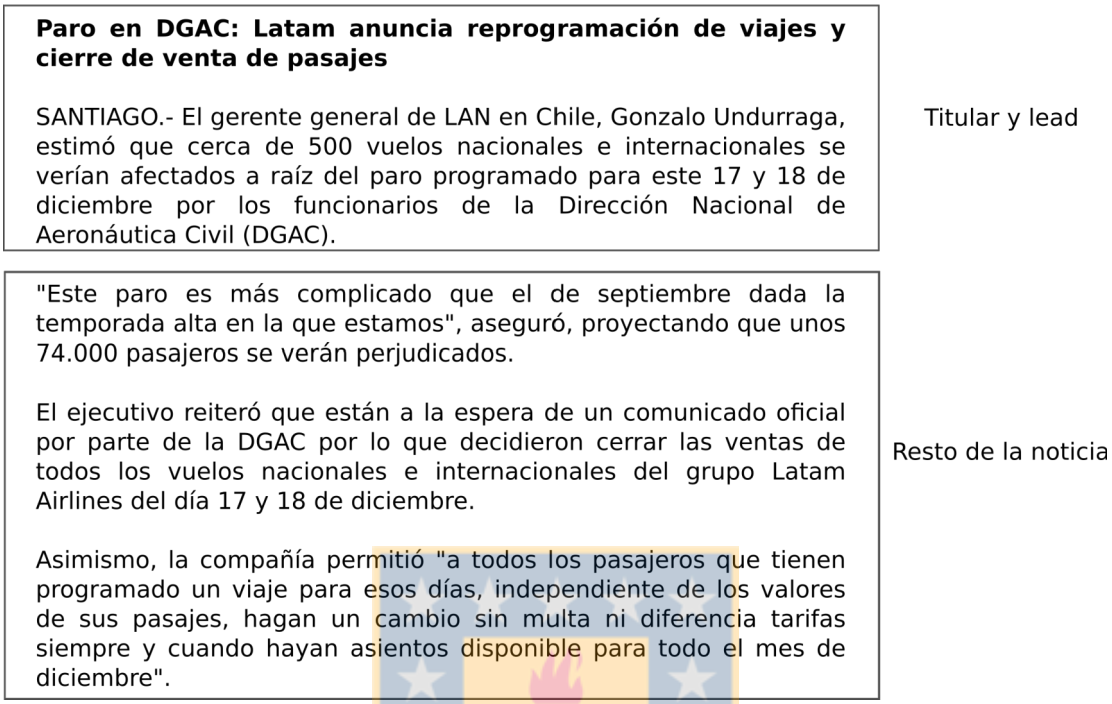

Figura 7: División de la noticia en dos partes.

De acuerdo a lo señalado por Van Dijk (1990) y Martínez Albertos (2004), la información fundamental de la noticia se debe incluir en la primera parte, esto es, en el titular y en el *lead*; he aquí el sustento de la división recién propuesta. De acuerdo con esto, a una persona que lee sólo esta parte de un texto, debe bastarle para hacerse una idea cabal de lo que trata el hecho descrito en la noticia. Los párrafos posteriores, que se incluyen en la segunda parte, constituyen información adicional y de menor importancia, que amplía lo dicho entre el titular y el *lead*, siguiendo el esquema planteado en este último. Por lo mismo, se afirma que el editor de una noticia, en caso de problemas de espacio en el diario -por ejemplo-, puede proceder sin temor a eliminar los párrafos de más abajo, teniendo la certeza de que no se afectará la correcta intelección del contenido del texto producido.

Por lo tanto, para escribir el titular y el *lead* de una noticia hay que tener claridad acerca del hecho completo de que tratará el texto (por ejemplo, un asesinato, el lanzamiento de un libro, la renuncia de un ministro de Estado, etc.). El periodista, por lo general toma nota del hecho y los datos que lo conforman en una libreta o, cuando es posible, graba a la fuente que entrega la información (también

podría basar parte de su texto en una fuente documental como, por ejemplo, un reporte policial). Una vez que cuenta con toda esta información y antes de sentarse a producir el texto, debe ser capaz de sintetizarlo y estructurar un punteo de ideas que constituyan un resumen efectivo de la situación descrita, con el fin de presentarlo en la menor cantidad de palabras posibles. Por lo mismo, el capítulo 2 de este trabajo conceptualiza a la noticia como un tipo particular de resumen, específico del género discursivo periodístico, que no se basa siempre en un texto escrito (como, por ejemplo, podría ser una noticia basada en un reporte policial), sino que también puede basarse en un hecho ocurrido (por ejemplo, un atropello), o en las declaraciones entregadas en una conferencia de prensa (por ejemplo, un entrenador de fútbol anuncia que dejará su cargo). Sea cual sea el caso, luego de esta etapa, el periodista debe ser capaz de ordenar sus datos, sintetizarlos y jerarquizarlos. Para esta tarea, la estructura de pirámide invertida y las preguntas fundamentales son una ayuda invaluable, ya que respondiendo a éstas de manera acertada es posible llegar a determinar cuáles son los datos más relevantes y, de esta forma, poder llegar a escribir de manera correcta la noticia.

Por lo anterior, realizando el proceso inverso es posible determinar si una noticia está bien construida, en otras palabras, a partir del conocimiento del hecho que genera la noticia y contrastando si las preguntas fundamentales están adecuadamente respondidas entre el titular y el *lead* se puede determinar si la estructuración de la pirámide invertida es adecuada en una noticia en particular. En la enseñanza del Periodismo Informativo en las universidades -género al que pertenecen las *noticias* (o *información periodística)*-, un ejercicio muy difundido para que los estudiantes aprendan a asimilar este modelo de escritura y a jerarquizar los datos, consiste en entregarles un punteo de ideas desordenadas, con el fin de que las jerarquicen y, a partir de éstas, puedan escribir una noticia. Ejemplos de lo anterior los podemos encontrar en el Isabel Project (2010), Oliva (2015) y en el mismo texto de Van Dijk (1990) que descompone la noticia para su análisis en la lista de ideas o datos que la conforman.

Por ello, partiendo de que se tiene como base un punteo de datos que incluye la información relevante y la de menor importancia, en el desarrollo del módulo 3 de evaluación de la pirámide invertida que se construyó, se aplica una idea similar a

la del módulo 1: se comparan las preguntas fundamentales en que se basa el texto con el titular y el *lead* de éste. Para ello se ocupa la misma fórmula que en el cálculo de la coherencia textual, esto es, comparar pares de unidades textuales, calculando el coseno del ángulo formado por los vectores que las representan, lo que da un puntaje que varía entre -1 y 1 (aunque normalmente fluctúa entre 0 y 1). Específicamente, las comparaciones que se realizan son:

- Comparar el titular (unidad textual 1) de la noticia procesada en el prototipo con un titular tipo (unidad textual 2) basado en las preguntas fundamentales.
- Comparar el *lead* (unidad textual 1) de la noticia procesada con las preguntas fundamentales (unidad textual 2). Para esto último, las preguntas fundamentales se unen en una sola cadena de texto, con el fin de que constituyan una sola unidad textual.
- Comparar, como conjunto, el titular y el *lead* (unidad textual 1) de la noticia procesada con las preguntas fundamentales (unidad textual 2). Para esto, el titular y el *lead* se combinan en una única cadena de texto, lo mismo que las preguntas fundamentales.
- Comparar los párrafos siguientes al *lead* (unidad textual 1) con los datos del punteo construido (unidad textual 2) que no forman parte de la información relevante proporcionada. En este caso, los párrafos siguientes al *lead* se funden en una sola cadena de texto, lo mismo que los datos adicionales con el fin de posibilitar su comparación.

Las anteriores son las cuatro comparaciones que se utilizaron como base para construir el tercer módulo. Las tres primeras se centran en la información relevante y, por ende, son las comparaciones más importantes; la última, por su parte, se enfoca en la información anexa, pero también tiene utilidad en vistas al análisis que debe realizar el prototipo.

6.2.2. Estructura del tercer módulo

Como se dijo anteriormente, el módulo 3, para evaluar la jerarquización de los datos de una noticia, debe funcionar conjuntamente con el módulo 1 -que

evalúa la coherencia textual- que fue descrito en el capítulo 4. Por lo mismo, no se entrará a detallar todo el código escrito, ya que muchas partes de éste ya fueron expuestas en el capítulo aludido al igual que las fórmulas matemáticas utilizadas.

De la misma forma que el módulo 1, éste se programó en Python 3, se utilizó NLTK y se empleó el Snowball Stemmer, incluido en NLTK, ya que trabaja con lengua española. En la Figura 8 se presenta la estructura del componente, incluyendo los tres módulos desarrollados.

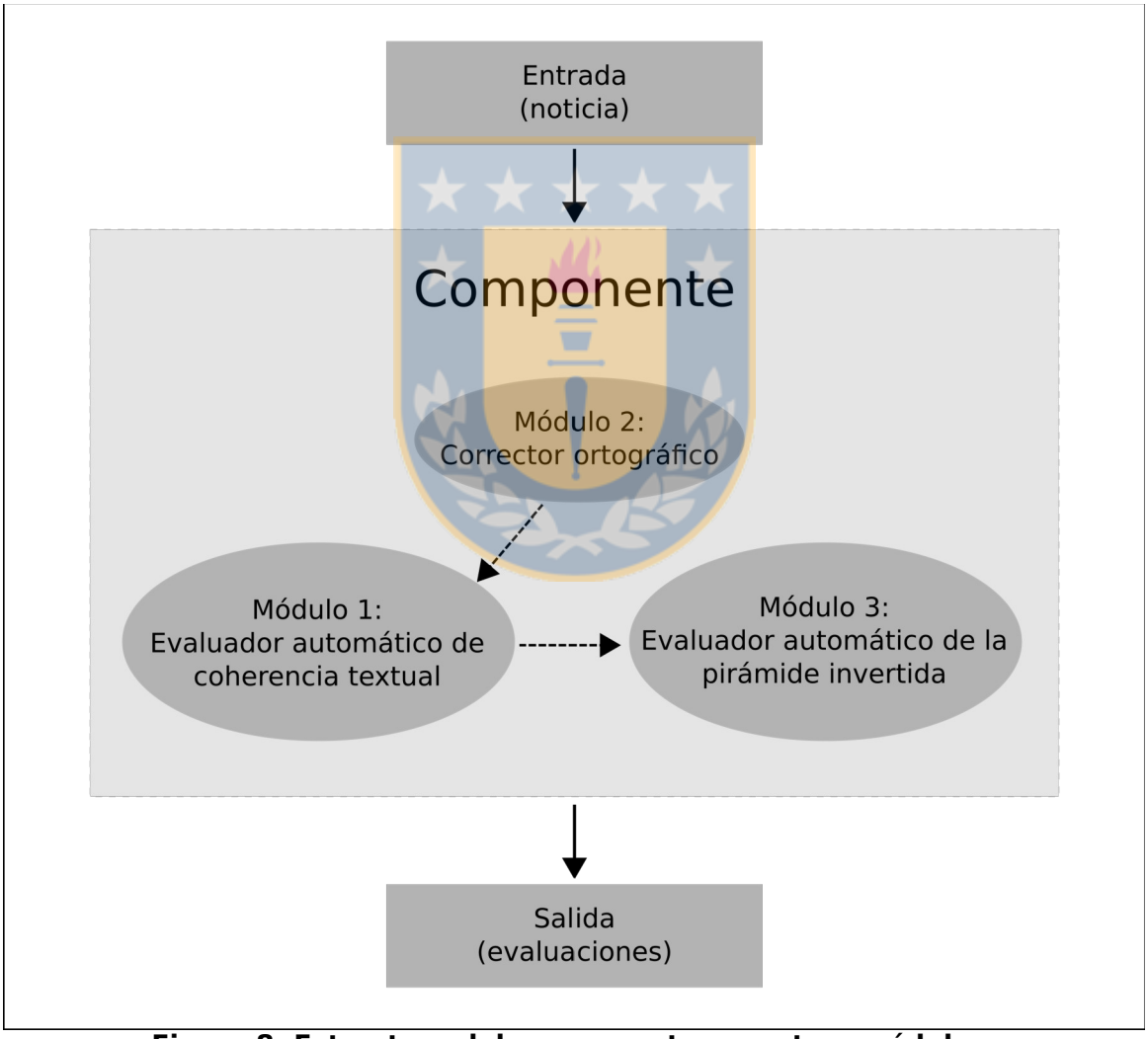

Figura 8: Estructura del componente y sus tres módulos.

La gran similitud entre un prototipo y otro está en el método de análisis que utilizan: comparar unidades textuales calculando el coseno de los ángulos que

forman los vectores que representan a cada una de ellas. La gran diferencia radica en el procesamiento que realizan de la noticia: el prototipo (módulo 1) que calcula la coherencia textual compara las unidades textuales adyacentes que conforman la noticia: el titular con el *lead*, el *lead* con el segundo párrafo, el segundo párrafo con el tercero y así sucesivamente. En el caso del prototipo (módulo 3) que evalúa la jerarquización de los datos en la noticia, la comparación como ya se dijo en el apartado anterior se realiza entre el titular de la noticia procesada con un titular tipo basado en las preguntas fundamentales, entre el *lead* y las preguntas fundamentales, entre el titular y el *lead* combinados con las preguntas fundamentales y entre el resto de la noticia con los datos adicionales del punteo elaborado. Por lo mismo, para el funcionamiento del módulo 3 se requiere que se diseñe una pauta de cotejo con las preguntas fundamentales y los datos adicionales, en base a la cual se realizarán las comparaciones requeridas. Por ello, pensando en una futura implementación de este módulo -y del componente en que se inserta- en un STI, se requerirá la elaboración de una pauta de cotejo por cada ejercicio que se integre en la máquina (lo que, en todo caso, no es una tarea extensa). A continuación se presenta la pauta de cotejo elaborada para el ejercicio que se detallará en 6.3; desde ya se aclara -después se explicará más a fondo- que los datos son ficticios, aunque basados en la contingencia nacional:

- **Qué:** Fue proclamado como candidato para las elecciones presidenciales de 2017 de la Nueva Mayoría.
- **Quién:** El exmandatario Ricardo Lagos Escobar.
- **Dónde:** En un acto efectuado en el Teatro Caupolicán de Santiago.
- **Cuándo:** A las 21 horas de hoy.
- **Por qué:** Fue ganador de las primarias de la Nueva Mayoría realizadas la semana pasada para definir el candidato de la coalición para ocupar la Moneda en el periodo 2018-2022.
- **Cómo:** Con una ovación de todos los presentes.
- **Titular tipo:** Ricardo Lagos proclamado candidato presidencial Nueva Mayoría.
- **Dato adicional (DA):** Lagos (PPD) se impuso en las primarias de la coalición a José Antonio Gómez (PRSD), Isabel Allende (PS) y Jorge Burgos (DC).
- **DA:** Lagos se enfrentará en las elecciones presidenciales que se realizarán el 19 de noviembre próximo a Manuel José Ossandón (RN) de Chile Vamos, Marco Enríquez Ominami (PRO), Alejandro Navarro (MAS), Lily Pérez (Amplitud) y Andrés Velasco (Ciudadanos).
- **DA:** Lagos fue Presidente de la República entre 2000 y 2006.

- **DA:** En su primera elección Lagos logró el triunfo tras derrotar en segunda vuelta a Joaquín Lavín por 51,31% contra 48,69% del militante UDI (enero de 2000).
- **DA:** Lagos fue Ministro de Educación en la presidencia de Patricio Aylwin entre 1990 y 1992; y Ministro de Obras Públicas en el mandato de Eduardo Frei Ruiz-Tagle entre 1994 y 1998.
- **DA:** Lagos fue derrotado por Eduardo Frei Ruiz-Tagle en las primarias de la Concertación en 1993 por 63,32% contra 36,68%.
- **DA:** En 1989 no fue elegido como candidato presidencial de la Concertación de Partidos por la Democracia y se presentó como candidato a senador, elección que perdió -pese a tener la segunda mayor votación- debido al sistema binominal, ya que su lista no pudo doblar a la lista adversaria.
- **DA:** Lagos fue uno de los fundadores del Partido por la Democracia (PPD) el 15 de diciembre de 1987.
- **DA:** Uno de los episodios más recordados de Ricardo Lagos ocurrió el 25 de abril de 1988, cuando en el programa "De cara al país" miró a la cámara y emplazó al entonces gobernante Augusto Pinochet Ugarte por la intención de éste de continuar ocho años más como Presidente de la República.

En 4.3 ya se presentó el código que se utiliza para lematizar los textos que se ingresen al módulo 1, procesar los textos en el Infomap-NLP con el fin de obtener el valor para los vectores y el código utilizado para calcular el coseno del ángulo formado por dos vectores. En este nuevo módulo, el código es esencialmente el mismo para efectuar dichas tareas. La mayor diferencia estriba en una nueva función que se definió, que se llama *analizador* y que realiza las tareas inherentes a las particularidades de este tercer módulo. El código se presenta a continuación:

```
def analizador(entrada, salida):
        filew = open(salida, 'w')
        with open (entrada) as filer1, open('pauta_analizador_DEF') as filer2:
                textos_preparados1 = preparar(filer1)
                coherencia_texto_completo=coherencia_texto(textos_preparados1)
                cohe = str(coherencia_texto_completo)
                promedio = np.mean(coherencia_texto_completo)
                preguntas = list(islice(filer2, 6))
                entrada = ''.join(preguntas)
                reps = {'Qué: ':'', 'Quién: ':'', 'Dónde: ':'', 'Cuándo: ':'', 'Por qué: ':'', 'Cómo: ':'', '\n': ' '}
                preguntas_fund = [reemplazar(entrada, reps)]
                titular_tipo_qq = list(islice(filer2, 1))
                datos_adicionales = [filer2.read().replace('\n',' ')]
                filer1.seek(0)
```

```
        filer2.seek(0)
        preguntas_que_quien = list(islice(filer2, 2))
        entrada_qq = ''.join(preguntas_que_quien)
        preguntas_qq_titular = [reemplazar(entrada_qq, reps)]
        titular = list(islice(filer1, 1))
        lead = list(islice(filer1, 1))
        titulead = [' '.join(titular + lead).replace('\n','')]
        resto = [filer1.read().replace('\n',' ')]
        textos_preparados2 = preparar(titular+preguntas_qq_titular)
        textos_preparados3 = preparar(lead+preguntas_fund)
        textos_preparados4 = preparar(titulead+preguntas_fund)
        textos_preparados5 = preparar(datos_adicionales+resto)
        textos_preparados6 = preparar(titular+preguntas_fund)
        coherencia_titular_preg_qq = coherencia_texto(textos_preparados2)
        coherencia_lead_pf = coherencia_texto(textos_preparados3)
        coherencia_titulead_pf = coherencia_texto(textos_preparados4)
        coherencia_resto_da = coherencia_texto(textos_preparados5)
        coherencia_titular_pf = coherencia_texto(textos_preparados6)
        filer1.seek(0)
        primera_linea = filer1.readline()
        filew.write(primera_linea + '\n' +
(((('\n'.join(cohe.split())).replace(',','')).replace('[','')).replace(']','') + '\n' + '_____' + '\n' +
str(promedio) + '\n\n\n' + 'Titular y qué-quién (tipo): ' + str(coherencia_titular_qq)+ '\n\n' + 'Titular y
qué-quién (preguntas): ' + str(coherencia_titular_preg_qq)+ '\n\n' + 'Titular y preguntas
fundamentales: ' + str(coherencia_titular_pf)+ '\n\n' + 'Lead y preguntas fundamentales: ' +
str(coherencia_lead_pf)+ '\n\n' + 'Titular+lead y preguntas fundamentales: ' +
str(coherencia_titulead_pf) + '\n\n' + 'Datos adicionales y resto de la noticia: ' +
str(coherencia_resto_da) + '\n\n' + 'Titular+lead y datos adicionales: ' + str(coherencia_titulead_da) +
'\n\n' + 'Preguntas fundamentales y resto de la noticia: ' + str(coherencia_resto_pf))
        filew.close()
```

En resumen lo que realiza la función anterior es invocar a la función que evalúa la coherencia textual de la noticia, es decir, llama al módulo 1, ya que también se incluye en el informe final que arroja el presente módulo el puntaje de coherencia textual de las unidades que componen la noticia. Luego prepara los textos de la noticia y de la pauta de cotejo, eliminando la información innecesaria en el caso de esta última (por ejemplo, los encabezados de las preguntas

fundamentales: “Qué:”, “Quién:”, etc.). Luego realiza la tarea de comparación de las unidades textuales ya descrita en 6.2.1. Finalmente, prepara el informe que arroja el componente (todos los módulos como conjunto) como salida, con los puntajes que resultan de las evaluaciones realizadas. A continuación se presenta un ejemplo de este archivo de salida:

Ricardo Lagos es proclamado candidato presidencial por la Nueva Mayoría

0.71952120605354775
0.7069877464131723
0.65005562021324181
0.64687712354379112

_____
0.680860424056

Titular y titular tipo (qué-quién): [1.0]
Lead y preguntas fundamentales: [0.87301447596956872]
Titular+lead y preguntas fundamentales: [0.91257160028190198]
Datos adicionales y resto de la noticia: [0.93928033077698991]

En la primera línea se presenta el titular de la noticia procesada. A continuación se entregan los resultados de la evaluación de la coherencia textual para cada par de unidades textuales adyacentes, para terminar con el puntaje de coherencia de la noticia como conjunto. En una segunda parte, se presentan las evaluaciones para: la comparación del titular con el titular tipo basado en las preguntas fundamentales; para la comparación del *lead* con las preguntas fundamentales; del titular y *lead* como conjunto con las preguntas fundamentales; y de los datos adicionales con el resto de la noticia.

Con los datos obtenidos es posible comenzar a realizar el análisis; para ello es necesario explicar primero cómo se recolectó el corpus de textos a analizar.

6.3. Diseño y aplicación de la tarea de escritura de una noticia

Con el fin de testear el módulo 3, se determinó que la mejor forma era con textos producidos por estudiantes de periodismo, quienes serían los usuarios del STI que se busca construir a futuro. Para recopilar estos textos, entonces, se conversó con los profesores de la Sección II del curso de Producción de Textos Periodísticos de

la Carrera de Periodismo de la Universidad de Concepción. El curso corresponde al segundo semestre (2015) del primer año de la carrera, por lo que los estudiantes están en una etapa temprana de su proceso formativo.

Como se dijo, el primer paso fue solicitar una entrevista a los dos docentes que dictan la asignatura. En la ocasión se les explicó en detalle para qué se quería realizar la experiencia y en qué podría consistir ésta. Ambos estuvieron de acuerdo y se acordó con ellos que se les entregaría una propuesta de actividad práctica a realizar por los alumnos del curso. Además, se les señaló que una parte fundamental era que ellos realizaran la corrección de los trabajos producidos (con el fin de después comparar sus evaluaciones con las del prototipo) y estuvieron de acuerdo, también. Posterior a esto, se solicitó el permiso correspondiente a Jefatura de Carrera para realizar la intervención.

Una semana después, se les envió la propuesta de actividad a los docentes. Luego, se acordó una nueva reunión con ellos para discutirla. Ambos estuvieron de acuerdo con la actividad planteada y propusieron realizarla como una actividad práctica real, esto es, incorporarla a las evaluaciones del curso, sobre todo considerando que ellos serían quienes corregirían los textos producidos.

A la actividad se le dio una forma similar a la estructura de las instrucciones de los trabajos prácticos de la asignatura y no se les advirtió previamente a los estudiantes de que se utilizarían sus escritos en el marco de una investigación, con el fin de que las condiciones en el aula fueran absolutamente reales. El trabajo se tomó en una sala habilitada con 60 computadores, por lo que hubo espacio suficiente para los 20 estudiantes que rindieron el trabajo práctico. El detalle de las instrucciones que recibieron los alumnos se puede revisar en el anexo 1. Sin embargo, aquí es importante mencionar lo siguiente:

- Como ya se indicó, los datos proporcionados a los estudiantes fueron ficticios, aunque basados en la contingencia nacional. Lo anterior con el fin de precaver cualquier intento de copia de los textos desde Internet, ya que mientras escribieron estaban conectados. Por otra parte, no es algo nuevo para ellos trabajar con datos ficticios en ejercicios de la asignatura.
- Relacionado con lo anterior, en las instrucciones del práctico se utilizó una

presentación que ubicaba a los alumnos en la situación ficticia en que se contextualizaba el ejercicio: “Imagine que estamos un par de años en el futuro y usted es redactor de un medio escrito, que se enfoca en el periodismo informativo. Como se avecinan las elecciones presidenciales de 2017, su editor le encarga escribir un texto sobre un hecho que ocurrió hace unos minutos. A partir de los datos expuestos más abajo -que son ficticios y se enfocan en el hecho mencionado-, elabore un texto informativo siguiendo las pautas vistas en clases (estructura de pirámide invertida, responder a las preguntas fundamentales, etc.). El texto debe cumplir con todos los requisitos para ser publicable en la sección de noticias de un medio escrito”.

- También se les indicó que el texto debía tener titular informativo, *lead* y cuerpo de la noticia (siguiendo la pirámide invertida); que no deberían incluirse bajada ni epígrafe en los elementos de titulación; que sólo debían utilizarse los datos presentados; que la extensión máxima del texto era de 2.500 caracteres y la mínima de 1.500 (caracteres con espacios); y que tenían un plazo de una hora y quince minutos para escribir el texto (desde 15:15 a 16:30).

Los datos que se les entregaron a los estudiantes son los que se presentan a continuación, respetando el orden en que ellos los recibieron.

- Lagos fue Presidente de la República entre 2000 y 2006.
- En 1989 no fue elegido como candidato presidencial de la Concertación de Partidos por la Democracia y se presentó como candidato a senador, elección que perdió - pese a tener la segunda mayor votación - debido al sistema binominal, ya que su lista no pudo doblar a la lista adversaria.
- La proclamación de Lagos se realizó en un acto efectuado en el Teatro Caupolicán de Santiago a partir de las 21 horas de hoy.
- Lagos fue derrotado por Eduardo Frei Ruiz-Tagle en las primarias de la Concertación en 1993 por 63,32% contra 36,68%.
- Lagos se enfrentará en las elecciones presidenciales que se realizarán el 19 de noviembre próximo a Manuel José Ossandón (RN) de Chile Vamos, Marco Enríquez Ominami (PRO), Alejandro Navarro (MAS), Lily Pérez (Amplitud) y Andrés Velasco (Ciudadanos).
- Lagos fue el ganador de las primarias de la Nueva Mayoría realizadas la semana pasada, que buscaban definir al candidato de la coalición para ocupar la Moneda en

el periodo 2018-2022.

- En su primera elección Lagos logró el triunfo tras derrotar en segunda vuelta a Joaquín Lavín por 51,31% contra 48,69% del militante UDI (enero de 2000).
- Lagos fue Ministro de Educación en la presidencia de Patricio Aylwin entre 1990 y 1992; y Ministro de Obras Públicas en el mandato de Eduardo Frei Ruiz-Tagle entre 1994 y 1998.
- Uno de los episodios más recordados de Ricardo Lagos ocurrió el 25 de abril de 1988, cuando en el programa “De cara al país” miró a la cámara y emplazó al entonces gobernante Augusto Pinochet Ugarte por la intención de éste de continuar ocho años más como Presidente de la República.
- Tras su proclamación Lagos fue ovacionado de pie por los asistentes al Teatro Caupolicán.
- Lagos fue uno de los fundadores del Partido por la Democracia (PPD) el 15 de diciembre de 1987.
- El exmandatario Ricardo Lagos Escobar fue proclamado como candidato de la Nueva Mayoría para las elecciones presidenciales de 2017.
- Lagos (PPD) se impuso en las primarias de la coalición a José Antonio Gómez (PRSD), Isabel Allende (PS) y Jorge Burgos (DC).

Una vez que terminaron sus trabajos, los estudiantes subieron una copia digital a la plataforma Infoda de la Universidad de Concepción, que se utiliza para organizar las actividades de las asignaturas y entregar las notas. Los profesores de la asignatura entregaron copia de los 20 archivos digitales para ser utilizados en esta tesis. Estos archivos, una vez recibidos, fueron procesados en el componente (módulos 1, 2 y 3). Si bien los estudiantes redactaron sus noticias en un procesador de textos, de todas formas el módulo 2 encontró algunos errores ortográficos que reparó. A continuación se presentan algunos ejemplos. Los errores detectados están resaltados en cursiva, negrita y subrayado:

> **Ejemplo 1:** Entre gritos y aplausos, el aspirante a la presidencia del PPD, Ricardo Lagos Escobar, proclamó su candidatura en el Teatro ***Caupolican*** de Santiago. El acto se llevó a cabo a las 21 horas del día de ayer, donde el exmandatario se consagró como la carta oficial a la presidencia por parte de la Nueva Mayoría, para el periodo 2018 – 2022.

> **Ejemplo 2:** El ex Ministro de Educación en la presidencia de Patricio Aylwin, deberá enfrentarse el próximo año a sus adversarios Manuel José ***Ossandon*** (RN) de Chile Vamos, Marco ***Enriquez*** Ominami (PRO), Alejandro Navarro (MAS), Lily Pérez (Amplitud) y Andrés Velasco (Ciudadanos).

> **Ejemplo 3:** En un acto celebrado a las 21 ***hras*** del día de ayer en las dependencias del Teatro Caupolicán, la Nueva Mayoría dio inicio a la carrera presidencial, que espera llevar a Ricardo Lagos a su segundo ciclo en la Moneda, ***ésto*** luego de que el exmandatario se

impusiera en las primarias de la coalición a Isabel Allende, Jorge Burgos y José Antonio Gómez.

Como se puede ver, en los ejemplos 1 y 2, los errores corresponden a omisiones de tildes en nombres propios, los que seguramente no estaban entre los patrones que manejaba el procesador de textos que los estudiantes utilizaron al escribir sus trabajos. En el caso del ejemplo 3, hay un error al escribir *horas* que no queda claro si es una escritura incorrecta de la abreviatura o la omisión de la vocal *o*; además, luego se tilda erradamente el demostrativo neutro *esto*.

Para el módulo corrector ortográfico no fue problema reconocer los errores en los nombres propios, ya que al construir el listado de palabras correctas del módulo 2, estos nombres estaban incluidos. En el caso de la omisión del tilde de esto, tampoco hubo mayor problema, ya que también había una regla de error para el caso. En relación a la mala escritura de *horas*, el módulo la calificó como palabra desconocida ya que no había una regla de error para ésta (luego se incorporó).

En relación a la tarea de corrección, los profesores de la asignatura participaron en proponer la pauta utilizada, que sirvió como base para afinar la pauta de cotejo para el prototipo. Lo que ellos evaluaron fue lo siguiente:

- Le otorgaron un puntaje al titular escrito (máximo cuatro puntos). Para ello determinaron que los elementos que debía contener un titular, a partir de los datos proporcionados, se basaban en las preguntas fundamentales *qué* y *quién*. Determinaron que un titular tipo correcto sería: “Ricardo Lagos proclamado candidato presidencial Nueva Mayoría”. Con sus diversas variantes como podría ser, por ejemplo, “Nueva Mayoría proclamó a Ricardo Lagos como su candidato presidencial”. En el fondo, se fijaron en que estuvieran presentes los elementos: *Ricardo Lagos*, *proclamar*, *candidato presidencial* y *Nueva Mayoría*.
- Determinaron que las preguntas fundamentales relevantes de las seis eran: qué, quién, cuándo y dónde. Es decir, que se proclamó a Ricardo Lagos como candidato presidencial de la Nueva Mayoría en un acto realizado a las 21 horas de hoy en el Teatro Caupolicán. Las preguntas *por qué* y *cómo* las

descartaron como información relevante, ya que la primera de éstas no era noticia pues las primarias habían sido la semana pasada y la segunda porque no era relevante que Lagos recibió una ovación tras ser proclamado. En consecuencia, cotejaron el *lead* con las cuatro preguntas fundamentales consideradas relevantes y otorgaron un total de cuatro puntos al estudiante que cumpliera con incluir todos estos elementos. Como en periodismo también se considera al titular en conjunto con el *lead*, realizaron la misma evaluación del titular y *lead* como un solo cuerpo con las cuatro preguntas fundamentales elegidas y también otorgaron un máximo de cuatro puntos a quienes incluyeran todos los elementos esperados.

- Por último, evaluaron los párrafos siguientes al *lead* verificando que cumplieran con concordar con la información menos relevante del punteo proporcionado. En este caso, otorgaron siete puntos al estudiante que cumpliera con todos los requisitos esperados.

Hay que señalar que el trabajo también tuvo otros puntajes (por ejemplo, ortografía y redacción), pero los expuestos aquí son sólo los que interesan para los fines de esta tesis y son los que se consideraron en la pauta de cotejo del prototipo.

Por último, antes de pasar a exponer los resultados, hay que señalar que los docentes de la asignatura realizaron la corrección de los textos en conjunto. En relación a sus antecedentes, ambos son periodistas. Uno de ellos, además, es escritor con varios libros publicados a nivel nacional y posee el grado de magíster; actualmente publica dos veces a la semana en un medio escrito de circulación regional. La otra persona es una periodista de extensa trayectoria en medios y, actualmente, desempeña el cargo de editora general en un medio escrito de circulación regional. Por lo tanto, ambos cumplen con los antecedentes y experiencia requerida para evaluar textos informativos, esto es, son profesionales expertos en Periodismo Informativo, tienen experiencia en medios de prensa escribiendo noticias y, lo más importante, son quienes imparten la docencia de pregrado en dicha materia en el curso en que se tomó la muestra.

Capítulo 7: Presentación y análisis de resultados

A continuación se presentarán los resultados obtenidos en las evaluaciones realizadas a los veinte textos en que se enfocó la experiencia y se efectuará un análisis de dichos resultados. Además, se detallarán algunas situaciones que se observaron en éstos y que son necesarias de destacar para una completa explicación de le experiencia.

7.1. Pilotaje del componente

Previo a presentar los resultados de la experiencia con estudiantes, es necesario señalar que en conjunto con el diseño de la experiencia práctica que se describió en el capítulo anterior, se realizó una prueba de pilotaje en un contexto acotado (dos sujetos) con el fin de ir observando si los pasos seguidos iban en el rumbo adecuado o no. Los resultados fueron satisfactorios y se presentan a continuación.

La actividad diseñada se aplicó a dos periodistas con más de cinco años de experiencia en la redacción de noticias. Ambos profesionales tienen estudios de postgrado: uno de ellos es doctor (c) en Lingüística y el otro PhD on Arts, School of English, Communications and Performance Studies. La particularidad de este pilotaje es que a uno de los dos participantes, al azar, se le entregó una versión de las instrucciones en que se le solicitaba cometer errores intencionados al escribir el *lead* de la noticia (el nudo más crítico en que se quería enfocar el pilotaje). Lo interesante del asunto es que sólo la persona a la que le correspondió cometer los errores sabía de esto, ya que se les pidió a ambos no comunicarlo, en caso de tocarles dicha responsabilidad. Por lo mismo, no se supo hasta más adelante, quien redactaría el texto con errores intencionados.

Los textos que produjeron se procesaron en el componente y los resultados obtenidos fueron los que se presentan en la Tabla III.

Tabla III: Resultados del pilotaje.

Número	Evaluación	Corpus	Sujeto 1	Sujeto 2
Comparación 1	Puntaje de coherencia del texto	DK1	0,61	0,55
		DK2	0,67	0,64
		DK1y2	0,64	0,59
		Promedio	**0,64**	**0,59**
Comparación 2	Comparación titular escrito con titular tipo	DK1	1	0,91
		DK2	1	0,91
		DK1y2	1	0,92
		Promedio	**1**	**0,91**
Comparación 3	Comparación *lead* con preguntas fundamentales	DK1	0,93	0,61
		DK2	0,95	0,66
		DK1y2	0,94	0,66
		Promedio	**0,94**	**0,64**
Comparación 4	Comparación titular más *lead* con preguntas fundamentales	DK1	0,94	0,67
		DK2	0,94	0,7
		DK1y2	0,94	0,7
		Promedio	**0,94**	**0,69**
Comparación 5	Comparación párrafos siguientes al *lead* con datos adicionales	DK1	0,98	0,92
		DK2	0,99	0,94
		DK1y2	0,99	0,92
		Promedio	**0,99**	**0,93**

En relación a los puntajes aquí presentados, hay que diferenciar de inmediato la evaluación de la coherencia del texto de las comparaciones realizadas. El puntaje de coherencia se mueve en rangos diferentes, tal como se señaló en 4.4 siguiendo a Hernández y Ferreira (2010): un puntaje de coherencia superior a 0,60 se considera medio-alto en el caso de la evaluación de coherencia textual (de hecho, el puntaje más alto que se obtuvo en el trabajo de los autores aludidos fue de 0,72). Por otra parte, puntajes cercanos a los obtenidos en las otras comparaciones (0,90 hacia arriba, como se verá en el presente capítulo) es muy poco probable que se den al evaluar la coherencia textual. Y he aquí que se comienza a vislumbrar lo interesante del método que se propone en el presente trabajo, como se expondrá en las páginas siguientes. Una explicación de por qué los puntajes de la medición de la coherencia textual y los de la evaluación de la jerarquización en la pirámide invertida se mueven en rangos diferentes, se entregará en las conclusiones de esta tesis.

Si se observan los puntajes de los sujetos, en lo relativo a la coherencia textual, se puede observar que el sujeto 1 alcanza puntajes mayores de coherencia

en cada uno de los espacios semánticos utilizados. Sin embargo, las diferencias con el sujeto 2 no son demasiado grandes. De hecho, si se promedian los puntajes de coherencia textual de la noticia del sujeto 1 para los tres espacios semánticos, se obtiene 0,64; en el mismo caso, para el sujeto 2, el valor es de 0,59. Estos datos nos hablan de que, según el componente, el texto del sujeto 1 presenta una coherencia textual algo mayor como conjunto que el texto del sujeto 2.

En el caso de la comparación del titular escrito con el titular tipo, el sujeto 1 presenta una evaluación de 1,00 en los tres espacios semánticos; lo anterior implica que su titular cumplió con lo esperado e incluyó los cuatro elementos que se buscaba que estuvieran presentes. En el caso del sujeto 2 en el mismo apartado, su puntaje promedio fue de 0,91, lo que implica que también su titular se acercó casi exactamente a lo esperado.

En el caso de las dos comparaciones siguientes -el *lead* con preguntas fundamentales y el titular más el *lead* con las preguntas fundamentales-, de inmediato llaman la atención las diferencias de puntaje, que son bastante más altos en el caso del sujeto 1. Hay que recordar que en el *lead* uno de los sujetos intencionadamente iba a cometer errores al redactarlo. Si vamos a los puntajes tenemos que en la comparación del *lead* con las preguntas fundamentales, el sujeto 1 obtiene un promedio de 0,94 y el sujeto 2 uno de 0,64. Para la comparación del titular más el *lead* con las preguntas fundamentales, el sujeto 1 obtiene un promedio de 0,94 y el sujeto 2 uno de 0,69. Lo anterior lleva a sospechar que el sujeto 2 fue quien cometió errores intencionados.

Por último, en el caso de la comparación final entre los párrafos siguientes al *lead* con los datos adicionales, el sujeto 1 obtiene un promedio de 0,99 y el sujeto 2 uno de 0,93. En este caso, ambos puntajes son altos y no demasiado diferentes.

Ahora bien, analizando la prueba como conjunto y enfocándose primero en las comparaciones 1, 2 y 5 -en las que no se indujeron errores intencionales-, se puede decir que ambas noticias presentan una coherencia textual de medio a medio-alta, lo que permite hablar de éstos como textos que no presentan grandes deficiencias en esos aspectos, asumiendo que la evaluación de la máquina es correcta, según lo validado en Hernández y Ferreira (2010). En el caso de las comparaciones 2 y 5 no hay grandes diferencias en los puntajes promedio de cada

sujeto y ambos son altos, por lo que se podrían asumir como correctos en caso de validarse este método de evaluación. Pero aún no es el momento de juzgar esto.

En el caso de las comparaciones 3 y 4, en las que se considera el *lead* en el que uno de los sujetos cometió errores intencionados en su construcción, de inmediato llama la atención -como se dijo- la diferencia de puntajes promedio. Dadas sus evaluaciones más bajas, se sospecha de inmediato que quien cometió los errores fue el sujeto 2. Luego de procesados los textos, se conversó con ambos sujetos y se corroboró que fue el sujeto 2 quien cometió los errores. Éste detalló que lo que hizo fue ignorar en la construcción del *lead* las preguntas fundamentales relativas al *dónde* y al *cuándo*, considerando sólo el *qué* y el *quién* de las que consideró relevantes y, además, incluyó el *por qué* a sabiendas de que no era relevante ya que trataba de que Ricardo Lagos había ganado las primarias de la Nueva Mayoría la semana pasada, por lo que el hecho ya no era noticioso.

Lo anterior, lleva a pensar preliminarmente, en el marco del pilotaje, que el componente podría tener un funcionamiento correcto y que su sensibilidad a los errores cometidos en la construcción de una parte fundamental de la noticia, como es el *lead*, opera de manera acertada y efectiva.

7.2. Presentación de los resultados de la actividad práctica con estudiantes

La exposición de los resultados obtenidos se estructura en base a las comparaciones realizadas. A saber:

- El titular (unidad textual 1) de la noticia procesada en el componente con un titular tipo (unidad textual 2) basado en las preguntas fundamentales.
- El *lead* (unidad textual 1) de la noticia procesada con las preguntas fundamentales (unidad textual 2).
- El titular y el *lead* como conjunto (unidad textual 1) de la noticia procesada con las preguntas fundamentales (unidad textual 2).
- Comparar los párrafos siguientes al *lead* (unidad textual 1) con los datos del punteo construido (unidad textual 2) que no forman parte de la información relevante proporcionada.

Además, hay que agregar, que en el caso de las evaluaciones realizadas por los humanos, tres de ellas son puntajes que van de 0 a 4 y el restante va de 0 a 7. Es necesario aclarar que no hay problema alguno con que sean escalas distintas, debido a que para el estadístico que se aplicará para comparar los datos -el coeficiente de correlación de Pearson-, la diferencia de escalas no es relevante. En el caso del componente, como ya se indicó, éste arroja sus puntajes de evaluación entre -1 y 1 (aunque en la práctica son entre 0 y 1). También, hay que señalar que para la presentación de los datos se utilizaron dos decimales.

Como se dijo en el párrafo anterior, se calculó el coeficiente de correlación de Pearson, con el fin de ver si las evaluaciones realizadas tanto por la máquina como por los humanos se comportaban de una forma análoga. Para calificar las correlaciones e interpretarlas, se utilizará la tabla propuesta por Fernández Collado, Hernández Sampieri y Baptista (2003).

Tabla IV: Interpretación de las correlaciones.

Valor	Interpretación de la correlación
-1.00	Correlación negativa perfecta.
-0.90	Correlación negativa muy fuerte.
-0.75	Correlación negativa considerable.
-0.50	Correlación negativa media.
-0.10	Correlación negativa débil.
0.00	No existe correlación alguna.
+0.10	Correlación positiva débil.
+0.50	Correlación positiva media.
+0.75	Correlación positiva considerable.
+0.90	Correlación positiva fuerte.
+1.00	Correlación positiva perfecta.

7.2.1. Comparación 1: el titular de la noticia procesada con un titular tipo

La primera comparación, como ya se adelantó, fue entre el titular que cada estudiante escribió para la noticia y un titular tipo que los docentes de la asignatura

elaboraron y consideraron como correcto. Este titular se basa, como ya se explicó, en las preguntas *qué* y *quién,* y es: “Ricardo Lagos proclamado candidato presidencial Nueva Mayoría”. Obviamente se consideraron correctas sus diferentes variantes, como por ejemplo, “Nueva Mayoría proclamó a Ricardo Lagos como su candidato presidencial”. En el fondo, los docentes se fijaron en que estuvieran presentes los elementos: Ricardo Lagos, proclamar, candidato presidencial y Nueva Mayoría. En la Tabla V se presentan los resultados arrojados por la máquina y los asignados por los humanos.

Tabla V: Resultados de la comparación 1.

Identificador	Espacio DK1	Espacio DK2	Espacio DK1y2	Humanos
Sujeto 1	1,00	1,00	1,00	4,00
Sujeto 2	1,00	1,00	1,00	4,00
Sujeto 3	0,71	0,81	0,77	1,50
Sujeto 4	0,88	0,90	0,87	2,50
Sujeto 5	0,74	0,78	0,71	2,00
Sujeto 6	0,29	0,40	0,35	0,50
Sujeto 7	0,61	0,74	0,66	1,50
Sujeto 8	0,69	0,75	0,65	1,50
Sujeto 9	0,72	0,80	0,72	2,00
Sujeto 10	0,89	0,93	0,91	3,00
Sujeto 11	0,97	0,98	0,98	3,50
Sujeto 12	0,89	0,92	0,88	2,50
Sujeto 13	1,00	1,00	1,00	4,00
Sujeto 14	0,96	0,96	0,96	3,00
Sujeto 15	0,83	0,80	0,80	1,50
Sujeto 16	0,66	0,78	0,69	1,50
Sujeto 17	0,60	0,71	0,67	1,00
Sujeto 18	1,00	1,00	1,00	4,00
Sujeto 19	0,94	0,96	0,94	3,00
Sujeto 20	1,00	1,00	1,00	4,00

Para presentar gráficamente los resultados de la tabla anterior se utilizó un gráfico de líneas apiladas, ya que lo importante es que se aprecie si los comportamientos de las distintas evaluaciones, para los diferentes sujetos, siguen tendencias similares. Por lo mismo, con el fin de facilitar la observación de estos resultados, para elaborar el gráfico se transformaron las evaluaciones de los humanos a una escala de 0 a 1, similar a la que ocupa el componente; lo anterior se

hizo mediante una regla de tres simple[11]. En el Gráfico 2 se presentan los resultados.

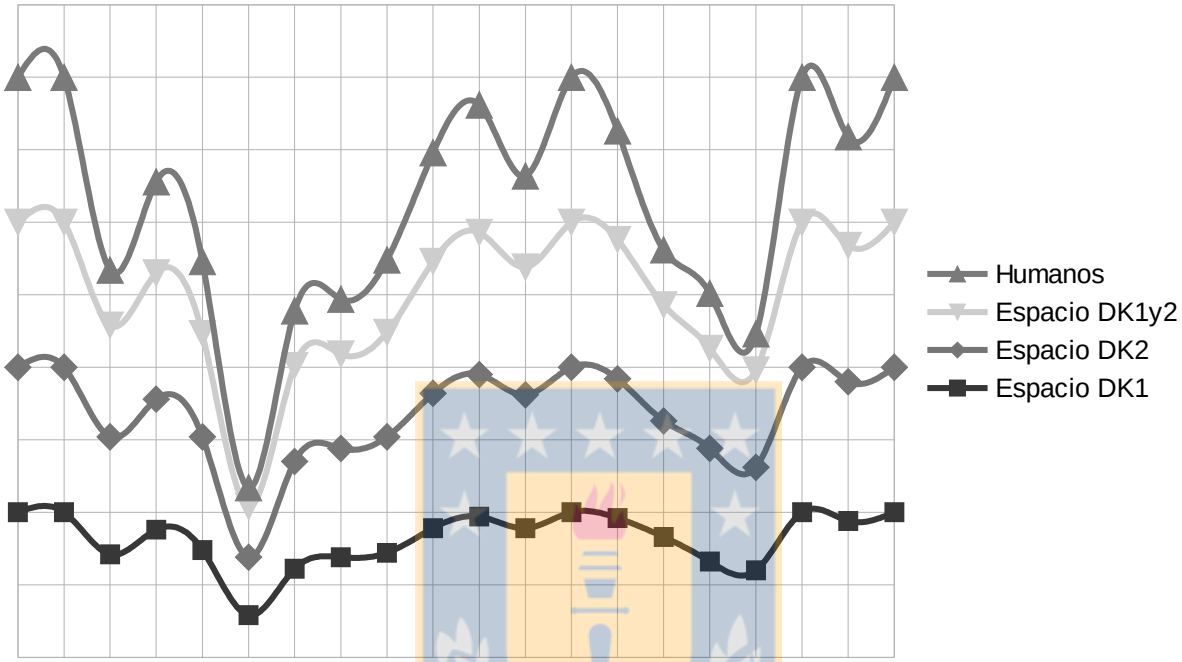

Gráfico 2: Comparación 1, titular de la noticia procesada con un titular tipo.

Al comparar la medición efectuada por los humanos para el titular, con la realizada por la máquina en el espacio semántico DK1, tenemos que el estadístico arroja que se correlacionan positivamente, que está correlación es fuerte o alta y que, además, es significativa (r=.916, p=.000). Para la comparación basada en espacio semántico DK2, los resultados también se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.906, p=.000). Por último, en la comparación basada en el espacio semántico DK1y2, los resultados también se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.924, p=.000). Sin entrar a analizar los resultados a fondo, cuestión que se realizará al final del presente capítulo, se puede adelantar que de la observación del gráfico se desprende que las cuatro líneas siguen tendencias

11 Aunque pueda considerarse una obviedad, es necesario decir que los resultados de las correlaciones que se realizaron no variarían en absoluto si se efectuaran con la escala modificada por la regla de tres. En todo caso, por una cuestión de rigor, las correlaciones se hicieron con la escala original, pese a ser esto indiferente para el resultado de las mismas. Las evaluaciones de los humanos transformadas a una escala de 0 a 1 sólo se realizaron para permitir una mejor visualización de las líneas del gráfico al construir éste.

similares en sus movimientos, lo que se ratifica al obtener tres correlaciones fuertes y significativas en la comparación de los humanos con la máquina para los tres espacios semánticos construidos.

7.2.2. Comparación 2: el *lead* de la noticia procesada con las preguntas fundamentales

La segunda comparación, según se dijo en 6.2.1, fue entre el *lead* de la noticia que escribieron los estudiantes y las preguntas fundamentales consideradas relevantes por los docentes: qué, quién, cuándo y dónde. En la Tabla VI se presentan los resultados arrojados por la máquina y los asignados por los humanos.

Tabla VI: Resultados de la comparación 2.

Identificador	Espacio DK1	Espacio DK2	Espacio DK1y2	Humanos
Sujeto 1	0,82	0,90	0,90	4,00
Sujeto 2	0,85	0,84	0,87	3,50
Sujeto 3	0,57	0,69	0,68	1,50
Sujeto 4	0,84	0,91	0,87	4,00
Sujeto 5	0,91	0,90	0,91	4,00
Sujeto 6	0,79	0,80	0,80	3,00
Sujeto 7	0,81	0,86	0,85	3,50
Sujeto 8	0,78	0,79	0,77	3,50
Sujeto 9	0,87	0,91	0,90	4,00
Sujeto 10	0,77	0,83	0,79	3,00
Sujeto 11	0,61	0,67	0,65	2,00
Sujeto 12	0,79	0,83	0,82	3,00
Sujeto 13	0,84	0,88	0,83	3,50
Sujeto 14	0,96	0,96	0,95	4,00
Sujeto 15	0,69	0,72	0,73	3,00
Sujeto 16	0,85	0,89	0,85	4,00
Sujeto 17	0,85	0,86	0,87	4,00
Sujeto 18	0,81	0,80	0,81	3,00
Sujeto 19	0,74	0,74	0,79	3,00
Sujeto 20	0,89	0,88	0,89	4,00

Al igual que en el caso anterior, para presentar gráficamente los resultados de la tabla se utilizó un gráfico de líneas apiladas y con el fin de facilitar la observación de los resultados en éste, para elaborar el gráfico se transformaron las evaluaciones de los humanos a una escala de 0 a 1, similar a la que ocupa el

componente; lo anterior se hizo mediante una regla de tres simple[12]. En el Gráfico 3 se presentan los resultados.

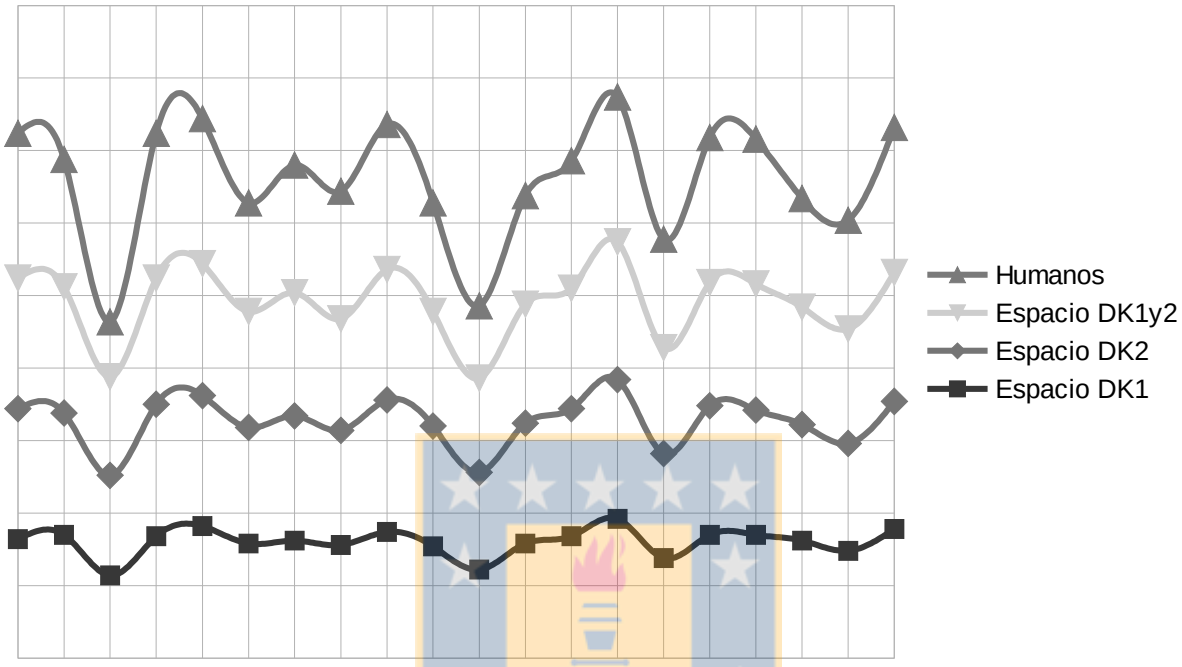

Gráfico 3: Comparación 2, *lead* de la noticia procesada con las preguntas fundamentales.

Al comparar la medición efectuada por los humanos para el *lead* y las preguntas fundamentales relevantes con la realizada por la máquina en el espacio semántico DK1, tenemos que el estadístico arroja que se correlacionan positivamente, que está correlación es fuerte o alta y que, además, es significativa (r=.913, p=.000). Para la comparación basada en espacio semántico DK2, los resultados también se correlacionan positivamente y la correlación es considerable y, además, es significativa (r=.891, p=.000). Por último, en la comparación basada en el espacio semántico DK1y2, los resultados también se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.903, p=.000). Al igual que en el punto anterior, sin entrar a analizar los resultados a fondo, se puede adelantar que de la observación del gráfico se desprende que las

12 Aunque pueda considerarse una obviedad, es necesario decir que los resultados de las correlaciones que se realizaron no variarían en absoluto si se efectuaran con la escala modificada por la regla de tres. En todo caso, por una cuestión de rigor, las correlaciones se hicieron con la escala original, pese a ser esto indiferente para el resultado de las mismas. Las evaluaciones de los humanos transformadas a una escala de 0 a 1 sólo se realizaron para permitir una mejor visualización de las líneas del gráfico al construir éste.

cuatro líneas siguen tendencias similares en sus movimientos. De las correlaciones realizadas, las correspondientes a los espacios semánticos DK1 y DK1y2 son fuertes y significativas; la correlación realizada a la luz del espacio semántico DK2 es significativa, también, aunque sólo es considerable, sin llegar a traspasar el umbral que permitiría considerarla como alta o fuerte; sin embargo, se encuentra muy cerca de serlo (.009).

7.2.3. Comparación 3: el titular y el *lead* (como conjunto) de la noticia procesada con las preguntas fundamentales

La tercera comparación, según se dijo en 6.2.1, fue entre el conjunto formado por el titular y el *lead* de la noticia que escribieron los estudiantes con las preguntas fundamentales consideradas relevantes por los docentes: qué, quién, cuándo y dónde. En la Tabla VII se presentan los resultados arrojados por la máquina y los asignados por los humanos.

Tabla VII: Resultados de la comparación 3.

Identificador	Espacio DK1	Espacio DK2	Espacio DK1y2	Humanos
Sujeto 1	0,89	0,93	0,94	4,00
Sujeto 2	0,90	0,89	0,91	4,00
Sujeto 3	0,60	0,73	0,72	1,50
Sujeto 4	0,88	0,90	0,88	4,00
Sujeto 5	0,87	0,89	0,88	4,00
Sujeto 6	0,77	0,81	0,78	3,00
Sujeto 7	0,82	0,87	0,85	3,50
Sujeto 8	0,77	0,79	0,76	3,00
Sujeto 9	0,87	0,90	0,89	4,00
Sujeto 10	0,86	0,90	0,86	4,00
Sujeto 11	0,68	0,72	0,71	2,00
Sujeto 12	0,81	0,84	0,81	3,50
Sujeto 13	0,88	0,91	0,88	3,50
Sujeto 14	0,94	0,95	0,94	4,00
Sujeto 15	0,72	0,76	0,75	3,00
Sujeto 16	0,84	0,91	0,86	4,00
Sujeto 17	0,86	0,87	0,87	4,00
Sujeto 18	0,86	0,85	0,86	3,50
Sujeto 19	0,81	0,82	0,84	3,50
Sujeto 20	0,92	0,92	0,91	4,00

Al igual que en los casos anteriores, para presentar gráficamente los

resultados de la tabla se utilizó un gráfico de líneas apiladas y con el fin de facilitar la observación de los resultados en éste, para elaborar el gráfico se transformaron las evaluaciones de los humanos a una escala de 0 a 1, similar a la que ocupa el componente; lo anterior se hizo mediante una regla de tres simple[13]. En el Gráfico 4 se presentan los resultados.

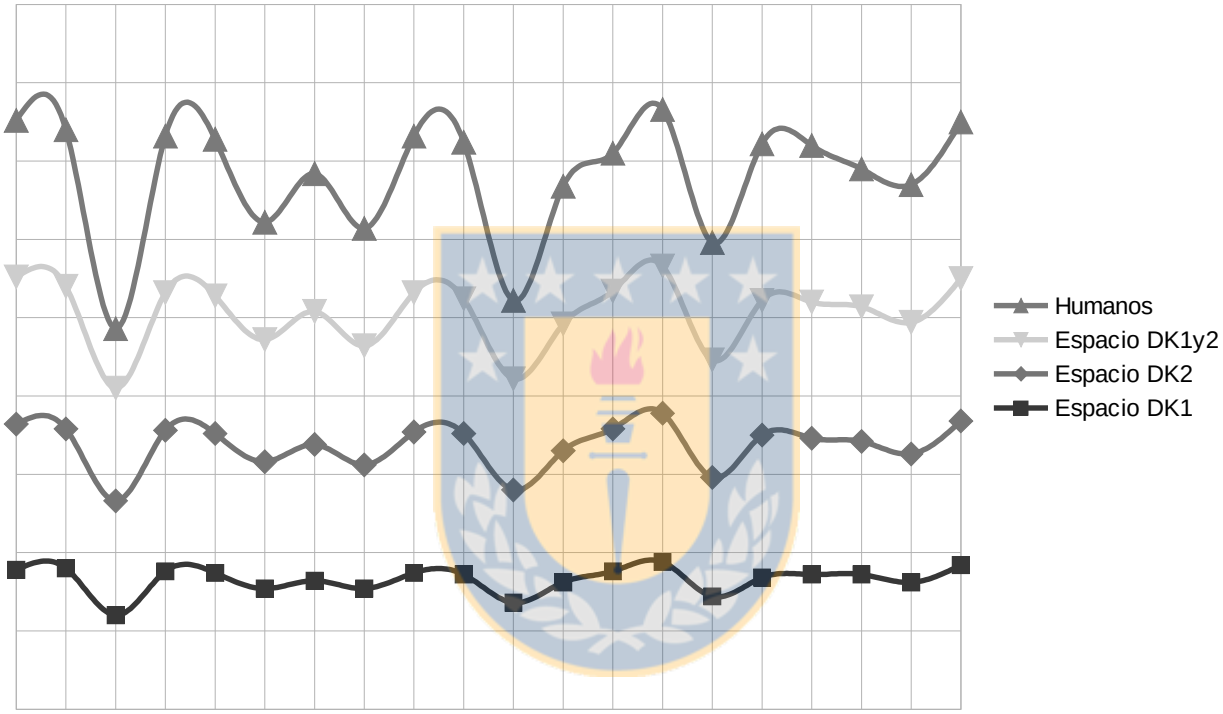

Gráfico 4: Comparación 3, titular y *lead* como conjunto de la noticia procesada con las preguntas fundamentales.

Al comparar la medición efectuada por los humanos para el conjunto titular más *lead* y las preguntas fundamentales relevantes con la realizada por la máquina en el espacio semántico DK1, tenemos que el estadístico arroja que se correlacionan positivamente, que está correlación es fuerte o alta y que, además, es significativa (r=.940, p=.000). Para la comparación basada en espacio semántico DK2, los resultados también se correlacionan positivamente y la correlación es

13 Aunque pueda considerarse una obviedad, es necesario decir que los resultados de las correlaciones que se realizaron no variarían en absoluto si se efectuaran con la escala modificada por la regla de tres. En todo caso, por una cuestión de rigor, las correlaciones se hicieron con la escala original, pese a ser esto indiferente para el resultado de las mismas. Las evaluaciones de los humanos transformadas a una escala de 0 a 1 sólo se realizaron para permitir una mejor visualización de las líneas del gráfico al construir éste.

fuerte o alta y, además, es significativa (r=.905, p=.000). Por último, en la comparación basada en el espacio semántico DK1y2, los resultados también se correlacionan positivamente y la correlación es considerable y, además, es significativa (r=.891, p=.000). Al igual que en el punto anterior, sin entrar a analizar los resultados a fondo, se puede adelantar que de la observación del gráfico se desprende que las cuatro líneas siguen tendencias similares en sus movimientos. De las correlaciones realizadas, las correspondientes a los espacios semánticos DK1 y DK2 son fuertes y significativas; la correlación realizada a la luz del espacio semántico DK1y2 es significativa, también, aunque sólo es considerable, sin llegar a traspasar el umbral que permitiría considerarla como alta o fuerte; sin embargo, se encuentra muy cerca de serlo (.009).

7.2.4. Comparación 4: párrafos siguientes al *lead* con los datos adicionales del punteo construido

La última comparación, según se dijo en 6.2.1, fue entre los párrafos siguientes al *lead* con los datos del punteo construido que no forman parte de la información relevante proporcionada. En la Tabla VIII se presentan los resultados arrojados por la máquina y los asignados por los humanos.

Tabla VIII: Resultados de la comparación 4.

Identificador	Espacio DK1	Espacio DK2	Espacio DK1y2	Humanos
Sujeto 1	0,96	0,96	0,97	7,00
Sujeto 2	0,95	0,95	0,94	6,50
Sujeto 3	0,93	0,95	0,94	6,50
Sujeto 4	0,85	0,90	0,86	6,00
Sujeto 5	0,89	0,90	0,89	6,50
Sujeto 6	0,94	0,96	0,95	7,00
Sujeto 7	0,87	0,91	0,88	6,50
Sujeto 8	0,92	0,95	0,93	7,00
Sujeto 9	0,94	0,96	0,95	6,50
Sujeto 10	0,89	0,90	0,89	6,50
Sujeto 11	0,82	0,86	0,82	5,50
Sujeto 12	0,95	0,97	0,95	7,00
Sujeto 13	0,97	0,98	0,97	7,00
Sujeto 14	0,97	0,97	0,97	7,00
Sujeto 15	0,95	0,97	0,95	7,00
Sujeto 16	0,93	0,96	0,94	7,00
Sujeto 17	0,87	0,95	0,90	6,50
Sujeto 18	0,87	0,94	0,90	6,00
Sujeto 19	0,94	0,97	0,95	7,00
Sujeto 20	0,90	0,95	0,92	6,50

Al igual que en los casos anteriores, para presentar gráficamente los resultados de la tabla se utilizó un gráfico de líneas apiladas y con el fin de facilitar la observación de los resultados en éste, para elaborar el gráfico se transformaron las evaluaciones de los humanos a una escala de 0 a 1, similar a la que ocupa el componente; lo anterior se hizo mediante una regla de tres simple[14]. En el Gráfico 5 se presentan los resultados.

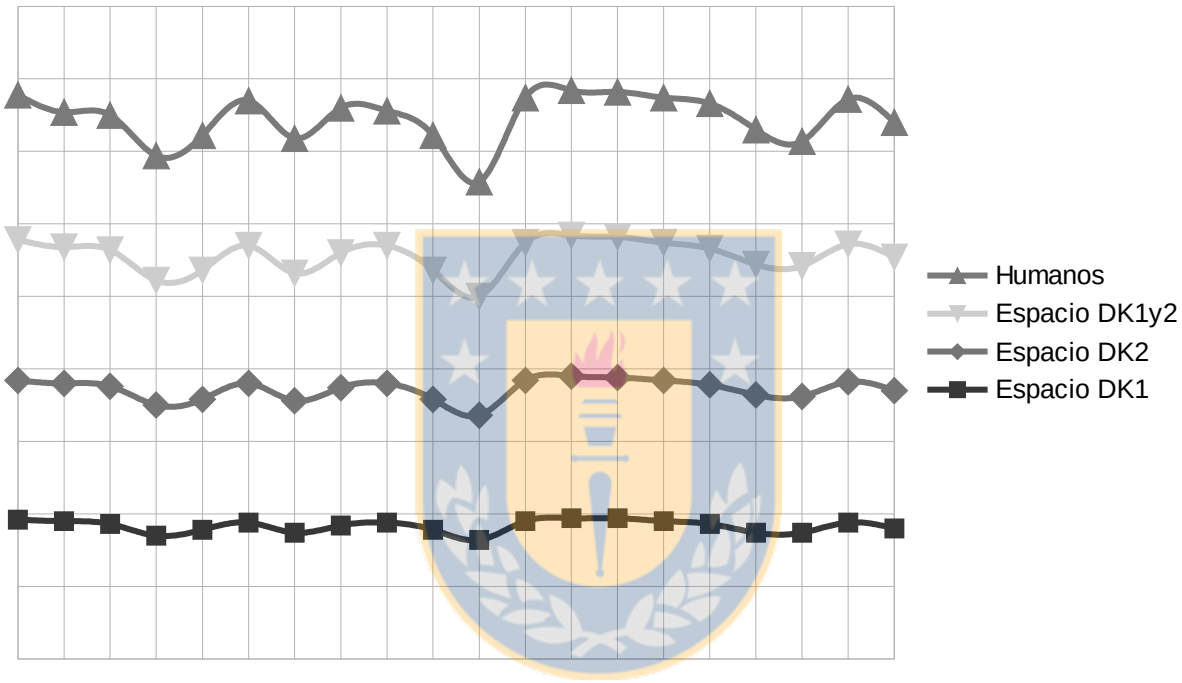

Gráfico 5: Comparación 4, párrafos siguientes al *lead* con los datos adicionales.

Al comparar la medición efectuada por los humanos para los párrafos siguientes al *lead* y los datos adicionales con la realizada por la máquina en el espacio semántico DK1, tenemos que el estadístico arroja que se correlacionan positivamente, que está correlación es considerable y que, además, es significativa (r=.855, p=.000). Para la comparación basada en espacio semántico DK2, los resultados también se correlacionan positivamente y la correlación es considerable y, además, es significativa (r=.807, p=.000). Por último, en la comparación basada en el espacio semántico DK1y2, los resultados también se correlacionan

14 Aunque pueda considerarse una obviedad, es necesario decir que los resultados de las correlaciones que se realizaron no variarían en absoluto si se efectuaran con la escala modificada por la regla de tres. En todo caso, por una cuestión de rigor, las correlaciones se hicieron con la escala original, pese a ser esto indiferente para el resultado de las mismas. Las evaluaciones de los humanos transformadas a una escala de 0 a 1 sólo se realizaron para permitir una mejor visualización de las líneas del gráfico al construir éste.

positivamente y la correlación es considerable y, además, es significativa (r=.864, p=.000). Al igual que en el punto anterior, sin entrar a analizar los resultados a fondo, se puede adelantar que de la observación del gráfico se desprende que las cuatro líneas siguen tendencias similares en sus movimientos, aunque el espacio DK1 tiene movimientos mucho menos pronunciados en su curva. De las correlaciones realizadas, todas son significativas y considerables, sin llegar ninguna a traspasar el umbral que permite considerarlas como altas o fuertes.

7.3. Análisis de los resultados de la actividad práctica con estudiantes

El análisis que se realizará de los resultados que se presentaron en el apartado anterior se enfocará en dos ejes: primero, en los resultados de la máquina a través de los tres espacios semánticos utilizados; y, segundo, en un análisis de la comparación realizada entre los resultados de los humanos y la máquina.

7.3.1. Análisis de los resultados del componente

Si bien la forma adecuada de comprobar el correcto funcionamiento de una aplicación computacional inteligente es comparar su rendimiento con evaluadores humanos, no deja de ser interesante el analizar los resultados de esta aplicación en sí misma, sobre todo en un caso como el del presente trabajo en que se realizaron las mismas comparaciones sobre tres espacios semánticos diferentes, construidos en base a tres corpus de textos de noticias sobre política: uno de 7.165 (DK1), otro de 6.794 (DK2) y otro -conformado por los textos de los dos primeros- de 13.959 (DK1y2).

Una forma interesante de observar cómo se comportan los desempeños del componente sobre los diferentes espacios semánticos, es correlacionar los puntajes de evaluación que arrojan entre ellos, en otras palabras, realizar una comparación de máquina con máquina. Para esto se seguirá el mismo esquema que en la comparación máquina y humano, esto es, se analizarán los mismos ítems evaluados.

Tabla IX: Correlaciones entre los espacios semánticos (comparación 1).

Comparación 1: titular con titular tipo	DK1	DK2	DK1y2
DK1	1	.984	.989
DK2	.984	1	.988
DK1y2	.989	.988	1

En la Tabla IX, como es obvio, se puede ver que al aplicar la correlación de Pearson y comparar una variable con sí misma, su resultado siempre será 1. Ahora lo interesante viene al aplicar la correlación a los resultados de la comparación entre el titular con el titular tipo obtenidos para los espacios DK1 y DK2: los resultados se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.984, p=.000). Para el caso de los espacios semánticos DK1 y DK1y2, los resultados también se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.989, p=.000). Por último, para el caso de los espacios DK2 y DK1y2, nuevamente los resultados se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.988, p=.000).

Tabla X: Correlaciones entre los espacios semánticos (comparación 2).

Comparación 2: *lead* con preguntas fundamentales	DK1	DK2	DK1y2
DK1	1	.930	.954
DK2	.930	1	.950
DK1y2	.954	.950	1

En la tabla anterior, los resultados de la comparación entre el *lead* con las preguntas fundamentales obtenidos para los espacios DK1 y DK2 se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.930, p=.000). Para el caso de los espacios semánticos DK1 y DK1y2, los resultados también se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.954, p=.000). Por último, para el caso de los espacios DK2 y DK1y2, nuevamente los resultados se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.950, p=.000).

Tabla XI: Correlaciones entre los espacios semánticos (comparación 3).

Comparación 3: titular más *lead* con preguntas fundamentales	DK1	DK2	DK1y2
DK1	1	.943	.950
DK2	.943	1	.961
DK1y2	.950	.961	1

En la tabla anterior, los resultados de la comparación entre el titular más el *lead* con las preguntas fundamentales obtenidos para los espacios DK1 y DK2 se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.943, p=.000). Para el caso de los espacios semánticos DK1 y DK1y2, los resultados también se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.950, p=.000). Por último, para el caso de los espacios DK2 y DK1y2, nuevamente los resultados se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.961, p=.000).

Tabla XII: Correlaciones entre los espacios semánticos (comparación 4).

Comparación 4: párrafos posteriores al *lead* con datos adicionales	DK1	DK2	DK1y2
DK1	1	.864	.973
DK2	.864	1	.940
DK1y2	.973	.940	1

Por último, en la Tabla XII, los resultados de la comparación entre los párrafos posteriores al *lead* con los datos adicionales obtenidos para los espacios DK1 y DK2 se correlacionan positivamente y la correlación es considerable y, además, es significativa (r=.864, p=.000). Para el caso de los espacios semánticos DK1 y DK1y2, los resultados también se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.973, p=.000). Por último, para el caso de los espacios DK2 y DK1y2, nuevamente los resultados se correlacionan positivamente y la correlación es fuerte o alta y, además, es significativa (r=.940, p=.000).

Ahora bien, analizando los resultados de las cuatro tablas anteriores como conjunto tenemos que todas las correlaciones son significativas y también altas. La única excepción la constituye el caso de la comparación 4, entre los párrafos posteriores al *lead* con los datos adicionales, en que la correlación entre los espacios DK1 y DK2 arroja una correlación considerable.

Visto todo esto, se puede afirmar que pese a ser corpus distintos, formados por textos diferentes pero pertenecientes al mismo dominio temático (noticias sobre política), el hecho de que las evaluaciones se correlacionen y que estas correlaciones sean altas, indica que la técnica del Análisis Semántico Latente es efectiva y funciona. De hecho, estamos hablando de nueve correlaciones, todas ellas significativas, y todas altas, excepto un caso que es algo menor y se la considera una correlación considerable.

Por supuesto, que el afirmar que la técnica del Análisis Semántico Latente funcione correctamente no es ninguna novedad y aporte, sino que sólo una constatación de algo que ya había sido probado por otros autores.

Eso sí, en el marco del presente trabajo de tesis, el hecho de comprobar que la técnica del LSA funciona correctamente es un indicador importante de que el componente construido funciona de manera adecuada y efectiva no sólo al medir coherencia textual, como ya se testeó en el capítulo 4 para el módulo 1, sino que también en la aplicación que se le quiere dar en el módulo 3: evaluar la jerarquización al construir la pirámide invertida en la producción escrita de noticias. Lo anterior, al menos al comparar la máquina con sí misma en su funcionamiento en diferentes espacios semánticos. En el apartado siguiente, se analizará el comportamiento de la máquina al comparar su desempeño con humanos.

7.3.2. Análisis de la comparación realizada entre los resultados de la máquina y los humanos

Antes de entrar al foco de este apartado propiamente tal, es oportuno señalar que podría pensarse que al unir los textos de los corpus DK1 y DK2 en un solo corpus, el espacio construido en base al corpus resultante (DK1y2) debería entregar, al realizar operaciones sobre él, puntajes que correspondieran al promedio de las mediciones que se hicieran sobre los espacios DK1 y DK2. Según lo que se

desprende del funcionamiento del Análisis Semántico Latente (LSA), esto no debiera necesariamente ser así, ya que el LSA es un método estadístico automático para la representación del significado de las palabras y pasajes de texto, basado en el análisis de extensos *inputs* de texto, a partir de los cuales se genera un espacio semántico y en éste las palabras, oraciones y el texto completo se representan a través de vectores (Kintsch et al., 2000). Analizando en más detalle la idea recién expuesta, se puede colegir que al variar el *input* debería variar el espacio semántico construido en que se representa el significado de las palabras, por lo tanto los puntajes arrojados no deberían ser iguales aunque el espacio DK1y2 sea, en rigor, un espacio creado a partir de la combinación de los corpus de los cuales se crean los espacios DK1 y DK2. Lo anterior se sostiene, a nuestro juicio, en que si bien el LSA es un método matemático estadístico, su materia prima, es decir, el *input* con el que trabaja no son números, sino que palabras; por ello, al combinarse dos corpus diferentes en uno solo, la representación de esas palabras en vectores no debiera necesariamente corresponder a una suerte de *promedio*. Para comprobar lo anterior, se presenta la Tabla XIII.

Tabla XIII: Resumen de las correlaciones (humanos-máquina).

Comparación	Correlación DK1	Correlación DK2	Correlación DK1y2
Comparación 1: titular con titular tipo	.916	.906	.924
Comparación 2: *lead* con preguntas fundamentales	.913	.891	.903
Comparación 3: titular más *lead* con preguntas fundamentales	.940	.905	.891
Comparación 4: párrafos posteriores al *lead* con datos adicionales	.855	.807	.864

Como se puede ver, en el caso de la comparación 1, el índice de correlación con DK1 es de .916; para el caso de DK2, es de .906. Si asumiéramos la idea del promedio, al ser DK1y2 un constructo creado a partir de la combinación de los corpus que originan DK1 y DK2 se podría creer que el índice de correlación de la

comparación 1 con DK1y2 debiera ser .911. Sin embargo, como se puede ver, el valor de éste, no sólo no corresponde a dicha cifra, sino que es mayor al índice de correlación con DK1 y DK2, al ser de .924.

En el caso de la comparación 2, el promedio de DK1 y DK2 sería .902, y la correlación con DK1y2 es de .903. Para la comparación 3 el promedio sería de .923, y la correlación con DK1y2 es de .891. Por último, en la comparación 4 el promedio sería de .831 , y la correlación con DK1y2 es de .864.

Por lo tanto, como se puede ver de lo recién expuesto, efectivamente el que dos corpus -a partir de los cuales se construyeron dos espacios semánticos distintos, DK1 y DK2- se combinen para construir a partir de ellos un espacio semántico multidimensional (DK1y2), no implica en absoluto que este nuevo espacio semántico vaya a arrojar en los resultados de las mediciones que se realicen sobre él, una suerte de promedio de los resultados de las mediciones realizadas en los dos espacios semánticos que se construyeron a partir de cada corpus original por separado. Zanjado este punto, ahora sí es posible entrar al análisis de los resultados propiamente tal.

En el Gráfico 6 se presentan los gráficos de las cuatro comparaciones realizadas entre la máquina y el componente, de una sola vez, con el fin de facilitar su visualización como conjunto y dar pie al análisis.

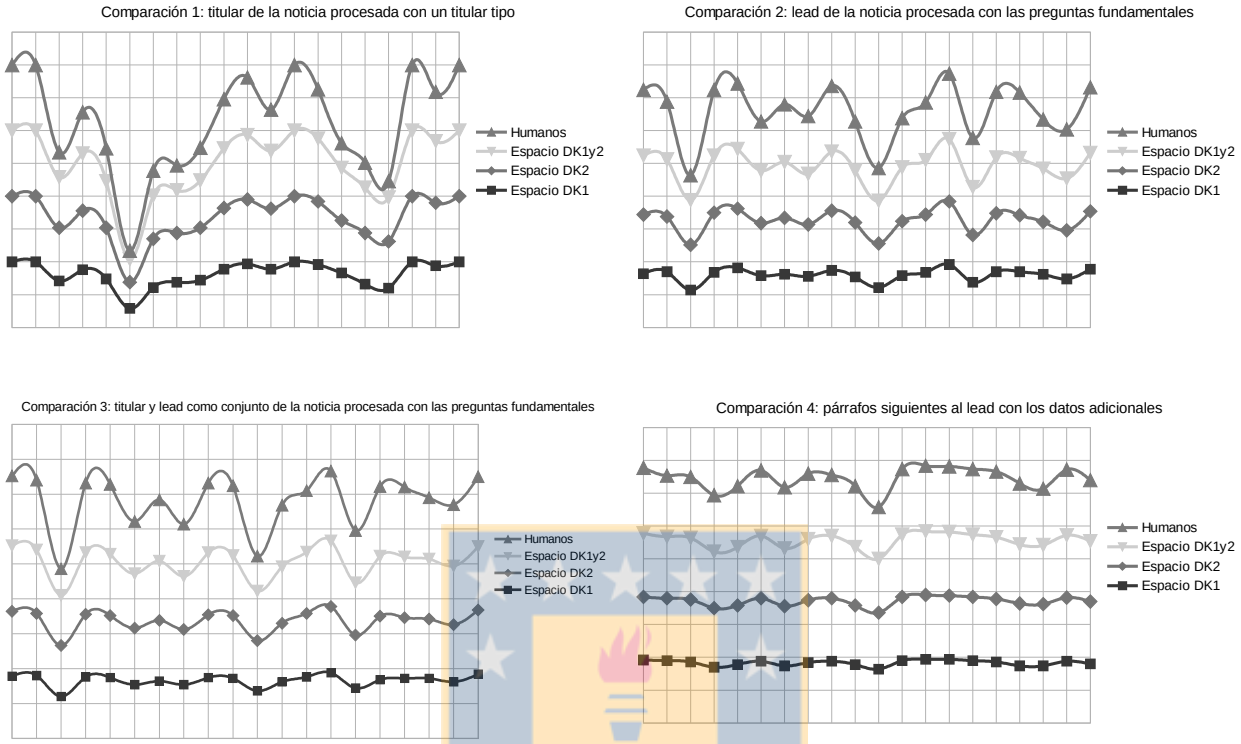

Gráfico 6: Gráficos de las cuatro comparaciones realizadas.

La idea de combinar los gráficos en una imagen fue para centrarse en el movimiento de las líneas que representan los puntajes asignados a cada una de las cuatro comparaciones por los humanos y la máquina (en los espacios DK1, DK2 y DK1y2). En este punto, es necesario recordar que estos gráficos se construyeron en base a las tablas presentadas en 7.2 que dan cuenta de los resultados de las comparaciones realizadas para cada uno de los 20 sujetos que participaron de la experiencia; dichos sujetos -los estudiantes- corresponden a cada uno de los puntos que hay marcados en las líneas de los gráficos.

Como se puede ver, las líneas presentan movimientos y tendencias similares en los cuatro gráficos, esto es, cuando una línea sube, las otras también lo hacen; y cuando una línea baja, las otras siguen la misma dirección.

Lo anterior indica, centrándonos sólo en el caso de la máquina, que cuando ésta asigna un puntaje en la medición realizada a la luz de un espacio semántico, las comparaciones efectuadas sobre los otros espacios semánticos arrojan puntajes

que siguen la misma tendencia, ya sea ésta al alza del valor a o la baja de éste. Sin embargo, a esta altura lo señalado ya no debiera ser sorpresa, sobre todo por el análisis de las correlaciones entre los distintos espacios semánticos presentadas en el apartado precedente.

Lo más interesante de esto es que en el caso de la línea que representa a los evaluadores humanos, la tendencia se repite. Dicho en otras palabras, cuando la máquina disminuye su evaluación para un sujeto determinado en un punto específico -en cualquiera de los tres espacios semánticos-, el puntaje asignado por el humano sigue una tendencia similar; y, en el caso contrario, cuando el puntaje asignado por la máquina aumenta, el asignado por los humanos sigue la misma tendencia.

Lo recién expresado es una señal clara de que el componente es sensible a lo que se busca medir en cada una de las comparaciones. Por lo mismo, centrándose, por ejemplo, en el caso del titular, si el humano detecta que éste tiene algún problema que implica que el puntaje asignado al estudiante debiera disminuir, el componente al analizar el texto escrito es capaz de evaluar lo anterior de la misma forma que el humano y, en consecuencia, también asignar un puntaje menor.

Pese a lo alentador de estos resultados, no hay que dejar de advertir que, aunque muy leves, hay excepciones. Éstas se aprecian con ojo atento en los gráficos y corresponden al caso del sujeto 18, en las comparaciones 2 y 3. La variación que presentan las líneas en estos casos son muy leves, pero existen: tienden a bajar un poco en el caso de la evaluación de los humanos.

Con el fin de intentar comprender lo anterior, primero hay que decir que las comparaciones 2 y 3 corresponden a la comparación entre el *lead* y las preguntas fundamentales, y a la combinación del titular más el *lead* con las preguntas fundamentales. Para analizar la situación, a continuación se presentan el titular y el *lead* de la noticia que escribió el sujeto.

Nueva Mayoría proclama a Ricardo Lagos como su candidato presidencial

En un acto celebrado a las 21 horas del día de ayer en las dependencias del Teatro Caupolicán, la Nueva Mayoría dio inicio a la carrera presidencial, que espera llevar a Ricardo Lagos a su segundo ciclo en la Moneda, esto luego de que el ex mandatario se impusiera en las primarias de la coalición a Isabel Allende, Jorge Burgos y José Antonio Gómez.

Primero, aunque no es el foco del problema, es posible darse cuenta de inmediato que en el titular están presentes todos los elementos del titular tipo y, por ende, responde a las preguntas *qué* y *quién*. Si revisamos, tanto la máquina como los humanos, le dieron puntuación máxima al estudiante en este apartado (comparación 1). En relación a las preguntas fundamentales, dado lo anterior, es posible afirmar que para la comparación 3, que combina titular con *lead*, ya estarían respondidas las preguntas *qué* y *quién*. Si buscamos en el *lead* la respuesta a otra de las preguntas fundamentales, se puede ver que aparece el *dónde*, que en este caso, es el *Teatro Caupolicán*. En relación al *cuándo*, tenemos que el estudiante señala que fue a las *21 horas de ayer*, aunque en los datos decía a las *21 horas de hoy*. En esta diferencia temporal puede estar la explicación de por qué los docentes le asignaron a la comparación 3 un puntaje de 3,5 sobre un máximo de cuatro. En el caso de la comparación 2, la posible deficiencia que se aprecia es que el estudiante no detalló claramente los elementos que responden a la pregunta *qué*, esto es, que el sujeto de la noticia -el *quién*, es decir, Ricardo Lagos- fue proclamado candidato presidencial de la Nueva Mayoría, sino que sólo relata que la Nueva Mayoría dio inicio a la carrera presidencial que busca llevar a Lagos a un segundo periodo en La Moneda. Por lo recién dicho, se estima que el estudiante perdió algo de puntaje en este apartado, lo que sumado a la respuesta un tanto inexacta al *cuándo*, habría llevado a los docentes a evaluar al alumno con tres puntos sobre un máximo de cuatro (hay que recordar que la comparación 2 no considera al titular, por lo que no vale lo que en éste se dice como respuesta al *qué* y *quién*). Teniendo claro lo anterior, podría intentar explicarse la leve tendencia al alza que manifiesta la evaluación de la máquina en los tres espacios semánticos para ambas comparaciones, basándose en la respuesta a la pregunta *dónde*, ya que para ésta, al medir la relación semántica entre *hoy* y *ayer*, obviamente que va a ser muy alta, como suele ocurrir entre los antónimos. Sin embargo, hay que decir que este intento de explicar la situación no sería válido, ya que *hoy* y *ayer* están entre las *stop words* de Infomap-NLP[15] y, por ende, no fueron consideradas en la evaluación,

15 Manning, Raghavan y Schütze (2008) señalan que hay palabras sumamente comunes que son de escaso valor en la tarea de seleccionar documentos que concuerden con la petición de un usuario. Estas palabras son excluidas del vocabulario de la aplicación que recupera información y se les llama stop words (traducido como palabras de parada, palabras de relleno o palabras de paso). En el Diccionario Tecnológico (2015) del sitio web chileno Usando.info, que trata sobre Arquitectura de la Información, Usabilidad y Accesibilidad de sitios web, se entrega otra definición de stop words y señala que “son aquellas palabras que por ser comunes, los buscadores

precisamente por el equívoco a que induciría la alta relación semántica que arrojarían al compararse. Por otro lado, todas éstas son sólo suposiciones, ya que se desconoce si efectivamente éste fue el criterio aplicado por los docentes. Por lo demás, se trata de sólo dos puntos entre todos los que se exponen en los gráficos presentados (80 ítems evaluados por los humanos) y la variación es muy leve. Sin embargo, la situación se retomará en las conclusiones de este trabajo (capítulo 8).

Así como es interesante analizar las tendencias de las líneas de las evaluaciones, también es importante profundizar en los resultados obtenidos al correlacionar el juicio de la máquina con el de los humanos. En la Tabla XIV se presentan los resultados de las correlaciones para la comparación 1.

Tabla XIV: Correlaciones humanos-máquina (comparación 1).

Humano	Correlación DK1	Correlación DK2	Correlación DK1y2
Comparación 1: titular con titular tipo	.916	.906	.924

Como ya se señaló todas estas correlaciones son significativas y altas, lo que permite afirmar que el componente es capaz de evaluar de una forma similar al juicio de los evaluadores humanos en el caso de la comparación del titular producido por el estudiante con el titular tipo que establecieron los docentes. De hecho, en esta comparación fue en la única de las cuatro en que las correlaciones en los tres espacios semánticos construidos fueron altas o fuertes. Por ello se puede señalar que, para el caso en cuestión, el método probado demostró ser eficiente y apropiado para emular el juicio que tendría un humano.

Tabla XV: Correlaciones humanos-máquina (comparación 2).

Humano	Correlación DK1	Correlación DK2	Correlación DK1y2
Comparación 2: *lead* con preguntas fundamentales	.913	.891	.903

En el caso de la comparación 2, si bien todas las correlaciones fueron

ignoran para asegurar la calidad de los resultados de lo que se busca. Normalmente entran en esta categorías las proposiciones y conjunciones". En el caso del presente trabajo, se modificó el código de Infomap-NLP para incluir en éste el mismo listado de stop words del español que se empleó en Hernández (2010) y Hernández y Ferreira (2010).

significativas, a diferencia del caso anterior, los puntajes obtenidos en la comparación sobre el espacio semántico DK2 arrojaron una correlación considerable, aunque muy cercana al umbral que permitiría considerarla como alta. Por lo mismo, también se puede señalar que, para el caso en cuestión, el método probado demostró ser eficiente y apropiado para emular el juicio que tendría un humano.

Tabla XVI: Correlaciones humanos-máquina (comparación 3).

Humano	Correlación DK1	Correlación DK2	Correlación DK1y2
Comparación 3: titular más *lead* con preguntas fundamentales	.940	.905	.891

Para la comparación 3, ocurrió una situación parecida a la anterior, ya que dos correlaciones fueron altas y una considerable. En el caso de esta última, para la medición realizada sobre el espacio DK1y2, también fue muy cercana al umbral que permitiría considerarla como alta (como en la comparación 2). Por ello, para esta comparación en particular, también se puede afirmar que el método probado demostró ser eficiente y apropiado para emular el juicio que tendría un humano.

Tabla XVII: Correlaciones humanos-máquina (comparación 4).

Humano	Correlación DK1	Correlación DK2	Correlación DK1y2
Comparación 4: párrafos posteriores al *lead* con datos adicionales	.855	.807	.864

El caso de la comparación 4, que se presenta en la tabla anterior, es el único de los cuatro en que todas las correlaciones resultan ser considerables y no hay ninguna alta. Hay que decir también que es el único caso diferente al resto, ya que no se centra en la jerarquización de los elementos de la noticia que involucran la estructura de pirámide invertida, esto es, el titular y el *lead*. Este caso se enfoca en comparar los párrafos posteriores al *lead* de la noticia construida por los estudiantes -como una sola unidad textual- con los datos adicionales proporcionados en la actividad práctica, es decir, los datos que no debían ser considerados para

responder las preguntas fundamentales. Por lo mismo, este caso no forma parte directamente del foco principal del presente trabajo, que pone énfasis en los dos estadios superiores de la pirámide invertida: titular y *lead*, como ya se señaló al explicar el modelo planteado para la evaluación de la estructura de pirámide invertida en una noticia, en 6.2.1. Sin embargo, como también se dijo en la ocasión, la construcción de los párrafos posteriores al *lead* no deja de tener importancia al también formar parte de la noticia, por lo mismo no se pretende minimizar estos resultados menos alentadores que en las tres primeras comparaciones.

Ahora bien, si se observa el gráfico en que se presentan las líneas de las distintas evaluaciones, como ya se adelantó, todas siguen tendencias similares, sin embargo las curvas son mucho menos marcadas en los casos de las evaluaciones en los espacios DK1 y DK2; la curva más marcada se observa en el espacio DK1y2. Por otra parte, si se atiende a la curva de las evaluaciones de los humanos para la comparación 4, tenemos que la curva es marcada y al espacio que más se asimila es, precisamente, al DK1y2. En el caso de las correlaciones, la más alta se da precisamente entre los humanos y el espacio DK1y2 (r= .864). En base a todo lo anterior, se podría dar como una posible explicación que al consistir la comparación realizada por la máquina en dos unidades textuales que tenían una cantidad de palabras mucho mayor cada una de ellas que las de las tres primeras comparaciones, el hecho de emplear un espacio semántico mayor en cuanto a número de palabras utilizadas en su construcción, implicaría que la comparación con los puntajes entregados por los humanos podría ser más precisa y, por ende, más cercana como fue en el caso de las comparaciones precedentes. Sin embargo, lo anterior es sólo una posible explicación y en ningún caso hay que considerarla como algo concluyente. Por otro lado, pese a no ser el foco principal de este trabajo, los resultados para la comparación 4 tampoco pueden ser considerados como malos, sino que simplemente como la parte en que el componente opera con menos acierto de las cuatro mediciones realizadas.

Por último, se puede señalar que finalizado el trabajo, es posible afirmar que el componente funcionó de una manera acorde a la esperada en las tres comparaciones en que se enfocaba principalmente la tarea. En la comparación restante, el componente operó de una forma adecuada, aunque se hubieran

esperado mejores resultados. Esto último, debe considerarse una motivación para perfeccionar esta parte del modelo en el futuro.

7.4. Valoración de los resultados de la actividad práctica con estudiantes

Una vez concluido el análisis de los resultados, se puede afirmar en relación al objetivo general del presente trabajo, que es “desarrollar un componente, a nivel de prototipo, que evalúe el proceso de escritura de noticias, enfocándose en la evaluación de la coherencia textual y en el aspecto semántico estructural de éstas (estructura de pirámide invertida), capaz de obtener un rendimiento equivalente o superior a la realizada por un evaluador humano experto”, que este objetivo se logró.

En relación al primer objetivo específico, que es “desarrollar un módulo, a nivel de prototipo, capaz de evaluar la coherencia textual en noticias”, se señala que también se logró, pues el módulo 1 cumple acertadamente la tarea para la cual fue concebido. Es más, dicho módulo podría ser utilizado en cualquier otro dominio temático, simplemente cambiando el corpus con el cual se construye el espacio semántico.

En el caso del segundo objetivo específico, “desarrollar un módulo, a nivel de prototipo, capaz de evaluar en forma automática el aspecto semántico estructural de una noticia, esto es, relevancia y jerarquización de la información presentada, en textos pertenecientes a un dominio temático específico”, se puede afirmar que también se cumplió: el módulo 3 es capaz de evaluar correctamente la estructura semántica de pirámide invertida en una noticia, atendiendo a la jerarquización que se realiza de los datos más relevantes al escribir el texto.

En lo referente al tercer objetivo específico del trabajo, que es “comparar la evaluación efectuada por el componente -de la relevancia y jerarquización de la información presentada en una noticia- con la realizada por evaluadores humanos”, este objetivo se cumplió, pues una vez diseñado el componente se realizó un actividad práctica con 20 sujetos, que escribieron 20 noticias y éstas fueron procesadas en el componente que las evaluó en tres espacios semánticos multidimensionales distintos. Dado que esta actividad práctica fue realizada como una actividad con nota en un curso de primer año de la Carrera de Periodismo de la

Universidad de Concepción, los 20 textos producidos también fueron evaluados por los dos docentes a cargo de la asignatura. Finalmente, las evaluaciones del componente y las de los humanos, para cada uno de los textos, fueron comparadas a través del coeficiente de correlación de Pearson.

Para la presente investigación, como ya se expuso al principio de esta tesis, se utilizó la hipótesis siguiente:

> H_1: La evaluación automática de la estructura semántica (pirámide invertida) de las noticias escritas, pertenecientes a un dominio temático específico, tiene un rendimiento equivalente o superior a la realizada por un evaluador humano.

La hipótesis nula es:

> H_0: La evaluación automática de la estructura semántica (pirámide invertida) de las noticias escritas, pertenecientes a un dominio temático específico, no tiene un rendimiento equivalente o superior a la realizada por un evaluador humano.

A la luz de los análisis realizados, se afirma que la hipótesis (H_1) se sostiene y, por lo tanto, se descarta la hipótesis nula (H_0). Esto pues, como ya se demostró, la evaluación automática de la estructura semántica (pirámide invertida) de las noticias escritas por los estudiantes de periodismo, es equivalente a la evaluación efectuada por humanos en el dominio formado por las noticias sobre política.

Se recuerda que, al definir ambas hipótesis, se conceptualizó *equivalente* no en el sentido de que la evaluación entregada por la herramienta prototipo (componente) fuera igual a la entregada por los evaluadores humanos. Es decir, en el presente trabajo, no se pretendió probar que si la máquina asignaba una evaluación X a un texto determinado, los evaluadores humanos también debían asignar X. Como se dijo, el sentido en que debe entenderse *equivalente* es que ambas evaluaciones -máquina y humanos- sigan tendencias similares y se correlacionen, situación que como recién se indicó, se sostiene a la luz del análisis de los datos obtenidos.

Se reitera que el caso de la comparación 4, si bien no era el foco principal del trabajo, fue una medición que se decidió realizar por la idea futura de integrar el prototipo en un sistema tutorial inteligente destinado a apoyar la enseñanza de

escritura de noticias en estudiantes de periodismo. El hecho de que los resultados de esta comparación no fueran tan alentadores como los de las tres primeras comparaciones, en ningún caso afecta el hecho de que la hipótesis se sostenga. Sin embargo, es un desafío a tener en cuenta para mejorar en el futuro trabajo.

Capítulo 8: Conclusiones

Este trabajo de tesis se enmarcó en el contexto disciplinar de la Lingüística Aplicada, específicamente en el ámbito de la interdisciplina denominada lingüística informática o computacional (siguiendo a Pastor, 2004). Desde allí se propuso el objetivo de desarrollar un componente, a nivel de prototipo, que evalúa el proceso de escritura del discurso especializado denominado noticias, enfocándose en la evaluación de la coherencia textual y en el aspecto semántico estructural de estas noticias, esto es, la estructura de pirámide invertida. Dicho componente se desarrolló con éxito y se construyó en base a tres módulos: uno encargado de la evaluación de la coherencia textual de las noticias (módulo 1, según se denominó en la introducción) y otro que evalúa la correcta construcción de la pirámide invertida (módulo 3); un tercer módulo (módulo 2) se encarga de analizar y corregir los errores ortográficos, con el fin de apoyar el funcionamiento de los primeros dos módulos mencionados. La construcción de este componente se realizó mediante el lenguaje de programación Python 3 y su funcionamiento se apoya en NLTK (Natural Language Toolkit) y, sobre todo, en el método del Análisis Semántico Latente. Para comprobar la efectividad del componente se utilizaron 20 noticias escritas por 20 estudiantes de Primer Año de Periodismo de la Universidad de Concepción, en una situación real de evaluación dentro del aula. Posteriormente, se correlacionaron las evaluaciones del componente y las de los profesores de la asignatura, con el fin de observar relaciones y equivalencias entre ambas.

Como primer punto de estas conclusiones es necesario resaltar uno de los aportes más interesantes del presente trabajo: tomar el método que se ha utilizado para medir la coherencia textual mediante Análisis Semántico Latente y, cambiando la forma de emplearlo, conseguir que opere con eficiencia en un ámbito diferente del que se utilizó en su concepción original. Gracias a esto, es posible evaluar la correcta construcción de la pirámide invertida en el caso de las noticias sobre política, lo que constituye un aporte original y novedoso que se podrá utilizar como base de futuros trabajos, tal cual se ha señalado a lo largo de esta tesis: la idea de construir un sistema tutorial inteligente destinado a apoyar la enseñanza de la producción escrita de noticias en estudiantes de periodismo. Teniendo esta tarea

como desafío futuro, nada impide pensar que con un corpus de un dominio temático diferente, por ejemplo, noticias policiales o sobre deporte, la técnica empleada deje de ser efectiva. En todo caso, si bien hay que constatarlo, la tarea realizada es una base importante en la que sustentarse para futuros desarrollos.

También es importante señalar que en la construcción de los módulos se utilizaron solamente herramientas de software libre ya disponibles. Lo más relevante de esto, es que dichas herramientas tienen la ventaja de que no se vulnera ninguna licencia comercial al utilizarlas, por lo que ambos prototipos son perfectamente legales. Además, al ser software libre está la posibilidad de modificar el código fuente del programa, como de hecho se hizo en una breve porción del código de Infomap-NLP, con el fin de adaptarlo a las necesidades específicas que se requerían. De esta forma, se logró que los módulos 1 y 3 trabajaran con un listado de *stop words* del español y no las originales en inglés. Relacionado con lo anterior, hay que señalar que si bien no estaba entre los objetivos del presente trabajo, surgió la necesidad de desarrollar un módulo corrector ortográfico con el fin de hacerse cargo de este tipo de error, ya que las incorrecciones ortográficas afectan las evaluaciones realizadas por los módulos 1 y 3; en la construcción de este módulo, también se utilizaron herramientas de software libre ya disponibles.

Continuando con las conclusiones, una explicación que quedó pendiente desde 7.1 y que se ofrece aquí luego de cerradas todas las mediciones que exigió esta tesis, es la que dice relación sobre por qué el puntaje de coherencia textual se mueve en rangos diferentes al de la evaluación de la jerarquización en la pirámide invertida. Tal como se señaló en 4.4, un puntaje de coherencia superior a 0,60 se considera medio-alto en el caso de la evaluación de coherencia textual; de hecho, el puntaje más alto que se obtuvo en Hernández y Ferreira (2010) fue de 0,72. En el caso de las evaluaciones de la pirámide invertida, los puntajes se mueven entre 0,80 a 1,00 en su mayoría. Obviamente, puntajes tan altos es muy improbable que se den al evaluar la coherencia textual. Lo anterior se explica porque en el caso de la evaluación de la coherencia textual, lo que se está midiendo al comparar el coseno de los ángulos formados por los vectores que representan a cada unidad textual, es la progresión en un texto, dicho de otro modo, se mide cómo un texto va progresando en su estructura de manera correcta, manteniendo la secuencia de lo

previo con la nueva información que se va introduciendo párrafo a párrafo, o sea, siguiendo a Jurafsky y Martin (2008), cómo el significado de diferentes unidades textuales se va combinando para construir un significado discursivo mayor. En cambio, en las mediciones realizadas para evaluar la correcta construcción de la pirámide invertida -jerarquización de datos-, se utiliza la técnica para comparar ya no unidades textuales en progresión, sino que unidades textuales que contienen información similar, es decir palabras que aunque sean diferentes tienen el mismo significado y, por ende, una relación semántica alta. Por lo mismo, como ya se dijo, al comparar una palabra con sí misma mediante el LSA el resultado será 1,00, por ello al comparar unidades textuales mayores que contengan información que se espera que sea similar, los puntajes obtenidos tenderán a aumentar su rango de oscilación en relación a la medición de coherencia textual. De ahí que los puntajes que arroja el segundo prototipo, por lo general, partan de 0,80 hacia arriba. Lo recién explicado es, como se señaló al inicio de estas conclusiones, uno de los aportes más interesantes del presente trabajo: tomar una técnica que se ha utilizado para medir la coherencia textual y, cambiando la forma de emplearla, conseguir que opere con eficiencia en un ámbito diferente del que se utilizó en su concepción original. Gracias a esto, es posible evaluar la construcción de la pirámide invertida en el caso de las noticias sobre política.

Dentro de las conclusiones es importante agregar que el trabajo de esta tesis doctoral, se relaciona con la investigación de Hernández y Ferreira (2010) -y con Hernández (2010)-, cuyo foco era la evaluación automática de coherencia textual en noticias policiales. Si bien esta tesis y el trabajo de 2010 son dos investigaciones independientes con un foco distinto, en la investigación de 2010 con todas sus limitaciones, se observaron algunas situaciones que quedaron planteadas en las conclusiones para mejorar en trabajos futuros. Por ello, dado que hay similitudes en algunas partes de ambos trabajos, es oportuno recordar en estas páginas las situaciones planteadas en aquella ocasión que se atendieron en la presente tesis, para ver cómo evolucionaron.

Una primera idea que quedó planteada en 2010 fue la siguiente: “La posibilidad de mejorar el *script* en que se basa el prototipo. Si bien el utilizado cumple a cabalidad con lo que se propone, seguramente un profesional experto en

informática podría perfeccionarlo aún más, aumentando su eficiencia en el sentido de tiempo de procesamiento y consumo de recursos de la máquina en que opera. Por otro lado, es seguro que hay lenguajes de programación que superan lo que el intérprete de órdenes bash empleado es capaz de realizar. Por supuesto que lo anterior, no pretende desmerecer en absoluto el trabajo realizado en el *script* que se construyó, pues el tiempo de procesamiento de éste es de alrededor de 10 segundos y el consumo de recursos es bajísimo; simplemente, lo que se señala, es la posibilidad de perfeccionarlo". En relación a la idea expuesta, en esta tesis también se construyó un analizador automático de coherencia textual (módulo 1), como se describió en el capítulo 4, ya que es necesario para el funcionamiento del módulo 3 que evalúa la estructura de pirámide invertida de las noticias. Sin embargo, pese a que los dos utilizan Infomap-NLP como herramienta de Análisis Semántico Latente, una gran diferencia entre una implementación y otra es que la primera se programó utilizando *scripts* basados en comandos bash; en esta segunda ocasión, para el desarrollo se utilizó Python 3 y NLTK. Un avance esencial surgido con este cambio es la mejora radical en la eficiencia del prototipo (módulo 1) en cuanto a consumo de recursos y tiempo de procesamiento, reduciéndose este último a menos de la décima parte del tiempo empleado en el prototipo anterior. Otra ventaja de esta nueva implementación, es que al operar el *script* que constituye este prototipo (módulo 1) en forma recursiva, permite la comparación de tantas parejas de unidades textuales adyacentes como se quiera y no tiene la limitación de 40 parejas que tenía el prototipo anterior. Por último, otra nueva función que tiene es que permite procesar más de un texto a la vez, a diferencia del prototipo de 2010 que sólo permitía trabajar con un único texto. Por todo lo anterior, a diferencia del trabajo de 2010, esta vez no quedó una sensación de que pudo utilizarse un lenguaje de programación mejor: Python 3 cumplió con todas las expectativas y, sin duda, es el lenguaje que se utilizará en el futuro proyecto que continúe el trabajo de la presente tesis.

Otro punto que surgió en la investigación de 2010 fue la necesidad de perfeccionar el corpus con que se construyó el espacio semántico. Se señaló que esta mejora debía ir en dos sentidos. El primero de ellos era equilibrar el número de textos que tratan sobre los diferentes temas incluidos dentro del dominio, esto con

el fin de evitar a futuro lo ocurrido con una distorsión que se detectó en la medición realizada por el prototipo de aquel entonces: debido a que el corpus en dicha ocasión fue recolectado manualmente y en diferentes medios de prensa en línea, hubo una tendencia a que predominaran noticias sobre atentados incendiarios en la Región de la Araucanía; esto ocasionó que al evaluar una noticia sobre un incendio común y corriente en una vivienda doméstica, la evaluación de la máquina no fuera precisa. Sobre el punto, en la ocasión se señaló que para el caso específico del dominio utilizado en aquel trabajo (noticias policiales), "los textos sobre homicidios, violaciones, abusos sexuales, atropellos, choques de vehículos, robos, hurtos, entre otros, debieran estar presentes en porcentajes similares dentro del corpus". En el presente trabajo, se buscó solucionar dicho punto, mediante el *script* de recolección automática de textos; de esta forma, al no haber recolección manual y provenir las noticias de un solo medio de prensa (La Tercera), el corpus lo formaron todas las noticias sobre política que dicho medio fue publicando día a día, con lo que se evitó que hubiera distorsiones producidas por la mayor presencia de un tema por sobre otro, en otras palabras, este equilibrio vino desde el propio medio que tiene la obligación de cubrir todo el espectro informativo en el dominio elegido (política). El segundo sentido en que se detectó una debilidad en el trabajo de 2010 fue en la cantidad total de textos que conformaban el corpus (1.505). En la ocasión se afirmó que "es obvio por el hecho de que, como ya se indicó, el LSA precisa para su óptimo funcionamiento de grandes corpus textuales, mientras más textos posea, más posibilidades de *aprendizaje* tiene el software de LSA". Es decir, mientras más textos haya en el corpus a partir de los cuales se construye el espacio semántico, mayor efectividad de las mediciones que realice el prototipo. Por ello, tras la construcción de tres espacios semánticos, basados en tres corpus (DK1, 7.165 textos; DK2, 6.794 texto; y DK1y2, 13.959 textos), es posible afirmar que las distorsiones producidas por el corpus pequeño y poco equilibrado de 2010 fueron solventadas a cabalidad por el tamaño de los nuevos corpus y el equilibrio que éstos ofrecen al incluir todas las noticias sobre política de un medio en particular en un periodo determinado.

Lo último que se recoge desde el trabajo de 2010 no es algo que se debía mejorar, pero se relaciona directamente con la investigación que ahora concluye.

Una de las incógnitas que hubo durante el desarrollo de aquel trabajo (2010), fue si el Análisis Semántico Latente, utilizado para evaluar coherencia textual, funcionaría correctamente en el tipo de texto denominado noticia, debido a la estructura particular que tiene ésta, es decir, la utilización de la pirámide invertida en su redacción. En las conclusiones de aquella investigación se dijo que “si bien, comprendiendo cómo opera el LSA pareciera que esta estructura no debiera afectar su correcto funcionamiento, la duda estuvo presente durante el desarrollo del trabajo, despejándose favorablemente”. Lo importante del caso, es que la duda que permeó transversalmente dicha investigación, fue el germen que motivó a la realización de la presente tesis, ya que tras la constatación de que el Análisis Semántico Latente podía evaluar en forma acertada la coherencia textual pese a las particularidades de la estructura en cuestión, surgió la motivación de emplear la técnica para evaluar la jerarquización de los datos de una noticia, en otras palabras, para evaluar la estructura de pirámide invertida en sí misma mediante el LSA. Cinco años después, es posible afirmar que además de operar correctamente el LSA para la evaluación de la coherencia textual en noticias, sin verse afectado por la estructura particular con que éstas se redactan, también es un método eficaz para evaluar la correcta jerarquización de los datos en la pirámide invertida, utilizando el mismo método empleado para evaluar la coherencia textual, como se señaló al inicio de estas conclusiones. Si bien lo anterior puede ser algo no tan relevante para quien lea este trabajo, se menciona por la importancia que tuvo dicha duda en la motivación y concreción de la presente tesis.

8.1. Limitaciones y proyecciones

Para cerrar este capítulo sobre las conclusiones, es interesante reflexionar sobre las limitaciones que presenta el trabajo realizado y que, como se demostró con lo recién expuesto sobre la investigación de 2010, son una fuente importante de motivación y desafío para el emprendimiento de futuros proyectos. Relacionado con lo anterior, también se presentarán las proyecciones que se vislumbran en base al presente estudio.

Una primera limitación detectada, exige retomar lo señalado en 7.3.2 al efectuar el análisis de la comparación realizada entre los resultados de la máquina y

los humanos, específicamente la revisión de las curvas de los gráficos. En la ocasión se advirtió que en el caso del sujeto 18, en las comparaciones 2 y 3, se podían apreciar variaciones muy leves en las líneas: tendencia a bajar un poco en el caso de la evaluación de los humanos. Si bien se señaló que eran casos marginales, ya que se trata de sólo dos puntos entre todos los que se exponen en los gráficos presentados (80 ítems evaluados por los humanos) y la variación es muy leve, permiten llamar la atención sobre una posible explicación en estas conclusiones. Como ya se dijo, la escala utilizada por la máquina para evaluar va de -1 a 1, aunque en rigor los puntajes estén siempre entre 0 y 1. Considerando la cantidad de decimales con que los módulos 1 y 3 arrojan cada evaluación, esto es, hasta 17 -aunque luego se reduzcan a dos aproximando-, es posible advertir que la máquina tiene innumerables opciones de evaluar los detalles que *perciba* en las mediciones realizadas en los textos. Por contrapartida, los humanos trabajaron con una escala muchísimo más limitada que no permite expresar tantos matices, ya que en el caso de las comparaciones 1, 2 y 3 asignaron un puntaje de 0 a 4, con un único decimal para expresar puntajes intermedios que siempre fueron 0,5 o 1,5 o 2,5 o 3,5; en la comparación 4 el puntaje fue de 0 a 7 (0,5 o 1,5 o 2,5 o 3,5 o 4,5 o 5,5 o 6,5). Por lo tanto, la escala empleada por los humanos sólo tenía nueve posibilidades (0 a 4) en un caso y 18 en el otro (0 a 7), contra las innumerables que tenía la máquina. Lo anterior, se reconoce como una debilidad de la medición realizada que pudo provocar las diferencias en el sujeto 18, aunque sean sólo dos entre el total de ítems evaluados por los humanos. Por ello, en futuros trabajos será necesario aplicar una escala de medición en los humanos que tenga más posibilidades de expresar matices con el fin de solventar cualquier diferencia por leve que sea. En todo caso, dicha escala tampoco puede ser abrumadora en posibilidades como la que maneja la máquina, ya que sería contraproducente con las posibilidades de procesamiento que presenta el cerebro humano.

Dentro de las limitaciones de esta tesis, también es importante retomar brevemente el caso de la comparación 4, pese a que ya fue explicado en detalle en 7.3.2. Como ya se dijo, éste es el único caso diferente al resto, ya que no se centra en la jerarquización de los elementos de la noticia que involucran la estructura de pirámide invertida, esto es, el titular y el *lead*, sino en comparar los párrafos

posteriores al *lead* de la noticia construida por los estudiantes con los datos adicionales proporcionados en la actividad práctica. Si bien esta comparación no forma parte directamente del foco del presente trabajo, no deja de tener importancia al formar parte de la noticia. Por lo mismo, constituye una limitación de este estudio el hecho de que los resultados de las correlaciones no hayan sido satisfactorios como para el caso de las tres primeras comparaciones realizadas. Por ello, un desafío futuro es, tomando las bases de lo aquí realizado, diseñar un método más efectivo para la evaluación de los párrafos de la noticia posteriores al *lead*, es decir, aquellos que entregan la información adicional y no se centran en dar respuesta a las preguntas fundamentales.

Otro aspecto que se puede considerar una limitación de la tarea realizada es que hubiera sido deseable poder tener acceso a una mayor cantidad de evaluadores humanos. Dado que se planteó la idea de trabajar en una situación real de evaluación en aula, debido a que el componente construido es para insertarlo en un STI destinado a apoyar la producción escrita de noticias en estudiantes de periodismo, se tomó la decisión -acertada, se sostiene- de que los evaluadores fueran los mismos docentes de la asignatura. Lo anterior, sin embargo, implicó restringir las posibilidades de acceder a más evaluadores, ya que se buscó que fueran profesionales expertos en Periodismo Informativo, que impartieran docencia de pregrado en dicha materia en el curso en que se tomó la muestra y que tuvieran experiencia en medios de prensa escribiendo noticias, lo que restringió drásticamente las posibilidades. Una alternativa a futuro podría ser, tomando como base la experiencia realizada, abrir la medición a otros evaluadores que si bien podrían no ser los docentes directos de la asignatura, sí cumplieran con los otros requisitos: expertos en Periodismo Informativo que ejerzan docencia en dicha materia y que tengan probada experiencia en medios de prensa redactando noticias. La anterior podría ser una buena forma de mejorar la medición efectuada en esta investigación.

Por último, para cerrar estas conclusiones, es bueno referirse a las proyecciones del presente trabajo, sobre todo, ya que desde la introducción se ha venido diciendo que un objetivo futuro es diseñar un sistema tutorial inteligente que permita apoyar la enseñanza de la escritura de noticias en estudiantes de

periodismo. Para ello, todo el trabajo desarrollado en la presente tesis será importantísimo, ya que es la base sobre la que debe sustentarse ese proyecto futuro. Hasta aquí se puede afirmar que es posible evaluar acertadamente y en forma automática la coherencia textual en noticias policiales y sobre política. Además, es posible evaluar en forma automática la correcta construcción de la estructura de pirámide invertida en noticias sobre política (como se dijo, nada impide pensar que esta técnica no funcione en otros dominios temáticos). Gracias a los módulos desarrollados, hay un trabajo importante adelantado en la parte del análisis semántico de los textos (coherencia textual y estructura de pirámide invertida) y, también, se tiene un módulo robusto de corrección ortográfica que, al ser dinámico, permite el mejoramiento de su precisión con relativa sencillez. Preliminarmente, y como un aporte a las proyecciones, se presenta un esquema preliminar de cómo debería insertarse el componente desarrollado dentro del módulo del tutor del futuro STI. En el planteamiento se sigue lo propuesto por Ferreira, Salcedo, Kotz y Barrientos (2012).

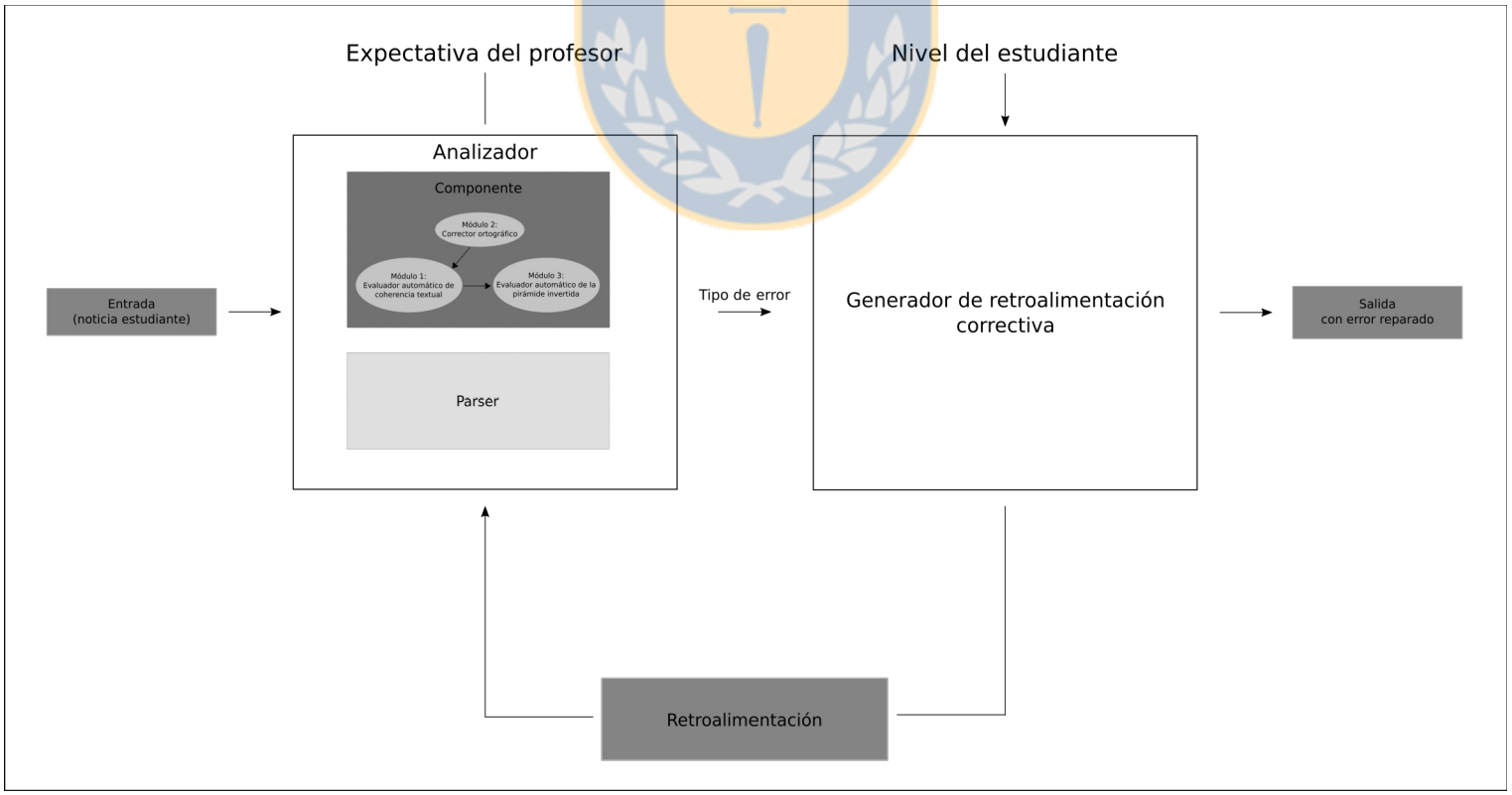

Figura 9: Esquema preliminar del componente dentro del módulo del tutor.

De manera sucinta, se puede señalar sobre la Figura 9, que ésta representa el componente incorporado dentro de la arquitectura hipotética del módulo del tutor de un STI. Como se indicó en el Capítulo 1, Ferreira et al. (2012) señalan que la arquitectura básica de un sistema tutorial inteligente incluye los módulos del tutor, el módulo del estudiante y el módulo del dominio, además de la interfaz. El módulo del tutor, en el cual se ubicaría el componente desarrollado, es el núcleo de un STI "que incluye suficiente conocimiento sobre un área particular para proporcionar respuestas ideales a preguntas y corregir no solo un resultado final sino cada pequeña etapa de razonamiento intermedio. Esto permite mostrar y modelar una forma correcta de resolver un problema". Dentro de la arquitectura expuesta, siguiendo el planteamiento de Ferreira et al. (2012), el componente desarrollado debería ubicarse dentro del analizador del módulo tutor, que es la parte del STI que trabaja sobre la entrada del humano, evaluando la misma y buscando errores (para el caso esta entrada sería el texto de una noticia). El componente construido realizaría las tareas que le son propias y su salida pasaría al generador de retroalimentación, con el fin de entregarle *feedback* al estudiante que le permita mejorar su texto y reparar los errores que hubiere cometido. Lo anterior, se recalca es un planteamiento hipotético, que sólo busca ilustrar de manera breve y gráfica cómo se podría insertar el componente desarrollado dentro del futuro STI. Se reafirma que dicha propuesta requiere un estudio más acabado y minucioso.

Visto lo que hay hasta aquí, una de las posibles primeras tareas que debería considerarse para continuar el camino hacia la construcción de un sistema tutorial inteligente (STI) sería trabajar en el desarrollo de un módulo de análisis sintáctico (o *parser*), a nivel de prototipo, que opere en conjunto con las herramientas ya finalizadas (por ello se incluyó en la Figura 9). Como una forma de ejemplificar esta necesidad, se puede señalar que en las noticias todos los titulares deben incluir un verbo, ya que es un requerimiento del género denominado Periodismo Informativo al cual se adscriben (otro género como el Periodismo Interpretativo no tiene esta exigencia para los titulares). En los trabajos que escribieron los 20 estudiantes en la actividad práctica, cuatro de ellos no incluyeron verbo en su titular (los docentes evaluaron esto en forma aparte, para no afectar las mediciones de la tesis). Obviamente, desde la técnica utilizada -Análisis Semántico Latente- no hay forma

de detectar la presencia o ausencia del verbo, por lo mismo, pese a ser un error en la construcción de la noticia, fue algo que no se pudo considerar. Por ello, es importante el desarrollo de una herramienta que pueda hacerse cargo de este tipo de errores y un módulo de análisis sintáctico (o *parsing*) es el camino correcto, para enfocarse en deficiencias como las del ejemplo u otras que requieran de este tipo de análisis. Sin embargo, no es la única tarea a considerar, ya que hay otras igualmente importantes. Someramente, se pueden mencionar, por ejemplo, desarrollar un modelo para presentar las tareas del STI a los estudiantes (en la presente tesis se utilizó plantearles la idea hipotética de que les habían encargado escribir la noticia sobre la proclamación del candidato presidencial); también hay que buscar una forma adecuada de retroalimentar a los estudiantes sobre los errores que cometan (tarea fundamental en un STI y en cualquier proceso de enseñanza); y, para no seguir más allá, diseñar una interfaz gráfica adecuada para una interacción amistosa entre los usuarios y el sistema tutorial inteligente. En resumen, hay mucho trabajo para continuar la tarea hasta aquí desarrollada y poder concretar -a futuro- la meta final que es construir un sistema tutorial inteligente, que sea capaz de apoyar la enseñanza de la escritura de noticias en quienes se estén formando para ser periodistas.

Referencias

- Allen, James. (1995). Natural Language Understanding. Redwood City (CA): The Benjamín/Cummings Publishing Company.
- Álvarez, Gerardo. 1995. *Textos y discursos*. Pp 12-13, 97-122. Concepción, Chile: Universidad de Concepción.
- Badia Cardús, Tony. 2003. Técnicas de procesamiento del lenguaje. En Martí, María A. (compiladora). Las tecnologías del lenguaje. Pp 193-248. Barcelona, España: Editorial UOC.
- Barrientos, Fernanda; Ferreira, Anita; y Salcedo, Pedro. 2012. Modelado del estudiante para el STI ELE-TUTOR: diseño de un componente adaptativo para apoyar la competencia lingüística del español como lengua extranjera. Boletín de filología, 47(1), 11-32. Disponible en: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-93032012000100001 [Consulta: 12-5-2013].
- Bereiter, Carl y Scardamalia, Marlene. 1983. *Does Learning to Write Have to Be so Difficult?* En Cassany, Daniel. 1999. *Construir la escritura*. Pp 11. Barcelona, España: Editorial Paidós Ibérica S.A.
- Boonthum, Chutima; Levinstein, Irwin y McNamara, Danielle. 2007. Evaluating self-explanations in iSTART: Word matching, latent semantic analysis, and topic models. En Kao, Anne y Poteet, Stephen (Eds.), *Natural Language Processing and Text Mining.* Pp. 91-106. Londres: Springer-Verlag UK.
- Carbonell, Jaime R. (1970). AI in CAI: An Artificial Intelligence approach to Computer Assisted Instruction. IEEE transaction on Man Machine System, 11(4), 190-202.
- Cassany, Daniel. 1988. *Describir el escribir*. Pp 27. Barcelona, España: Editorial Paidós Ibérica S.A.
- Cassany, Daniel. 1999. *Construir la escritura*. Pp 11, 12. Barcelona, España: Editorial Paidós Ibérica S.A.
- Charaudeau, Patrick. 2003. *El discurso de la información*. Pp 166 – 167. Barcelona, España: Editorial Gedisa.

- Centro Virtual Cervantes. 2016. Diccionario de términos clave de ELE. [En línea]. Disponible en: http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/linguisticaaplicada.htm [Consulta: 24-2-2016].
- Cervera, Ángel; Hernández, Guillermo; Pichardo, Coronada y Sánchez, Jesús. 2006. *Saber escribir*. Pp 433-440. Buenos Aires, Argentina: Santillana Ediciones Generales e Instituto Cervantes.
- Coseriu, Eugenio. 1981. Lecciones de Lingüística General. Pp 272. Madrid: Gredos.
- De Vega, Manuel. 1998. *La psicología cognitiva: ensayo sobre un paradigma en transformación*. Anuario de Psicología. Vol. 29, N°2. Pp 21-44. Barcelona: Universitat de Barcelona.
- De Vega, Manuel; Díaz, José Miguel y León, Inmaculada. 1999. Procesamiento del discurso. En Cuetos, Fernando y De Vega, Manuel (compiladores). *Psicolingüística del español*. Pp 271-306. España: Trotta.
- Deerwester, Scott; Dumais, Susan; Furnas, George; Landauer, Thomas y Harshman, Richard. 1990. Indexing by Latent Semantic Analysis. [En línea]. Disponible en: http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf [Consulta: 29-7-2011].
- Diccionario Informático. 2016. Alegsa.com. [En línea]. Disponible en: http://www.alegsa.com.ar/Dic/procesamiento%20de%20lenguajes %20naturales.php [Consulta: 24-2-2016].
- Diccionario Tecnológico. 2015. Usando.info. [En línea]. Disponible en: http://usando.info/diccionario/s#Stopwords [Consulta: 20-12-2015].
- El Mercurio de los estudiantes. 2011. [En línea]. Disponible en: http://www.elmercuriodelosestudiantes.cl [Consulta: 23-04-2011].
- Ferreira, Anita. 2006. Estrategias de *feedback* positivo y correctivo en el español como lengua extranjera. Rev. signos, vol.39, n.62. Pp. 379-406. [En línea]. Disponible en: http://www.scielo.cl/scielo.php?pid=s0718-09342006000300003&script=sci_arttext [Consulta: 11-1-2012].
- Ferreira, Anita. 2007. Estrategias efectivas de feedback correctivo para el aprendizaje de lenguas asistido por computadores. Rev. signos, vol.40, n.65.

Pp. 521-544 . [En línea]. Disponible en: http://www.scielo.cl/scielo.php?pid=S0718-09342007000300007&script=sci_arttext [Consulta: 19-10-2011].

- Ferreira, Anita. 2015. El *feedback* correctivo escrito directo e indirecto en la adquisición y aprendizaje del español como lengua extranjera. Proyecto Conicyt, Fondecyt regular, 1140651. En ejecución desde el 1 de marzo de 2014.
- Ferreira, Anita; Moore, Johanna y Mellish, Chris. 2007. A study of feedback strategies in foreign language classrooms and tutorials with implications for Intelligent Computer-Assisted Language Learning Systems. En *International Journal of Artificial Intelligence in Education*, vol. 17, no. 4, pp. 389-422.
- Ferreira, Anita y Kotz, Gabriela. 2010. ELE-Tutor Inteligente: Un analizador computacional para el tratamiento de errores gramaticales en Español como Lengua Extranjera. Rev. signos, vol.43, n.73. Pp. 211-236. [En línea]. Disponible en: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342010000200002&lng=es&nrm=iso&tlng=es [Consulta: 11-1-2012].
- Ferreira, Anita; Salcedo, Pedro; Kotz, Gabriela y Barrientos, Fernanda. 2012. La arquitectura de ELE-TUTOR: un Sistema Tutorial Inteligente para el Español como Lengua Extranjera. Rev. signos, vol.45, n.79. Pp. 102-131. [En línea]. Disponible en: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342012000200001&lng=pt&nrm=iso&tlng=es [Consulta: 9-3-2013].
- Ferreira, Anita; Elejalde, Jessica y Vine, Ana (2014). Análisis de Errores Asistido por Computador basado en un Corpus de Aprendientes de Español como Lengua Extranjera. Rev. signos, vol.47, n.86. Pp. 385-411. [En línea]. Disponible en: http://www.scielo.cl/pdf/signos/v47n86/a03.pdf [Consulta: 20-11-2015].
- Foltz, Peter; Kinstch, Walter y Landauer, Thomas. 1998. The measurement of textual coherence with Latent Semantic Analysis. [En línea]. Disponible en: http://lsa.colorado.edu/papers/dp2.foltz.pdf [Consulta: 20-01-2015].
- Fundéu.es. 2015. La pirámide invertida como estructura textual. [En línea]. Disponible en: http://www.fundeu.es/escribireninternet/la-piramide-invertida-como-estructura-textual-12/ [Consulta: 20-12-2015].
- Goldman, Susan; Golden, Richard y van den Broek, Paul. 2007. *¿Por qué son*

útiles los modelos computacionales de comprensión de textos? Revista Signos, vol. 40, N°65. Pp.545-572. ISSN 0718-0934.

- Graesser, Arthur; VanLehn, Kurt; Rosé, Carolyn; Jordan, Carolyn y Harter, Derek. 2001. Intelligent Tutoring Systems with Conversational Dialogue. AI Magazine Volume 22 Number 4. [En línea]. Disponible en: http://www.public.asu.edu/~kvanlehn/Stringent/PDF/01AIM_AG_KVL.pdf [Consulta: 19-10-2011].
- Graesser, Arthur; Penumatsa, Phanni; Ventura, Matthew; Cai, Zhiqiang y Hu, Xiangen. 2007. Using LSA in Autotutor: Learning Through Mixed-Initiative Dialogue in Natural Language. En Landauer, Thomas, McNamara, Danielle, Dennis, Simon y Kintsch, Walter. (editores). *Handbook of Latent Semantic Analysis*. Pp. 263-264. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Hernández Sampieri, Roberto, Fernández Collado, Carlos y Baptista, Pilar. 2014. Metodología de la investigación. 6ª edición. México: McGraw-Hill Interamericana.
- Hernández, Sergio. 2010. Evaluación automática de coherencia textual en noticias policiales. Tesis para optar al grado de Magíster en Lingüística, Universidad de Concepción, dirigida por la Dra. Anita Ferreira y codirigida por el Dr. Bernardo Riffo.
- Hernández, Sergio y Ferreira, Anita. 2010. Evaluación Automática de Coherencia Textual en noticias policiales utilizando Análisis Semántico. Revista de Lingüística Teórica y Aplicada 48(2). Pp 211-236.
- Irrázabal, Natalia; Saux, Gastón; Burin, Débora y León, José Antonio. 2006. El resumen: Evaluación de la comprensión lectora en estudiantes universitarios. Anu. investig. 2006, vol.13. [En línea]. Disponible en: http://www.scielo.org.ar/scielo.php?pid=S1851-16862006000100035&script=sci_arttext [Consulta: 18-10-2011].
- Isabel Project. 2010. Cómo escribir un artículo periodístico. Ejercicios prácticos. [En línea]. Disponible en: https://virtuallearningbuses.files.wordpress.com/2012/04/mat3_taller-formativo_cc3b3mo-escribir-un-artc3adculo-periodc3adstico.pdf
- Jurafsky, Daniel y Martin, James. 2008. Speech and language processing, 2ª

edición. Pp. 683 – 695, 787 - 807. Usa: Prentice Hall.

- Kintsch, Walter y Van Dijk, Teun 1975. Comment on se rappelle et on résume des histories. Langages, 40, 98-116. Citado en Irrázabal, Natalia; Saux, Gastón; Burin, Débora y León, José Antonio. 2006. El resumen: Evaluación de la comprensión lectora en estudiantes universitarios. Anu. investig. 2006, vol.13. [En línea]. Disponible en: http://www.scielo.org.ar/scielo.php?pid=S1851-16862006000100035&script=sci_arttext [Consulta: 18-10-2011].
- Kintsch, Eileen; Steinhart, Dave; Stahl, Gerry & LSA research group. 2000. Developing Summarization Skills through the Use of LSA-Based Feedback. [En línea]. Disponible en: http://lsa.colorado.edu/papers/ekintschSummaryStreet.pdf [Consulta: 09-10-2011].
- Landauer, Thomas y Dumais, Susan. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. [En línea]. Disponible en: http://lsa.colorado.edu/papers/plato/plato.annote.html [Consulta: 22-7-2011].
- Landauer, Thomas; Foltz, Peter y Laham, Darrell. 1998. “An Introduction to Latent Semantic Analysis. Discourse Processes”. Discourse Processes.25(2&3), 259-284. [En línea]. Disponible en: http://lsa.colorado.edu/papers/dp1.LSAintro.pdf [Consulta: 07-09-2014].
- Landauer, Thomas. 2002. On the computational basis of learning and cognition: Arguments from LSA. [En línea]. Disponible en: http://lsa.colorado.edu/papers/Ross-final-submit.pdf [Consulta: 22-07-2011].
- Lavid, Julia. (2005). Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI. Madrid: Ediciones Cátedra (Grupo Anaya S.A.).
- Longa, Víctor y Lorenzo, Guillermo. 2008. What about a (really) minimalist theory of language acquisition? [En línea]. Disponible en: http://www.unioviedo.es/biolang/pdf/ling.2008.018.pdf [Consulta: 20-7-2011].
- Louwerse, Max. 2004. Un modelo conciso de cohesión en el texto y coherencia en la comprensión. Rev. signos, vol.37, n.56. Pp. 41-58. [En línea]. Disponible en: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-

09342004005600004&lng=es&nrm=iso [Consulta: 11-07-2011].

- Manning, Christopher; Raghavan, Prabhakar y Schütze Hinrich. 2008. Introduction to Information Retrieval, Cambridge University Press. [En línea]. Disponible en: http://nlp.stanford.edu/IR-book/information-retrieval-book.html [Consulta: 20-12-2015].
- Martí, María A. 2003. Las tecnologías del lenguaje. Pp 9-11. Barcelona, España: Editorial UOC.
- Martín Vivaldi, Gonzalo. 1993. *Géneros periodísticos*. Pp 369. Madrid, España: Editorial Paraninfo.
- Martín Vivaldi, Gonzalo. 2004. Curso de redacción. Pp 367-371. Madrid, España: Editorial Paraninfo.
- Martínez Albertos, José Luis. 1989. *El lenguaje periodístico*. Pp 19 a 23. Madrid, España: Editorial Paraninfo.
- Martínez Albertos, José Luis. 2004. *Curso general de redacción periodística*, 5ª edición, 3ª reimpresión. Pp 21. Madrid, España: Editorial Paraninfo.
- McCarthy, Philip; Briner, Stephen; Rus, Vasile y McNamara, Danielle. 2007. Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. En Kao, Anne y Poteet, Stephen (Eds.), Natural language processing and text mining. Pp. 107-122. Londres: Springer-Verlag.
- McNamara, Danielle; Kintsch, Eileen; Butler Songer, Nancy y Kintsch, Walter. 1996. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction. [En línea]. Disponible en: http://www.colorado.edu/ics/sites/default/files/attached-files/93-04.pdf [Consulta: 26-10-2014].
- McNamara, Danielle. 2004. Aprender del texto: Efectos de la estructura textual y las estrategias del lector. Rev. signos, vol.37, n.55. Pp. 19-30. [En línea]. Disponible en: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342004005500002&lng=es&nrm=iso [Consulta: 26-10-2011].
- Oliva, Carlos. 2015. Ejercicios de escritura de noticias: la jerarquización de

datos (material de clases). Carrera de Periodismo, Universidad de Concepción.

- Palincsar, Annemarie y Brown, Ann. 1984. Reciprocal teaching of comprehension- fostering and comprehension-monitoring activities. Cognition and Instruction, 1 (2), 117-175. Citado en Irrázabal, Natalia; Saux, Gastón; Burin, Débora y León, José Antonio. 2006. El resumen: Evaluación de la comprensión lectora en estudiantes universitarios. Anu. investig. 2006, vol.13. [En línea]. Disponible en: http://www.scielo.org.ar/scielo.php?pid=S1851-16862006000100035&script=sci_arttext [Consulta: 18-10-2011].
- Parodi, Giovanni. (Ed.) (2005). Discurso especializado e instituciones formadoras. Valparaíso: Ediciones Universitarias de Valparaíso.
- Parodi, Giovanni. (2007). El discurso especializado escrito en el ámbito universitario y profesional: Constitución de un corpus de estudio. Revista signos, 40(63), 147-178. [En línea]. Disponible en: http://www.scielo.cl/scielo.php?pid=S0718-09342007000100008&script=sci_arttext [Consulta: 24-2-2015].
- Parodi, Giovanni. (2009). Géneros discursivos y lengua escrita: Propuesta de una concepción integral desde una perspectiva sociocognitiva. Letras, 51(80), 19-56. [En línea]. Disponible en: http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S0459-12832009000300001 [Consulta: 24-2-2015].
- Pastor, Susana. 2004. Aprendizaje de segundas lenguas: lingüística aplicada a la enseñanza de idiomas. Pp 21-94. Valencia, España: Universidad de Alicante.
- Platón. 1960. Diálogos socráticos. Buenos Aires: W. M. Jackson.
- Real Academia Española. 2014. Diccionario de la lengua española. 23° edición. Madrid, España: Espasa.
- Singer, Murray y Zwaan, Rolf. 2003. Text comprehension. En Graesser, Arthur; Gernsbacher, Morton Ann y Goldman, Susan (compiladores). *Handbook of discourse processes*. Pp 83-121. Mahwah, USA: Lawrence Erlbaum Associates.
- Smith, Neil. 2001. Chomsky: ideas e ideales. Pp 63-67. Madrid: Cambridge University Press, sucursal en España.
- Steinhart, David. 2001. Summary Street: an intelligent tutoring system for

improving student writing through the use of latent semantic analysis. Institute of Cognitive Science, University of Colorado, Boulder. [En línea]. Disponible en: http://lsa.colorado.edu/papers/daveDissertation.pdf [Consulta: 09-10-2011].

- Van Dijk, Teun. 1990. *La noticia como discurso*. Pp 26-28. Barcelona, España: Ediciones Paidós Ibérica.
- Venegas, René. 2003. *Análisis Semántico Latente: una panorámica de su desarrollo*. Revista Signos, vol. 36, N°53. Pp.121-138. ISSN 0718-0934.
- Venegas, René. 2005. Las relaciones léxico-semánticas en artículos de investigación científica: Una aproximación desde el análisis semántico latente (Tesis para optar al grado de Doctor en Lingüística). Pp 166-172. Valparaíso: Pontificia Universidad Católica de Valparaíso.
- Venkatesh, R., Naganathan, E. R. y Uma Maheswari, N. 2010. Intelligent Tutoring System Using Hybrid Expert System With Speech Model in Neural Network. International Journal of Computer Theory and Engineering, Vol. 2, No. 1 February, 2010. [En línea]. Disponible en: http://www.ijcte.org/papers/108-G225.pdf [Consulta: 19-10-2011].
- Villayandre Llamazares, Milka. 2010. Aproximación a la lingüística computacional. Tesis para optar al grado de doctor dirigida por los Salvador Gutiérrez Ordóñez y Manuel Iglesias Bango. Departamento de Filología Hispánica y Clásica, Universidad de León, España.
- Warren, Carl. 1975. *Géneros periodísticos informativos*. Pp 90-110. Barcelona, España: Editorial A.T.E.

Anexo 1: Práctico aplicados a los estudiantes de periodismo

UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS SOCIALES
DEPTO. DE COMUNICACIÓN SOCIAL
CARRERA DE PERIODISMO.

Práctico

INSTRUCCIONES:

Imagine que estamos un par de años en el futuro y usted es redactor de un medio escrito, que se enfoca en el periodismo informativo. Como se avecinan las elecciones presidenciales de 2017, su editor le encarga escribir un texto sobre un hecho que ocurrió hace unos minutos. A partir de los datos expuestos más abajo -que son ficticios y se enfocan en el hecho mencionado-, elabore un texto informativo siguiendo las pautas vistas en clases (estructura de pirámide invertida, responder a las preguntas fundamentales, etc.). El texto debe cumplir con todos los requisitos para ser publicable en la sección de noticias de un medio escrito. Las exigencias son:

1) El texto debe tener: titular informativo, *lead* y cuerpo de la noticia (siguiendo la pirámide invertida). No debe incluirse bajada ni epígrafe en los elementos de titulación.
2) Sólo deben utilizarse los datos presentados.
3) La extensión máxima del texto es de 2500 caracteres, la mínima de 1500 (caracteres con espacios).
4) Letra Times News Roman, tamaño 12, interlineado 1.5, tabulado y justificado. El nombre del estudiante va a la derecha antes del inicio del texto.
5) Plazo para subir al INFODA: 16:30 horas.
6) Se premia la claridad de ideas, se castigan severamente los errores de ortografía y redacción. En la corrección se tendrán en cuenta la selección y jerarquización de datos, así como la ortografía y la redacción del texto.

Advertencia: esto es sólo un ejercicio con fines didácticos, que busca medir su capacidad para jerarquizar datos y redactar correctamente un texto informativo para un medio impreso. El periodismo no consiste en ficción ni en inventarse los datos.

DATOS:

- Lagos fue Presidente de la República entre 2000 y 2006.
- En 1989 no fue elegido como candidato presidencial de la Concertación de

Partidos por la Democracia y se presentó como candidato a senador, elección que perdió - pese a tener la segunda mayor votación - debido al sistema binominal, ya que su lista no pudo doblar a la lista adversaria.

- La proclamación de Lagos se realizó en un acto efectuado en el Teatro Caupolicán de Santiago a partir de las 21 horas de hoy.
- Lagos fue derrotado por Eduardo Frei Ruiz-Tagle en las primarias de la Concertación en 1993 por 63,32% contra 36,68%.
- Lagos se enfrentará en las elecciones presidenciales que se realizarán el 19 de noviembre próximo a Manuel José Ossandón (RN) de Chile Vamos, Marco Enríquez Ominami (PRO), Alejandro Navarro (MAS), Lily Pérez (Amplitud) y Andrés Velasco (Ciudadanos).
- Lagos fue el ganador de las primarias de la Nueva Mayoría realizadas la semana pasada, que buscaban definir al candidato de la coalición para ocupar la Moneda en el periodo 2018-2022.
- En su primera elección Lagos logró el triunfo tras derrotar en segunda vuelta a Joaquín Lavín por 51,31% contra 48,69% del militante UDI (enero de 2000).
- Lagos fue Ministro de Educación en la presidencia de Patricio Aylwin entre 1990 y 1992; y Ministro de Obras Públicas en el mandato de Eduardo Frei Ruiz-Tagle entre 1994 y 1998.
- Uno de los episodios más recordados de Ricardo Lagos ocurrió el 25 de abril de 1988, cuando en el programa “De cara al país” miró a la cámara y emplazó al entonces gobernante Augusto Pinochet Ugarte por la intención de éste de continuar ocho años más como Presidente de la República.
- Tras su proclamación Lagos fue ovacionado de pie por los asistentes al Teatro Caupolicán.
- Lagos fue uno de los fundadores del Partido por la Democracia (PPD) el 15 de diciembre de 1987.
- El exmandatario Ricardo Lagos Escobar fue proclamado como candidato de la Nueva Mayoría para las elecciones presidenciales de 2017.
- Lagos (PPD) se impuso en las primarias de la coalición a José Antonio Gómez (PRSD), Isabel Allende (PS) y Jorge Burgos (DC).