Revista Signos
ISSN: 0035-0451
revista.signos@ucv.cl
Pontificia Universidad Católica de Valparaíso
Chile

Sabaj, Omar
Hacia una matriz de rasgos lingüísticos con impacto textual: Un estudio exploratorio
Revista Signos, vol. 40, núm. 63, 2007, pp. 197-218
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

Revista Signos
2007, 40(63)
197-218

Hacia una matriz de rasgos lingüísticos con impacto textual: Un estudio exploratorio*

Omar Sabaj
Pontificia Universidad Católica de Valparaíso
Chile

Resumen: Gracias al uso de medios informáticos, en la actualidad, existen diversas líneas de investigación dentro de la denominada Lingüística de Corpus que abordan la descripción de textos auténticos a partir de un conjunto de rasgos lingüísticos. La crítica que se ha realizado a estos estudios es que solo describen aspectos lingüísticos a nivel de la palabra que difícilmente logran tener un impacto en la descripción de los corpus a nivel textual. En este estudio, se revisan críticamente las distintas matrices de rasgos y las metodologías que se han aplicado tanto para el inglés como para el español con el objetivo de proponer una matriz que atienda a algunas de las críticas a los estudios anteriores. La propuesta total, en concreto, se basa en los siguientes componentes: los modos de organización del discurso (MOD), las secuencias o *lexical bundles*, el modelo de indexación de eventos y la teoría de la valoración. Además, se incluirá un componente asociado a categorías gramaticales y, por último, un componente derivado de un proyecto anterior (las dimensiones establecidas en el proyecto Fondecyt 1020786). Una parte de la matriz, basada en estos componentes, se aplica de modo exploratorio a un pequeño conjunto de textos de características diversas con el objetivo de comprobar su potencial descriptivo.

Palabras Clave: Registros diversificados, matriz de rasgos, estudios multidimensionales.

Recibido:
23-VI-2006
Aceptado:
3-XI-2006

Correspondencia: Omar Sabaj (omar.sabaj@ucv.cl). Tel.: (56-32) 2273388. Pontificia Universidad Católica de Valparaíso. Av. Brasil 2830, piso 9, Valparaíso, Chile.

* Investigación realizada en el marco de los proyectos Fondecyt 1060440 y DI PUCV 184.720/2006.

Towards a linguistic feature matrix with textual impact: An exploratory study

Abstract: Thanks to the use of computational means, there are different lines of research nowadays in the field of the so called corpus linguistics, which describe authentic texts from linguistic features. The criticism that has been made to such studies is founded on the fact that they only reach the word level with scarce impact on the textual level. In this work, we critically revised different feature matrices and methodologies that have been used for studying Spanish and English, in order to propose a matrix that can deal with part of the criticism made to previous researches. Concretely, the proposal is based on the following components: the organization modes of discourse, the sequences or lexical bundles, the Model of Event indexing and the Appraisal Theory. In addition, a component associated with grammatical categories, a component of syntactic units, and a component of statistics measures will be included. Finally we will also consider a component derived from a previous project (the dimensions established in the project Fondecyt 1020786). In order to verify its descriptive potential, the matrix, based on these components, is applied in an exploratory way to a small set of texts.

Key Words: Diversified registers, feature matrix, multidimensional studies.

INTRODUCCIÓN

En los últimos años, se han desarrollado grandes avances en la descripción de corpus, a través de los denominados enfoques multirasgos (AMR) y multidimensiones (AMD). Estos estudios han proporcionado una forma exhaustiva de abordar la descripción de textos reales y la variación de estos en registros diversificados de una lengua determinada. A pesar de sus enormes fortalezas y de la innovadora forma de abordar la descripción de corpus multiregistros en distintas lenguas, estas investigaciones no han estado exentas de problemas. El objetivo de este trabajo tiene dos orientaciones. Se expondrán primero, algunas críticas que se han realizado al enfoque multirasgos y multidimensional. En segundo lugar, se propone y prueba una matriz de rasgos con el objetivo de abordar el estudio de registros diversificados utilizando una propuesta metodológica alternativa.

En los antecedentes teóricos, señalaremos, brevemente, en qué consisten los estudios multirasgos y multidimensionales. Propondremos, a continuación, una nueva orientación en estos estudios, a partir de la presentación de las bases teóricas que nos conducirá a la determinación de una matriz de rasgos con impacto textual y discursivo. Por último, aplicaremos de modo exploratorio una parte de la propuesta a cinco registros de origen diverso y discutiremos sus ventajas y debilidades. Concluimos con un breve resumen de la investigación y las proyecciones de la misma.

1. Antecedentes teóricos

1.1. El enfoque multirasgos y multidimensional

En este punto, introducimos sucintamente las principales características del denominado enfoque multirasgos y multidimensional (AMR y AMD). Señalamos, además, bibliografía pertinente para una descripción más detallada del enfoque.

El análisis multidimensional de la variación de registros se basa en el método estadístico multifactorial, que Biber (1986, 1988, 1993, 1994, 2003) ha venido aplicando desde la década de los ochenta para la descripción de la variación lingüística de registros diversificados. El método multifactorial para la descripción de registros sigue distintos pasos:

a) Determinación y definición comunicativa funcional de un conjunto de rasgos lingüísticos;
b) Construcción de un corpus de registros diversificados;
c) Conteo y normalización de los rasgos en el conjunto de registros;
d) Aplicación de un análisis multifactorial a través del cual se agrupan en factores aquellos rasgos lingüísticos que co-ocurren sistemáticamente;
e) Interpretación funcional comunicativa de los factores en dimensiones.

Cabe destacar que una de las mayores fortalezas de los estudios multidimensionales radica en que la aplicación del método a diferentes lenguas (Biber & Ventura, 2007) produce las mismas dimensiones. En otras palabras, pareciera ser que la existencia de estas dimensiones (interpretación funcional de un factor o grupo de rasgos que co-ocurren sistemáticamente) son características universales que permiten describir los registros de cualquier lengua.

Para una descripción técnica más detallada del enfoque AMR y AMD, véase Biber (1988), Hair, Anderson, Tatham y Black (2001) y Parodi (2005a).

1.2. Problemas de los estudios basados en rasgos

Presentamos, a continuación, algunas críticas al enfoque AMR y AMD. La primera crítica se basa en un estudio de Lowerse, McCarthy, McNamara y Graesser (2004). Luego, se presentan otros aspectos problemáticos basados en la experiencia de la utilización del enfoque en distintas investigaciones desarrolladas en los Programas de Postgrado en Lingüística de la Pontificia Universidad Católica de Valparaíso, Chile.

1.2.1. Buscando más allá de la palabra: la crítica de Lowerse et al. (2004) y los índices de *Coh-metrix*

En un breve artículo, Lowerse et al. (2004) replicaron el estudio de Biber (1988). Para ello, utilizaron, entre otros, los rasgos de la herramienta *Coh-metrix* (McNamara, Lowerse, & Graesser, 2004). En el trabajo, los autores critican el estudio de Biber (1988), sosteniendo que los rasgos lingüísticos estudiados no superan el nivel de la palabra y que, por lo tanto, no permiten la caracterización textual de los registros:

> "but despite these impressive results, the theoretical question that remains lurking is to what extent these linguistic features fully capture the nature of a text and thereby the nature of a register" (Lowerse et al., 2004: 1).

Es interesante destacar que, a pesar de la crítica, los autores (Lowerse et al., 2004) consideran que el método multirasgos y multidimensional del estudio de Biber (1988) arroja resultados sorprendentes. Aun así, sostienen que el estudio realizado por ellos permite distinguir entre registros orales y escritos, asunto que no aparece en las dimensiones determinadas por Biber (1988). Pero ¿cuáles son los rasgos que utilizan estos investigadores para describir la variación entre registros? ¿Superan ellos realmente el nivel de la palabra? Para responder a estas preguntas debemos entender el funcionamiento de la herramienta *Coh-metrix*, que pasamos a describir a continuación.

Coh-metrix es una herramienta computacional que genera índices de la representación lingüística y discursiva de un texto. La herramienta ayuda, según sus autores, a proporcionar un índice de la lecturabilidad de un texto a partir de la determinación de las marcas explícitas de cohesión. La cohesión se entiende en este marco como las características explícitas del texto que ayudan al lector a conectar mentalmente las ideas del texto. Cabe señalar que, según los autores, la herramienta permite, además, de forma subsidiaria, describir la variación lingüística de registros a pesar de que ese no es su propósito original.

La herramienta *Coh-metrix* establece 60 índices que están agrupados en categorías mayores. Estos índices se presentan en un anexo disponible en http//:omarsabaj.wordpress.com.

Las categorías generales que agrupan los 60 índices de *Coh-metrix* son:

1. Información general de identificación y referencia
2. Lecturabilidad
3. Información general sobre el texto y las palabras
4. Índices sintácticos
5. Índices semánticos y referenciales
6. Dimensiones del modelo de situación[1]

Si se analizan detenidamente estas categorías y se estudia en detalle el anexo disponible en http//:omarsabaj.wordpress.com, se podrá establecer que, si bien es cierto que algunas de estas categorías sí tienen un impacto textual (2, 5 y 6), todas ellas también se basan en índices al nivel de la palabra. En este sentido, el trabajo de Lowerse et al. (2004) puede ser demasiado crítico respecto del trabajo de Biber (1988) y la única diferencia entre ambos sería que los resultados del estudio de Lowerse et al. (2004) permiten distinguir dimensiones (la oralidad y la escritura) que no se logran diferenciar en Biber (1988), pero ni los rasgos de uno ni de otro estudio se basan en niveles claramente textuales o discursivos. Lo anterior debido a que tanto el texto como el discurso se cristalizan en haces de rasgos. Por lo tanto, determinar si un rasgo tiene *per se* un carácter textual no es de interés. Lo que sí parece relevante es determinar si un conjunto de rasgos permite o no detectar las diferencias textuales o discursivas que existen entre diferentes registros.

1.2.2. Otras críticas

Exponemos en este punto otras críticas que hemos construido a partir de la experiencia del trabajo con análisis de rasgos.

1.2.3. ¿Es posible la integración de los estudios?

Tal como señala Ciapuscio (2005) para la noción de género, en lingüística, la proliferación de términos, aproximaciones y enfoques provoca una falta de integración en las investigaciones. En nuestra opinión creemos que también sucede lo mismo en los estudios de corpus. Debido a esto, en esta investigación, incorporaremos los resultados obtenidos en un proyecto anterior (Fondecyt 1020786), incluyendo en la matriz de rasgos que proponemos, datos, resultados o rasgos ya abordados en dicho proyecto.

1.2.4. Los factores, su determinación y su interpretación en dimensiones

Otro problema que se puede señalar a las investigaciones basadas en rasgos, específicamente a los modelos multidimensionales, radica en la complejidad metodológica de la etapa de la interpretación de los factores en dimensiones.

Como se sabe, en estos trabajos, se determinan en primer término un conjunto de rasgos lingüísticos que son definidos funcionalmente. Luego, el análisis multifactorial genera factores en los cuales se agrupan los rasgos que co-ocurren sistemáticamente. Algunos de estos factores son desechados porque no son susceptibles de ser interpretados funcionalmente. En general, la interpretación de los factores en dimensiones no es una tarea fácil. Para la conformación de los factores, se puede optar por mantener los rasgos que aparecen en diferentes factores (Lowerse et al., 2004) o bien, omitir los rasgos que se repiten en más de un

factor (Biber, 1988). En este segundo caso, es un tanto más fácil la interpretación del factor en una dimensión, pero al omitir un rasgo en un factor si ya ha aparecido en factor anterior, se reduce la naturalidad del factor ya que sabemos que los mismos rasgos pueden tener funciones variadas.

En esta investigación, innovaremos respecto de estas investigaciones, puesto que la dimensión estará determinada *a priori*. Para ello, cada dimensión estará definida con antelación y estará compuesta por un conjunto de rasgos que según la bibliografía se relacionan con esa dimensión. Por ejemplo, determinaremos una dimensión argumentativa a partir de los rasgos lingüísticos que se han determinado, en otros trabajos, para la argumentación. Esta innovación tiene una ventaja y una desventaja. La ventaja es que no necesitaremos hacer una interpretación de los rasgos, pues ya estará establecido, a partir de otros trabajos empíricos, que ellos son parte de una dimensión determinada. La desventaja es que las dimensiones que utilizaremos no serán derivadas de las características propias de los textos analizados, lo que puede implicar un cierto alejamiento de una descripción fiel de los rasgos que realmente aparecen en los textos investigados. Sin duda, nuestra decisión se basa en la intuición de que la ventaja tiene mayor peso práctico que el de la desventaja antes expuesta.

2. Hacia una matriz de rasgos con impacto textual

La matriz que proponemos está compuesta por cuatro unidades de niveles jerárquicos: una dimensión, una categoría, una subcategoría y un rasgo. La dimensión corresponde a un aspecto general de los registros, los cuales en menor o mayor medida corresponden a aspectos discursivos de los mismos y se adscriben a distintas disciplinas o teorías específicas, por ejemplo, a los modos de organización del discurso. Las categorías y subcategorías son unidades de análisis que se establecen dentro de esas teorías y los rasgos son manifestaciones lingüísticas de esas categorías. Entendemos manifestación lingüística en un sentido amplio ya que los rasgos que proponemos pueden ser unidades morfológicas, gramaticales, sintácticas o bien una secuencia de morfemas sin una unidad sintáctica. Aunque algunos de estos rasgos no tienen identidad discursiva, es decir, no son unidades de análisis que se apliquen directamente a ese nivel (y, por lo tanto, este estudio tampoco superaría las críticas de Lowerse et al. 2004), su concurrencia sistemática en una dimensión, sí da cuenta de aspectos relevantes de los corpus a nivel textual.

Dicho de otra forma, aunque no todos los rasgos utilizados en nuestro trabajo superan el nivel de la palabra, la aparición conjunta de ellos sí pueden dar cuenta de algunos aspectos textuales de los registros.

El modelo multidimensional que describimos a continuación está compuesto por las siguientes dimensiones:

1. La teoría de la valoración
2. Los modos de organización del discurso
3. El modelo de indexación de eventos
4. Las dimensiones del Proyecto Fondecyt 1020786
5. Secuencias léxicas o *lexical bundles*
6. La distribución de categorías gramaticales y unidades de medición estadística generales.

Los primeros tres componentes se estudiarán a partir de marcas o rasgos predeterminados por la bibliografía. Los rasgos de estas tres dimensiones se encuentran en un anexo disponible en http//:omarsabaj.wordpress.com. Los otros componentes, en cambio, se determinan caso a caso, es decir, su determinación está dada por los datos.

Cabe señalar que, aunque los rasgos que incluiremos en cada una de las dimensiones no han sido tomados siempre de estudios del español, muchos de ellos, pertenecen a dimensiones que han sido consideradas universales a las lenguas, como la causa, la temporalidad y la espacialidad.

2.1. Bases teóricas de la matriz

Definimos concisamente las categorías de cada una de las dimensiones que investigaremos. Para una revisión detallada de la matriz, esto es, los rasgos específicos y su definición, revísese http//:omarsabaj.wordpress.com. Cabe especificar que nuestra investigación sigue solo en parte los estudios realizados anteriormente. Esto, en el sentido de que, primero, se estudian una serie de dimensiones a partir de rasgos, pero la forma de determinar las dimensiones es *a priori* y no emergen necesariamente de los registros investigados, como sucede en los estudios de Biber (1988) y Lowerse et al. (2004). Por otro lado, se sigue al estudio de Lowerse et al. (2004), en tanto un rasgo puede aparecer en más de una dimensión y no en una sola, como en el estudio de Biber (1988); lo que según nuestra opinión es más representativo de la realidad empírica de los registros.

Por otra parte, debido a que nos ubicamos dentro de los lineamientos de la lingüística de corpus, es decir, utilizamos herramientas computacionales para el análisis, hemos intentado seleccionar rasgos que se adecuan a las posibilidades de búsqueda a nuestro alcance. En otras palabras, hemos considerado la viabilidad de las búsquedas según las herramientas a las que tenemos acceso, en tanto que, hasta el momento solo contamos con corpus no marcados morfosintácticamente.

2.1.1. La teoría de la valoración

La presencia del lenguaje evaluativo ha sido considerada como un factor importante al momento de diferenciar registros. Debido a esto, en esta investigación incorporaremos algunas de las categorías de la teoría de la valoración.

Siguiendo a Martin y White (2005) y a White (1999), entendemos que la teoría de la valoración se ocupa de los recursos lingüísticos por medio de los cuales los textos/hablantes llegan a expresar, negociar y naturalizar determinadas posiciones intersubjetivas y en última instancia, ideológicas. Un esquema del modelo propuesto por Martin y White (2005) es el siguiente:

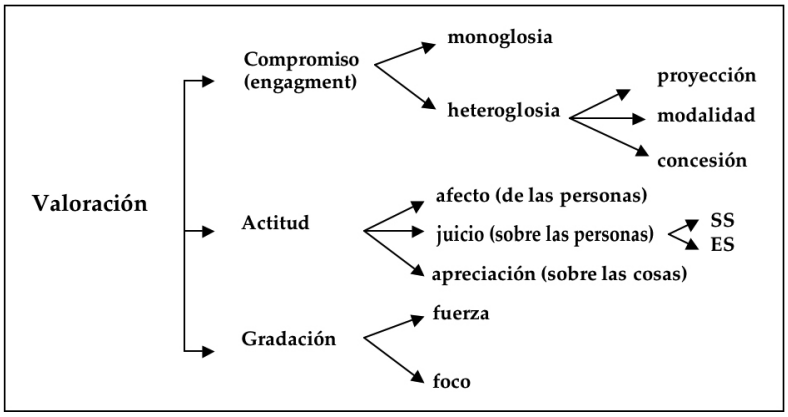


Figura 1. Síntesis del modelo de la teoría de la valoración (Martin & White, 2005).[2]

Según los autores, esta teoría permite elaborar un modelo para describir el lenguaje como una expresión de la valoración o evaluación presente en el discurso. La teoría propone que en cada una de las categorías de la Figura 1 se puede establecer una polaridad positiva o negativa. Así, podemos establecer marcas de afecto positivo o afecto negativo, juicio positivo o negativo. Debido a que no utilizaremos todo el modelo definimos solo aquellas categorías que integran nuestra matriz.

El compromiso es el grado en que se manifiestan distintas voces en el discurso. Cuando solo aparece una sola voz, esto es, en los discursos monoglósicos solo hay un punto de vista sobre el mundo que se representa. En los discursos heteroglósicos, en cambio, aparecen distintos puntos de vista sobre el mundo que se representa (Martin & White, 2005). Para reconocer enunciados mono o heteroglósicos se debe tener en cuenta, entre otros aspectos, la estructura sintáctica y el tipo de acto de habla.

La actitud, el segundo componente de la teoría, tiene relación con la expresión del campo semántico de la emoción, el juicio sobre las personas y su comportamiento y la apreciación sobre las cosas del mundo. El afecto es la expresión de la emoción y la emotividad en el lenguaje. El juicio es la expresión de la valoración, a un nivel social, sobre las personas y sus comportamientos: consta de dos subsistemas, la sanción social (SS) y la estimación social (ES). Su manifestación lingüística corresponde, generalmente, a adjetivos calificativos y apreciativos. La apreciación es la expresión de nuestra evaluación de las cosas o productos (sustantivos concretos) y se manifiesta en adjetivos calificativos (Martin & White, 2005).

Aunque la gradación no se considera directamente en este estudio, la entendemos como la expresión de la intensificación y del debilitamiento de los enunciados. Decimos que sí consideramos la gradación, aunque indirectamente, porque creemos, a diferencia de lo que consideran los autores del modelo, que la gradación no es un sistema independiente sino que, en un sentido más amplio, opera dependiente de los otros sistemas como la actitud y el compromiso. Así, en la existencia de lo que denominamos grados del adjetivo (bueno, mejor, buenísimo) es, según nuestra opinión, un mecanismo de gradación dependiente de la expresión de la apreciación.

2.1.2. Los modos de organización del discurso: lo descriptivo, lo narrativo y lo argumentativo

Los modos de organización del discurso son formas de organización prototípica de los textos según sus propósitos. Esta organización se materializa en formas lingüísticas específicas. La mayoría de los autores (Charaudeau, 1992; Bassols & Torrent, 1997; Calsamiglia & Tusón, 1999; Charaudeau & Maingueneau, 2002) reconocen tres modos de organización: la descripción, la narración y la argumentación.

La descripción se ha asociado a la representación lingüística del mundo real o imaginado. A través de la descripción expresamos la manera de percibir el mundo por medio de los sentidos y se puede aplicar a estados o procesos (Charaudeau, 1992).

La narración es la materialización lingüística de la experiencia humana y se asocia generalmente a la sucesión cronológica de eventos. La causalidad, la unidad temática, la transformación y la unidad de acción han sido considerados elementos constitutivos del modo de organización narrativo. La argumentación, por su parte, ha sido uno de los modos más estudiados por su larga tradición en los estudios de retórica. Se han propuesto distintos modelos para explicar la argumentación, entre ellos, la lógica factual de Toulmin (1958) y la nueva retórica de Perelman y Olbrechts-Tyteca (1994) son los más reconocidos.

Siguiendo la idea de Venegas (2005a), diremos que los modos de organización del discurso no

son directamente observables en los textos. Lo que observamos es la manifestación lingüística de los modos de organización del discurso y, por lo tanto, no hablamos de 'la descripción', 'la narración' y 'la argumentación' sino que preferimos los términos 'lo descriptivo', 'lo narrativo' y 'lo argumentativo'. La predominancia de estas manifestaciones en los textos se relaciona directamente con la finalidad de los discursos y además se puede establecer un tipo de texto prototípico para cada una de estas expresiones. Así, 'lo narrativo' se presenta prototípicamente en un cuento o novela; 'lo descriptivo' en un manual en que se explican las partes de un sistema y 'lo argumentativo' sería característico de un tipo de texto específico como el editorial o el foro.

Respecto de las marcas, generalmente se han asociado los adjetivos, los sustantivos y los verbos copulativos a la descripción. En la narración, priman los verbos en pasado y verbos de tipo psicológico, así como todo tipo de indicadores deícticos y los marcadores discursivos temporales. La argumentación, por su parte, se caracteriza sobre todo por los marcadores y conectores discursivos de carácter concesivo, causal, contrafactual, etc.

En este trabajo, recopilamos de distintos autores marcas de estos modos. El detalle de las marcas seleccionadas para cada modo de organización del discurso se encuentra, junto con sus referencias, en un anexo disponible en http//:omarsabaj.wordpress.com.

2.1.3. El modelo de indexación de eventos

El modelo de indexación de eventos (Zwaan, Langston & Graesser, 1995; Zwaan & Radvansky, 1998) es una teoría propuesta para explicar la comprensión de textos narrativos. Si bien ha sido postulado para explicar la comprensión, el modelo incluye componentes que resultan interesantes para la exploración de registros. Los cinco componentes que se incluyen en dicho modelo son: tiempo, espacio, causalidad, motivación/intención, protagonista.

Al igual que en el trabajo de Lowerse et al. (2004), en el presente estudio no incluiremos las cinco dimensiones, aquellas que según la teoría son más transversales a otros tipos de textos y que no solamente caracterizan a los textos narrativos. Estas dimensiones son la causalidad, el espacio y el tiempo. Cada una de estas dimensiones se cristalizan en distintos tipos de unidades, morfológicas, léxicas y sintácticas en un anexo disponible en http//:omarsabaj.wordpress.com.

2.1.4. Las dimensiones del proyecto Fondecyt 1020786

El proyecto Fondecyt 1020786, titulado "El análisis del discurso científico y su comprensión lingüística en la formación técnico-profesional", se ocupó de estudiar la variación de diferentes registros, a saber, discurso técnico profesional (en forma de manuales, guías didácticas,

etc.), entrevistas orales y literatura latinoamericana (Parodi, 2005a). El aporte del proyecto en cuestión es haber proporcionado una descripción compleja y multidimensional del registro técnico-profesional (textos utilizados en los colegios técnico-profesionales de la ciudad de Valparaíso, Chile), en comparación con el corpus de literatura latinoamericana y el corpus de entrevistas orales.

Para la investigación se siguió la metodología de Biber (1998), en la cual, a partir de un conjunto de rasgos, se establecen los factores que son luego interpretados en funciones. Las dimensiones que se construyeron son:

Dimensión 1: Foco contextual e interactivo
Dimensión 2: Foco narrativo
Dimensión 3: Foco compromiso
Dimensión 4: Foco modalizador
Dimensión 5: Foco informativo

Cada dimensión se caracteriza por la co-ocurrencia sistemática de un cierto tipo de rasgos. Además, estas dimensiones explican, en mayor o menor medida (entre más rasgos tiene una dimensión más explica), la variación entre los registros estudiados. La Dimensión 1: Foco contextual e interactivo caracteriza al registro oral. Los rasgos más característicos en esta dimensión son las subordinadas adverbiales de causa-efecto, los adverbios de tiempo, los adverbios de negación, los pronombres de segunda persona y los de primera singular. La Dimensión 2: Foco narrativo caracteriza al registro literario y sus rasgos dan cuenta de la sucesión cronológica de eventos. Sus rasgos prominentes son los pronombres de tercera persona singular, los pronombres de primera persona plural, el futuro perifrástico, el pretérito imperfecto, los pronombres de tercera persona plural, el modo indicativo, las desinencias de primera persona plural, los verbos modales de volición, el pretérito indefinido y pronombres de negación, entre otros. La Dimensión 3: Foco compromiso caracteriza a aquellos textos literarios donde la aparición de personas, junto con sus intenciones, es prominente. Sus rasgos característicos son los verbos privados, los pronombres de primera persona singular, el pretérito indefinido, los verbos de volición, las desinencias de primera persona singular, el modo indicativo y el pretérito imperfecto, entre otros. La Dimensión 4: Foco modalizador caracteriza a aquellos textos donde aparecen explícitamente las actitudes de los hablantes. En el estudio realizado, son las entrevistas orales las que puntúan más alto en esta dimensión. Sus rasgos más importantes son las formas activas con ser, los atenuadores, los verbos modales de posibilidad, los adverbios modales, los adjetivos predicativos y las desinencias de tercera persona plural. La Dimensión 5: Foco informativo se manifiesta en unos pocos rasgos, a saber, los verbos modales de obligación, el modo subjuntivo, las nominalizaciones, los participios en función adjetiva y las frases preposicionales. Esta dimensión caracteriza al registro técnico-

profesional y resulta de especial importancia porque permite distinguir además los registros orales (donde aparece con puntajes negativos) y los registros escritos. Para un detalle del resto de los rasgos, así como también, para la relación entre las dimensiones y los registros estudiados, revísese Parodi (2005a).

2.1.5. Secuencias léxicas o *lexical bundles*

Una secuencia, paquete léxico, *cluster* o n-grama es una cadena de elementos que se repiten juntos frecuentemente en un corpus determinado. Estos paquetes pueden o no cumplir una función sintáctica ya que pueden estar formados por elementos que no configuran una unidad sintáctica (e.g. “de la”).

Los paquetes léxicos (Biber, 2005) han sido utilizados como indicadores para explorar la variación interdisciplinar de registros. Luego que se determina su extensión, frecuencia y variabilidad (por ejemplo, que sean de 2 a 4 unidades que aparezcan más de 20 veces cada un millón de palabras y que aparezca al menos una vez en cada uno de los registros estudiados) se pueden clasificar, si así lo requiere la investigación, funcionalmente. Los paquetes léxicos han sido buenos indicadores de la variabilidad funcional entre distintos registros asociados a áreas disciplinares (Biber, 2005). Algunos de los paquetes léxicos estudiados por Biber (2005) son: ‘al mismo tiempo’, ‘así como’, ‘el hecho de que’, etc.

2.1.6. La distribución de categorías gramaticales y unidades de medición estadística

En este estudio, consideraremos como otra fuente de información lingüística, la distribución porcentual de las categorías gramaticales. Hemos seleccionado aquellas categorías que transmiten el contenido de los textos (sustantivos, verbos copulativos, verbos no copulativos, adjetivos, adverbios de lugar, adverbios de modo, adverbios de tiempo), descartando otras categorías que no transmiten contenido léxico como las conjunciones, las subjunciones y las preposiciones.

Sobre estas categorías se aplicarán además elementos de medición estadística generales, los que generalmente se aplican en los estudios de corpus. El primero de ellos es la tasa de tipo por caso, la que proporciona un indicador de variabilidad de todas las palabras del corpus. Este índice se puede aplicar además a algunas de las categorías gramaticales, por ejemplo, podemos estudiar la variabilidad verbal (Sabaj, 2004) o adjetival.

2.2. Bases metodológicas de la construcción de la matriz

Exponemos en este punto algunos aspectos metodológicos de la construcción de la matriz. En un primer momento, como ya se señaló, exploramos distintos estudios de corpus e incorporamos de ellos aquellos rasgos que pudieran incluirse dentro de una teoría de corte discursivo.

Luego, en una segunda revisión bibliográfica, buscamos estudios que, aunque no estuvieran basados en corpus, proponían algunos rasgos asociados a dimensiones discursivas. En una tercera fase, incorporamos el resto de los rasgos encontrados a nuestra matriz, intentando siempre que nuestros componentes respondieran a factores relacionados con aspectos textuales.

Tal y como mostramos anteriormente, nuestro estudio en general está orientado por la teoría, ya que los rasgos a estudiar están todos predeterminados. Como dijimos, esto puede ser una debilidad puesto que no está probado por los datos que lo que denominamos componente responde al concepto de dimensión de los estudios multifactoriales (en el sentido que estos emergen de los datos). Aun así, creemos que esta forma de estudiar lengua en uso (i.e., a partir de componentes predeterminados) es una manera innovadora de acercarse a los estudios multirasgos y tiene las ventajas metodológicas que ya expusimos.

Ahora bien, el estudio y la construcción de la matriz, propiamente tal, es una exploración, puesto que se busca analizar el potencial de la matriz para caracterizar los registros y detectar diferencias entre ellos. Dicho en otras palabras, se busca analizar cuáles son los rasgos de la matriz que permiten visualizar la variación teórica entre los registros. Es importante señalar que, aunque la caracterización y el análisis de variación están intrínsecamente relacionados, todos los rasgos tienen potencial descriptivo (puesto que son parte de la lengua en uso real), pero no todos resultan ser críticos en la diferenciación de registros. Así, este estudio puede ser entendido como una búsqueda de rasgos críticos para, en una futura investigación, contar con una matriz de rasgos que ya han sido detectados como importantes para describir la variación inter-registro. Este punto, también, pretende ser innovador respecto de otras metodologías en los estudios de rasgos: lo que buscamos *a priori* es la determinación de diferencias y no la mera caracterización.

Debido a lo anterior, lo que nos interesa en el estudio exploratorio es determinar cuán variable es el comportamiento de los datos. De este modo, entre más varíe la ocurrencia de los rasgos, serán más críticos para determinar la variación de los registros.

Ahora bien, en esta exploración no determinaremos si las diferencias entre los porcentajes de ocurrencia de un componente entre dos registros es significativa estadísticamente; solo analizaremos cuáles son los conjuntos de rasgos que más varían y solo en la matriz definitiva se calculará si las diferencias son o no significativas desde un punto de vista estadístico. Además, se debe considerar que un conjunto de rasgos no siempre va a diferenciar a todos los registros, por lo tanto, determinaremos qué rasgos distinguen a qué registros.

3. El estudio exploratorio

Como una forma de probar la utilidad de nuestra matriz, aplicaremos una parte de ella a un conjunto pequeño de textos de características variadas. Los componentes que aplicaremos para este estudio exploratorio son algunos de los rasgos de los modos de organización del discurso, de la teoría de la valoración y del modelo de indexación de eventos. Los rasgos seleccionados aparecen operacionalizados en un anexo disponible en http//:omarsabaj.wordpress.com, denominado "cadenas de búsqueda".

3.1. Corpus

Para este estudio exploratorio, hemos seleccionado un corpus compuesto por cinco registros de contextos diversos, como forma de contar con un conjunto heterogéneo. Hemos seleccionado estos registros para establecer el potencial diferenciador de los rasgos escogidos. El corpus por analizar ha sido recolectado en el marco de diversas investigaciones de corpus realizadas por los investigadores de los Programas de Postgrado en Lingüística de la Pontificia Universidad Católica de Valparaíso, Chile. Su conformación, más una breve descripción y procedencia, se presenta en el siguiente cuadro.

Cuadro 1. Conformación del corpus.

Registro	Descripción	Procedencia	Número de palabras
Artículos sobre literatura	Artículos críticos sobre obras literarias latinoamericanas escritos por expertos chilenos	Internet	20.023
Artículos científicos Biología	Artículos científicos de biología de revistas indexadas en Scielo, Chile	Venegas (2005b)	149.495
Manuales de Ingeniería en construcción	Manuales de estudio de la carrera de Ingeniería en Construcción de la Pontificia Universidad Católica de Valparaíso	Fondecyt 1060440	313.244
La Biblia	Texto religioso judeo-cristiano	Internet	1.049.643
Oralidad	Conjunto de breves interacciones cotidianas entre personas cercanas	Corpus oral de la Universidad Autónoma de Madrid	873.669
Total			**2.406.074**

3.2. Procedimientos de producción de información

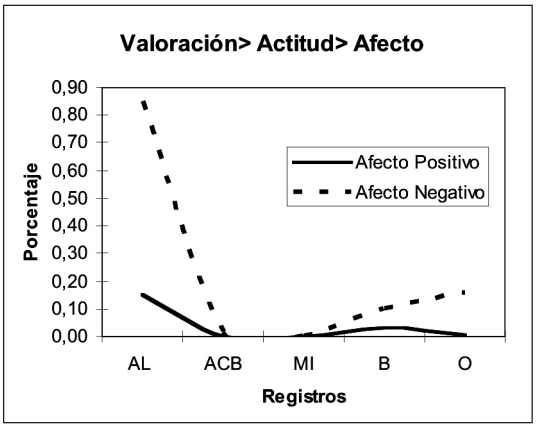
Debido a que no contamos con el total del corpus marcado sintácticamente, solo utilizamos cadenas de búsqueda de formas, tratando de crear un sistema de análisis simple y disponible gratuitamente a toda la comunidad investigadora. Para ello, se utilizó el programa de concordancias gratuito Antconc 3.2 desarrollado por Laurence Anthony (2006) en la Universidad de Waseda, Japón. Los resultados fueron normalizados en porcentajes y fueron traspasados a planillas de cálculo, a partir de las cuales se graficaron.

4. Resultados

4.1. La teoría de la valoración

4.1.1. Afecto

En el Gráfico 1, aparece el porcentaje de los marcadores de afecto en los registros estudiados.



(AL = Artículos sobre literatura; ACB = Artículos científicos de biología; MI = Manuales de ingeniería; B = La Biblia; O = Oralidad).

Gráfico 1. Marcas de afecto.

Tal como se aprecia en el Gráfico 1, el afecto tiene una muy baja ocurrencia en todos los registros estudiados, ya que no supera el 1% de las ocurrencias. Prima, además, el afecto negativo sobre el positivo. A primera vista, el afecto positivo no pareciera ser un buen indicador de diferencias entre los registros. El afecto negativo, en cambio, permite diferenciar los AL de los restantes registros. Los marcadores de afecto no resultan buenos indicadores de

las diferencias teóricas entre los registros MI, B y O. Esto se debe, entre otras razones, a sus bajos porcentajes.

4.1.2. Juicio

En el Gráfico 2, se muestra los porcentajes de ocurrencia de los marcadores de juicio en los registros estudiados.

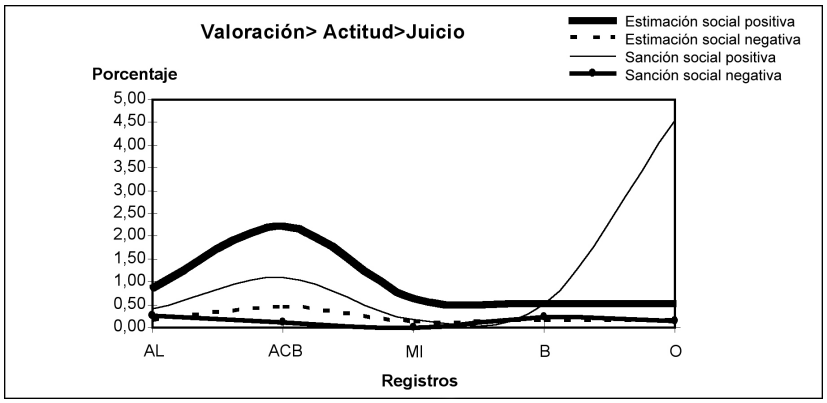


Gráfico 2. Marcadores de juicio.

En el Gráfico 2 podemos visualizar que la estimación social positiva en los ACB y la sanción social positiva en la O marcan puntos de diferencia entre los registros estudiados. La estimación social positiva y la negativa, junto con la sanción social negativa tienen un comportamiento muy plano en los registros AL, MI, B y O. No solo su comportamiento es plano entre los registros sino que también en cada uno de ellos presentan un porcentaje de aparición muy similar.

Las diferencias entre los porcentajes de aparición de la sanción social positiva en la Oralidad y en el resto de los registros son las más marcadas. La estimación social negativa y la positiva y la sanción social negativa no son, entonces, buenos indicadores de las diferencias de los registros MI, B y O.

4.1.3. Apreciación

En el Gráfico 3, aparecen los porcentajes de los marcadores de apreciación en el corpus.

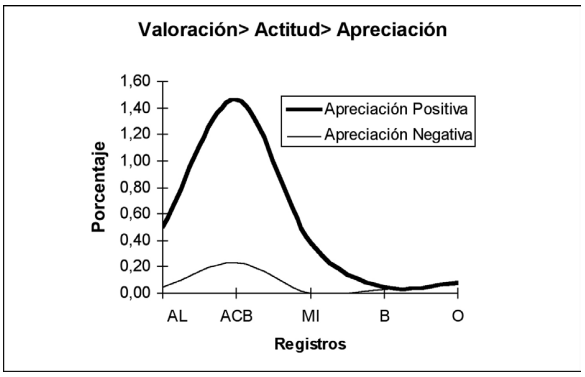


Gráfico 3. Marcas de Apreciación.

En el Gráfico 3, se puede observar que la apreciación positiva aparece como un componente marcado en los ACB. Esta ocurrencia, relativamente alta, puede deberse a la caracterización de procesos biológicos y permite distinguir a ese registro, en particular, del resto. Las marcas de apreciación en la B y la O son muy escasas en relación al total de palabras respectivo. En los ACB la diferencia entre los dos tipos de apreciación aparece más marcada que en los otros casos. Del Gráfico 3, se puede desprender que la apreciación negativa no parece ser un factor que refleje las diferencias macro-contextuales de los registros y no sería, por lo tanto, un factor crítico.

4.2. Los Modos de organización del discurso

En el Gráfico 4, se muestran los porcentajes de aparición de las marcas asociadas a los modos de organización del discurso.

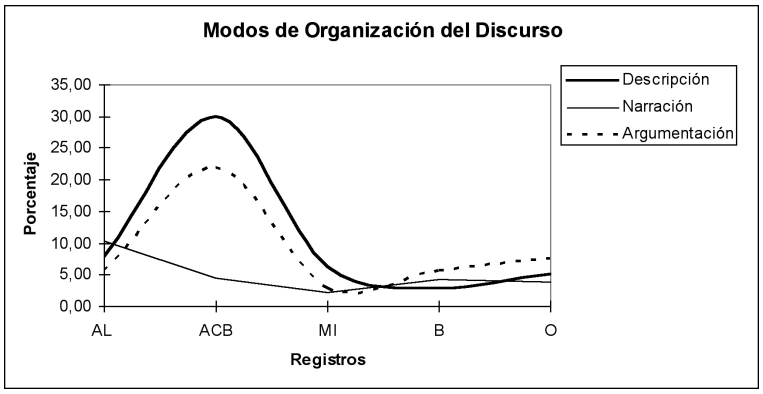


Gráfico 4. Los modos de organización del discurso.

A diferencia de los casos anteriores, los modos de organización del discurso tienen un comportamiento más heterogéneo a través de los registros, lo que se puede constatar en el rango porcentual (de 2% a 30%) más amplio. Esto permite sostener que algunas de las diferencias porcentuales de estos componentes en los distintos registros resultan ser críticos para constatar la variación teórica entre los registros. Tal es el caso, principalmente, de la descripción y la narración en los ACB donde presentan porcentajes mucho más prominentes que en los otros registros. Consecuentemente con lo esperado, la narración es más relevante en los AL que en cualquier otro registro. Al igual que en el caso de las categorías de la teoría de la valoración, los tres modos tienen un comportamiento llano en los registros MI, B y O.

4.3. El modelo de indexación de eventos

En el Gráfico 5, se presentan los porcentajes de aparición de los marcadores de tres dimensiones del modelo de indexación de eventos.

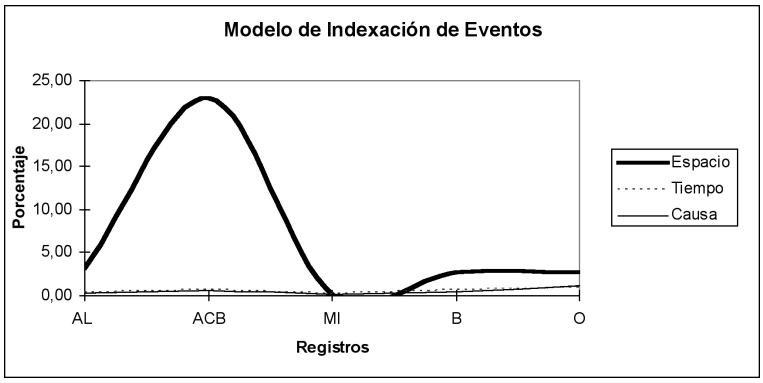


Gráfico 5. El modelo de indexación de eventos.

En el Gráfico 5, se puede observar que una de las tres dimensiones estudiadas puede ser utilizada para el estudio de las diferencias entre registros. El espacio aparece como un aspecto importante en el registro ACB. Es interesante destacar que tanto la causa como el tiempo no parecen ser muy sobresalientes en ninguno de los registros. A pesar de lo anterior, el espacio sí puede ser considerado como un rasgo crítico en la diferenciación del registro ACB del resto de los registros estudiados.

COMENTARIOS Y PROYECCIONES

En el presente trabajo, hemos propuesto una matriz de rasgos para probar su potencial para la determinación de la variación de registros teóricamente diferentes. Una parte de la matriz propuesta ha sido aplicada para el estudio empírico de la variación de cinco registros. Las tendencias generales muestran que son los modos de organización del discurso, junto con la dimensión espacial del modelo de indexación de eventos, las que resultan ser más importantes como dimensiones que ayudan a determinar la variación entre los registros analizados.

Como en todo estudio exploratorio, el presente se ha encontrado con diversas dificultades que pueden ser consideradas en trabajos futuros. La debilidad que se puede establecer en esta investigación es que:

> Los rasgos estudiados pueden cumplir diversas funciones por lo que su asignación a una dimensión no siempre es directa y debe ser revisado uno a uno. Esta polifuncionalidad de los rasgos puede provocar una interpretación errada de los datos.

A pesar de lo anterior, creemos que el estudio es un aporte por las siguientes razones: a) la metodología es simple y se utilizan herramientas disponibles de forma gratuita a la comunidad investigadora; b) a pesar de que no todos los rasgos incluidos en la matriz señalan una diferencia entre los registros, todos ellos tienen un alto potencial descriptivo, ya que provienen de ejemplos de lengua auténtica. En este sentido, hemos descubierto que algunos de los rasgos seleccionados en nuestra matriz sí tienen un impacto en un nivel textual, sobre todo aquellos que permiten distinguir registros (los modos de organización del discurso y la dimensión espacial del modelo de indexación de eventos).

Las proyecciones del estudio tienen al menos dos orientaciones. La primera dice relación con la superación de la debilidad del mismo. En este sentido, la matriz final debería presentar un mecanismo que determinara qué elementos son más probablemente polifuncionales, de forma de tener precaución en la interpretación de los datos. Una segunda proyección se basa en el desarrollo del modelo. Este debe incluir la aplicación del resto de los elementos estudiados en la matriz que no se incluyeron en esta primera exploración. Así también, el desarrollo de la investigación requiere la aplicación de un modelo estadístico más sofisticado que permita ver correlaciones simples y múltiples entre todos los componentes. Con este estudio podríamos saber, por ejemplo, si las marcas de argumentación se asocian a afecto positivo o si la descripción tiene alguna correlación con la aparición de marcas de la dimensión espacial.

REFERENCIAS BIBLIOGRÁFICAS

Adam, J. (1992). *Les textes: Types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris: Nathan.

Alarcos Llorach, E. (1999). *Gramática de la lengua española*. Madrid: Espasa-Calpe.

Anthony, L. (2006). *Antconc3.2.0w* [en línea]. Disponible en: http://www.antlab.sci.waseda.ac.jp/software/antconc3.2.0w.exe

Bassols, M. & Torrent, A. (1997). *Modelos textuales, teoría y práctica*. Barcelona: Eumo Editorial.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, (62), 384-414.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational linguistics*, (19), 243-258.

Biber, D. (1994). Using register-diversified corpora for general language studies. En S. Armstrong (Ed.), *Using large corpora* (pp. 180-201). Cambridge, Massachusetts: MIT Press.

Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. En Leistyna, P. & Meyer, Ch. (Eds.), *Corpus analysis. Language structure and language use* (pp. 47-70). Amsterdam: Rodopi.

Biber, D. (2005). Paquetes léxicos en textos de estudio universitario: Variación entre disciplinas académicas. *Revista Signos, 38*(57), 19-29.

Biber, D. & Tracy-Ventura, N. (2007). Dimensions of register variation in Spanish. En G. Parodi (Ed.), *Working with Spanish corpora* (en prensa). London: Continuum.

Calsamiglia, H. & Tusón, A. (1999). *Las cosas del decir, manual de análisis del discurso*. Barcelona: Ariel.

Charaudeau, P. (1992). *Grammaire du sens et de l'expression*. Paris: Hachette Éducation.

Charaudeau, P. & Maingueneau, D. (Dirs.) (2002). *Dictionnaire d'analyse du discours*. Paris: Seuil.

Ciapuscio, G. (2005). La noción de *género* en la Lingüística Sistémico Funcional y en la Lingüística Textual. *Revista Signos, 38*(57), 31-48.

Fellbaum, C. (Ed.) (1998). *Wordnet: An electronical lexical database*. Cambridge: MIT Press.

Hair, J., Anderson, R., Tatham, R. & Black, W. (2001). *Análisis multivariante*. Madrid: Prentice Hall.

Lo Cascio, V. (1998). *Gramática de la argumentación*. Madrid: Alianza.

Louwerse, M., McCArthy, P., McNamara, D. & Graesser, A. (2004). *Variation in language and cohesion across written and spoken registers* [en línea]. Disponible en: http://www.autotutor.org/publications/AddPaper/LouwerseMcCarthyMcNamaraGraesser2004.pdf

Marinkovich, J. & Cademartori, Y. (2004). Foco narrativo y foco informativo: Dos dimensiones para una descripción de los manuales en la formación técnico-profesional. *Revista Signos, 37*(55), 31-40.

Martin, J. & White, P. (2005). *The language of evaluation: Appraisal in English*. London: Palgrave.

McNamara, D., Lowerse, M. & Graesser, A. (2004). *Coh-metrics: Automated cohesion and coherence scores to predict text readability and facilitate comprehension* [en línea]. Disponible en: http://cohmetrix.memphis.edu/cohmetrixpr/archive/*Coh-metrix*Grant.pdf

Munguía, I., Munguía, M. & Rocha, G. (2000). *Gramática de la lengua española. Reglas y ejercicios*. México D.F.: Larousse.

Parodi, G. (2005a). Lingüística de corpus y análisis multidimensional: Exploración de la variación en el Corpus PUCV-2003. En G. Parodi (Ed.), *Discurso especializado e instituciones formadoras* (pp. 83-126). Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. (Ed.) (2005b). *Discurso especializado e instituciones formadoras*. Valparaíso: Ediciones Universitarias de Valparaíso.

Perelman, Ch. & Olbrecht-Tyteca, L. (1994). *Tratado de la argumentación. La nueva retórica*. Madrid: Gredos.

Rojas, C. (1988). *Verbos locativos en español*. México: Universidad autónoma de México.

Sabaj, O. (2004). Especificidad, especialización y variabilidad verbal: Una aproximación computacional en estadística léxica. *Revista Signos, 37*(56), 75-89.

Toulmin, S. (1958). *The uses of arguments*. Londres: Cambridge.

Venegas, R. (2005a). Hacia una identificación automatizada de rasgos argumentativos en corpus. En G. Parodi (Ed.), *Discurso especializado e instituciones formadoras* (pp. 127-158). Valparaíso: Ediciones Universitarias de Valparaíso.

Venegas, R. (2005b). *Las relaciones léxico-semánticas en artículos de investigación científica: Una aproximación desde el análisis semántico latente*. Tesis doctoral, Pontificia Universidad Católica de Valparaíso, Chile.

White, P. (1999). *Un recorrido por la teoría de la valoración* [en línea]. Disponible en: www.grammatics.com/valoración/

Zwaan, R., Langston, M. & Graesser, A. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6, 292-297.

Zwaan, R. & Radvansky, G. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123, 162-185.

NOTAS

[1] El modelo de situación (Zwaan & Radvansky, 1998) tiene 5 dimensiones, pero en *Coh-metrix* no se empleó la dimensión relativa al protagonista.

[2] La síntesis es nuestra, se han traducido los términos siguiendo las propuestas en español de White (2005).