LINGÜÍSTICA DE CORPUS Y ANÁLISIS MULTIDIMENSIONAL: EXPLORACIÓN DE LA VARIACIÓN EN EL CORPUS PUCV-2003 *

GIOVANNI PARODI
Pontificia Universidad Católica de Valparaíso

INTRODUCCIÓN

Los corpus de textos naturales computarizados han mostrado un importante impacto en los análisis lingüísticos en las últimas dos o tres décadas. En particular, las investigaciones acerca de la lengua inglesa, algunas europeas y ciertas asiáticas han revelado que los estudios lingüísticos basados

* Este trabajo se basa en el análisis comparativo de una parte de los corpora de textos naturales incluidos el denominado *Corpus El Grial PUCV-2003*, compilado por el equipo de investigadores en lingüística de corpus que coordino en la Pontificia Universidad Católica de Valparaíso, Chile. Dentro del conjunto de corpus contenidos en *El Grial,* hemos seleccionado el Corpus PUCV-2003 compuesto por un Corpus Técnico-científico (CTC), un Corpus de Entrevistas Orales (CEO) y un Corpus de Literatura Latinoamericana (CLL), por representar una muestra de textos prototípicamente escritos (especializados y no especializados) y orales interactivos no especializados. *El Grial,* además de ser el nombre del conjunto de *corpora,* también constituye el sitio *web* que aloja estos textos y a través del cual es factible acceder a ellos digitalmente y realizar en línea y gratuitamente una serie de consultas. Para ello v. www.elgrial.cl.

En términos metodológicos, en esta investigación se lleva a cabo un tipo de análisis de datos lingüísticos y de registros discursivos que solo cuenta con antecedentes en la indagación de la lengua inglesa. Nos referimos al análisis multirasgos (AMR), en el que se exploran un amplio conjunto de rasgos gramaticales y de orden lingüístico (65 en este caso) y se cuantifica su ocurrencia en cada texto, y luego, con propósitos comparativos, se normalizan dichas cifras. Junto a este, también opera en el análisis multiregistros (AMR), el cual consiste en comparar por medio de procedimientos estadísticos (análisis factorial) la ocurrencia sistemática de algunos de estos conjuntos de rasgos en determinadas variedades de textos que comparten aspectos lingüísticos, cognitivos y sociales: oral, escrito, especializado, no-especializado.

Revista Española de Lingüística, 35, 1, 2005, págs. 45-76.

en grandes muestras de textos digitales no siempre corroboran las intuiciones iniciales de los investigadores. El uso de corpus con sustento computacional así como la disponibilidad de programas computacionales que permiten su tratamiento y posterior análisis han impulsado las indagaciones lingüísticas hacia fronteras antes imposibles.

A pesar de ello, el empleo conjunto de aproximaciones lingüísticas basadas en corpus y de enfoques computacionales con tecnología digital no ha sido una combinación habitual en el ámbito de las ciencias del lenguaje (Castel 2004a y b). Este hecho ha redundado en que un reducido número de textos y solo algunas variables sean analizadas manualmente sin aprovechar los beneficios de la digitalización de los corpus y de los robustos programas computacionales de análisis automático hoy en día disponibles.

Por otro lado, se detectan escasos o nulos estudios acerca del español que describan las características lingüísticas del discurso especializado didáctico escrito en lengua española. La mayoría de estos trabajos focaliza el llamado discurso divulgativo de especialidad (Ciapuscio 2000, 2003, Ciapuscio y Kuguel 2002, Calsamiglia 2000, Cassany, López y Martí 2000, López 2002, Cademártori 2003) o aborda los marcadores discursivos (Portolés 1998, Martín Zorraquino y Portolés 1999, Montolío 2001); otros, se concentran en algunos pocos rasgos lingüísticos (Ciapusccio 1992, Harvey 2002); los menos, estudian la terminología especializada (Cabré 1999, 2000, 2002, Lorente 2002). Resulta aún mucho más difícil encontrar investigaciones del español en que se describa rasgos lingüísticos de registros especializados a partir de textos que se entregan como lectura obligatoria a alumnos de liceos técnico-profesionales. De hecho, los estudios que comparan registros escritos en diversas áreas del conocimiento técnico-científico especializado son casi inexistentes. Como resultado, los textos de registros escolares de nivel especializado han recibido sorprendemente poca atención (Menéndez 1999, Parodi y Gramajo 2003, Parodi 2004 y 2005) o han sido virtualmente ignorados. Investigaciones en esta línea y que sigan principios de lingüística de corpus y utilicen apoyo computacional y estadístico también resultan singularmente poco frecuentes.

Ahora bien, en el presente artículo se indaga la variabilidad lingüística y funcional en el interior del corpus PUCV-2003 (90 textos con un total de 1.466.744 palabras). Con el objetivo final de describir acuciosamente un subcorpus técnico-científico (PUCV-CTC), se realiza un análisis multirasgos y multidimensiones con apoyo de técnicas estadísticas multivariadas (Análisis Factorial de Componentes Principales). A partir de un total de 65

rasgos lingüísticos sobresalientes, explorados en el total del corpus, se busca determinar —sobre la base de la alta co-ocurrencia de patrones de rasgos significativos— las «dimensiones» subyacentes a dichos rasgos.

De modo sintético, los objetivos de esta investigación son: (a) identificar los patrones lingüísticos relevantes y en co-ocurrencia en el total del corpus PUCV-2003, desde una perspectiva empírica cuantitativa, (b) comparar sistemáticamente tres *corpora:* uno especializado técnico-científico escrito, otro no especializado literario escrito y un tercero oral no especializado de entrevistas; todo ello a partir de dimensiones funcionales, determinadas a base de aquellos patrones lingüísticos de co-ocurrencia, y (c) identificar similitudes y diferencias —en términos dimensionales— entre tres áreas de especialización del subcorpus técnico-científico.

En la primera parte del trabajo se reúnen antecedentes relevantes para el tema del análisis multirasgos y multidimensiones. En el marco metodológico de la investigación, se detallan los *corpora,* los rasgos lingüísticos explorados, los procedimientos de etiquetaje morfológico automático de los textos y de técnicas estadísticas ejecutadas. Posteriormente, se presentan los resultados, entre los que destacan los cinco factores de agrupamientos de rasgos, interpretados en cinco dimensiones relevantes y factibles de análisis funcional. La dimensión denominada «Foco Informacional» resulta ser la que —estadísticamente— mejor da cuenta del corpus técnico-científico escrito.

I. Marco de Referencia

Biber 1994 defiende dos ideas significativas respecto a la representatividad de los estudios lingüísticos basados en corpus que me parecen relevantes: 1) todo corpus debe tener una amplitud importante, y 2) un corpus debe contener registros o categorías textuales diversificadas. Ello conlleva mayor validez de las conclusiones y permite la comparación y la generalización. La primera de estas ideas ha sido destacada en la bibliografía acerca de lingüística de corpus en los últimos veinte años (Francis 1979, Leech 1991, 1992, Sinclair 1982, 1991, Johansson 1991; Stubbs 1996). La segunda resulta ser más original y constituye una de sus importantes aportaciones en torno a la cual Biber ha desarrollado sus investigaciones en diversas lenguas: la variación lingüística a través de diferentes registros orales y escritos.

Biber 1988 y 2003 y Biber, Conrad y Reppen 1998 han dado cuenta de interesantes variaciones sistemáticas de orden gramatical y léxico en diversos registros del inglés oral y escrito. Dos hallazgos entre muchos de los presentados parecen relevantes: por un lado, los rasgos lingüísticos individuales presentan una ocurrencia diversa en variados registros; por otro, los mismos o similares rasgos lingüísticos pueden tener funciones diferentes en textos pertenecientes a diversos registros. En este sentido, una de las ventajas de este enfoque metodológico se funda en un principio lingüístico comunicativo que resulta extremadamente sensato: la variación entre registros no se explica únicamente por un solo parámetro o dimensión, lo que equivale a sostener que existen múltiples distinciones situacionales entre registros. Dicho de otro modo, no es posible pensar que un rasgo lingüístico o, incluso, unos pocos de ellos puedan explicar exclusivamente una determinada variación entre registros (por ejemplo: oral/escrito, formal/informal). Las investigaciones en que se ha aplicado el análisis multivariado han revelado que diferentes dimensiones se construyen a partir de conjuntos diferentes de rasgos lingüísticos co-ocurrentes, reflejando así diversas interpretaciones funcionales subyacentes (por ejemplo: objetividad, abstracción de información, modalización). Del mismo modo, las tradicionales distinciones de índole más dicotómica (interactivo/no-interactivo), se ven desafiadas por los estudios con análisis multidimensional ya que se ha llegado a demostrar que existe un *continuum* de variación lingüística a lo largo de los registros. Por supuesto, esto último es concordante con las investigaciones que se adscriben a la idea de categorías de límites difusos *(fuzzy categories)* y que hoy en día tienen gran aceptación entre la comunidad científica (Lakoff 1972).

Un supuesto teórico fundamental del enfoque multivariado lo constituye el principio de que la co-ocurrencia de rasgos lingüísticos (determinada mediante procedimientos estadísticos) refleja funciones comunicativas compartidas, es decir, que estos patrones de co-ocurrencia de rasgos se interpretan en términos de funciones situacionales, sociales y cognitivas comunes. En otras palabras, los rasgos lingüísticos co-ocurren en determinados textos porque ellos muestran funciones compartidas específicas. Por ejemplo, las oraciones pasivas, las nominalizaciones y los verboides se relacionan todos con «informatividad» y «focalización del objeto». Del mismo modo que los pronombres de primera y segunda persona singular, el tiempo presente, el modo imperativo y los adverbios de lugar y de tiempo están ligados a la «interactividad».

En este sentido, se asume que un grupo de rasgos co-ocurre frecuentemente en ciertos textos porque ellos son usados para expresar un conjunto

determinado de funciones comunicativas; esto es, no pueden determinarse *a priori* estas funciones. Por ello, resulta crucial —inicialmente— contar con un análisis individual de rasgos en términos funcionales, pues desde allí se establecen los fundamentos para determinar las interpretaciones funcionales subyacentes al conjunto de rasgos co-ocurrentes.

Tal como se adelantó en la introducción, en esta investigación seguimos —en líneas generales— la propuesta metodológica de Biber 1988 respecto del Análisis Multirasgos (AMR) y Multidimensiones (AMD). Ello, dado que tal enfoque fue creado originalmente como un método analítico para el estudio detallado de las variaciones entre registros, como se aprecia en nuestro caso, nos resulta de utilidad para describir y comparar los textos de los tres córpora. Este enfoque metodológico fue inicialmente desarrollado, según Biber 1988, 1994 para (1) determinar los patrones lingüísticos sobresalientes y en co-ocurrencia en una lengua, desde una perspectiva empírica cuantitativa; y (2) comparar registros orales y escritos en un espacio lingüístico definido por aquellos patrones en co-ocurrencia.

El AMR y AMD utilizan las herramientas metodológicas de la lingüística de corpus según las últimas tendencias (Sinclair 1991, Leech 1991, Svartvik 1992), esto es, requiere de un diseño de corpus representativo, programas computacionales automáticos para etiquetar morfosintácticamente los textos y programas para analizar y cuantificar la ocurrencia de los rasgos lingüísticos previamente marcados en el corpus. Biber, Reppen, Clark y Walter 2001 proponen cuatro ventajas importantes para adoptar una aproximación basada en lingüística de corpus. Parafraseándolas, ellas pueden resumirse como sigue:

1. La adecuada representación del discurso en su forma de ocurrencia natural, a través de muestras amplias y representativas compiladas a partir de textos originales.
2. El procesamiento lingüístico (semi)automático de los textos mediante el uso de ordenadores, que permiten un análisis mucho más amplio y profundo de los textos mediante un vasto conjunto de rasgos lingüísticos caracterizadores.
3. Mucha mayor confiabilidad y certeza en los análisis cuantitativos de los rasgos lingüísticos de grandes muestras de textos.
4. La posibilidad de resultados acumulativos y replicables. Posteriores investigaciones pueden utilizar los mismos *corpora* u otros *corpora* pueden ser analizados con las mismas herramientas computacionales.

II. Metodología

2.1. *Los objetivos de la investigación y los* corpora

En esta investigación se pretende determinar estadísticamente —mediante análisis factorial— los patrones lingüísticos sobresalientes y co-ocurrentes en el corpus PUCV-2003 y realizar un estudio comparativo a partir de los tres diferentes grupos de textos recolectados, según las dimensiones a determinar e interpretar funcionalmente. Por último, a través del enfoque multidimensional, se efectúa una comparación de las tres áreas técnicas del CTC. Para ello se hace uso de programas computacionales que apoyan la labor de etiquetaje y de interrogación de los textos. También se utilizan técnicas estadísticas para definir los factores y comparar los resultados estadísticos.

La conformación general de Corpus PUCV-2003 se desglosa en 90 textos que equivalen a 1.466.744 palabras. Este corpus general está dividido en tres registros o subcórpora (Corpus Técnico-Científico -CTC, Corpus de Literatura Latinoamericana -CLL, y Corpus de Entrevistas Orales -CEO). Inicialmente se recolectó el CTC y, posteriormente, con el objetivo de llevar a cabo procedimientos comparativos entre diversos registros que brinden una profunda y certera descripción del CTC y cumpliendo procedimientos de rigurosidad en el marco de la lingüística de corpus, se recolectaron otros dos *corpora,* a saber, el CLL y el CEO. La siguiente tabla muestra su distribución por número de textos y palabras.

Tipo de Corpus	Número de archivos o textos	Total de Palabras
Corpus PUCV-CTC	74 (82%)	626.790 (42%)
Corpus PUCV-CLL	12 (13%)	459.860 (32%)
Corpus PUCV-CEO	04 (5%)	380.094 (26%)
Totales	90 (100%)	1.466.744 (100%)

Tabla 1. Constitución Global del Corpus PUCV-2003

2.1.1. *Corpus de textos Técnico-Científicos (CTC)*

Tal como se muestra en la Tabla 2, el Corpus Técnico-Científico (CTC) está compuesto por setenta y cuatro textos con un total de 626.790 palabras, recolectado en establecimientos secundarios técnico-profesionales de la

ciudad de Valparaíso, Chile, en tres áreas técnicas. Estas tres diferentes áreas del conocimiento técnico especializado tienen relación con la formación de tres diferentes profesionales técnicos, a saber, sector marítimo (Especialidad Operación Portuaria), sector metalmecánico (Especialidad Mecánica Industrial), y sector de administración y comercio (Especialidad Contabilidad). Los textos recopilados corresponden a aquellos que se entregan a los alumnos como parte de lecturas obligatorias o complementarias en cada área técnica, esto es, son parte importante del acceso de estos estudiantes al conocimiento especializado.

El desglose de esta información se entrega en la Tabla 2:

Área Técnica CTC	Número de textos	Número de palabras
Marítima (Operación Portuaria)	36 (49%)	155.160 (25%)
Industrial (Mecánica)	18 (24%)	246.374 (39%)
Administración y Comercio (Contabilidad)	20 (27%)	225.256 (36%)
Totales	74 (100%)	626.790 (100%)

Tabla 2. Constitución del CTC

Como se aprecia, no existe una relación directa entre número de textos por ámbito de especialidad y número de palabras. Así, en el ámbito marítimo de operación portuaria se registra la mayor cantidad de textos (49% del total), pero el menor corpus de palabras (25% del total). Por el contrario, y de manera interesante, en el área técnica de mecánica industrial se recolectó el grupo más reducido de textos (24%), pero ellos conforman la muestra más grande respecto al número de palabras (39%). Por su parte, el área de administración y comercio (Contabilidad) arroja cifras similares a la anteriormente descrita. En ella se obtuvo un total de 20 textos (27% del total) y un número elevado de palabras (36% del total). Estas cifras revelan una cierta heterogeneidad respecto a la configuración del corpus de acuerdo a cada ámbito de especialización, y también muestran que no existe una relación directa entre área técnica y porcentaje de textos y palabras. En todo caso, el número de palabras por ámbito demuestra no ser relevante en términos estadísticos.

Mediante un análisis multiniveles se detectaron doce tipos textuales diferentes en el CTC, los cuales fueron rastreados cualitativamente a partir de los *corpora* (para un estudio detallado de esta determinación tipológica, ver Parodi y Gramajo 2003).

2.1.2. *Corpus de textos de Literatura Latinoamericana (CLL)*

La selección de los doce textos que componen el Corpus de Literatura Latinoamérica escrita se ejecutó basándose en entrevistas con los profesores de la asignatura de Lengua Castellana y Comunicación de los tres establecimientos técnico-profesionales. En ellas se les solicitó un listado de obras literarias que ellos dieran como lectura a sus alumnos de 4° año de Enseñanza Media en las tres áreas de especialización de las cuales se recogió el CTC. Luego de una comparación de los listados de textos obtenidos, se decidió —estrictamente sobre la base de un criterio de homogeneidad— construir este corpus según las obras literarias que coincidían entre los tres establecimientos educacionales. Es decir, el corpus se conformó a partir de las obras comunes para todos los alumnos. Ello derivó en este grupo de autores y las correspondientes obras. En la Tabla 3, se presenta los detalles descriptivos de cada texto junto al número parcial y total de palabras.

Obra literaria (CLL)	Número de palabras
CLL 1 (PUCV 75)	27.853 (6,5%)
CLL 2 (PUCV 76)	1.414 (0,5%)
CLL 3 (PUCV 77)	30.797 (7%)
CLL 4 (PUCV 78)	56.491 (12%)
CLL 5 (PUCV 79)	47.173 (10%)
CLL 6 (PUCV 80)	33.405 (7%)
CLL 7 (PUCV 81)	94.779 (21%)
CLL 8 (PUCV 82)	50.704 (11%)
CLL 9 (PUCV 83)	12.974 (3%)
CLL 10 (PUCV 84)	24.467 (5%)
CLL 11 (PUCV 85)	4.628 (1%)
CLL 12 (PUCV 86)	75.175 (16%)
Totales	459.860 (100%)

Tabla 3. Constitución del CLL

La cantidad de palabras de este corpus presenta un número inferior en cerca de ciento setenta mil palabras al del CTC. Este hecho no constituye en sí un problema para comparaciones ya que las cifras se utilizan normalizadas en textos de 1.000 palabras. Además, a pesar de ser un número inferior al otro corpus, su cantidad es significativa para los estándares empleados actualmente en lingüística de corpus y permite sin dificultades la aplicación

de programas computacionales estadísticos como los que se requieren para el análisis factorial.

2.1.3. *Corpus de textos Entrevistas Orales (CEO)*

Este tercer corpus está formado por dos entrevistas orales realizadas a un total de setenta y cinco alumnos de 4° año de Enseñanza Media de establecimientos técnicos y no técnico-profesionales de la ciudad de Valparaíso (educación diferenciada y no-diferenciada). La primera entrevista, de tipo entrevista en profundidad semi-dirigida, consistió en una conversación acerca de técnicas de estudio y estrategias de lectura y comprensión. La segunda entrevista se estructuró según algunas de las temáticas abordadas en la primera conversación y tuvo un carácter más abierto y menos dirigido que la primera. Las entrevistas se realizaron por alumnos y alumnas de último año de la carrera de Pedagogía en Castellano de la PUCV. Se recurrió, en parte, a este perfil de entrevistador con el objetivo de crear un ambiente de confianza y distensión en la conversación.

Solo por razones de organización interna y mejor acceso de procesamiento técnico, se decidió dividir las ciento cincuenta entrevistas en cuatro archivos computacionales. Por ello, la distribución y cuantificación de este corpus oral dialógico se presenta del siguiente modo en la Tabla 4.

Número Archivo (CEO)	Número total de palabras
CEO 1 (PUCV 87)	86.616 (22%)
CEO 2 (PUCV 88)	89.199 (24%)
CEO 3 (PUCV 89)	102.092 (27%)
CEO 4 (PUCV 90)	102.187 (27%)
Totales	380.094 (100%)

Tabla 4. Constitución del CEO

2.2. *Análisis Multirasgos y Multidimensiones: pasos metodológicos*

La realización acuciosa de un completo análisis MR y MD implica una serie de decisiones y etapas rigurosas y, en algunos casos, conlleva alta complejidad técnica computacional y dominio estadístico especializado. Biber 1988, Biber y otros 1998 y Conrad y Biber 2001 proporcionan una serie pasos metodológicos para su aplicación. Ellos pueden ser resumidos en los siguientes doce puntos nucleares:

1. Diseño, recolección, organización y digitalización del corpus.
2. Selección de un conjunto de rasgos lingüísticos sobre la base de bibliografía especializada de acuerdo con los registros involucrados que serán considerados en el análisis.
3. Caracterización funcional de los rasgos lingüísticos seleccionados.
4. Disponibilidad de programas computacionales capaces de analizar automáticamente los textos en formato plano (ASCII o txt).
5. Marcaje estructural o etiquetado morfológico y/o sintáctico de los textos del corpus.
6. Interrogación manual o (semi)automática de cada uno de los textos a partir de los rasgos en estudio para determinar su ocurrencia.
7. Construcción de bases de datos normalizadas, dado el número de palabras divergente entre los textos.
8. Aplicación, con asistencia de programas computacionales, del análisis factorial a las frecuencias de ocurrencia de los rasgos. Ello con el fin de reducir las variables involucradas y determinar patrones de co-ocurrencia entre los rasgos lingüísticos.
9. Establecimiento de un conjunto de factores (cada factor queda conformado por un conjunto de rasgos lingüísticos) mediante el análisis factorial, con algún tipo de rotación (Varimax, Cuantrimax, Oblimin, etc.).
10. Interpretación funcional de los factores, producto del análisis factorial, a partir de la co-ocurrencia de rasgos, constituyendo así una dimensión subyacente de variación.
11. Confirmación o refutación de la interpretación de los factores mediante el cálculo de los puntajes factoriales.
12. Cálculo de los puntajes de dimensión para cada texto respecto de cada una de las dimensiones. En esta fase, se comparan los puntajes de cada registro en cada dimensión y se estudian similitudes y/o diferencias lingüísticas y funcionales.

2.3. *Rasgos lingüísticos*

Respecto a los rasgos lingüísticos a indagar, se llevó a cabo inicialmente un rastreo bibliográfico con el fin de identificar categorías gramaticales representativas que mostraran relevancia funcional en español. Con esta información disponible, que reveló un cierto grado de dificultad en su detec-

ción, se construyó una matriz con un total de sesenta y cinco rasgos lingüísticos caracterizadores del español. En la Tabla 5 se presentan los sesenta y cinco rasgos, agrupados en torno a quince categorías más generales.

Rasgos Lingüísticos Proyecto Corpus PUCV-2003	
A. Marcadores de tiempo verbal	H. Formas estativas activas
1. Pretérito indefinido (indicativo)	33. Ser
2. Pretérito imperfecto (indicativo)	34. Estar
3. Pretérito perfecto (indicativo y subjuntivo)	I. Tipos verbales
	35. Públicos
4. Presente (indicativo y subjuntivo)	36. Privados
5. Futuro (indicativo y subjuntivo)	37. Persuasivos
6. Futuro perifrástico	38. Perceptivos
B. Marcadores de modo verbal	J. Verbos modales
7. Indicativo/imperativo	39. Posibilidad
8. Subjuntivo/imperativo	40. Necesidad
9. Modo indicativo	41. Obligación
10. Modo subjuntivo	42. Volición
11. Modo imperativo	K. Marcadores de modalidad
C. Desinencias verbales de persona	43. Atenuadores
12. Primera singular	44. Enfáticos
13. Segunda singular	L. Adverbios
14. Tercera singular	45. De lugar
15. Primera plural	46. De tiempo
16. Segunda plural	47. De modo
17. Tercera plural	48. De cantidad
D. Pronombres personales	M. Marcadores de subordinación
18. Primera persona singular	49. Subordinadas sustantivas con *que*
19. Primera persona plural	50. Subordinadas adjetivas pron. relativo
20. Segunda persona singular	51. Subordinadas adverbiales de razón o c/e
21. Segunda persona plural	
22. Tercera persona singular	52. Subordinadas adverbiales de concesión
23. Tercera persona plural	
24. Demostrativos	53. Subordinadas adverbiales condicionales
E. Formas nominales	
25. Nominalizaciones	54. Subordinadas adverbiales de tiempo
26. Sustantivos (comunes y propios)	55. Frases infinitivo en función nominal
F. Formas Pasivas	N. Frases preposicionales y adjetivos
27. Pasivas con «se»	56. Frases prep. (compl. del nombre)
28. Pasivas con ser sin agente	57. Adjetivos atributivos (calificativo)
29. Pasivas con ser con agente	58. Adjetivos predicativos

30. Pasivas con estar	59. Adjetivos demostrativos
G. Especificidad léxica	60. Participios función adjetiva
31. Relación *type/token* por forma	Ñ. Marcadores de Coordinación
32. Relación *type/token* por lema	61. Conjunciones adversa., adit. y disyun.
	O. Marcadores de negación
	62. Adverbio de negación
	63. Adverbios de negación temporal
	64. Conjunción de negación
	65. Pronombres de negación

Tabla 5. Rasgos lingüísticos Proyecto Corpus PUCV-2003

Una descripción funcional de cada uno de estos rasgos junto a su relación con ciertos registros y a un detallado apoyo bibliográfico referencial, se encuentra en Parodi 2005.

2.4. *Etiquetaje lingüístico automático*

En un intento por buscar una complementación de lo lingüístico y lo computacional y debido a la insuficiente capacidad de manejar y realizar estudios comparativos más abarcadores a base de la totalidad del corpus recolectado con tecnologías de vanguardia, adquirimos un paquete de programas computacionales para el español. Estas herramientas *(Programa Connexor)* permiten etiquetar morfosintácticamente los textos de acuerdo a ciertos principios establecidos internacionalmente y construir así una base de datos marcada. También se debió construir una interfaz computacional, accesible a través de Internet, la cual permitió la interrogación de cada texto del corpus (www.elgrial.cl). El procedimiento, aplicado a la totalidad de los textos del corpus PUCV-2003, consistió en:

1. Codificación SGML (*Standard Generalized Mark Up Language*)
2. Partidor o separador de oraciones *(Spliter* o *chunker)*
3. Marcaje morfológico
4. Desambiguador lingüístico y estocástico

La interfaz *El Grial* brinda un acceso directo y expedito a la totalidad de los textos recopilados y a una diversidad de alternativas de interrogación con resultados cuantificados y ejemplificados en cada caso, limitados parcialmente por el tipo de marcaje disponible que permite el programa *Connexor.*

Una vez concluidos los procedimientos señalados, se procedió a la interrogación de los textos a través de la interfaz *El Grial* para obtener la ocurrencia de los sesenta y cinco rasgos. Para ello se utilizó diversos accesos a la interfaz, ya sea la interrogación directa a través de búsquedas de ocurrencias simples o a través de comandos en CQP *(Corpus Query Program)* destinados a aquellos rasgos de alcance sintáctico y no solo morfológico. Para procesar estadísticamente las frecuencias de cada rasgo lingüístico en cada uno de los 90 textos, los cómputos de frecuencias fueron sometidos a un proceso de normalización.

2.5. *Análisis factorial*

Como se sabe, el análisis factorial es un procedimiento estadístico que permite identificar agrupamientos de rasgos lingüísticos que co-ocurren frecuentemente en los textos. Este análisis identifica correlaciones entre un número amplio de variables (sesenta y cinco rasgos lingüísticos) y aquellas que se distribuyen de modo similar. La estructura factorial en que se agrupa conjuntamente a las variables que tienden a co-ocurrir es producto de una matriz correlacional de todas las variables inicialmente involucradas. Cada grupo de variables co-ocurrentes resulta ser un factor, el cual es posteriormente interpretado en términos de categorías funcionales como una dimensión de variación. Este procedimiento tiende —entre otros— a la reducción de las variables involucradas en virtud de su co-ocurrencia significativa (Oakes 1998, Hair, Anderson, Tathan y Black 2001).

En el caso de esta investigación, una vez que se realizó el análisis factorial (Factores Principales), se determinaron siete factores principales con una rotación tipo Oblimin (Oakes 1998, Hair y otros 2001). Estos factores fueron confirmados a través de los puntajes factoriales. Se obtuvo así siete posibles dimensiones de las cuales solo fue factible interpretar funcionalmente cinco; por tanto, dos factores no permiten esbozar una dimensión consistente debido tanto al reducido número de rasgos constitutivos como a su naturaleza heterogénea. En general, se requiere al menos cinco pesos relevantes para alcanzar una interpretación significativa del constructo subyacente al factor (Biber 1988).

Dentro de cada factor, se presentá el listado de rasgos lingüísticos determinados y frente a ellos aparece —en cada caso— un número, normalmente con decimales. Esta cifra es el peso factorial que indica una medida de fuerza de la relación entre el rasgo en cuestión y el factor como un todo.

En otras palabras, este número muestra cuan representativo es el rasgo lingüístico del constructo funcional que subyace al factor. Como se sabe, las cargas o pesos factoriales fluctúan entre +1 y -1. Un valor más cercano a +1 refleja mayor representatividad del rasgo dentro del factor.

III. Resultados

En este apartado, se entregan dos tipos de resultados empíricos a partir de los siguientes procedimientos: (1) se identifican los factores y se determinan las funciones comunicativas compartidas por los conjuntos de rasgos en co-ocurrencia (interpretación funcional), y (2) se analiza la distribución de estos grupos de rasgos a través de los tres registros (CTC, CLL y CEO) y de los tres ámbitos técnico-científico del CTC (marítimo, industrial y comercial). En otras palabras, se procede a interpretar funcionalmente los parámetros estadísticos encontrados y se estudia su incidencia en cada uno de los registros y áreas de especialización profesional con el fin de identificar relaciones en el nivel lingüístico y funcional.

3.1. *Factores y dimensiones*

Tal como se verá a continuación, la solución final al análisis factorial concluyó con cinco factores óptimos. En ellos se excluyó todos aquellos rasgos que puntuaban con un valor absoluto menor a 0,40 dado que normalmente en este tipo de estudios se estiman sin importancia relativa frente a la interpretación, incluso si fueran estadísticamente significativos. Otras investigaciones utilizan como valor de referencia la cifra de 0,30 (Biber 1988); no obstante ello, en este estudio se decidió usar un puntaje de corte superior.

Solo los rasgos destacados o importantes deberían ser interpretados como parte de cada factor. Una carga negativa o positiva no influencia la relevancia de un peso, sino que, más bien, releva grupos de rasgos que se encuentran distribuidos en los textos de un modo complementario. Así, en los cinco factores siguientes, ciertos rasgos con valor positivo co-ocurren con una alta frecuencia en los textos del corpus, mientras que otro grupo de características co-ocurrentes pero con pesos negativos generan fuertes lazos entre ellos. Ambos grupos de rasgos presentan una especial relación. Ellos se distribuyen en un patrón complementario de ocurrencia, es decir, algunos

rasgos (con peso positivo) con alta frecuencia en un texto tienden a denotar la ausencia de ciertos rasgos (con peso negativo) en los mismos textos y viceversa (distribución complementaria). Una explicación pormenorizada de cada uno de las cinco dimensiones se puede encontrar en Parodi 2005.

Como se adelantó, los rasgos lingüísticos, agrupados en cada factor, son interpretados funcionalmente en una dimensión textual a través de un análisis evaluativo de las funciones comunicativas más ampliamente compartidas por esos rasgos en cuestión. Los resultados, presentados en las Tablas siguientes, muestran los rasgos y los correspondientes pesos positivos y negativos en cada uno de las cinco dimensiones.

FACTOR 1
Dimensión 1: Foco Contextual e Interactivo

Subordinadas adverbiales de causa - efecto	0,945
Adverbios de tiempo	0,934
Adverbio de negación	0,928
Pronombres segunda persona singular	0,911
Pronombres primera persona singular	0,823
Desinencias de segunda persona singular	0,813
Pronombre de negación	0,731
Adverbios de lugar	0,723
Modo indicativo	0,693
Desinencias primera persona singular	0,668
Futuro perifrástico	0,662
Enfatizadores	0,652
Formas activas «ser»	0,637
Verbos modales de volición	0,630
Pronombres demostrativos	0,592
Pronombres de segunda persona plural	0,531
Subordinadas adverbiales condicionales	0,523
Adverbios de negación temporal	0,503
Subordinadas sustantivas	0,497
Subordinadas adverbiales de tiempo	0,487
Verbos privados	0,474
Frases infinitivas en función nominal	0,466
Presente	0,424

Frases preposicionales complemento del nombre	-0,545
Nominalizaciones	-0,479
Sustantivos	-0,443
Participios en función adjetiva	-0,437

Los rasgos que constituyen la Dimensión 1 Foco Contextual e Interactivo son los más numerosos y ostentan los pesos estadísticos más altos. De los 23 rasgos iniciales con peso estadístico sobre 0,40, se desatacan quince rasgos con pesos superiores a 0,60. Cabe señalar que los rasgos positivos que co-ocurren en esta dimensión tienen los pesos más altos en comparación con todos los rasgos que caracterizan a las otras seis dimensiones. Los rasgos con pesos negativos son relativamente menores en número y sólo cuatro presentan pesos sobre 0,40. Los rasgos que se reúnen en este factor denotan una gran relación funcional entre la mayoría de ellos. Su interpretación no resulta compleja. De este modo, la estructura de la Dimensión 1 no constituye un producto exclusivo de la técnica de extracción factorial, sino que se basa en cuestiones lingüísticas y comunicativas subyacentes. Ésta resulta ser una dimensión poderosa que representa un patrón de variación sustancial entre los textos orales y escritos, especializados y no-especializados del español.

Para interpretar esta dimensión se debe evaluar sistemáticamente las posibles funciones comunicativas compartidas por los rasgos co-ocurrentes. En este caso, comparativamente, los rasgos con peso negativo son menos e indican una clara interpretación (nominalizaciones, frases preposicionales con función de complemento del nombre, sustantivos y participios pasivos en función adjetiva). Ellos son clásicamente considerados portadores de la carga referencial del texto, permiten la integración y precisión de grandes cantidades de información y una alta frecuencia de ellos apunta a una fuerte densidad informacional. Por su parte, los rasgos con altos puntajes de ocurrencias positivas revelan una referencia directa al contexto físico y temporal, determinan marcos de orden de la sucesión de hechos, establecen una vinculación con la acción y expresan motivos y consecuencias. También, a través de estos rasgos, se expresa referencia directa a los participantes y existen suficientes evidencias de que la marca de lo situado está presente. Esta dimensión se concreta en la acción, en la sucesión de acontecimientos y en las relaciones interpersonales de tipo dialógico. Los rasgos lingüísticos involucrados, en su conjunto, permiten suponer que los textos caracterizados por esta dimensión no contienen información altamente abstracta; sino por el contrario, la alta frecuencia de ocurrencia de rasgos tipificadores y estadísticamente positivos se asocian con un foco en la explicitud y dependencia del contexto y en la activa participación de los interlocutores, rasgos clásicos del discurso oral y dialógico.

Esta Dimensión 1 representa un parámetro fundamentalmente importante de variación, compuesto de un conjunto amplio de rasgos lingüísticos y definitorios de una clara distinción entre dos polos opuestos como son, por un lado, lo contextual e interactivo (oralidad: conversaciones) y, por otro, lo informacional, altamente planificado y cohesionado (escritura: exposición).

FACTOR 2 Dimensión 2: Foco Narrativo	
Pronombres de segunda persona plural	0,842
Pronombres de primera persona singular	0,828
Futuro perifrástico	0,823
Pretérito imperfecto	0,820
Pronombre tercera persona plural	0,708
Modo indicativo	0,686
Desinencias primera persona plural	0,667
Verbos modales de volición	0,651
Pretérito indefinido	0,614
Pronombre de negación	0,590
Verbos privados	0,577
Adverbios de lugar	0,533
Pronombres segunda persona singular	0,529
Verbos perceptivos	0,496
Adverbio de negación	0,493
Adverbios de negación temporal	0,482
Formas activas «estar»	0,460
Verbos públicos	0,445
Pronombres primera persona plural	0,431
Desinencias tercera persona singular.	0,423
Conjunciones adver., disy. y aditivas	0,411
Frases infinitivas en función nominal	0,405
Conjunción *ni*	0,402

Nominalizaciones	-0,581
Frases preposicionales complemento del nombre	-0,562
Adjetivos atributivos	-0,442

En oposición, la Dimensión 2 Foco Narrativo denota una evidente orientación hacia un determinado tipo de trama textual de orden narrativo. Ello se refleja en un amplio espectro de rasgos con peso positivo (de igual número al de la Dimensión 1), tales como, algunos pronombres personales y las respectivas desinencias verbales; ellos denotan un marcado acento en

la identificación de las personas del discurso. En directa relación con lo anterior, se aprecia la co-ocurrencia de los tiempos verbales del pasado: el pretérito imperfecto y el pretérito indefinido. En suma, esta dimensión se asocia con una sucesión de acontecimientos, que implica la precisión de circunstancias de tiempo y lugar, como también la participación de las personas del discurso. Ella permite identificar textos literarios orales o escritos, a diferencia de textos altamente especializados.

FACTOR 3 Dimensión 3: Foco Compromiso	
Verbos privados	0,824
Pronombres primera persona singular	0,789
Pretérito indefinido	0,705
Verbos modales de volición	0,655
Desinencias primera persona singular	0,640
Modo indicativo	0,630
Pretérito imperfecto	0,604
Pronombre de negación	0,569
Pronombres segunda persona singular	0,563
Desinencias segunda persona plural	0,562
Frases infinitivas en función nominal	0,518
Subordinadas sustantivas	0,467
Subordinadas adverbiales de concesión	0,452
Formas activas *estar*	0,435
Pronombres segunda persona plural	0,427
Adverbio de negación temporal	0,411
Pronombres primera persona plural	0,402
Frases preposicionales complemento del nombre	-0,457

La Dimensión 3 es interpretada como Foco Compromiso pues la alta ocurrencia de verbos tales como privados *(decidir, adivinar, sentir, determinar, demostrar, estimar, reconocer)* y de volición (*querer* + infinitivo) y de los pronombres personales y las desinencias verbales de primera persona constituyen marcas relevantes de la expresión del «yo». La clara identidad de quien escribe o habla queda manifiesta en el texto de manera explícita y quien participa se compromete e involucra con lo que dice y hace. Este compromiso con el discurso y su contenido revela los afectos y los propósitos del escritor/hablante. Esta dimensión está asociada a textos en los que sobresale la intención y la actitud del emisor que revela su voluntad de in-

volucrarse en el discurso de manera explícita y de asumir un rol preponderante. Dicho de otro modo, esta dimensión caracteriza a textos en los que aparecen participantes reales que expresan intenciones y actitudes proposicionales frente a lo dicho.

FACTOR 4 Dimensión 4: Foco Modalizador	
Formas activas *ser*	0,671
Atenuadores	0,656
Verbos modales de posibilidad	0,641
Adverbios de modo	0,606
Adjetivos predicativos	0,565
Desinencias tercera persona plural	0,549
Subordinadas adjetivas	0,514
Desinencias tercera persona singular	0,405

Sustantivos	-0,494

Para construir la interpretación de la cuarta Dimensión, Foco Modalizador, se atendió de manera especial a la alta co-ocurrencia significativa de atenuadores *(parecer que, creer, tal vez, a lo mejor, quizás, quizá),* verbos modales de posibilidad *(poder)* y adverbios modales *(probablemente, posiblemente).* Ellos revelan un parámetro funcional muy preciso: la regulación y atenuación de la información entregada, es decir, la expresión de la probabilidad y la incertidumbre de los hechos o acontecimientos descritos o narrados. La conjunción sistemáticas de estas marcas lingüísticas tienden a darse en textos con énfasis en cómo *(modus)* se dicen las cosas, más que en lo dicho *(dictum).* Esta distribución de rasgos presenta el contenido de un discurso como incierto y abierto a la verificación; por el contrario, se aleja de la supuesta «objetividad» que otros grupos de rasgos típicamente tienden a representar, tales como aquellos que enfatizan lo referencial. En estos últimos, las restricciones impuestas llevan a textos con alta precisión léxica y densidad informativa.

La Dimensión 4 agrupa rasgos característicos de textos orales y escritos, clásicamente ligados a la narración y descripción. No obstante ello, también puede marcar textos expositivos y argumentativos, más representativos de la alta especialización.

FACTOR 5 Dimensión 5: Foco Informacional	
Verbos modales de obligación	0,496
Modo subjuntivo	0,494
Nominalizaciones	0,456
Participios en función adjetiva	0,413
Frases preposicionales complemento del nombre	0,413
Desinencias tercera persona singular	-0,632
Pretérito indefinido	-0,630
Forma estativa activa *estar*	-0,595
Verbos privados	-0,575
Pronombre de negación	-0,572
Verbos modales de volición	-0,503

Finalmente, la última y quinta dimensión, denominada Foco Informacional, se constituye a partir de once rasgos: cinco positivos y seis negativos. Los rasgos positivos en co-ocurrencia como verbos modales de obligación revelan la necesidad y certeza de los juicios expresados; el modo subjuntivo remite a organizaciones de mayor complejidad sintáctica. La presencia de nominalizaciones junto a participios en función de adjetivo, sustantivos (comunes y propios) y frases preposicionales son todos rasgos indicadores de integración y compactación de información altamente abstracta, típica del discurso especializado escrito. En resumen, los rasgos positivos agrupados en torno a esta última dimensión se encuentran básicamente orientados hacia la informatividad, entendida ésta como la concentración de información en unidades y estructuras lingüísticas compactas, que presentan los datos lo más concisa y precisamente posible. Por otra parte, los seis rasgos negativos aquí presentes (tercera persona singular, pretérito indefinido, verbo *estar* estativo activo, verbos privados, pronombres de negación y verbos modales de volición) apuntan, preferentemente, hacia una contextualización de eventos señalados en el discurso, tendencia atenuada en textos de alta informatividad.

Ahora bien, cada una de estas cinco dimensiones es producto de un conjunto distintivo de rasgos lingüísticos co-ocurrentes; cada una de ellas define potencialmente un grupo divergente de similitudes y diferencias entre registros y áreas de especialización. Cabe señalar que producto del tipo de rotación seleccionada (Oblimin), la cual resulta más apropiada en el manejo de datos lingüísticos y arroja una mayor certeza de los hechos analizados, la

conformación de un factor puede presentar rasgos repetidos con otro factor, ya que estos no son eliminados del análisis una vez incluidos en un primer factor. Ello implica posibles complejidades llegado el momento de la interpretación funcional; no obstante, su conformación resulta ser más realista en cuanto al lenguaje humano y su relación con una determinada dimensión. Como se sabe, en la lengua un tipo de rasgo caracterizador es fácilmente detectable como parte constitutiva en más de una función comunicativa; es el conjunto de rasgos en co-ocurrencia sistemática lo que revela un patrón de variación singular con posibilidad de interpretación comunicativa.

Las dimensiones 1, 2 y 5 (Foco Contextual e Interactivo, Foco Narrativo y Foco Informacional) se muestran bastante distintivas, dado que la mayoría de sus rasgos positivos son diferentes. Por ello, es posible establecer una clara separación entre las funciones que buscan representar y los tipos de textos que distinguen. Esta distinción —aguda y fina— está fundamentada en que muchos de los rasgos negativos presentan también heterogeneidad y por ello se apunta a registros o tipos de textos muy estereotipados con escasos entrecruzamientos. No obstante, las dimensiones 3 y 4 (Foco Compromiso y Foco Modalizador) no parecen apuntar a categorías tan finamente establecidas y es evidente que tienden a resultar similares en algunos aspectos. Este hecho se explica en que, por una parte, comparten un número de rasgos lingüísticos cuyas funciones subyacentes pueden ser muy similares y, por otra, en que algunos de sus rasgos aunque no idénticos tienden a una interpretación funcional similar. Ello no ha de extrañar ya que —tal como se advirtió anteriormente— la rotación tipo Oblimin de los datos conlleva este tipo de resultados; resultados que normalmente dan cuenta de lenguas naturales.

3.2. *Variabilidad en el corpus PUCV-2003*

En este apartado, las similitudes y diferencias entre los registros y áreas de especialización técnico-científicas son estudiadas con respecto a cada una de las cinco dimensiones y en cuanto a la totalidad de las mismas de manera simultánea. Los tres registros en estudio pueden ser similares en algunos aspectos del *continuum* de los patrones especificados, pero también pueden variar grandemente entre una dimensión y otra.

En el Gráfico 1 se presentan los puntajes promedio por dimensión (calculado por medio del puntaje factorial o *factor score)* para cada uno de los

3 registros del corpus PUCV-2003 (técnicos, literarios y orales) con respecto a las cinco dimensiones en cuestión. Recordemos que con ello se busca —a través de las dimensiones ya determinadas— comparar los registros y especificar relaciones específicas en el español escrito especializado y no especializado, así como con la oralidad no técnico-científica; de modo más preciso, se intenta identificar posibles diferencias o similitudes lingüísticas y funcionales entre los textos en estudio.

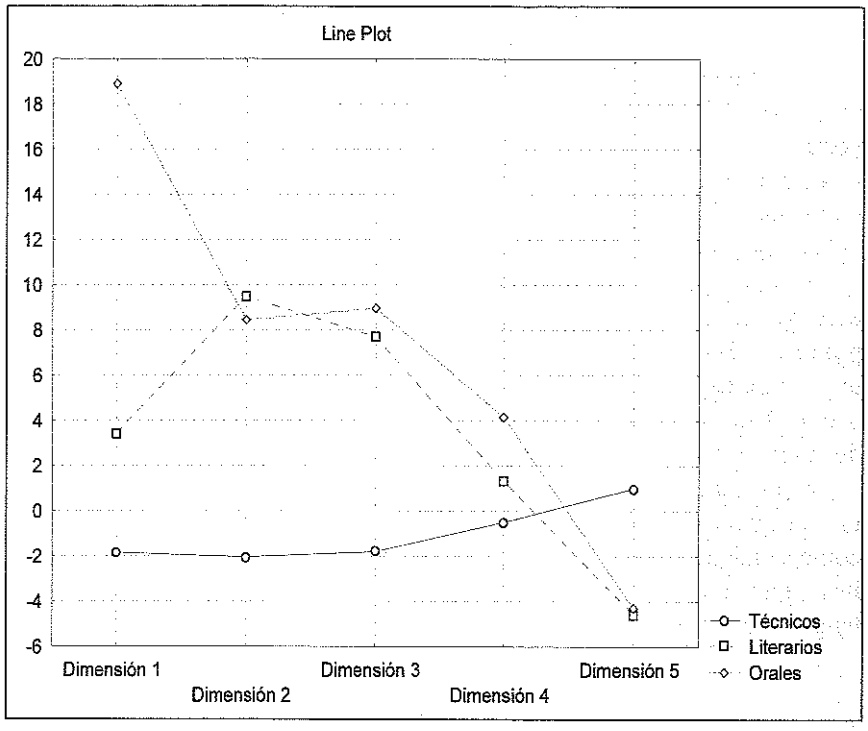

Gráfico 1. Dimensiones y registros.

Resulta interesante comprobar que los textos del CTC obtienen un puntaje factorial negativo muy similar a lo largo de las cuatro primeras dimensiones. Esta homogeneidad revela que, por un lado, los rasgos que identifican estas cuatro dimensiones deben presentar una ocurrencia similar en los textos de este corpus; según estos datos, la interactividad, contextualización, modalización y compromiso no serían rasgos característicos ni de alta ocurrencia en los textos del ámbito técnico y científico, comparado con los

otros dos registros. Por otra parte, es esclarecedor que la Dimensión 5 Foco Informacional se presente con un puntaje promedio positivo (0,9), lo que permite identificar y distinguir significativamente el CTC del CLL y CEO. Las nominalizaciones, frases preposicionales como complemento del nombre y participios pasivos adjetivos, entre otros, son rasgos de alta relevancia en estos textos escritos. Estos rasgos contribuyen al cumplimiento del dispositivo informacional, el cual se aleja claramente de contenidos interpersonales y afectivos. Al mismo tiempo, como se observa, los textos de literatura latinoamericana escrita y los textos basados en entrevistas orales poco planificadas, con un grado importante de espontaneidad, muestran puntajes negativos en la Dimensión 5 y, en cambio, obtienen los más altos puntajes promedio positivos en el resto de las otras cuatro dimensiones. Entre ellos, cabe destacar el elevado puntaje factorial que los textos orales alcanzan en la Dimensión 1 Foco Contextual e Interactivo (19); se hace evidente, entonces, que en las entrevistas orales cara a cara los rasgos como pronombres personales y desinencias verbales, adverbios de tiempo, lugar y modo, tiempo presente y pronombres y adjetivos demostrativos se constituyen en prototípicos.

Este análisis comparativo arroja similitudes y diferencias muy elocuentes. Permite identificar las áreas de intersección entre la oralidad de las entrevistas y la narratividad de la literatura latinoamérica. También ayuda a separar distintivamente los textos de alta especialización en los que se detectan construcciones gramaticales más complejas y de mayor empaquetamiento y reducción de información de aquellos que involucran detalladamente a los participantes y sus relaciones interpersonales. En estos últimos textos (orales y escritos) se detecta y se expresa —de modo más explícito— el involucramiento del autor a través de marcas lingüísticas específicas (ciertos tipos de verbos, adverbios, pronombres, etc.). Hecho que también puede caracterizar el discurso especializado escrito, pero que muchas veces se implica a través de otros recursos.

Como se aprecia, la variación entre estos registros enfrenta un *continuum,* identificado en este caso por medio de las dimensiones y los rasgos lingüísticos que ellas capturan, y permite identificarlas y caracterizarlas prototípicamente. Las implicancias derivadas de ello son múltiples, en particular, en lo que respecta al discurso especializado de divulgación didáctico, foco de interés de esta investigación. El diseño de materiales didácticos que aborden el discurso especializado en español escrito y quienes se concentren en su enseñanza deben capitalizar estas descripciones.

Finalmente, en este último gráfico, se entregan los puntajes factoriales por dimensión para cada uno de las tres áreas de especialización del corpus CTC-PUCV-2003 (comercial, industrial y marítimo).

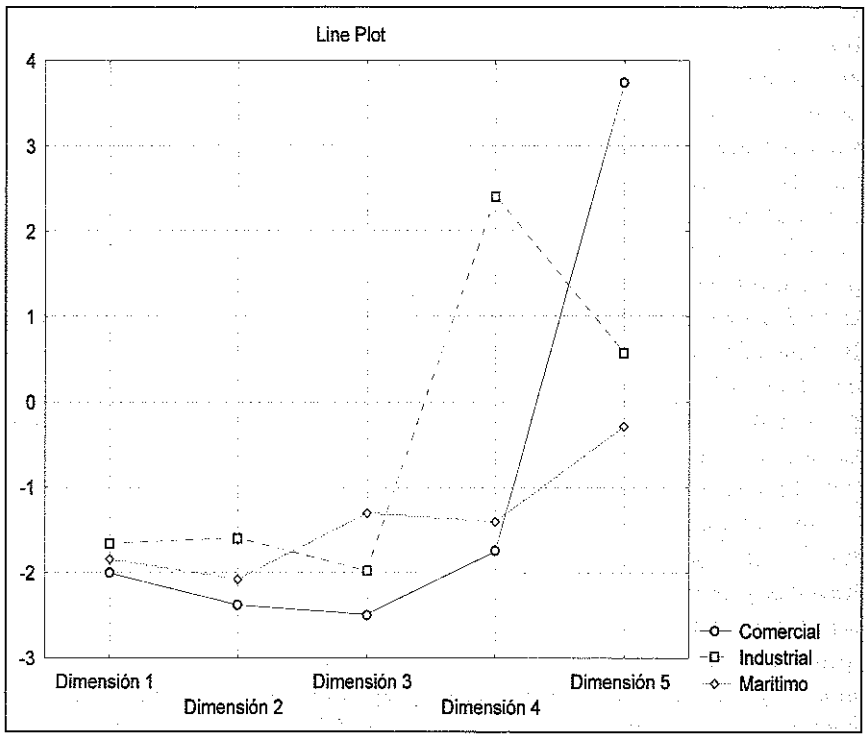

Gráfico 2. Dimensiones y ámbitos de especialización

Es interesante observar que también en el Gráfico 1 el CTC muestra una clara diferenciación en cuanto a la Dimensión 5 Foco Informacional del resto de los otros dos registros (CLL y CEO); en este nuevo análisis, más pormenorizado, es justamente esta quinta dimensión la que parece revelar la mayor distinción entre las áreas técnico-científicas. Según se aprecia, el ámbito comercial presenta el mayor puntaje promedio positivo en esta dimensión (3,8), hecho que muestra su mayor carga informacional a través de alta densidad léxica y complejidad sintáctica.

Parodi 2004 ya había detectado, a través de un estudio descriptivo simple, cierta variabilidad interna en el comportamiento de los sesenta y cinco

rasgos lingüísticos entre las tres áreas técnicas. Esta investigación preliminar muestra un patrón distintivo entre el área marítima y las otras dos. El mismo dato encuentra apoyo certero en estos análisis dimensionales con técnicas tipo *factor score*. Según se comprueba, el mayor puntaje negativo en cuanto a la carga informacional lo obtiene el ámbito de especialización marítima (-0,1), el mismo que apareciera distante de los otros dos en el trabajo que comentamos. Considerando este resultado, la dimensión cinco parece distinguir entre textos de un ámbito y otro. Será necesario un estudio cualitativo profundo para dar cuenta con mayor detalle de qué clases textuales del área comercial se alinean en torno a este patrón dimensional.

Seguidamente en la jerarquía, la Dimensión 4 también aporta a la distinción entre especialidades. Es ahora el área industrial la que revela mayor puntaje promedio positivo en la Dimensión Foco Modalizador (2,5), mientras que el marítimo y el comercial obtienen puntajes negativos muy cercanos y sin significatividad estadística entre ellos. Estos datos permiten inferir que los textos del ámbito industrial contienen mayor regularidad en los patrones sistemáticos de ocurrencia en torno a los rasgos distintivos de la atenuación e incertidumbre.

En el resto de las dimensiones (1, 2 y 3), las tres áreas del CTC presentan cifras negativas y relativamente similares. Hechos que indican que estas dimensiones no aportarían gran explicación diferenciadora en la descripción del discurso técnico-científico de divulgación didáctico escrito. También se puede sugerir que los textos de estos ámbitos especializados no destacan por la ocurrencia de estos rasgos que denotan interacción, relaciones interpersonales e involucramiento de los participantes en el discurso. Todo ello es congruente con otros antecedentes disponibles acerca de este tipo de textos.

Conclusiones

En este artículo se exploró la distribución de un grupo de rasgos gramaticales relevantes funcionalmente en la descripción de tres corpus de lenguaje natural a través de registros diversificados. También, de modo más específico, se buscó llevar a cabo una descripción inicial del corpus especializado de corte técnico-científico (CTC) desde la perspectiva del análisis multidimensional. Para ello se combinó la utilidad de los recursos de la lin-

güística de corpus y de los avances en el desarrollo de programas tecnológicos de vanguardia para el español en la construcción e interrogación de una base de datos digital con etiquetaje lingüístico automático. Se buscaba de este modo incorporar las tecnologías emergentes a la investigación lingüística acerca del español.

En cuanto a la determinación e identificación de rasgos relevantes y dimensiones lingüísticas, desde una óptica general, el análisis multidimensiones probó ser una metodología poderosa en el marco de la lingüística de corpus. Los beneficios y ventajas del AMD implican más que la mera descripción de los rasgos superficiales de los textos sino que conduce a la determinación de regularidades sistemáticas que develan funciones comunicativas y aproximaciones hacia una descripción profunda de la función que tal o cual texto aporta en un contexto determinado de uso. La ventaja que radica en el análisis multirasgos y la muestra relativamente grande de textos permite generar conclusiones robustas y con mayores implicancias. Estos aspectos llevan a la lingüística —en esta línea— no solo a contar con una alternativa metodológica de investigación, sino a delinear un paradigma que brinda renovados bríos a las indagaciones venideras en español.

El hallazgo de las cinco dimensiones, identificadas a partir de un análisis cuantitativo de la distribución de los sesenta y cinco rasgos lingüísticos en los noventa textos, resultó en un recurso de extraordinario potencial descriptivo. Los aportes, a partir de la visión multidimensional, entregan una distinción fundamental entre los textos estudiados de la modalidad oral y escrita: los primeros aparecen anclados en la interactividad, contextualización y las relaciones interpersonales, esto es, fuertemente descritos por la Dimensión 1 Foco Contextual e Interactivo; al mismo tiempo, la oralidad aquí descrita (entrevistas) no se relaciona directamente con una prosa informacionalmente densa, tanto desde el punto de vista léxico y sintáctico. Si bien es cierto que el registro literario escrito coincide con el oral en las Dimensiones Foco Narrativo, Foco Compromiso y Foco Modalizador, la distinción manifiesta en la Dimensión 1 es suficiente como para establecer una diferencia importante. Evidentemente, esto no permite hablar de una fuerte polarización entre oralidad y escritura, lo que reafirma la adscripción a la idea de un desplazamiento continuo entre los registros, a modo de categorías prototípicas y también difusas. Por otra parte, el AMD también ha permitido identificar registros más expositivos y técnicos escritos de otros claramente narrativos literarios escritos.

Otro posible hallazgo relevante que se desprenden de los resultados de este estudio es la similitud detectada en varias lenguas en cuanto a la determinación de dimensiones lingüísticas que ayuden a la descripción y distinción entre registros (Kittredge 1982, Biber 1994, Aijmer 2002, Reppen, Fitzmaurice y Biber 2002). El español, en términos generales, al igual que el inglés, el somalí y el coreano presenta múltiples dimensiones que reflejan —al menos— la interactividad, el foco informacional y el compromiso. En todas estas lenguas —gracias al AMD— se ha podido establecer distinciones entre oralidad y escritura y diversos registros específicos en cada modalidad de lengua. En este sentido, la variabilidad registrada en usos lingüísticos al interior de las lenguas particulares indagadas, aunque existentes en este sentido de patrones, parecen tender a una cierta universalidad.

La descripción del CTC se desarrolló parcialmente en este trabajo. Es necesario abordar un análisis más pormenorizado de las distinciones entre las tres áreas de especialización al interior del corpus técnico-científico (comercial, marítima e industrial) y se deberá continuar con el cálculo de los puntajes factoriales entre las diversas clases textuales que componen este corpus con el fin de indagar el aporte del AMD a la diferenciación y descripción de las mismas. Los antecedentes empíricos hasta ahora aportados dan cuenta de una interesante congruencia a lo largo de las Dimensiones 1, 2 y 3 entre las tres áreas técnico-científicas: estos textos no parecen tener una fuerte marca de narratividad, de involucramiento de los participantes ni de interactividad. Por el contrario, los textos del área industrial y marítimo sí presentan diferencias significativas en las Dimensiones Foco Modalizador y Foco Informativo. El dominio industrial muestra una tendencia hacia la modalización y el dominio marítimo se distingue de los textos de las otras dos áreas en un grado positivo mayor de densidad informativa.

Estimo que de este estudio se desprenden implicancias relevantes para el diseño de pruebas de diversa índole tales como de evaluación de contenidos técnicos y de comprensión textual propiamente. También para la elaboración de material didáctico. Ello, ya que los análisis de los registros describen las características del tipo de uso lingüístico empleado en el material al que se exponen estos alumnos de establecimientos técnico-profesionales en las áreas mencionadas. De acuerdo a nuestros resultados, estos estudiantes requieren desarrollar la habilidad para manejar una variedad del español especializado escrito muy particular: prosa académicamente densa en términos léxicos, morfológicos y sintácticos pero también textos con marcas de modalización.

Los resultados aportados revelan la significatividad del estudio del uso lingüístico en entornos técnico-científicos escolares del ámbito profesional y confirman la necesidad de crear materiales especializados de instrucción sobre la base de descripciones empíricas de los registros en estudio. Esta evidencia también sugiere que existen razones significativas para atender diferencialmente al análisis, comprensión y producción del registro técnico-científico que circula y se genera tanto en el ámbito educativo como en el ambiente laboral. Todo estudiante debería practicar con una amplia variedad de registros ya que en su vida profesional no solo encontrará los técnicos especializados. Tampoco parece recomendable que alumnos de estos liceos técnico-profesionales secundarios se enfrenten de manera automática a este tipo de textos marcados por una prosa informacional extremadamente compleja. El material debería contener una organización progresiva desde lo más divulgativo didáctico hasta aquellos textos más típicos del ámbito profesional.

Por último, cabe recordar que el procesamiento del corpus fue exitosamente realizado gracias a la tecnología desarrollada en la Pontificia Universidad Católica de Valparaíso. El funcionamiento del proceso de interrogación de los corpus a través de la interfaz *El Grial* en Internet no presentó mayores problemas en su ejecución. De este modo, contamos con un corpus etiquetado que posibilita múltiples combinaciones de indagación. Como una forma de facilitar la interrogación misma y de contar con otras herramientas (tal como N-Gramas) y factibilidad de graficación de los datos obtenidos, se construyó el programa BUCÓLICO (Buscador de Concordancias Lingüísticas en Corpus); éste, no solo cuenta con los noventa textos del Corpus PUCV-2003, sino que permite la incorporación de nuevos textos para su estudio (al respecto, ver Parodi y Venegas 2004)

Referencias bibliográficas

Aijmer, K. 2002: «Modal adverbs of certainty and uncertainty in an English-Swedish perspective», en Hasselgard, H., Johansson, S., Behrens B. y Fabricius-Hansen, C. (eds.), *Information structure in a cross-linguistics perspective*, Amsterdam, John Benjamins, págs. 97-113.

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

Biber, D. 1994: «Using register-diversified corpora for general language studies», en Armstrong, S. (ed.), *Using large corpora,* Cambridge, The MIT Press, págs. 180-201.

Biber, D. 2003: «Variation among university spoken and written registers: A new multi-dimensional analysis», en Leistyna, P. y Meyer, Ch. (eds.), *Corpus analysis. Language structure and language use,* Amsterdam, Rodopi, págs. 47-70.

Biber, D., Conrad, S. y Reppen, R. 1998: *Corpus linguistics. Investigating language structure and use,* Cambridge, Cambridge University Press.

Biber, D., Reppen, R., Clark, V., y Walter, J. 2001: «Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus», en Simpson, R. y Swales, J. (eds.), *Corpus Linguistics in North America. Selections from the 1999 Symposium,* Ann Arbor, The University of Michigan Press, págs. 48-57.

Cabré, M. 1999: «Hacia una teoría comunicativa de la terminología: aspectos metodológicos», *Revista Argentina de Lingüística* 15, págs. 24-38.

— 2000: *La terminología: Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos,* Barcelona, Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.

— 2002: «Textos especializados y unidades de conocimiento: metodología y tipologización», en García, J. y Fuentes, M. (eds.), *Texto, terminología y traducción,* Barcelona, Almar, págs. 122-187.

Cademártori, Y. 2003: «La inscripción de las personas en textos de divulgación científica», *Revista Latinoamericana de Estudios del Discurso,* 3,1, págs. 9- 27.

Calsamiglia, E. 2000: «Decir la ciencia: Las prácticas divulgativas en el punto de mira», *Revista Iberoamericana del Discurso y Sociedad,* 2, 2, págs. 3-8.

Cassany, D., López, C. y Martí, J. 2000: «La transformación divulgativa de redes conceptuales científicas. Hipótesis, modelo y estrategias», *Revista iberoamericana de Discurso y Sociedad,* 2, 2, págs. 73-103.

Castel, V. 2004a: «Towards a Generation-oriented Grammar of the Research Paper Abstract», en *Actas del Primer encuentro latinoamericano de Lingüística Sistémica Funcional: La lengua y la educación,* Universidad Nacional de Cuyo, Mendoza, Argentina.

— 2004b: «La catalogación electrónica en español: método y herramienta de desarrollo», en *Neue Romania, Número especial: Kontrastive Textsorten: Alltags- und Fachkommunikation in Argentinien und Deutschland,* Institut für Romanische Philologie der Freien Universität Berlín.

Ciapuscio, G. 1992: «Impersonalidad y desagentivación en la divulgación científica», *Lingüística española actual,* 2, págs. 183-205.

— 2000: «Hacia una tipología del discurso especializado», *Revista Iberoamericana del Discurso y Sociedad,* 2, 2, págs. 39-71.

— 2003: *Textos especializados y terminología,* Barcelona, Instituto Universitario de Lingüística Aplicada, UPF.

— y Kuguel, I. 2002: «Hacia una tipología del discurso especializado», en García, J. y Fuentes, M. (eds.), *Texto, terminología y traducción,* Salamanca, Almar, págs. 37-73.

Conrad, S. y Biber, D. 2001: «Multi-dimensional methodology and the dimensions of register variation in English», en Conrad, S. y Biber, D. (eds.), *Variation in English. Multidimensional Studies,* págs. 13-42.

Francis, N. 1979: «A tagged corpus: problems and prospects», en Greenbaum, S., Leech, G. y Svartvik, J. (eds.), *Studies in English linguistics for Randolph Quirk,* Londres, Longman, págs. 192-209.

Galán, C. 1999: «La subordinación causal y final», en Bosque, I. y Demonte, V. (coords.), *Gramática Descriptiva de la Lengua Española,* Madrid, Espasa Calpe, págs. 3597-3642.

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

Harvey, A. 2002: «Representación e imagen del quehacer científico en los Medios de Comunicación», en Parodi, G. (ed.), *Lingüística e Interdisciplinariedad,* Valparaíso, Ediciones Universitarias de Valparaíso, págs. 335-353.

Johansson, S. 1991: «Times change, and so do corpora», en Aijmer, K. y Altenberg, B. (eds.), *English Corpus Linguistics. Studies in honor of Jan Svartvik,* Londres, Longman, págs. 305-314.

Kittredge, R. 1982: «Variation and homogeneity of sublanguages», en Kittredge, R. y Lehrberger, J. (eds.), *Sublanguages. Studies of language in restricted semantic domains,* Berlín, Walter de Gruyter, págs. 145-189.

Lakoff, G. 1972: «A study in meaning criteria and the logic of fuzzy concepts», *Chicago, Linguistics Society* 8, págs. 183-288.

Leech, G. 1991: «The state of the art in corpus linguistics», en Aijmer, K. y Altenberg, B. (eds.), *English Corpus Linguistics. Studies in honor of Jan Svartvik,* Londres, Longman, págs. 8-29.

— 1992: «Corpora and theories of linguistic performance», en Svartvik, J. (ed.), *Directions in corpus linguistics: proceeding of Nobel symposium,* Berlín, Mouton de Gruyter, págs. 105-122.

López, C. 2002: «Aproximaciones al análisis de los discursos profesionales», *Revistas Signos* 35, 51-52, Págs. 194- 215.

Lorente, M. 2002: «Verbos y discurso especializado». [En línea]. Disponible en: http://elies.rediris.es/elies16/Lorente.html

Martín Zorraquino, M. y Portolés, J. 1999: «Los marcadores del discurso», en Bosque, I. y Demonte, V. (coords.), *Gramática descriptiva de la lengua española,* Vol. III, Madrid, Espasa Calpe, págs. 4051-4213.

Menéndez, S. 1999: «El discurso del libro de texto: Una propuesta estratégico-pragmática», *Revista Iberoamericana de Discurso y Sociedad* 1, 2, págs. 85-104.

Montolío, E. 2001: *Conectores de la lengua escrita,* Barcelona, Ariel Practicum.

Oakes, M. 1998: *Statistics for corpus linguistics,* Edinburgo, Edinburgh University Press.

Parodi, G. (ed.) 2005: *Discurso especializado e instituciones formadoras. Aproximación a los textos de comunidades técnico-profesionales,* Valparaíso, Ediciones Universitarias de Valparaíso.

— 2004: «Textos de especialidad y comunidades discursivas técnico-profesionales: una aproximación basada en corpus computarizado», *Estudios Filológicos* 39, págs.7- 36.

— y Gramajo, A. 2003: «Los tipos textuales del corpus PUCV-2003: una aproximación multiniveles», *Signos* 36, 54, págs. 207-223.

— y Venegas, R. 2004: «BUCÓLICO: Programa para el análisis de corpus lingüístico», *Revista de Lingüística y Literatura* 15, págs. 21-47.

Portolés, J. 1998: *Marcadores del discurso,* Barcelona, Ariel Practicum.

Reppen, R., Fitzmaurice, S. y Biber, D. 2002: *Using corpora to explore linguistic variation,* Amsterdam, John Benjamins.

Sinclair, J. 1982: «Reflections on computer corpora in English language research», Johansson, S. (ed.), *Computer corpora in English language research,* Bergen, Norwegian Computing Centre for the Humanities, págs. 1-6.

— 1991: *Corpus, concordance, collocation,* Oxford, Oxford University Press.

Stubbs. M. 1996: *Text and corpus analysis. Computer-assisted studies of language and culture,* Malden, Massachusetts, Blackwell.

Svartvik, J. 1992: «Corpus linguistics comes of age», en Svartvik (ed.), *Directions in corpus linguistics: proceeding of Nobel symposium,* Berlín, Mouton Gruyter, págs. 7-13.