VERBOS Y SUSTANTIVOS EN TEXTOS CIENTÍFICOS. ANÁLISIS DE VARIACIÓN EN UN CORPUS DE TEXTOS DE CIENCIAS APLICADAS, NATURALES, SOCIALES Y HUMANIDADES[1]

Guillermo Soto
Ricardo Martínez
Scott Sadowsky
Universidad de Chile

The present investigation proposes that the nominalization index (the ratio of nouns to verbs in a given text) successfully predicts the academic-scientific character of a text. The results of a computer-assisted analysis of a 12.7 million word corpus show that this index is sensitive to the opposition between scientific/academic texts and letters to the editor/journalistic texts. It is likewise sensitive to the differences between applied sciences, natural sciences, social sciences and humanities texts, although there is a certain amount of overlap between these last two. When contrasting expert and novice texts, the index is only sensitive in the case of applied and natural sciences. It is suggested that the greater relative occurrence of nouns could be related to the construction of an object-centric depersonalized discourse in which the vicissitudes of actors' mental processes are believed to be irrelevant in the interpretation of discourse-coherence relationships.

1. Introducción

1.1. La determinación de géneros discursivos y registros lingüísticos en la lingüística computacional

La determinación de géneros y registros discursivos constituye una tarea tradicional de los estudios del discurso (Ciapuscio, 1997; De Beaugrande y Dressler, 1997; Calsamiglia y Tuson, 1999; Van Dijk, 1997). En términos muy generales, pueden distinguirse dos grandes perspectivas en esta materia: de un lado, estudios trascendentes que intentan caracterizar los géneros atendiendo a la relación entre los textos y sus contextos sociales y cognitivos; del otro, indagaciones inmanentes que persiguen dar cuenta de los registros considerando sus rasgos lingüísticos formales. Mientras los primeros predominan en estudios de corte retórico y literario, los segundos son desarrollados en su mayoría por lingüistas. Ciertamente, ambas orientaciones no son incompatibles, toda vez que, en primer término, la determinación de rasgos lingüísticos presupone, normalmente, alguna referencia al valor en uso de estos, y, en segundo lugar, la relación entre texto y contexto se manifiesta en

[1] Este trabajo forma parte del proyecto de investigación DID SOC-01/01-2 de la Universidad de Chile. Guillermo Soto: gsoto@uchile.cl; Ricardo Martínez: ricardomartinezg@gmail.com; Scott Sadowsky: ssadowsky@gmail.com.

determinados fenómenos lingüísticos. De ahí, pues, que en general la determinación de los géneros o los registros obedezca a alguna combinación de rasgos inmanentes y trascendentes[2].

En las últimas décadas ha habido un importante avance en la caracterización y determinación de géneros o registros a partir de análisis cuantitativos automáticos sobre corpora extensos. Gran parte de estos estudios se centran en propiedades inmanentes puramente formales, esto es, no operan sobre categorías léxico-gramaticales dotadas de expresión y contenido sino sobre unidades de expresión carentes de significado[3]. Si bien estos trabajos pueden resultar útiles para tareas prácticas como la clasificación automática de documentos, desde el punto de vista estrictamente lingüístico tienen un interés limitado precisamente por no analizar el discurso y sus constituyentes como hechos de lenguaje. Más pertinentes, en este sentido, pueden ser indagaciones que, empleando las herramientas de la lingüística computacional y de corpus, intenten operar sobre categorías lingüísticas. Este último tipo de investigaciones podría aportar información tanto sobre el funcionamiento textual de las categorías consideradas como sobre el funcionamiento global de los textos como hechos de lenguaje.

Entre los estudios que intentan combinar fenómenos inmanentes y trascendentes mediante análisis automático computarizado de categorías léxico-gramaticales, destaca el trabajo seminal de Biber (1988). En este, a partir de un estudio multidimensional en que se emplea el análisis factorial, se determinan seis dimensiones de covariación de rasgos lingüísticos. Estas, interpretadas sociocognitivamente, sirven para proponer una categorización de registros escritos y orales del inglés. El análisis propuesto por Biber consta de tres momentos. Primero, una etapa cualitativa de determinación de las unidades por considerar (variables y textos); segundo, un análisis cuantitativo que considera tanto la frecuencia de las variables como sus correlaciones; y tercero, una interpretación cualitativa de los resultados globales. Desde el punto de vista metodológico, se trata de un estudio mixto en que tanto la primera etapa como la tercera dependen de estudios previos que han establecido relaciones funcionales específicas o generales a partir de indagaciones que han considerado otros métodos y otros corpora.

El método propuesto por Biber se ha aplicado, con mayores o menores ajustes, a diversos estudios sincrónicos, diacrónicos y de desarrollo del lenguaje en distintas lenguas (Sardinha, 2000), entre ellas el español (Parodi, 2005; Soto/Sadowsky/Martínez, en preparación). La magnitud de rasgos considerados en este tipo de análisis y el ingenio diseñado para conformar constelaciones explica, probablemente, su adopción como modelo por un número creciente de investigadores.

La aplicación del método de Biber en el análisis de corpus relativamente extensos del orden de cientos de miles de palabras ha tenido como resultado, primero, una complejización de dicotomías clásicas como oralidad-escritura[4]; segundo, un apoyo cuantitativo a estudios funcionales previos que caracterizaban fenómenos léxico-

[2] Para una caracterización más detenida de la oposición entre géneros y registros, v. MacDonald (2002), quien, además, discute el concepto relacionado de 'estilo'. La perspectiva social de los géneros se presenta en Bajtín, 1986; el enfoque retórico, en Miller, 1984; una propuesta lingüístico-retórica puede encontrarse en Swales, 1990; una aproximación a los géneros como tecnologías cognitivas se expone en Soto (2005b).

[3] Un listado de este tipo de herramientas se encuentra en http://tinyurl.com/3qw3q.

[4] Esta complejización no es exclusiva del trabajo de Biber (cfr., entre otros, Tannen, 1982a, 1982b; Chafe y Tannen, 1987; Ludwig, 1989).

gramaticales; y, tercero, una organización de registros textuales en un espacio multidimensional estructurado a partir de vectores interpretados funcionalmente.

No obstante sus importantes logros, este tipo de metodología presenta algunos problemas. Como en el estudio de Biber no se analizan los datos del corpus desde el punto de vista de un análisis del discurso que considere el funcionamiento efectivo de unidades dotadas de expresión y contenido, esto es, signos en uso, no existe garantía con respecto a la conexión propuesta entre rasgos y funciones. Este problema, evidente en el nivel de las construcciones gramaticales, que no se reducen de manera simple e indiscutible a rasgos formales explícitos, está presente, también, en todo tipo de signo, dado que la relación forma-función no es biunívoca y la frontera entre semántica y pragmática es difusa. Proyectado al análisis biberiano, lo anterior significa que si bien puede determinarse que hay rasgos con mayor frecuencia en textos en que predominan ciertas funciones, no se puede aseverar que el rasgo cumpla siempre esa función ni que determinada función se instrumente sólo en determinados rasgos. De hecho, incluso cuando el rasgo se manifiesta predominantemente en discursos con cierta función no hay garantía de una relación funcional directa. Un ejemplo extremo se advierte en el caso de la argumentación. Como ha expuesto Hyland (2002), el hecho de que rasgos asociados a ciertos tipos de argumentación no aparezcan en determinados textos no significa que en ese texto la argumentación no sea relevante, menos aun cuando esos rasgos no instrumentan directamente la argumentación sino que son rasgos que estarían presentes en unidades léxico-gramaticales o discursivas que realizan la argumentación[5].

En síntesis, el estudio de Biber, más allá de sus innegables méritos, presenta las siguientes limitaciones: emplea caracterizaciones funcionales previas que no son sometidas a un análisis discursivo; pretende operar con unidades gramaticales complejas que, no obstante, son identificadas sólo en su polo expresivo; no indaga en recursos específicos que, en determinados géneros, podrían instrumentar ciertas funciones; y propone interpretaciones funcionales globales que no se detienen en el funcionamiento textual. En general, estas restricciones obedecen tanto a la ambición del proyecto de Biber, que se propone un análisis variacionista automático de categorías gramaticales complejas, como a la ausencia en este de un análisis del discurso cualitativo que observe el funcionamiento efectivo de las unidades en el texto-discurso.

En términos más amplios, la presente discusión intenta mostrar que una adopción acrítica de la lingüística de corpus de matriz biberiana puede llevar a creer que es posible el análisis puramente empírico del lenguaje a partir del conteo, más o menos sofisticado, de recurrencias de datos presentes en un corpus. En una posición extrema, el analista se limitaría a generar inducciones a partir de un estudio cuantitativo de la realidad del lenguaje, tal y como esta se manifiesta en el corpus, sin la necesidad, incluso, de una teoría del lenguaje. Sin embargo, como se ha esbozado, el análisis automático de corpus presenta una serie de limitaciones. Como señala Weigand (2004), el texto es solo parte del objeto de estudio del lingüista interesado en el funcionamiento real del lenguaje, toda vez que no da cuenta directa del aspecto cognitivo y del contextual implicados en los procesos de producción y comprensión del discurso (van Dijk, 1997). La determinación del funcionamiento efectivo tanto del texto globalmente considerado como de las unidades que lo constituyen requiere normalmente de acceso al contexto extralingüístico y al

[5] Una discusión más amplia de esta crítica en Soto (2004).

conocimiento de mundo, dos dimensiones que los corpora no registran y que, por tanto, no pueden ser directamente investigadas en ellos. Más específicamente, una misma forma puede tener distintos significados o funciones dependiendo de su papel en la actividad comunicativa concreta, lo que restringe, desde el punto de vista lingüístico, el alcance de los análisis puramente cuantitativos que operan sobre expresiones y no sobre signos. Además de lo anterior, las funciones, esto es, los significados o sentidos en los textos, pueden implementarse de manera indirecta a través de procesos complejos que implican no solo el empleo de recursos gramaticales y léxicos sino también la generación de inferencias de distinto tipo y diverso alcance (Tomlin et al, 2000), cuestiones, estas últimas, externas al texto explícito.

Las críticas formuladas hasta aquí no apuntan a un rechazo de los proyectos de análisis automático computarizado de categorías léxico-gramaticales, sino, más bien, a una precisión de su alcance y, consecuentemente, de su valor para el análisis del discurso. En este sentido, sugieren una actitud cautelosa con respecto a los análisis de este tipo, en especial cuando, como en el caso de Biber (1988), se proponen interpretaciones funcionales no observadas directamente en los discursos y se consideran categorías gramaticales que, aun en el análisis manual, son difíciles de precisar como ocurre con frecuencia, por ejemplo, con las construcciones pasivas.

1.2. Verbos y sustantivos en el discurso científico

Considerando lo anterior, parece haber espacio para estudios que intenten analizar automáticamente la variación de fenómenos gramaticalmente relevantes sin proponer, por otra parte, un nivel interpretativo funcional que trate de caracterizar de modo automático el funcionamiento de dichas unidades en los textos (o, más ampliamente, el funcionamiento discursivo-textual de las unidades léxico-gramaticales), lo que requeriría de una metodología de análisis cualitativo. Una investigación de este tipo debería centrarse, primero, en el examen cuantitativo computarizado de elementos léxico-gramaticales de aceptación general y baja discutibilidad en su determinación, y, segundo, a partir de ello proponer una interpretación mínima que permitiera distinguir de manera sencilla grupos de textos. En otro trabajo, actualmente en preparación, los autores del presente estudio han realizado una exploración en que, siguiendo el modelo de análisis cuantitativo de Biber, se simplifican los rasgos considerados, se aumenta el corpus en un orden de magnitud, y se minimizan las interpretaciones. Por su parte, en el presente trabajo se explora otra alternativa de análisis que, haciéndose cargo también de las críticas ya expuestas, compara, en el mismo corpus, la frecuencia relativa de dos categorías léxico-gramaticales básicas, de aceptación general y altamente relevantes desde el punto de vista lingüístico. A partir de este análisis, el trabajo propone tanto una distinción entre textos académico-científicos y textos no académicos como un ordenamiento de textos científicos de distintas áreas disciplinarias: ciencias aplicadas, ciencias naturales, ciencias sociales y humanidades. Adicionalmente, el trabajo considera la eventual incidencia que el expertizaje del autor puede tener en la proporción relativa de verbos y sustantivos.

Las categorías léxico-gramaticales seleccionadas para el análisis son las de verbo y sustantivo, dos nociones fundamentales que, además de su aceptación extendida, son portadoras, normalmente, de los significados extralingüísticos relativos a las entidades y los procesos, por lo que contribuyen sustancialmente al contenido informativo de los textos. Si bien la aseveración de Gutiérrez Rodilla (1998) en el sentido de que el vocabulario es “el

elemento caracterizador del lenguaje científico" (p. 37) puede discutirse, ella parece especialmente atendible, como lo indica la propia autora, en lo que respecta a sustantivos, verbos y, en menor grado, al parecer, adjetivos.

Además de las razones ya aportadas, la decisión de centrar el análisis en estas dos categorías radica, primero, en que ambas, como es de conocimiento general, constituyen las piezas básicas de la organización de cláusulas, oraciones y enunciados gramaticales[6], y, segundo, en que diversos estudios realizados en corpora de distintas lenguas muestran que los textos científicos, y más ampliamente la prosa académica, presentan un fuerte incremento de sustantivos y sintagmas nominales con respecto al común de los textos (cfr. Biber, 1988; Halliday y Martin, 1993; Gutiérrez Rodilla, 1998; Gross y Harmon, 1999; Wolters y Kirsten, 1999; Albentosa y Moya, 2000).

El estilo fuertemente nominalizado de los textos científicos ha recibido diversas interpretaciones. Aunque típicamente los estudios se han realizado en inglés, estas también parecen ser, en mayor o menor medida, atendibles para textos escritos en otras lenguas. Según algunos, dicho estilo obedecería a la construcción discursiva de una realidad fija y determinada en que predominarían los objetos (Halliday y Martin, 1993); para otros, respondería a la capacidad entificatoria de los especialistas, que se manifestaría en una elevada presencia de nominalizaciones y sintagmas nominales complejos (Zenteno, 1996). Por su parte, Gutiérrez Rodilla (1998) relaciona este fenómeno con la presencia de tecnicismos, un factor observado, asimismo, por Halliday y Martin. También se ha propuesto que daría cuenta del privilegio que este tipo de discurso otorga a los objetos por sobre los procesos (Gross y Harmon, 1999); o que funcionaría como un recurso de la despersonalización del discurso científico (Albentosa y Moya 2000). Soto y Zenteno (2004) muestran, en todo caso, que, al menos con respecto a los sintagmas nominales, existen variadas funciones que estas estructuras pueden desempeñar en los artículos de investigación científica, las que van desde las referenciales hasta las de organización textual y de estructuración jerárquica de la información en el discurso. En este sentido, es posible que la elevada presencia de sustantivos obedezca a más de una razón simple.

Más ampliamente, Heylighen y Dewaele (2002) plantean que los sustantivos comunes son más frecuentes en textos con baja dependencia de contexto (instancias de expresión "formal"[7]) mientras que los verbos son más frecuentes en estilos fuertemente contextuales. Como recuerda Anderson (1996), la mayor frecuencia de sustantivos comunes respecto de verbos finitos suele indicar mayor longitud de las oraciones, lo que podría, a su vez, connotar una mayor distancia entre los participantes del discurso. Desde esta perspectiva, la razón entre sustantivos y verbos podría indicar cierta posición en el continuo distancia-compromiso. Finalmente, Wolters y Kirsten (1999) proponen que una gran frecuencia de verbos finitos es característica de textos narrativos o de "ficción" y de ciertos textos periodísticos, mientras que textos académicos, legales y políticos presentan baja frecuencia de estos verbos. En su conjunto, los planteamientos anteriores pueden relacionarse ya con el estilo epistémico del discurso académico y científico, como señala MacDonald (2002), ya

[6] El concepto de 'enunciado gramatical' se entiende, en el presente trabajo, en el sentido propuesto por Daneš (1966).

[7] Heylighen y Dewaele (2002) emplean 'formal' en el sentido que tiene en frases como 'lenguaje formal' o 'teoría formal', no en el sentido superficial de 'propio de una ceremonia o convención'.

con la dicotomía entre el "modo de pensamiento" narrativo y el paradigmático, propuesto por Bruner (1986), aunque se requiere de mayor investigación para asentar estas relaciones[8].

Estamos conscientes de que, más allá de los argumentos entregados hasta aquí, puede discutirse aun la existencia de una distinción clara entre la categoría sustantivo y la categoría verbo, problema que, en última instancia, tiene que ver con el estatus de las categorías lingüísticas en general. Contra el enfoque tradicional de categorías discretas, se han desarrollado, en particular desde la década de 1970, posturas que plantean tanto categorías de fronteras difusas como de estructura prototípica, radial o cohesionadas por la semejanza de familia de sus miembros. En este marco, se ha propuesto la existencia de grados de "sustantividad" ("nouniness") y se han explorado, entre otras, funciones verbales de tipo nominal y nominales de tipo verbal, todo ello con alcances semánticos y de forma. No obstante, ni el reconocimiento de fenómenos de prototipicidad ni la evidencia de un desempeño funcional complejo fuerzan a la adopción de un modelo de categorización que deseche la dicotomía verbo/sustantivo (Lakoff, 1987; Langacker, 1987). Una discusión amplia del estatus de las categorías lingüísticas puede hallarse en los trabajos recogidos por Aarts et al (2004); y con respecto a los conceptos en general, Margolis y Laurence (1999).

1.3. La presente investigación

La presente indagación se origina en la idea de que la relación entre sustantivos y verbos no solo permite distinguir los textos académicos y científicos de los que no lo son, sino que, junto con ello, en los propios textos académicos y científicos dicha relación es sensible al área disciplinaria a la que pertenece el texto. Más específicamente, se propone que mientras más se aproxima un texto a las ciencias naturales y a las ciencias aplicadas y la tecnología, más aumenta la proporción de sustantivos respecto de la de verbos. De acuerdo con esta hipótesis, si se estableciera una secuencia de textos que, partiendo de los no científicos, concluyera en los de ciencias naturales y tecnologías, pasando por los de humanidades y de ciencias sociales, esta secuencia sería correlativa con una gradiente en que la proporción de sustantivos respecto de la de verbos iría aumentando. Por otra parte, en cada área disciplinaria los expertos deberían mostrar una mayor proporción de sustantivos que los sujetos que aún no alcanzan expertizaje.

La proporción de sustantivos respecto de verbos en un texto determinado se denominará 'índice de nominalización'. La presente investigación propone que el índice de nominalización es un buen predictor tanto del carácter académico-científico de un texto como de su pertenencia a un área disciplinaria específica. Es importante tener claro que el 'índice de nominalización', tal y como aquí se define, no es una medida de la densidad informativa o referencial del texto, toda vez que se limita a comparar sustantivos con verbos sin establecer la proporción de sustantivos o de sintagmas nominales con respecto al total de elementos léxico-gramaticales. Tampoco puede entenderse como un índice de procesos morfológicos de nominalización ni de procesos semánticos de entificación, aunque intuitivamente pareciera haber una relación entre todos ellos. Finalmente, este estudio es agnóstico respecto de las interpretaciones funcionales propuestas en relación con la presencia de sustantivos en los discursos académicos y científicos, tarea para la que sería

[8] En Soto (2005) se explora la posibilidad de que las exigencias del modo de pensamiento paradigmático propio del discurso científico expliquen ciertas opciones gramaticales que supondrían una disminución del grado de narratividad en los artículos científicos.

necesario un análisis cualitativo que observara directamente el funcionamiento de las unidades en el texto/discurso.

2. Método

2.1. Criterios de elaboración del corpus

Los criterios de elaboración del corpus tuvieron por objeto, primero, cubrir un amplio espectro de campos disciplinarios que permitiera vislumbrar las propiedades del discurso académico especializado escrito en el español de Chile; segundo, dar cuenta de las propiedades de textos escritos tanto por sujetos expertos como por aprendices; y tercero, garantizar la calidad de los textos como ejemplares de discurso académico escrito. Con estos fines, se seleccionó un corpus constituido por textos académicos auténticos escritos por investigadores y por estudiantes chilenos en las áreas de ciencias aplicadas, naturales, sociales y humanidades. En el caso de los investigadores, se seleccionaron artículos científicos publicados en revistas especializadas, mientras que en el caso de los estudiantes, se seleccionaron tesis de grado. Para identificar el área disciplinaria específica de los artículos de investigadores, se consideró la publicación en que aparecía el texto (fuente). Así, por ejemplo, los artículos publicados en la Revista Geológica de Chile se agruparon en el área geología-oceanografía. Los textos de estudiantes se organizaron atendiendo al programa de estudios que estos cursaban. Posteriormente, las áreas disciplinarias se ordenaron en cuatro grandes categorías de aceptación general: ciencias aplicadas, ciencias naturales, ciencias sociales y humanidades. Para esta tarea, se consideró el estándar propuesto por el Fondo Nacional de Desarrollo Científico y Tecnológico de Chile, Fondecyt[9], el que se adaptó a los objetivos de la presente investigación reduciendo la cantidad de agrupaciones y reordenando algunas fuentes. Así, los textos de música se incluyeron en la categoría humanidades, al igual que los de teología, que no se consideran en el estándar Fondecyt; los de lingüística-filología, por su parte, se trasladaron desde la categoría humanidades a la de ciencias sociales, donde, además, se incluyó la fuente análisis del discurso; por último, en la agrupación ciencias aplicadas se incluyeron fuentes que el estándar Fondecyt subsume en el área de tecnología y ciencias silvoagropecuarias.

Como control del análisis, se utilizaron artículos periodísticos y cartas dirigidas a los medios de comunicación. Con la primera variedad de textos, se intentó dar cuenta de una situación de escritura profesional no académica, mientras que con la segunda, de una situación de escritura no profesional y no académica. El cuadro 1 muestra las variedades textuales básicas incluidas en el corpus, ordenadas de acuerdo con el grado de expertizaje de los autores respecto de la tarea de escritura específica (variable "profesional") y el carácter académico o no académico de los textos.

	[+ Académicos]	[- Académicos]
[+ Profesionales]	Artículos científicos	Artículos periodísticos
[- Profesionales]	Tesis	Cartas a medios de prensa

Cuadro 1: Variedades textuales básicas incluidas en el corpus

[9] http://www.fondecyt.cl/DOCUMENTOS/DISCIPLINAS%20FONDECYT.xls (15 de marzo de 2003).

2.2. Elaboración del corpus

A partir de los criterios anteriores, se elaboró un corpus de 12,7 millones de palabras, seleccionado del megacorpus de lengua escrita Corpus Dinámico del Castellano de Chile (Codicach), creado por Sadowsky entre los años 2001-2003. La selección del corpus se realizó de manera automática, de forma que no hubo un análisis cualitativo más allá de las categorizaciones expuestas en el apartado anterior. Con esto, se intentó eliminar los potenciales sesgos interpretativos de los investigadores. Al momento del análisis (noviembre de 2003) el megacorpus Codicach estaba compuesto de 830 millones de palabras de texto continuo[10] en 1,4 millones de archivos. Los textos extraídos del Codicach fueron sometidos a un proceso de depuración, poda y estandarización que permitió su análisis como texto plano.

En los cuadros que siguen, se muestra la distribución de los diferentes grupos de textos completos que constituyen el corpus, ordenados por variedad (académica o no académica), expertizaje de sus autores, área disciplinaria y fuente de los textos (tesis, artículos de investigación, cartas, artículos periodísticos). Además, en cada caso se indica el número de palabras. El Cuadro 2a presenta el corpus de textos académicos escritos por estudiantes; el Cuadro 2b, el de textos académicos de profesionales; el Cuadro 2c, el de textos no académicos de no profesionales; y el Cuadro 2d, el de textos no académicos escritos por profesionales. El corpus total considerado para el análisis estuvo compuesto de 12.683.960 palabras (6.231 archivos).

Area	**Fuente**	**Nº palabras**
CIENCIAS APLICADAS		
Agricultura	Tesis	20.185
Gestión ambiental	Tesis	53.750
CIENCIAS NATURALES		
Biología-Zoología-Botánica	Tesis	243.716
Medicina	Tesis	117.590
Química	Tesis	148.003
CIENCIAS SOCIALES		
Economía	Tesis	265.463
Lingüística-Filología	Tesis	38.977
Periodismo	Tesis	54.147
HUMANIDADES		
Literatura	Tesis	163.969
TOTAL		1.105.800

Cuadro 2a: Corpus: textos académicos, autores no profesionales

Area	**Fuente**	**Nº palabras**
CIENCIAS APLICADAS		
Agricultura	Agricultura Técnica	469.539

[10] *Running words.*

Nutrición	Revista Chilena de Nutrición	29.139
CIENCIAS NATURALES		
Biología-Zoología-Botánica	Biological Research	74.561
Biología-Zoología-Botánica	Gayana Botánica	86.185
Biología-Zoología-Botánica	Gayana Concepción	91.768
Biología-Zoología-Botánica	Revista Chilena de Historia Natural	374.839
Geología-Oceanografía	Investigaciones Marinas	157.300
Geología-Oceanografía	Revista Geológica de Chile	160.890
Medicina	Archivos de Medicina Veterinaria	428.508
Medicina	Boletín Chileno de Parasitología	41.260
Medicina	Parasitología al día	70.627
Medicina	Revista Chilena de Anatomía	232.646
Medicina	Revista Chilena de Enfermedades Respiratorias	55.379
Medicina	Revista Chilena de Infectología	169.207
Medicina	Revista Chilena de Obstetricia y Ginecología	45.190
Medicina	Revista Chilena de Pediatría	611.391
Medicina	Revista Médica de Chile	1.231.355
Química	Boletín de la Sociedad Chilena de Química	180.997
CIENCIAS SOCIALES		
Análisis del discurso	Misceláneos	32.748
Economía	Cuadernos de Economía	71.379
Lingüística-Filología	Estudios Filológicos	285.268
Lingüística-Filología	Literatura y Lingüística	261.976
Lingüística-Filología	Misceláneos	111.440
Sociología-Antropología	Chungará	65.561
Sociología-Antropología	Misceláneos	114.858
Urbanismo	EURE	166.851
HUMANIDADES		
Historia	Historia Santiago	291.132
Historia	Revista de Estudios Histórico-Jurídicos	1.011.546
Literatura	Acta Literaria	82.356
Música	Revista Musical Chilena	678.170
Teología-Filosofía	Teología y Vida	668.281
TOTAL		8.352.347

Cuadro 2b: Corpus: textos académicos, autores profesionales

Area	**Fuente**	**Nº palabras**
Control	Cartas a medios de prensa	566.649
TOTAL		566.649

Cuadro 2c: Corpus: textos no académicos, autores no profesionales

Area	**Fuente**	**Nº palabras**
Control	Artículos periodísticos	2.659.164

TOTAL	2.659.164

Cuadro 2d: Corpus: textos no académicos, autores profesionales

2.3. Procesamiento del corpus

Se realizó un proceso de etiquetado con el software MS-Tools, desarrollado por la Universitat Politècnica de Catalunya y la Universitat de Barcelona, utilizando los parámetros sp -S -G none (Carreras y Padró, 2002). Este proceso permitió aplicar 355 etiquetas al corpus, las que fueron disgregadas empleando un script ad hoc que posibilitó el análisis individual de los rasgos gramaticales. Las ocurrencias de formas (tokens) y rasgos gramaticales se cuantificaron meditante el software kfNgram[11], seleccionando 1-gramas con un piso de 1. Se efectuó este procedimiento en cada uno de los documentos estudiados y se trasladaron los resultados a una hoja de cálculo para su procesamiento estadístico posterior. La información se representó ordenada por ocurrencia absoluta en el documento, ocurrencia relativa respecto del total de palabras del documento, y ocurrencia relativa cada mil palabras.

2.4. Análisis de variación de verbos y sustantivos

Se cuantificó la ocurrencia de las categorías léxico-gramaticales sustantivo común (SC) y verbo (V) en términos de frecuencia relativa por texto (T), por fuente (F) y por agrupación (G). Posteriormente, se compararon las categorías tanto por fuente como por agrupación y se determinó, en cada caso, el porcentaje en que cada una de ellas contribuía al total de ambas (SC + V). No se consideró pertinente realizar un análisis específico por área disciplinaria. Los resultados fueron, en cada caso, sometidos al test ANOVA con el fin de establecer su significatividad estadística.

3. Resultados

3.1. Análisis por fuente

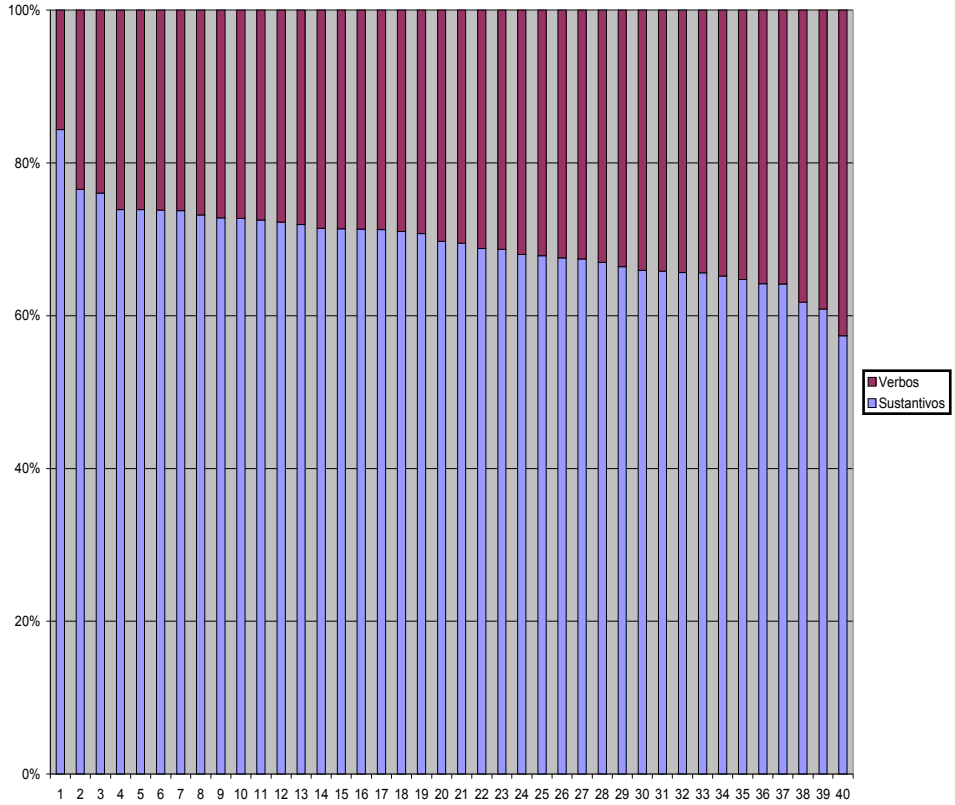
El Gráfico 1 presenta una comparación, por fuente, del porcentaje de participación de sustantivos comunes y verbos en el universo SC + V. Las distintas columnas corresponden a distintas fuentes textuales. El segmento superior de cada columna (granate) indica el porcentaje relativo de verbos en la fuente textual, mientras que el segmento inferior (azul) indica el porcentaje relativo de sustantivos. Las columnas se ordenan en una gradiente de mayor a menor porcentaje relativo de sustantivos en las fuentes textuales[2].

---

[11] Los autores agradecen a William H. Fletcher (fletcher@usna.edu) de la United States Naval Academy por su gentileza en permitir el uso de este programa y por desarrollar una versión especial del software para cumplir con los requisitos del presente estudio.

[12] Los tipos de fuentes textuales correspondientes a las columnas son los siguientes: 1, estudiantes de ciencias aplicadas (ECA); 2-4, profesionales en ciencias naturales (PCN); 5, profesionales de ciencias aplicadas (PCA); 6, PCN; 7, PCA; 8-12, PCN; 13, estudiantes de ciencias naturales (ECN); 14, PCN; 15, ECN; 16, estudiantes de ciencias sociales (ECS); 17-21, PCN; 22-23, profesionales de ciencias sociales (PCS); 24, ECS; 25, profesionales de humanidades (PH); 26, PCS; 27, ECS; 28, ECA; 29, estudiantes de humanidades (EH); 30, ECN; 31, PH; 32-33, PCS; 34, PH; 35, PCN; 36, PH; 37-38, PCS; 39, PH; 40, PCS; 41, cartas publicadas en periódicos, y 42, noticias periodísticas de diario.

Gráfico 1. Comparación de sustantivos con verbos por fuente



Como se advierte, la gradiente que ordena las fuentes de acuerdo con el mayor o menor aporte de los sustantivos comunes al total se correlaciona de manera bastante sistemática con el área disciplinaria de la fuente, si bien existen algunos traslapos. En efecto, el porcentaje de sustantivos comunes decrece progresivamente desde las áreas de ciencias aplicadas y naturales (donde llega a sobrepasar el 80%) hasta los textos control correspondientes a cartas y escritos periodísticos (donde incluso desciende del 60%), pasando por las ciencias sociales y las humanidades, que se ubican con porcentajes intermedios en la secuencia. En general, aunque no de manera estricta, existe cierta tendencia a que los textos profesionales de cada área estén mejor rankeados que los no profesionales. Las veintiún primeras fuentes, con excepción de la 16, corresponden a textos de agricultura, medicina, geología-oceanografía, nutrición, biología-zoología-botánica y química, esto es, textos de ciencias aplicadas y naturales. Si bien el primer texto de una fuente de ciencias sociales, específicamente de economía, aparece en el lugar 16, los textos de esta área disciplinaria tienden a ocurrir desde el número 22 en adelante, alternándose con los de humanidades, que empiezan a aparecer a partir del lugar 25. Dos fuentes de ciencias naturales aparecen situadas en lugares bastante bajos en la lista. En un caso, posición 30, se

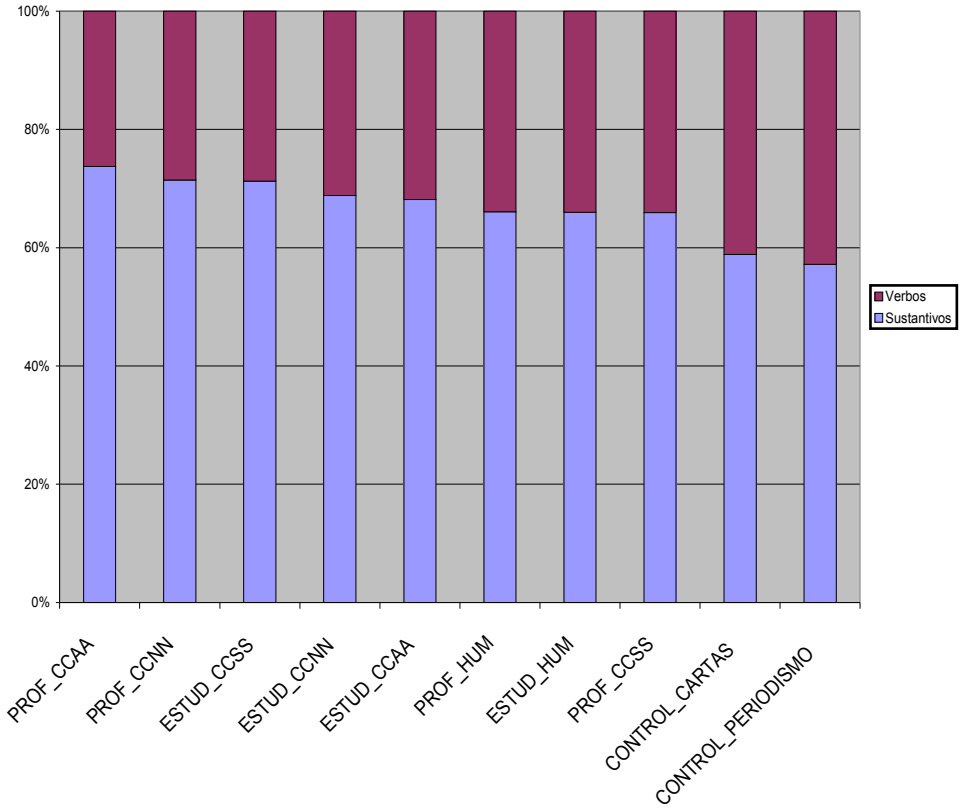
trata de tesis de estudiantes de biología y en el otro, lugar 35, de textos extraídos de la revista Biological Research. Finalmente, los lugares 40 y 41 son ocupados por cartas al director y textos periodísticos.

Considerando que la ubicación en la lista de los textos de la fuente Biological Research se aleja enormemente de la tendencia observada en los textos profesionales de ciencias naturales, se realizó un análisis manual de esta revista. Este arrojó que la gran mayoría de los artículos de investigación científica publicados en ella están escritos en inglés, lo que significa que no fueron recogidos por el megacorpus Codicach. Los textos incluidos en el megacorpus corresponden, principalmente, a la sección editorial, de entrevistas y de noticias de interés para los investigadores, es decir, textos de naturaleza periodística.

Por otra parte, cabe destacar que, específicamente en el caso de los textos de ciencias sociales, se altera dramáticamente la tendencia general según la cual los textos de profesionales presentan un índice de nominalización mayor que los de estudiantes. En efecto, los escritos mejor rankeados en este campo disciplinario corresponden a tesis de estudiantes de economía (N 16), y los más bajos en el ranking, a artículos de investigación de profesionales de sociología y antropología (N 40). Una interpretación posible de este fenómeno es que la categorización que se empleó como entrada para el análisis haya sido muy gruesa. De hecho, los trabajos de lingüística y filología, y análisis del discurso se incluyeron en este campo, pudiendo haber sido considerados ejemplares del campo de las humanidades. Esta explicación, sin embargo, no da cuenta del hecho de que el mismo fenómeno se advierte tanto en economía como en lingüística y filología. Se requieren investigaciones ulteriores para poder explicar esta distribución. Una hipótesis que podría ser sondeada es que las tesis de estudiantes de economía tengan un fuerte componente cuantitativo o teórico, mientras que los artículos de investigación se orienten a dominios más cualitativos, como el de la economía política, en otras palabras, que existan distinciones más finas al interior de las disciplinas. En el caso de los trabajos de lingüística y filología, podría explorarse la posibilidad de que las tesis de estudiantes desarrollen indagaciones lingüísticas propiamente tales, mientras que las revistas especializadas incluyan trabajos filológicos de naturaleza más humanista.

Anova
F 33,48729

Gráfico 2. Comparación de sustantivos con verbos por agrupación



3.2. Análisis por agrupación mayor

En el Gráfico 2 se presenta la comparación, por agrupación mayor, del porcentaje de participación de sustantivos comunes y verbos en el universo SC + V.

Sig. 3,89E-58

Tal y como en el caso anterior, se observa una gradiente de mayor a menor índice de nominalización, que, partiendo de los textos profesionales en ciencias aplicadas, concluye en los textos de control (cartas y trabajos periodísticos). Esta gradiente confirma, en términos generales, la ordenación presente en el Gráfico 1. Claramente, es posible establecer una distinción entre los textos de investigación, donde el índice de nominalización es siempre superior al 60%, y los escritos de control, donde este se ubica bajo dicho valor. Más compleja resulta la interpretación de los resultados en el dominio de los textos de investigación, donde pueden distinguirse tres situaciones. En primer término, los artículos de profesionales en ciencias aplicadas presentan el mayor índice de nominalización (73,8%), seguidos de los de ciencias naturales (71,4%), orden que se invierte, aunque por escaso margen, en los estudiantes, donde los de ciencias naturales

tienen un índice mínimamente superior (68,8%) al de los de ciencias aplicadas (68,2%). Este resultado es, en general, convergente con la hipótesis de nuestro trabajo, en cuanto los textos de ciencias más "duras" y, entre estos, los de profesionales presentan mayor índice de nominalización. Distinto es el caso de los de ciencias sociales. Si bien estos se ubican en una posición más baja que la de los textos de ciencias aplicadas y naturales escritos por expertos, tal y como se discutió en la sección anterior en este caso los estudiantes tienen un índice global más alto que los profesionales (71,3% contra 66%). Todavía más, las tesis en ciencias sociales presentan el mayor índice de nominalización entre los trabajos de estudiantes, fenómeno que aún no podemos interpretar. Finalmente, los textos de humanidades replican, en el otro extremo de la gradiente de trabajos de investigación, y de modo más tenue, la situación observada en ciencias aplicadas y naturales. En efecto, los trabajos de profesionales de humanidades presentan un índice más alto (66,1%) que los de estudiantes (66%). Esta diferencia, con todo, es bastante baja, lo que se explica, posiblemente, por la distribución heterogénea de los trabajos de profesionales en este campo, tema discutido en la sección anterior.

4. Discusión

Los resultados de la presente indagación muestran que la relación entre sustantivos y verbos permite distinguir, en español, entre, por una parte, textos académicos y científicos, y, por otra, textos de control del campo periodístico. En efecto, tal y como muestran tanto el Gráfico 1 como el Gráfico 2, las cartas al director de medios de comunicación y los trabajos periodísticos de medios de comunicación de masas se ubican en la posición final de una gradiente que ordena los textos de mayor a menor índice de nominalización. Por el contrario, los trabajos académicos y científicos se disponen en niveles más altos de la gradiente. Este resultado converge con la amplia literatura especializada, sobre todo en inglés, aunque también en español, que afirma el carácter nominalizado del discurso académico y científico.

Tal y como se precisó en la introducción, varias interpretaciones son consistentes con este tipo de resultados. Si bien el presente análisis es independiente y, en estricto sentido, agnóstico respecto de ellas, es posible sugerir un marco interpretativo global en que los resultados de esta investigación, y los de otras de corte cualitativo, adquieren un sentido, a nuestro juicio, unitario y culturalmente significativo. En efecto, aun cuando existen importantes diferencias entre las interpretaciones propuestas por los diversos autores, en términos generales estas son compatibles con la noción de que el mayor empleo relativo de sustantivos se relacionaría con la construcción de un discurso despersonalizado centrado en los objetos, en que los avatares de la vida mental de los actores, y, por tanto, el contexto sociocultural y emocional de las acciones por ellos ejecutadas, serían irrelevantes para la interpretación de las relaciones de coherencia del discurso. Ciertamente, la viabilidad de esta interpretación supone, como hemos señalado con anterioridad, análisis cualitativos específicos que muestren la manera en que esto efectivamente ocurriría; no obstante, más allá de los estudios mencionados en la Introducción de este trabajo, existen algunas bases que apoyan la presente propuesta. En primer lugar, el estudio de retórica histórica de Bazerman (1988) muestra cómo ciertos discursos científicos de la física fueron evolucionando desde un formato narrativo a uno expositivo-argumentativo, en respuesta a presiones sociocomunicativas y en pos de una organización que centrara la discusión en el

aparato teórico y la experimentación presentados en el escrito, dejando en segundo plano a los actores; en segundo término, el trabajo de Soto (2005a), tras analizar la sección de método de artículos de ciencias naturales y sociales, muestra cómo el proceso de detransitivización y esquematización del agente, manifiesto, entre otros recursos, en el amplio empleo de cláusulas pasivas-reflejas e impersonales-reflejas, recibe una caracterización significativa al proponer un esquema de desfocalización del agente en esta sección de los artículos.

Los dos estudios mencionados pueden relacionarse, hasta cierto punto, con la distinción planteada por Bruner (1986), y ya indicada en este trabajo, entre un modo de pensamiento narrativo y otro paradigmático. Para Bruner, la distinción entre las dos formas de pensamiento radica en que, mientras la primera se centra en relaciones humanas donde priman las causaciones mentalistas contextualmente situadas, la segunda privilegia las relaciones conceptuales y la causación física. Nuestra propuesta, más desarrollada en Soto (2005a), es que los artículos de investigación científica se organizarían de modo de potenciar, en el proceso de comprensión, lo que Bruner ha denominado pensamiento paradigmático. Es posible que esta forma de organización descanse en una concepción subyacente en los científicos, de acuerdo con la cual esta sería la manera adecuada de comprender y evaluar las proposiciones en ciencias. En los textos de control, por el contrario, habría una mayor presencia del formato narrativo toda vez que ellos tratarían de eventos y problemáticas cotidianos en que las acciones humanas y sus motivaciones sociales e individuales jugarían un rol destacado.

La interpretación hasta aquí avanzada sugiere que el índice de nominalización también debería ser un buen predictor de la diferencia entre ciencias naturales y ciencias humanas y humanidades, en el sentido de que las primeras deberían tener un mayor índice relativo de sustantivos que las segundas. Esta sugerencia descansa en el supuesto de que las ciencias humanas y humanidades no pueden prescindir completamente de un nivel de explicación mentalista y contextualmente situado mientras que las ciencias naturales no precisan de este nivel para caracterizar los fenómenos en estudio. En términos de Dennett (1987), las ciencias de la naturaleza operarían en el nivel de la actitud física mientras que las humanas agregarían a este el de la actitud intencional, es decir, mientras las primeras se preocuparían de dar cuenta de los fenómenos del mundo físico operando con relaciones causales y de propiedad de naturaleza física (sea cual fuere la manera en que estas propiedades y relaciones se definan de modo específico), las segundas se preocuparían tanto de fenómenos físicos como sociales y mentales, y operarían tanto con relaciones físicas como sociales y mentales. En otras palabras, la naturaleza del problema enfrentado por ambos campos disciplinarios y la manera en que este se encara se reflejaría en el índice de nominalización. Los resultados del presente trabajo son en gran medida compatibles con esta interpretación, toda vez que, tanto en el Gráfico 1 como en el 2, la gradiente ubica en el extremo de mayor índice de nominalización los textos de profesionales de ciencias naturales y aplicadas, y en el extremo de menor índice entre los escritos académicos, los de humanidades y ciencias sociales.

En lo que dice relación con el contraste entre textos de profesionales y de estudiantes, los resultados muestran que, al menos en ciencias naturales y aplicadas, la hipótesis de la investigación es correcta, esto es, los de profesionales presentan mayor índice de nominalización que los de estudiantes. El marco interpretativo propuesto sugiere que lo anterior podría relacionarse con el hecho de que los estudiantes no dominan por completo

las propiedades de esta variedad de discurso, en otras palabras, no construyen un discurso tan centrado en objetos y relaciones de propiedad y causación físicas como el producido por los profesionales[13]. La no diferenciación en el caso de los textos de profesionales y estudiantes de humanidades podría relacionarse, por su parte, con la mayor proximidad relativa que este campo tiene con el formato narrativo.

Con todo, no resulta claro el que, entre los textos de estudiantes, los de ciencias sociales se ubiquen en un lugar más alto del ranking que los de ciencias naturales y aplicadas, toda vez que nuestro marco interpretativo sugiere para los trabajos de los primeros un índice más bajo que el de los segundos. Específicamente, y como ya se indicó en la sección de resultados, el índice de nominalización en las tesis de estudiantes de ciencias sociales es inesperadamente alto. En efecto, estas presentan un índice mayor no solo al de los otros trabajos de estudiantes sino que, incluso, al de los de profesionales de ciencias sociales. Estos últimos, por su parte, son prácticamente iguales a los de los artículos de humanidades, lo que no va en contra de la interpretación sugerida. Si bien en la sección anterior se esbozaron algunas posibles explicaciones de este fenómeno, es claro que este requiere de una indagación específica de corte cualitativo.

Más allá de las limitaciones expuestas, pensamos que este trabajo muestra, de modo convincente, que el índice de nominalización es sensible a la oposición entre, por una parte, textos científicos y académicos y, por otra, escritos periodísticos y cartas a medios de comunicación. Asimismo, se relaciona directamente con una escala descendente que, partiendo con los textos de ciencias aplicadas, sigue con los de ciencias naturales, de ciencias sociales y de humanidades, aunque con importantes traslapos entre estos últimos. En cuanto al contraste entre expertos y novatos, el estudio muestra que el índice es sensible solo en el caso de las ciencias aplicadas y naturales. Junto con ello, los resultados son concordantes con el marco interpretativo sugerido en este y en otros trabajos de nuestro equipo. Si bien investigaciones posteriores con corpora aun más diversificados podrían debilitar algunas de las interpretaciones de este trabajo, ampliando, por ejemplo, el alcance del índice de nominalización más allá del dominio académico y científico, la noción misma de índice de nominalización –que, hasta donde nuestro conocimiento alcanza, no había sido propuesta hasta ahora—, parece ser lo suficientemente simple y poderosa como para ser de utilidad en el análisis computarizado de géneros textuales. Además de los estudios ya sugeridos, tanto los resultados como el marco interpretativo de esta investigación permiten proyectar diversas indagaciones. En primer término, la incorporación de nuevas variedades textuales, como las leyes y los documentos burocráticos, permitiría precisar el alcance del índice de nominalización. Por otro lado, la indagación en recursos lingüísticos como los verbos mentales podría aportar más apoyo a la interpretación propuesta, en el sentido de que esta sugiere que dichos verbos deberían correlacionarse de manera inversa con el índice de nominalización. También parece interesante, por último, realizar una indagación más profunda en la relación entre las creencias sobre la actividad científica y la expresión lingüística de los discursos, especialmente en el campo de las ciencias humanas. A partir de lo ya expuesto, puede especularse que textos de autores que siguen programas de investigación vinculados más fuertemente con concepciones de las ciencias naturales, como el conductismo, tendrían un mayor índice de nominalización que trabajos de autores menos asociados con estas concepciones.

[13] Subyace a esta interpretación el supuesto de que el formato narrativo tendría un estatus más básico que el paradigmático, idea de aceptación generalizada (cfr. Bocaz y Soto 2000).

REFERENCIAS BIBLIOGRÁFICAS

AARTS, Bas., DENISON, David, KEIZER Evelien y POPOVA, Gergana (eds.), *Fuzzy grammar. A reader*, Oxford, Oxford University Press, 2004.

ALBENTOSA, José Ignacio y MOYA, Arsenio Jesús, "La reducción del grado de transitividad de la oración en el discurso científico en lengua inglesa", *Revista Española de Lingüística*, 30 (2), pp. 445-468, 2000.

ANDERSON, RICHARD, "'Look at all those nouns in a row!' –Authoritarianism and the iconicity of political Russian", *Political Communication*, 13, pp. 145-164, 1996.

BAJTÍN, Mijail, "The problem of speech genres", *Speech genres and other late essays*, Austin, University of Texas Press, 1986 (original: "Problema rechevykh zhanrov", en *Estetika slovesnogo tvorchestva*, Moscú, Iskusstvo, 1979).

BAZERMAN, Charles, *Shaping written knowledge. The genre and activity of experimental article in science*, Madison, WI, University of Wisconsin Press, 1988.

BIBER, Douglas, *Variation across speech and writing*, Cambridge, Cambridge University Press, 1988.

BOCAZ, Aura y SOTO, Guillermo, "Narrar y exponer: el tratamiento del discurso en la reforma educacional", *El Mercurio*, suplemento *Artes y Letras*, 26 de noviembre, pp. 20-21, 2000.

BRUNER, Jerome, *Actual minds, posible worlds*, Cambridge, Mass., Harvard University Press, 1986.

CALSAMIGLIA, Helena y TUSON, Amparo, *Las cosas del decir. Manual de análisis del discurso*, Barcelona, Ariel, 1999.

CARRERAS, Xavier y PADRÓ, Luis, "A flexible distributed architecture for natural language analyzers", *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, LREC, 2002.

CHAFE, Wallace y TANNEN, Deborah, "The relation between written and spoken language", *Annual Review of Anthropology*, 16, pp. 383-407, 1987.

CIAPUSCIO, Guiomar, *Tipos textuales*, Buenos Aires, Universidad de Buenos Aires, 1997.

DANEŠ, Frantisek, "A three-level approach to syntax", *Travaux linguistiques de Prague*, 1, "L'école de Prague d'aujourd'hui", pp. 225-240, 1966.

DE BEAUGRANDE, Robert y DRESSLER, Wolfgang, *Introducción a la lingüística del texto*, Barcelona, Ariel, 1997. (Traducido de la versión inglesa: *Introduction to text linguistics*, Londres, Longman, 1981.)

DENNETT, Daniel, *The intentional stance*, Cambridge, Mass., MIT Press, 1987.

GROSS, Alan y HARMON, Joseph "What's right about scientific writing", *The Scientist*, 13 (24), p. 20, 1999.

GUTIÉRREZ RODILLA, Bertha M., *La ciencia empieza en la palabra. Análisis e historia del lenguaje científico*, Barcelona, Península, 1998.

HALLIDAY, Michael, KIRKWOOD Alexander y MARTIN, James, *Writing science: Literacy and discursive power*, Londres, Falmer Press, 1993.

HYLAND, Ken, *Teaching and Researching Writing*, Londres, Longman/Pearson, 2002.

HEYLIGHEN, Francis y DEWAELE, Jean-Marc "Variation in the contextuality of language: An empirical measure", *Foundations of Science*, 7, pp. 293-340, 2002.

LAKOFF, George, *Women, fire, and dangerous things: What categories reveal about the mind*, Chicago, University of Chicago, 1987.

LANGACKER, Ronald, "Nouns and Verbs", *Language*, 63, pp. 53-94, 1987.

LUDWIG, Ralph, "L'oralité des langues créoles – agrégation et integration". En *Les créoles français entre l'oral et l'écrit*, Tübingen, Gunter Narr Verlag, 1989.

MACDONALD, Susan Peck, "Prose styles, genres, and levels of analysis", *Style*, 36 (4), pp. 618-639, 2002.

MARGOLIS, Eric y LAURENCE, Stephen (eds.), *Concepts: Core readings*, Cambridge, Mass., MIT Press, 1999.

MILLER, Carolyn, "Genre as social action", *Quarterly Journal of Speech*, 70, pp. 151-167, 1984.

PARODI, Giovanni, "Lingüística de corpus y análisis multidimensional: exploración de la variación en el corpus PUCV-2003". En G. Parodi (ed.), *Discurso especializado e instituciones formadoras*, Valparaíso, Ediciones Universitarias de Valparaíso, 2005.

SARDINHA, Tony Berber, "Análise multidimensional", *Delta*, 16 (1), pp. 99-127, 2000.

SOTO, Guillermo, "Una propuesta de caracterización del artículo científico", ponencia presentada en el *III Encuentro Nacional de Análisis del Discurso*, Valdivia, Universidad Austral de Chile, 28 de septiembre a 1 de diciembre, 2004.

SOTO, Guillermo, "Construcciones de agente degradado en la sección método de los artículos científicos". En A. M. Harvey (comp.), *En torno al discurso: estudios y perspectivas*, Santiago, Ediciones de la Universidad Católica de Chile, 2005a.

SOTO, GUILLERMO, "Los géneros discursivos como tecnologías cognitivas", *RASAL, Revista de la Sociedad Argentina de Lingüística*, 1, pp. 37-51, 2005b.

SOTO, Guillermo, Scott SADOWSKY y Ricardo MARTÍNEZ, "Análisis factorial de rasgos gramáticales en artículos de investigación en ciencias naturales, aplicadas y humanas en escritos en español", en preparación.

SOTO, Guillermo y Carlos ZENTENO, "Los sintagmas nominales en textos científicos escritos en español", *Estudios de Lingüística* (Universidad de Alicante), nº 18, pp. 275-292, 2004.

SWALES, James, *Genre análisis: English in academic and research settings,* Cambridge, Cambridge University Press, 1990.

TANNEN, Deborah, "The myth of orality and literacy", en W. Frawley (ed.), *Linguistics and literacy*, Nueva York – Londres, Plenum Press, 1982a.

TANNEN, DEBORAH, "Oral and written strategies in spoken and written narratives", *Language*, 58 (1), pp. 1-21, 1982b.

TOMLIN, Russell S., FORREST, Linda Ming MING PU y Myung HEE KIM, "Semántica del discurso". En T. van Dijk (comp.), *El discurso como estructura y proceso*, Barcelona, Gedisa, 2000. (original: "Discourse semantics", en T. Van Dijk (comp.), *Discourse as structure and process*, Londres, Sage, 1997.)

VAN DIJK, Teun, "El estudio del discurso", en T. van Dijk (comp.), *El discurso como estructura y proceso*, Barcelona, Gedisa, 2000. (original: "The study of discourse", en T. Van Dijk (comp.), *Discourse as structure and process*, Londres, Sage, 1997.)

WEIGAND, EDDA, "Possibilities and Limitations of Corpus Linguistics". En K. Aijmer (ed.), *Understanding and Misunderstanding in Dialogue. Selected papers from the 8th international conference on Dialogue Analysis*, Tübingen, Niemeyer (Beiträge zur Dialogforschung), 2004.

WOLTERS, Maria y Mathias KIRSTEN, "Exploring the use of linguistic features in domain and genre classification", *Proceedings of the Meeting of the European Chapter of the Association forComputational Linguistics*, Bergen, Noruega, 1999.

ZENTENO, Carlos, "La tendencia entificatoria en el discurso del especialista", en M. Rodríguez y M. A. Farías (eds.) *Investigación multidisciplinaria. Estrategias integradas de investigación en lingüística, literatura y disciplinas afines*, Santiago, Universidad de Santiago de Chile, 1997.