



Literatura y Lingüística

ISSN: 0716-5811

literaturalinguistica@ucsh.cl

Universidad Católica Silva Henríquez

Chile

Parodi S., Giovanni; V., René

Bucólico: aplicación computacional para el análisis de textos. (hacia un análisis de rasgos de la informatividad)

Literatura y Lingüística, núm. 15, 2004, p. 0

Universidad Católica Silva Henríquez

Santiago, Chile

Disponible en: <http://www.redalyc.org/articulo.oa?id=35201514>

- Cómo citar el artículo
- Número completo
- Más información del artículo
- Página de la revista en [redalyc.org](http://www.redalyc.org)

[redalyc.org](http://www.redalyc.org)

Sistema de Información Científica

Red de Revistas Científicas de América Latina, el Caribe, España y Portugal

Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

LITERATURA Y LINGÜÍSTICA

Bucólico: aplicación computacional para el análisis de textos. (hacia un análisis de rasgos de la informatividad)

Giovanni Parodi S. y René Venegas V.

Pontificia Universidad Católica de Valparaíso. Chile

Resumen

La aplicación computacional BUCÓLICO (buscador de concordancias Lingüísticas en Corpus) permite organizar tres corpora lingüísticos previamente procesados y etiquetados morfológicamente o analizar otros corpora, no etiquetados computacionalmente, a través del procesamiento de lenguaje natural. Desde un enfoque metodológico de la lingüística de Hábeas Computacional, se indaga, comparativamente a partir de los datos obtenidos con bucólico, en la ocurrencia de cuatro estructuras lingüísticas de alta densidad, compactación o integración informacional.

Palabras claves: - Rasgos lingüísticos de la información - Lingüística de corpus - Análisis de textos

Abstract

Bucólico (Buscador de Concordancias Lingüísticas en Corpus) is a computational tool which is helps organize three linguistic corpora, previously processed and automatically tagged morfosyntactically, and it analices other linguistic corpora, not computationally tagged though the precessing of natural language. The methodological principles of the modern computational corpus linguistics reserches about the occurrence of four linguistic features, typically associated with high density, integration, compactness and informational load.

Key words: - Linguistics informational feature - linguistic corpus - text analysis

1. Introducción

No hay dudas de que la irrupción de las actuales tecnologías de la información ha afectado el mundo contemporáneo y ha determinado un cambio progresivo de enormes proporciones en prácticamente todos los ámbitos: la economía, la educación, la cultura, el entretenimiento, la ciencia y la política. Hoy en día es difícil imaginar alguna actividad humana que no haya sido impactada por las técnicas digitales. Es un hecho que la computación, sus diversas proyecciones multimediales y las tecnologías de la comunicación han obligado a reestructurar el modo en que interactuamos en la vida cotidiana y en la laboral y, más importante aún, están impactando fuertemente la manera en que pensamos. Hay quienes incluso -arriesgadamente- afirman que estamos en presencia de un nuevo ser humano. En este artículo estamos lejos de tan osada afirmación, sobre todo, en lo que a cuestiones ontológicas y epistemológicas se refiere.

Como es de esperar, las ciencias del lenguaje no han estado distantes a estos desarrollos. Aunque posiblemente con lentitud, se han producido avances tecnológicos de magnitud. En esta línea, por ejemplo, cabe consignar que los traductores automáticos (no importando aquí su calidad), los diccionarios electrónicos monolingües y multilingües, diversos atlas lingüísticos computarizados, variados software de tipo educativo para el desarrollo de competencias

LITERATURA Y LINGÜÍSTICA

lingüísticas específicas, ya han hecho su aparición y muchos se encuentran disponibles gratuitamente en línea a través de Internet.

Las diversas (inter)disciplinas más clásicas de la lingüística tales como la psicolingüística, la neurolingüística, la dialectología, la sociolingüística y la lexicografía, entre otras, se proyectan renovadamente gracias a los instrumentos digitales que han venido en su complemento. A la vez, se ha dado origen a otras inter o transdisciplinas, por ejemplo, la psicolingüística computacional y, otras no necesariamente con base inicial en lingüística, pero sí fuertemente arraigadas en la tecnología computacional: procesamiento del lenguaje natural; recuperación, extracción y explotación de datos; tecnologías del habla, etc.

Al mismo tiempo, en la mayoría de los países hispanohablantes se ha venido tomando conciencia de la necesidad de realizar reformas educacionales centradas en formar un ciudadano capaz de construir sus propios aprendizajes, más preocupado por adquirir nuevas y mejores competencias y no tanto en acumular conocimientos específicos. Las reformas en marcha apuntan a la contextualización del conocimiento y a la colaboración en el aprendizaje.

Estas dos fuerzas, la del cambio tecnológico y la de la reforma educacional, nos han llevado a reflexionar sobre las prácticas lingüísticas en nuestros países hispanoparlantes, y muy en particular, sobre el rol de las TIC's (tecnologías de la información y de la comunicación) en la investigación en el ámbito de la lingüística y sus inter y transdisciplinas.

En este contexto, presentaremos aquí los lineamientos generales y expondremos los componentes esenciales de un programa computacional desarrollado en el marco de un proyecto de investigación mayor ([PARODI Y GAMAJO, 2003](#); [PARODI, 2004a](#) y [b](#); [MARINKOVICH y CADEMARTORI, 2004](#)), así como también daremos cuenta de una investigación en el marco de la lingüística de corpus. En la primera parte, se entrega una breve revisión del ámbito de la lingüística y los computadores. En lo que sigue se aborda la descripción y ejemplificación del programa computacional BUCOLICO (Buscador de Concordancias Lingüísticas en Corpus). En la última parte, con el doble objetivo tanto investigativo descriptivo como ejemplificador del use de la herramienta computacional, se indaga el Corpus PUCV-2003 y sus tres subcorpus: corpus técnico-científico (CTC), corpus de entrevistas orales (CEO), y corpus de literatura escrita latinoamericana (CLL). El objetivo específico de este análisis es describir comparativamente en los tres subcorpus la ocurrencia de estructuras lingüísticas tradicionalmente identificadas de manera importante con una alta carga informacional, a saber, sustantivos (comunes y propios), nominalizaciones, frases preposicionales como complemento del nombre y participios en función adjetiva.

2. Lingüística y desarrollos tecnológicos

Se debe reconocer que a pesar de grandes esfuerzos a importantes inversiones, en lo que respecta a tecnologización de la investigación en Chile -en el ámbito de la lingüística aún se está lejos de los estándares de los centros internacionales de avanzada. A pesar de ello, se realizan importantes proyectos con apoyo gubernamental (Fondecyt, Fondef, entre otros), cuyo objetivo transversal es impulsar y elevar el nivel de la investigación y de la transferencia tecnológica. No obstante lo anterior, es innegable que la tecnologización de la investigación en lingüística no se encuentra en el nivel deseado y que se debe impulsar un cambio progresivo y sostenido que permita posicionarse en el nivel internacional. Si no se logra esta actitud, se corre el riesgo de quedar postergados de manera exponencial en el breve tiempo.

En el ámbito aplicado y educacional, no es una novedad que los niveles de comprensión y producción escrita que nuestros alumnos revelan hoy en día están por debajo de lo esperado. Existe una abundante bibliografía que da cuenta de los escasos logros ([ARNOUX, NOGUEIRA y SILVESTRI, 2002](#); [PERONARD, GÓMEZ, PARODI Y NÚÑEZ, 1998](#); [PARODI, 2003](#)). Como se sabe, los programas computacionales (ya sea en CD ROM o a través de Internet) pueden ser una alternativa metodológica emergente para resolver algunas de las dificultades lingüísticas de nuestros alumnos y, al mismo tiempo, convertirse en una herramienta que posibilite cambios

LITERATURA Y LINGÜÍSTICA

metodológicos en las prácticas áulicas ([FERREIRA, CAMPOS Y RUGGERI, 1998](#); [ECHEVERRÍA, 2002](#); [ECHEVERRÍA Y RAMOS, 2002](#); [VÉLIZ, 2002](#); [FERREIRA-CARBRERA Y ATKINSON-ABUDITRY, 2002](#); [PARODI, 2002](#); [PARODI, NÚÑEZ Y GRAMARJO, 2003](#); [CHAPELLE, 2001](#); [GRAESSER, VAN LEHN, ROUSE, JORDAN Y HARTER, 2002](#); [CABOT, 2000](#); [WARSCHAUER Y KERN, 2000](#)).

Por otro lado, en el área de investigación más pura, los programas computacionales necesarios para llevar a cabo investigaciones en lo que se ha denominado *lingüística de corpus* y *lingüística computacional* ([MORENO, 1998](#)), solo se han desarrollado para el español en algunos centros académicos de Europa y, en algunos casos, en países no hispanoparlantes ([PARODI Y GRAMAJO, 2003](#); [PARODI, 2004a](#)). En efecto, la mayor disponibilidad de ellos es en lenguas diferentes al castellano.

Los avances desde las ciencias computacionales y desde áreas conexas como son las tecnologías del habla, el procesamiento del lenguaje natural, la recuperación y tratamiento de información ([JURAFSKY Y MARTIN, 2000](#); [JACKSON Y MOULINIER, 2002](#); [MANNING Y SCHÜTZE, 1999](#); [BOD, 2003](#); [JURAFSKY, 2003](#)), entre otros, hacen imprescindible que en la investigación lingüística propiamente ya no sea posible trabajar con textos o corpus ejemplares de reducido número de palabras, a no ser para investigaciones de índole exclusivamente cualitativa; es vital contar con grandes muestras de textos (de mayor poder explicativo), que requieren ser digitalizadas y procesadas con programas computacionales de alto poder descriptivo. Para estos efectos, es imprescindible contar con etiquetadores morfológicos (*taggers*) y sintácticos (*parsers*), con lematizadores y desambiguadores lingüísticos y estocásticos. También se deben construir bases de datos a interfaces que permitan la interrogación de los corpora digitalizados y marcados según las medidas estandarizadas internacionalmente (Códigos SGML: Standard Generalized Mark-Up Language). Todo ello asociado a paquetes estadísticos que permitan las correlaciones pertinentes.

3. BUCÓLICO: una propuesta en el marco de la lingüística de corpus

El programa computacional BUCOLICO (Buscador de Concordancias Lingüísticas en Corpus) tiene como propósito principal administrar los datos obtenidos a partir del estudio descriptivo realizado sobre el Corpus PUCV-2003, que se describirá más adelante, recolectados por el proyecto FONDECYT 1020786.

El procedimiento de creación de este programa incluye el modelamiento de una base de datos constituida por las frecuencias normalizadas de rasgos lingüísticos en el corpus PUCV-2003 obtenidas por medio de la interfaz de interrogación de corpus BwanaNet del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra, Barcelona, España. Esta base de datos fue diseñada utilizando el programa Microsoft Access, en tanto que el lenguaje utilizado para la programación fue "Visual Basic" .

El programa permite llevar a cabo diversas tareas de consulta sobre hábeas. Así, por una parte, es posible obtener información respecto de la constitución de cada corpus ingresado; esto posibilita conocer las áreas, las clases textuales y las siglas utilizadas en la investigación. Además, como se muestra en la [figura 1](#), el usuario tiene acceso a cada uno de los textos utilizados, así como a los sesenta y cinco rasgos lingüísticos, agrupados en 16 categorías, seleccionados para el análisis multirasgos ([PARODI, 2004b](#)).

LITERATURA Y LINGÜÍSTICA



Figura 1. Información respecto del corpora, de las áreas y las clases textuales.

Por otra parte, el programa fue diseñado para llevar a cabo análisis cuantitativo de los textos incluidos en el Corpus PUCV-2003; de esta manera, es posible conocer la frecuencia de una palabra objetivo en un texto dado o en todo el corpus. Además, junto con la frecuencia es posible conocer el cotexto oracional (concordancia), es decir, se pueden conocer las palabras a la derecha y a la izquierda que acompañan la palabra que esta siendo buscada (ver [Figura 2](#)).

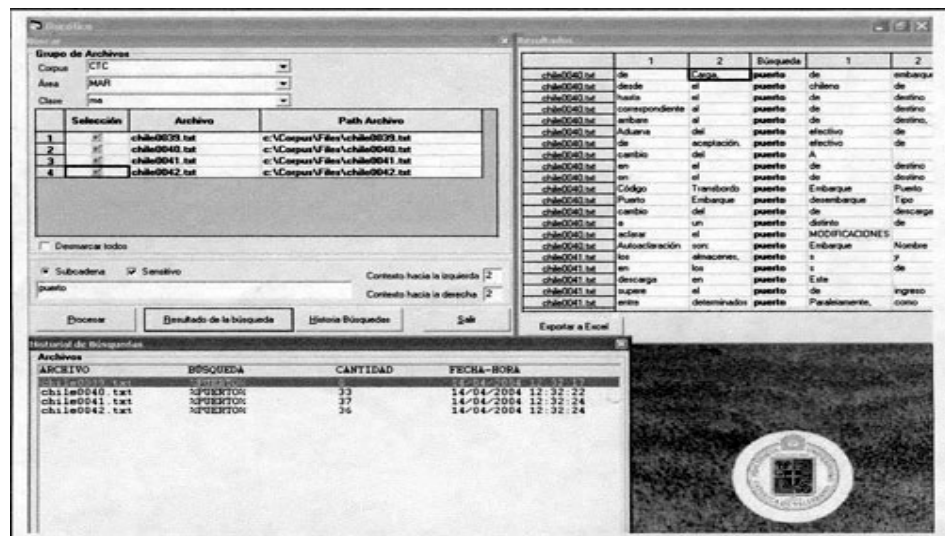


Figura 2. Herramienta de búsqueda de palabras y el conteo de frecuencia.

LITERATURA Y LINGÜÍSTICA

A modo de ejemplo, en la [Figura 2](#), se representa la búsqueda de la palabra -puerto- en los cuatro manuales del Área Marítima del corpus Técnico-Científico. Podemos observar en la pantalla *Buscar* la identificación de la clase textual, el área y el corpus en el cual se realizara la búsqueda de la palabra a investigar. La palabra objetivo puede ser buscada de forma sensitiva, es decir, tal cual como ha sido escrita, o en forma no sensitiva, esto es, se incluirán en la búsqueda las distintas maneras en las cuales ha sido escrita la palabra (mayúsculas y/o minúsculas), además se puede determinar cuantas palabras se requieren del contexto oracional. Al seleccionar *Procesar* se despliega la ventana *Resultados*: en ella se destaca la palabra objetivo y se presentan las palabras entre las cuales aparece en cada texto de la clase textual seleccionada. Por ultimo, al seleccionar Resultado de la búsqueda se despliega la ventana *Historial de Búsquedas*, en la cual se presenta el archivo en el que fue buscada la palabra, la palabra objetivo, la frecuencia de aparición de la palabra en cada texto y la fecha y hora de búsqueda. En esta ventana se va registrando la ultima búsqueda realizada, en tanto que en la ventana *Historia de Búsquedas* (no desplegada en el ejemplo) se muestran todas las búsquedas de palabras realizadas en los textos.

Otra tarea posible de llevar a cabo sobre textos no marcados es la caracterización estadística de los textos; esto es, se puede obtener información correspondiente al número de palabras, al largo promedio de palabras, al número de oraciones y al largo promedio de oraciones de cada uno de los textos a analizar, datos que son fundamentales para aplicaciones posteriores, por ejemplo, en formulas de lecturabilidad (ver [figura 3](#)).

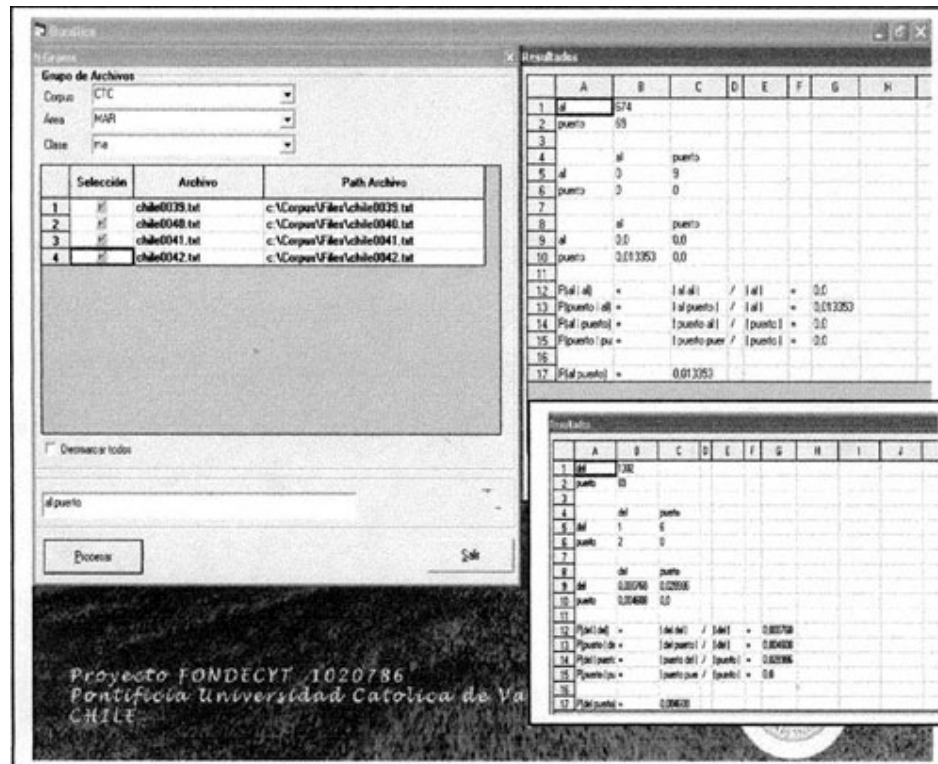
| | Selección | Archivo | Path Archivo | N° Palabras | Promedio Largo Palabra | N° Oraciones | Promedio Largo Oración |
|----|-------------------------------------|---------------|-------------------------------|-------------|------------------------|--------------|------------------------|
| 1 | <input checked="" type="checkbox"/> | chile0021.txt | c:\Corpus\Files\chile0021.txt | 10660 | 5,093743 | 1011 | 10,46409 |
| 2 | <input checked="" type="checkbox"/> | chile0022.txt | c:\Corpus\Files\chile0022.txt | 1683 | 4,935235 | 93 | 10,09670 |
| 3 | <input checked="" type="checkbox"/> | chile0023.txt | c:\Corpus\Files\chile0023.txt | 4923 | 4,968109 | 323 | 15,24149 |
| 4 | <input checked="" type="checkbox"/> | chile0024.txt | c:\Corpus\Files\chile0024.txt | 14130 | 5,277565 | 851 | 16,60399 |
| 5 | <input checked="" type="checkbox"/> | chile0025.txt | c:\Corpus\Files\chile0025.txt | 3750 | 5,0696 | 194 | 19,3299 |
| 6 | <input checked="" type="checkbox"/> | chile0026.txt | c:\Corpus\Files\chile0026.txt | 2485 | 4,810511 | 126 | 19,72222 |
| 7 | <input checked="" type="checkbox"/> | chile0027.txt | c:\Corpus\Files\chile0027.txt | 89540 | 4,934832 | 4303 | 20,57634 |
| 8 | <input checked="" type="checkbox"/> | chile0028.txt | c:\Corpus\Files\chile0028.txt | 5787 | 4,97287 | 214 | 27,04206 |
| 9 | <input checked="" type="checkbox"/> | chile0029.txt | c:\Corpus\Files\chile0029.txt | 32469 | 5,021392 | 1717 | 18,92196 |
| 10 | <input checked="" type="checkbox"/> | chile0030.txt | c:\Corpus\Files\chile0030.txt | 7429 | 5,25212 | 478 | 15,54184 |
| 11 | <input checked="" type="checkbox"/> | chile0031.txt | c:\Corpus\Files\chile0031.txt | 8970 | 4,619175 | 426 | 21,05634 |
| 12 | <input checked="" type="checkbox"/> | chile0032.txt | c:\Corpus\Files\chile0032.txt | 15019 | 4,733671 | 824 | 18,22694 |
| 13 | <input checked="" type="checkbox"/> | chile0033.txt | c:\Corpus\Files\chile0033.txt | 10055 | 5,16191 | 659 | 15,25797 |
| 14 | <input checked="" type="checkbox"/> | chile0034.txt | c:\Corpus\Files\chile0034.txt | 9967 | 4,910304 | 483 | 20,63561 |
| 15 | <input checked="" type="checkbox"/> | chile0035.txt | c:\Corpus\Files\chile0035.txt | 18349 | 5,299362 | 1649 | 11,12735 |

Otro análisis posible de realizar es de tipo probabilístico, N-Grams, en el cual se calcula un conjunto finito de variables aleatorias encadenadas (secuencia de palabras) que tienen una probabilidad conjunta. Es un modelo que intenta predecir la ocurrencia de una palabra dada una secuencia de palabras que la preceden.

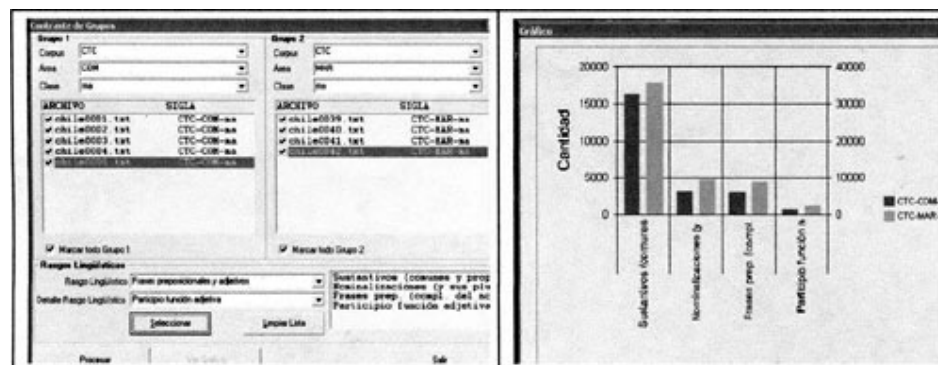
Por último, en relación a los textos no marcados, se pueden llevar a cabo análisis de frecuencias asociadas a palabras con y sin repetición, en uno o más archivos de texto seleccionados, esto es conocido como análisis Type/Token.

LITERATURA Y LINGÜÍSTICA

En la [Figura 4](#), se presenta un breve ejemplo del use de la herramienta N-Grams, la cual determina la probabilidad de aparición de una secuencia de palabras. En este caso, se buscaron en los manuales del Área Marítima la secuencia -al puerto- y -del puerto-, obteniéndose una probabilidad mayor para la primera secuencia (0,0133) que para la segunda (0,0046).



Por otra parte, el programa almacena una matriz de frecuencias, obtenida a partir del análisis de frecuencia realizado de los rasgos lingüísticos en cada uno de los textos constituyentes del corpus de textos marcados morfológicamente. Esta matriz se construyó utilizando la interfaz BwanaNet del Instituto Universitario de Lingüística Aplicada (IULA), dependiente de la Universidad Pompeu Fabra, Barcelona. Los resultados fueron almacenados en el programa en forma de una base de datos, a partir de la cual se pueden comparar y graficar las frecuencias de aparición de rasgos lingüísticos entre los distintos corpus y/o textos, tal como lo muestra la [figura 5](#).



LITERATURA Y LINGÜÍSTICA

Con esta herramienta es posible comparar la frecuencia de una selección de rasgos en dos grupos de textos distintos. En la [Figura 5](#), se presenta una comparación de las frecuencias de los rasgos sustantivos, nominalizaciones, frases preposicionales y participio en función adjetiva en los manuales del Área Comercial versus el Área Marítima. Como se observa en el gráfico resultante, los manuales del Área Marítima presentan mayor frecuencia de los rasgos seleccionados que los manuales del Área Comercial. Es relevante comentar que todos los resultados obtenidos a partir de estas herramientas son exportables al programa Microsoft Excel para posteriores análisis y graficación.

En síntesis, BUCOLICO es un programa que permite administrar la información obtenida a partir de textos etiquetados morfológicamente, en forma de matriz de frecuencias, y procesar textos en lenguaje natural. Cabe señalar la importancia que tiene la integración multidisciplinar en el desarrollo de aplicaciones tecnológicas de este tipo (lingüistas, ingenieros computacionales, estadísticos, diseñadores gráficos, entre otros), y que permitió desarrollar una herramienta comparativamente rudimentaria con los avances tecnológicos detectados fuera de Latinoamérica, pero que se constituye en un primer avance en el diseño de instrumentos computacionales elaborados en el marco del programa de investigación en desarrollo en la Pontificia Universidad Católica de Valparaíso, Chile.

4. La investigación PUCV

Tal como se dijo en la introducción de este artículo, en los apartados precedentes se ha presentado el programa computacional que ha permitido el desarrollo de la siguiente investigación empírica, descriptiva y comparativa de tres corpus de textos originales digitalizados. En lo que sigue, se entregan tres tipos de resultados empíricos preliminares: (1) una descripción global del Corpus PUCV-2003, (2) una breve descripción de los rasgos Lingüísticos a indagar en el corpus, y (3) comparación de frecuencias de los rasgos distintivos de la información - sustantivos, nominalizaciones, frases preposicionales como complemento de nombre y participios en función adjetiva - en los corpus, las áreas del corpus CTC y en la clase textual Manual Técnico del corpus CTC.

4.1. Descripción del Corpus PUCV-2003

La conformación general del Corpus PUCV-2003 se desglosa en 90 textos que equivalen a 1.466.744 palabras. Este corpus general está dividido en tres grandes registros o subcorpora (Corpus Técnico-Científico -CTC-, Corpus de Literatura Latinoamericana Escrita -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 recolectó otros dos corpora, a saber, el CLL y el CEO. La siguiente [tabla](#) muestra su distribución en número de textos y palabras.

| Tipo de Corpus | Número archivos o textos | de 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.

LITERATURA Y LINGÜÍSTICA

4.1.1. Corpus de textos Técnico-Científicos (CTC)

Tal como se muestra en la [Tabla 1](#), 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 diferentes orientaciones profesionales. Estas tres diferentes áreas del conocimiento técnico especializado dicen relación con la formación de tres diferentes profesionales técnicos, a saber, sector marítimo (Especialidad Operación Portuaria), sector metalmeccánico (Especialidad Mecánica Industrial), y sector de administración y comercio (Especialidad Contabilidad). 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. Los tipos de textos que conforman el corpus CTC corresponden a:

| | | Nº Textos | Nº Palabras |
|---|---------------------|-----------|-------------|
| 1 | Artículo Técnico | 1 | 9.346 |
| 2 | Descripción Técnica | 8 | 25.170 |
| 3 | Diagrama | 2 | 116 |
| 4 | Formulario | 3 | 2.287 |
| 5 | Guía Didáctica | 14 | 20.063 |
| 6 | Glosa Legal | 2 | 4.142 |

LITERATURA Y LINGÜÍSTICA

| | | | |
|----------------|----------------|-----------|----------------|
| 7 | Glosario | 4 | 9.747 |
| 8 | Instructivo | 9 | 17.864 |
| 9 | Leyes | 3 | 67.905 |
| 10 | Manual Técnico | 24 | 463.468 |
| 11 | Reglamento | 2 | 4.485 |
| 12 | Tablas | 2 | 2.197 |
| Totales | | 74 | 626.790 |

Tabla 3. Clases Textuales del Corpus Técnico-Científico

Su distribución porcentual a través del corpus CTC es la que se presenta en el siguiente gráfico:



Gráfico 1. Porcentaje de distribución de palabras entre las clases textuales del Corpus CTC

Como se observa, los manuales técnicos concentran cerca del 74% del total de palabras de la muestra de textos, haciendo de esta clase textual la más representativa del corpus CTC. Es debido a esto último, que se optó por utilizar esta clase textual por área de especialización (Marítima, Comercial a Industrial) como submuestra de las clases textuales en la investigación que se presentara más adelante.

Según lo expuesto anteriormente, a continuación se entrega una definición operacional de Manual Técnico, así como la distribución de los textos y cantidad de palabras de los manuales por área (ver [Tabla 4](#)).

***“Manual Técnico:** Es un tratado de carácter didáctico enmarcado dentro de un área profesional técnico-científica. Es rico en ejemplos, tablas y recursos multimodales, lo que facilita su comprensión y el acceso a información especializada. Su función primordial es la referencial, pudiendo tener*

LITERATURA Y LINGÜÍSTICA

secundariamente una función apelativa. Su estructura textual predominante es expositiva-normativa". (PARODI Y GRAMAJO, 2003: 218).

| Manuales por Área | Número de textos | Número de palabras |
|---|------------------|-----------------------|
| Marítima (Operación Portuaria) | 4 (16%) | 84.484 (18,2%) |
| Industrial (Mecánica) | 15 (62.5%) | 242.244 (52,2%) |
| Administración y Comercio (Contabilidad) | 5 (20,8%) | 137.225 (29,6%) |
| Totales | 24 (100%) | 463.953 (100%) |

Tabla 4. Distribución de textos y palabras en los Manuales Técnicos del CTC

4.1.2. Rasgos lingüísticos

Respecto a los rasgos lingüísticos indagados, se elaboró un conjunto inicial de dieciséis categorías representativas de características gramaticales y funcionales del español (PARODI, 2004a y b). Estos rasgos lingüísticos fueron rastreados a partir de bibliografía relevante en el tema y atendiendo a las posibilidades comunicativa-funcionales de estos. A partir de estas dieciséis categorías iniciales, se procedió a construir una matriz más específica de los rasgos caracterizadores del español según cada una de estos grandes lineamientos preliminares; de este modo, se llega a un total de sesenta y cinco rasgos lingüísticos de importancia gramatical y funcional. A continuación, en la [Tabla 5](#), se presenta el total de estos rasgos, agrupados según las categorías iniciales.

LITERATURA Y LINGÜÍSTICA

| 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) | 1. Tipos verbales |
| 4. Presente (indicativo y subjuntivo) | 35. Públicos |
| 5. Futuro (indicativo y subjuntivo) | 36. Privados |
| 6. Futuro perifrástico | 37. Persuasivos |
| B. Marcadores de modo verbal | 38. Perceptivos |
| 7. Indicativo/imperativo | J. Verbos modales |
| 8. Subjuntivo/imperativo | 39. Posibilidad |
| 9. Modo indicativo | 40. Necesidad |
| 10. Modo subjuntivo | 41. Obligación |
| 11. Modo imperativo | 42. Volición |
| C. Desinencias verbales de persona | K. Marcadores de modalidad |
| 12. Primera singular | 43. Atenuadores |
| 13. Segunda singular | 44. Enfáticos |
| 14. Tercera singular | L. Adverbios |
| 15. Primera plural | 45. De lugar |
| 16. Segunda plural | 46. De tiempo |
| 17. Tercera plural | 47. De modo |
| D. Pronombres personales | 48. De cantidad |
| 18. Primera persona singular | M. Marcadores de subordinación |
| 19. Primera persona plural | 49. Subordinadas sustantivas con "que" |
| 20. Segunda persona singular | 50. Subordinadas adjetivas pron. relativo |
| 21. Segunda persona plural | 51. Subordinadas adverbiales de razón o c/e |
| 22. Tercera persona singular | 52. Subordinadas adverbiales de concesión |
| 23. Tercera persona plural | 53. Subordinadas adverbiales condicionales |
| 24. Demostrativo | 54. Subordinadas adverbiales de tiempo |
| E. Formas nominales | 55. Frases infinitivas en función nominal |
| 25. Nominalizaciones | N. Frases preposicionales y adjetivos |
| 26. Sustantivos (comunes y propios) | 56. Frases prep. (compl. del nombre) |
| F. Formas Pasivas | 57. Adjetivos atributivos (calificativos) |
| 27. Pasivas con "se" | 58. Adjetivos predicativos |
| 28. Pasivas con ser sin agente | 59. Adjetivos demostrativos |
| 29. Pasivas con ser con agente | 60. Participio función adjetiva |
| 30. Pasivas con estar | N. Marcadores de Coordinación |
| G. Especificidad lexical | 61. Conjunciones adversa., adit. y disyun. |
| -37-. Relación type/token por forma | O. Marcadores de negación |
| 32. Relación type/token por lema | 62. Adverbio de negación |
| | 63. Adverbios de tiempo |
| | 64. Conjunción de negación |
| | 65. Pronombres de negación |

Tabla 5. Rasgos lingüísticos caracterizadores del español

LITERATURA Y LINGÜÍSTICA

5. Resultados

5.1. Estudio de los rasgos distintivos de la información

En lo que sigue se da cuenta de un estudio de frecuencias en base a cuatro rasgos lingüísticos, a saber: sustantivos, nominalizaciones, frases preposicionales y participios en función adjetiva. El objetivo de este estudio es describir las ocurrencias de estos rasgos en los corpus, en las áreas del corpus CTC y en la clase Manual Técnico de cada área del corpus CTC. Ello con el objetivo específico de explorar la relevancia y posible carácter diferenciador de un conjunto de rasgos caracterizadores de informatividad en los textos y áreas técnico-científicas en comparación

La presencia de estos rasgos en los textos permite identificar, según diversos autores ([PARODI, 2004b](#), [BURDACH, 2000](#), [PICALLO, 1999](#), [CHAFE, 1982, 1985](#); [JANDA, 1985](#); [BIBER, 1986, 1988](#)), la presencia de integración y compactación de información altamente abstracta, propia de un discurso de tipo referencial.

En términos funcionales más específicos, podemos decir que los sustantivos son los portadores principales del significado referencial en un texto y una alta frecuencia de sustantivos en un texto es un indicador de alta densidad de información ([BIBER, 1988](#)). En tanto las nominalizaciones permiten integrar información en pocas palabras ([CHAFE, 1982, 1985](#)) y reducir oraciones completas en series más compactas y eficientes de frases nominales ([BURDACH, 2000](#); [JANDA, 1985](#)), además tienen la función de transportar información altamente abstracta ([BIBER, 1986](#)). De la misma manera las frases preposicionales, en nuestro caso como complemento del nombre, sirven para integrar altas cantidades de información en un texto. Además se establece que las preposiciones son un dispositivo importante por condensar altas cantidades de información ([BIBER, 1988](#)). [CHAFE \(1982, 1985\)](#) especifica que las preposiciones funcionan como dispositivos para la integración de la información en unidades de sentido y para la expansión de la cantidad de información contenida dentro de una unidad de sentido. Por otra parte, los estudios que tratan acerca del participio lo sitúan preferentemente en el discurso escrito más que en el oral y su interpretación usual es que son usados para la integración y elaboración estructurales ([BIBER, 1988](#); [CIASPUSCIO, 1992](#)). [JANDA \(1985\)](#) establece que se utilizan en la toma de apuntes porque son más compactos e integrados y por ello sirven para la producción de un discurso altamente informativo cuando el tiempo es limitado.

En síntesis, los rasgos lingüísticos que hemos agrupado aquí están fundamentalmente orientados a cumplir una función distintiva respecto de la información en los textos. Esto es, los rasgos distintivos de la información (RDI) permiten reconocer en los textos una alta concentración de la información en unidades y estructuras lingüísticas más pequeñas, que presentan los datos lo más concisa y precisamente posible ([BIBER, 1988](#); [HALLIDAY Y MARTIN, 1993](#); [BURDACH, 2000](#)), revelando una alta carga informativa fundamentalmente de carácter referencial.

El estudio de frecuencias que a continuación presentamos, utilizando los datos proporcionados por el programa BUCOLICO y graficados en Excel, tiene por objetivo ejemplificar la utilidad del programa y, además, describir la distribución porcentual de los RDI en tres niveles de análisis. Un primer nivel de análisis es el que corresponde a la comparación de los RDI entre los tres grupos de textos del Corpus PUCV2003; un segundo nivel de análisis toma en cuenta la comparación de los RDI entre las tres áreas técnicas del corpus CTC y un tercer nivel de análisis corresponde a la comparación de los RDI entre los manuales técnicos del corpus CTC. Para hacer posible las comparaciones de frecuencias entre los diversos niveles se han normalizado las frecuencias de ocurrencias multiplicando cada frecuencia por 1000 y luego dividiendo el resultado por el número de palabras de cada texto ($FrN = FrR \times 1000 / Tpal.$). Así, por ejemplo, si queremos estimar la frecuencia normalizada de rasgo sustantivo en los tres corpus:

LITERATURA Y LINGÜÍSTICA

| Corpus | FrR (sustantivo) | Base de normalización | Tpal | FrN |
|--------|------------------|-----------------------|---------|--------|
| CTC | 143508 | X 1000 | 626.790 | 17730 |
| CLL | 85052 | X 1000 | 459.860 | 2255 |
| CEO | 44219 | X 1000 | 380.094 | 466,72 |

Tabla 5. Ejemplo de normalización

Si observamos los datos no normalizados (FrR) de la [Tabla 6](#), podemos establecer que si los comparamos -sin considerar el número total de palabras- la diferencia existente entre el CTC y el CLL es apenas de 1,7 veces, en tanto que la diferencia entre el CTC y el CEO es de 3,2 veces. En tanto que si normalizamos las frecuencias (FrN) la diferencia entre el CTC y el CLL es de 7,9 veces y entre el CTC y el CEO es de 38 veces. Esto nos demuestra que las diferencias son mucho más significativas de lo que era posible observar sin considerar el total de palabras de cada corpus, siendo los resultados normalizados mucho más significativos y confiables.

Nivel I: RDI por corpus

En este primer nivel de análisis, describiremos las ocurrencias de los RDI en los tres corpus de PUCV2003. De esta manera, en el [gráfico 2](#) se presenta la distribución porcentual de los RDI agrupados.

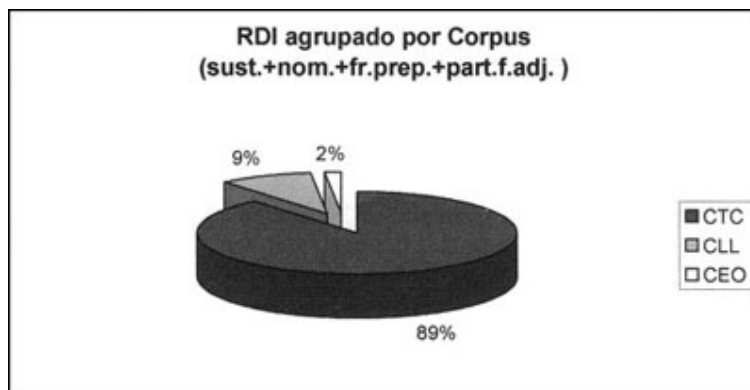


Gráfico 2. Porcentaje de los RDI en el Corpus PUCV-2003

Según se observa en el [gráfico 2](#), la mayor presencia de RDI se encuentra en el corpus CTC, con un 89%, lo que da cuenta del carácter distintivo de la compactación e integración de la información en los textos de tipo técnico-científico. Los textos de índole narrativo-literario y de entrevistas orales no destacan por una alta ocurrencia de estos rasgos, lo que los posiciona en otro eje del continuum informatividad/no-informatividad. PARODI (2004b) muestra, por medio de un estudio estadístico de análisis factorial, que estos dos grupos de textos se caracterizan principalmente por lo que se ha denominado un Foco Narrativo y un Foco Contextual a Interactivo, cuyos rasgos lingüísticos caracterizadores son los adverbios de tiempo y lugar, el tiempo presente y pasado, la incrustación de personas de primera persona singular, entre otros.

LITERATURA Y LINGÜÍSTICA

En este sentido, este nuevo estudio de frecuencias comparativas más detallado confirma las distinciones ya detectadas y nos arroja nuevos datos mas precisos.

En el [gráfico 3](#) se observa que, en los tres corpus, el sustantivo se presenta como el rasgo distintivo de la información más frecuente. Sin embargo, y en consonancia con el [gráfico 2](#), podemos determinar que el sustantivo es 7,8 veces más frecuente en el CTC que en el CLL y 37,9 veces más frecuente que en el CEO. Por otra parte, se detecta que la frase preposicional secunda en frecuencia al sustantivo en el CTC con una frecuencia normalizada de 3731 ocurrencias, teniendo una frecuencia 4,8 veces menor a la del sustantivo en el CTC. Además, la frase preposicional en el CTC es 17 veces más frecuente que en el CLL y 141 veces más que en el CEO. La nominalización, con una frecuencia muy similar a la frase preposicional en el CTC, se ubica en tercera posición, siendo 24 veces más frecuente que en el CLL y 134 veces más frecuente que en el CEO. Por último, el participio en función adjetiva, es el RDI que se presenta con menor frecuencia en el CTC con un frecuencia de apenas 829, siendo 21,3 veces menos frecuente que el sustantivo. Sin embargo, este rasgo es 12 veces más común en el CTC que en el CLL y es 200 veces más frecuente que en el CEO.

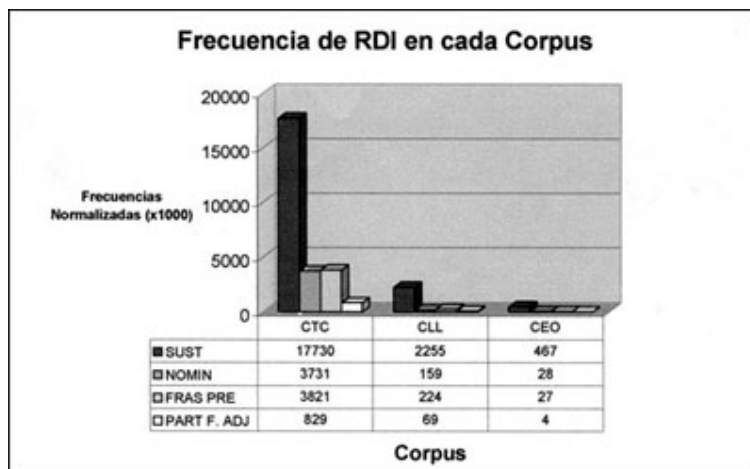


Gráfico 3. Los RDI en los tres tipos de corpus del PUCV-2003

En síntesis, es posible establecer fehacientemente que los RDI se presentan con mucha mayor frecuencia en el corpus CTC, reafirmando con esto la función distintiva de estos rasgos en relación a la concentración y densidad de la información, planteada ya por diversos investigadores. En este sentido, podemos aseverar que los textos incluidos en el corpus CTC presentan una alta densidad informativa y una mayor compactación e integración de la información que en los textos de tipo literarios y de entrevistas orales.

LITERATURA Y LINGÜÍSTICA

Nivel 2: RDI por área de especialización

En lo que viene, se indagan los cuatro RDI en las tres áreas de especialización técnico-científica del corpus PUCV-2003. El siguiente gráfico muestra las cifras porcentuales:

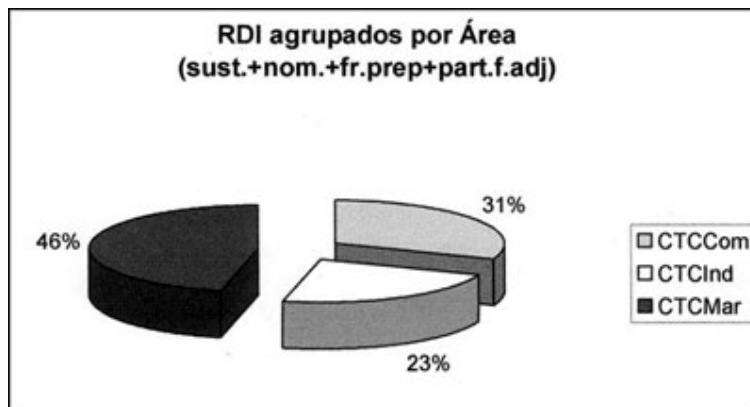


Gráfico 4. Los RDI en conjunto en las tres áreas del CTC.

En el [Gráfico 4](#) se aprecia que la mayor frecuencia porcentual de RDI se concentra en el Área Marítima del CTC con 46%, seguida en porcentaje por el Área Comercial del CTC y finalmente por el Área Industrial del CTC. Es interesante comprobar, al menos en este punto, que los textos de las tres áreas técnico-científicas no presentan una gran homogeneidad en cuanto a la ocurrencia de estos rasgos.

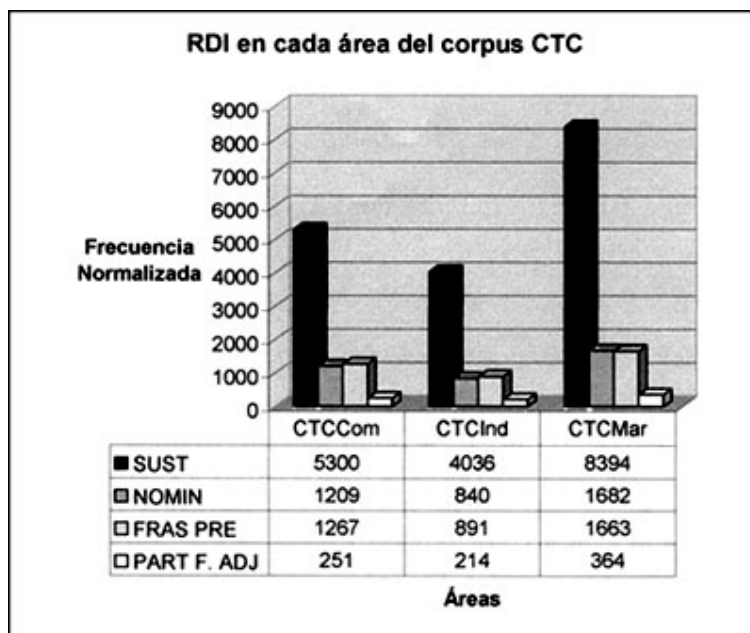


Gráfico 5. Los RDI en las tres áreas técnicas del CTC.

LITERATURA Y LINGÜÍSTICA

En este gráfico se puede constatar que el rasgo sustantivo presenta la mayor frecuencia entre los cuatro RDI en estudio al interior de cada una de las tres áreas y que es nuevamente el Área Marítima la que aparece con una mayor frecuencia normalizada de 8.394 sustantivos, siendo 1,6 veces más frecuente que en el Área Comercial que presenta una frecuencia normalizada de 5.300 y 2,1 veces más frecuente que en el Área Industrial, que muestra una frecuencia de 4036. En cuanto al rasgo nominalización, es posible determinar que predomina en el Área Marítima, presentando una frecuencia de 1.682 ocurrencias, esto es, 1,4 veces más que en el Área Comercial y el doble más frecuente que en el Área Industrial. El rasgo frase preposicional, levemente menos frecuente que la nominalización, también predomina en el Área Marítima por sobre las otras áreas, siendo apenas 1,3 veces más frecuente que en el Área Comercial y 1,9 veces más que el Área Industrial. Por último, el rasgo participio en función adjetiva, que se presenta como el rasgo menos frecuente en todas las áreas, predomina en el Área Marítima con 364 ocurrencias, siendo 1,5 veces más frecuente que en el Área Comercial y 1,7 veces más frecuente que en el Área Industrial.

Cabe señalar que en todas las áreas los rasgos presentan una distribución de frecuencias muy similar, esto es, predomina el sustantivo, seguido ya sea por la nominalización o la frase preposicional, normalmente con diferencias muy bajas y finalmente, seguido por el participio en función adjetiva.

En vista de estas cifras, es posible establecer que entre las áreas la distribución de los rasgos es muy similar, existiendo una alta densidad informacional, así como compactación a integración de la información en estos textos. Sin embargo, cabe señalar que el Área Marítima presenta una mayor frecuencia de estas características, confirmandose lo representado en el [gráfico 4](#). Resulta interesante comprobar que son los sustantivos, en primera instancia, los que se constituyen en el rasgo lingüístico caracterizador de la informatividad en estos corpus; no obstante ello, se debe recordar que una determinada función como la informatividad se compone de un conjunto de rasgos que co-ocurren sistemáticamente y por ello la co-ocurrencia significativa de estos cuatro elementos gramaticales dan fuerza a la indagación específica.

Nivel 3: RDI en los Manuales Técnicos del CTC

A continuación, se indagan los cuatro RDI en los veinticuatro Manuales Técnicos del CTC que, en términos de palabras, dan cuenta del 74% del corpus PUCV-2003.

En el [Gráfico 6](#), se presenta la distribución de los RDI en los Manuales Técnicos (MT), agrupados por cada área del CTC. Podemos observar que los MT del Área Industrial son los que reúnen el mayor porcentaje de frecuencias, 58% del total. Por otra parte, los MT del Área Comercial presentan la segunda mayor frecuencia porcentual, con un 26%, seguido los MT del Área Industrial con un 16% del total.



Gráfico 6. Los RDI en los Manuales Técnicos del CTC.

LITERATURA Y LINGÜÍSTICA

En el siguiente gráfico se entrega un análisis más pormenorizado de los datos y cifras de los cuatro RDI en el MT por cada ámbito de especialización.

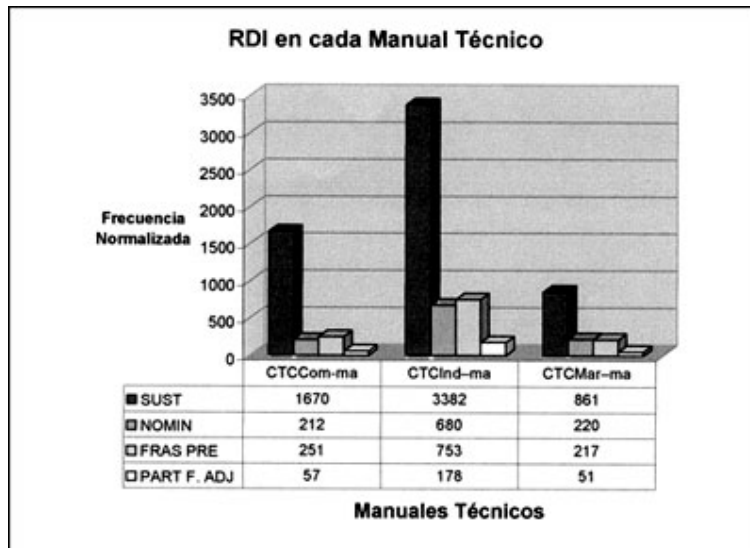


Gráfico 7. Los RDI en los Manuales Técnicos por área de especialización.

En este gráfico se detecta que el sustantivo es el rasgo que predomina en todos los manuales, sin embargo, es mucho más frecuente en los manuales del Área Industrial con una frecuencia de 3.382 ocurrencias, siendo el doble más frecuente que en los manuales del Área Comercial y 3,9 veces más frecuente que en los manuales del Área Marítima. Por otra parte, la frase preposicional es el segundo más frecuente de los rasgos en las áreas industrial y comercial, en tanto que en la marítima se ubica bajo la nominalización con una diferencia mínima entre ambos rasgos. En este sentido, la frase preposicional es 3 veces más frecuente en el Área Industrial que en la comercial y 3,4 veces más frecuente que en la marítima. La nominalización se presenta como el tercer rasgo más frecuente en las áreas industrial y comercial y el segundo más frecuente en el Área Marítima. La presencia de nominalización en el Área Industrial es 3,2 veces más frecuente que en el Área Comercial y 3 veces más frecuente que en el Área Marítima. Finalmente, el rasgo participio en función adjetiva se presenta como el rasgo menos frecuente en todos los manuales, siendo 3,1 veces más frecuente en los manuales del Área Industrial que en el Área Comercial y 3,4 veces más frecuente que en el Área Marítima.

En suma, podemos establecer que si se observa la presencia de los RDI en los Manuales Técnicos de las tres áreas del CTC, la mayor frecuencia de estos rasgos se presenta en los manuales del Área Industrial y que el rasgo sustantivo es el más frecuente en todas las áreas, seguido por la frase preposicional, y la nominalización en proporciones muy similares. Esto nos permite pensar que los manuales del Área Industrial tienen una mayor densidad informacional y que compactan a integrar información con mayor frecuencia que los manuales de las otras áreas. Al respecto, [MARINKOVICH Y CADEMARTORI\(2004\)](#) por medio de un estudio de tipo cualitativo muestran que algunos de estos manuales técnicos contienen secuencias de tipo narrativo en su carácter de textos académicos de divulgación didáctica. Es factible, desde esta óptica, sugerir que la posible diferencia porcentual de los RDI tenga alguna relación con lo anteriormente comentado, esto es, que los manuales de las áreas comercial y marítima pudieran presentar un estilo más didáctico acudiendo a cierto tipo de estrategias de reformulación divulgativas y, por ende, podrían revelar una comparativa y relativa menor cantidad de RDI.

LITERATURA Y LINGÜÍSTICA

5.2. Las nominalizaciones y los sufijos derivacionales nominales en el Manual Técnico

Dado que el estudio ha permitido distinguir entre sustantivos propiamente tales y las denominadas nominalizaciones que se constituyen en sustantivos de tipo derivacional a partir de otra categoría gramatical, resulta interesante explorar las frecuencias de ocurrencias de las nominalizaciones y los diversos mecanismos de sufijación derivacional. Para ello, focalizaremos el análisis de las nominalizaciones por área de especialización en la clase textual Manual Técnico (MT), ya que -como se ha dicho anteriormente- esta clase textual representa el 74% de las palabras del Corpus CTC. La cuantificación de las formas nominalizadas en cada una de las tres áreas de especialización arrojó los siguientes resultados:



Gráfico 8. Ocurrencia de las nominalizaciones en los manuales.

De acuerdo a las cifras entregadas en este gráfico, se puede apreciar la relativamente homogénea distribución porcentual de la ocurrencia de las nominalizaciones. Se comprueba así que su ocurrencia es sumamente pareja en las tres áreas de especialización (comercial, industrial y marítima), solo con un leve aumento de 7 puntos porcentuales en el Área Marítima. Ello indica que -en cuanto a este recurso lingüístico- esta clase textual se presenta con bastante homogeneidad, siendo su ocurrencia una estructuración gramatical importante como modo de presentación a integración de la información que pretende llegar a una audiencia no experta en las temáticas tratadas.

Las nominalizaciones constituyen un tipo de las llamadas metáforas gramaticales, término acuñado por [HALLIDAY \(1993\)](#) desde la lingüística sistémica funcional, mediante el cual se realiza un doble proceso de compactación de información y de recategorización de unidades lingüísticas. Esta transformación es un desplazamiento desde una estructura lingüística a otra en que la semántica suele permanecer intacta o escasamente alterada, pero el modo lingüístico de su expresión ha variado, muchas veces comprimiéndose cierta información y ejecutándose un proceso de reducción de piezas lingüísticas. En el siguiente gráfico, se entregan cifras de la ocurrencia de sustantivos formados por sufijos derivativos, típicos de la constitución de grupos nominales del tipo que nos interesa. En particular, hemos seleccionado un grupo de los sufijos derivativos más característicos del español.

Con el fin de profundizar en estos resultados, a continuación, se explora un conjunto de sufijos derivacionales que destacan por constituir sustantivos típicos de núcleos en nominalizaciones del español. El [Gráfico 9](#) muestra la ocurrencia en los MT.

LITERATURA Y LINGÜÍSTICA

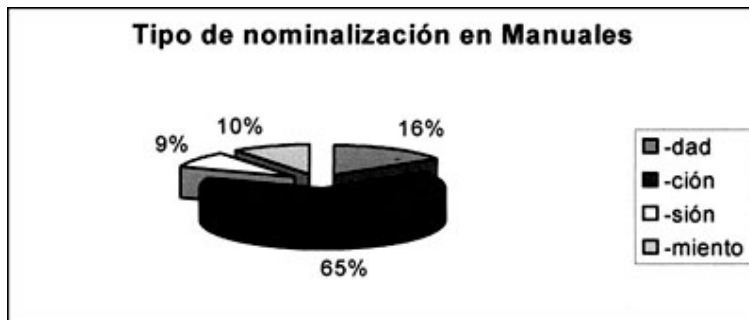


Gráfico 9. ocurrencia de los tipos de nominalizaciones en los manuales de CTC.

Según se aprecia en este gráfico, la mayor ocurrencia del tipo de derivación nominal se presenta con el sufijo *-ción* (65%), en tanto con porcentajes bastante bajos se observan las ocurrencias de *-dad*, *-miento* y *-sión* (16%, 10% y 9%, respectivamente). Si recordamos que según el DRAE (1992) el sufijo *-ción* se utiliza para formar sustantivos verbales, que expresan acción y efecto y que además de su significado abstracto, *-ción* y sus variantes pueden denotar objeto, lugar, etc., es posible sostener que la nominalización terminada en *-ción* es un tipo de agrupación nominal, cuya función principal es *-empaquetar-* procesos verbales a través de los cuales se expresan acciones (con agentes, procesos, meta) y efectos (de la acción realizada por alguien sobre algo y su(s) correspondientes consecuencias).

Como se sabe, las nominalizaciones son un recurso que, entre otros, sirve para establecer la cadena correferencial, ya que a través de ellas se puede retomar y resumir lo expresado en otra parte del texto; de este modo, las metáforas gramaticales o nominalizaciones son un medio eficiente para dar garantía del establecimiento de cohesión textual por parte de un lector experto. Una alta ocurrencia de nominalizaciones, tal como la detectada en los Manuales Técnico-Científicos del corpus PUCV-2003, es prueba de un tipo de discurso que destaca por complejos encadenamientos de razonamientos típicos de textos científicos ([HALLIDAY, 1993](#); [MARTIN, 1993](#)) y que apunta a un alto grado de abstracción, hechos que, seguramente, implican una serie de dificultades para su comprensión por parte de lectores semi-legos y no poseedores de estrategias de lectura eficientes. Desde este punto de vista, si bien los manuales técnicos son un tipo de texto que debe apoyar al inexperto en la materia a acercarse a ella de modo casi iniciático, su organización compleja -según los datos aquí entregados- exigiría un manejo lingüístico experto.

De acuerdo a los distintos tipos de derivación nominal; seleccionados en esta investigación, podemos establecer que el sufijo *-ción* es el que predomina ante los otros tipos de derivación nominal. Es, por tanto, el tipo de nominalización más utilizada en los Manuales Técnicos del corpus PUCV-2003.

6. Conclusiones

En primer lugar, cabe señalar que la herramienta computacional BUCOLICO mostró ser un recurso tecnológico poderoso para la interrogación y análisis de corpus marcados y no marcados morfológicamente. Parte de los datos aquí estudiados no han requerido un etiquetaje lingüístico, sino que se han trabajado como formas gramaticales en su ocurrencia normalizada a partir de los textos del corpus PUCV-2003. Por una parte, esto muestra la utilidad y posibilidad de agregar nuevos corpus al programa BUCOLICO sin la necesidad de contar con tecnología que previamente implique su marcaje automático. Por otra, se revelan las proyecciones de investigaciones, apoyadas en este tipo de recursos digitales, que trabajen sobre extensos corpus de textos lingüísticos dando así origen a resultados mucho más robustos y confiables. También se debe destacar que esta herramienta, relativamente rudimentaria, construida sin altos costos

LITERATURA Y LINGÜÍSTICA

monetarios ni conocimientos de alta especialización, pero sí a partir de un equipo multidisciplinario, se constituye en una prueba concreta de que en nuestros ámbitos académicos es factible efectuar desarrollos entre lingüística e informática de manera contundente.

Ahora bien, en cuanto al estudio de los cuatro rasgos de distinción informativa (RDI), una vez descritas las frecuencias porcentuales y normalizadas (x1000) de los RDI en cada uno de los corpora (CTC, CLL, CEO), en cada una de las áreas del CTC (Comercial, Marítima e Industrial) y en los manuales de cada área del CTC, es posible concluir que los RDI estudiados son mucho más frecuentes en el CTC que en los otros dos corpus del PUCV-2003, siendo el sustantivo el rasgo que predomina en el CTC, así como en los otros corpus. En un segundo orden de frecuencias aparece la nominalización y la frase preposicional, resultando mucho mayor la presencia de estos rasgos en el CTC que en el CLL y en el CEO. Estas amplias diferencias en la ocurrencia en el CTC nos permite establecer que este corpus se distingue por la alta densidad informativa, así como por la compactación e integración de la información, en congruencia con otro tipo de resultados como los aportados por [PARODI \(2004b\)](#).

Más específicamente, respecto a las áreas técnico-científicas del CTC, es posible precisar que la mayor frecuencia de RDI ocurre en el Área Marítima, replicándose la distribución de los rasgos vista en los corpus a través de cada una de las áreas. Esto permite pensar que los textos incluidos en esta área presentan una función referencial muy marcada, asociada a una alta densidad informativa.

En relación al tercer nivel de análisis podemos establecer que, comparativamente, los Manuales Técnicos del Área Industrial presentan una frecuencia de RDI correspondiente a más del doble de estos rasgos en los Manuales Técnicos de las áreas comercial y marítima. Se reconoce nuevamente el patrón de distribución de los cuatro rasgos en donde el sustantivo ocupa la primera posición en todos los manuales, destacando considerablemente la muy alta frecuencia de estos en el Área Industrial; así también aparecen la frase preposicional y la nominalización, esta vez algo más diferenciadas entre sí en favor de la frase preposicional. Y, finalmente, aparece el participio en función adjetiva ocupando el último lugar en la distribución de frecuencias, tal como ha aparecido en todos los niveles estudiados. Estos datos confirman la definición de Manual Técnico presentada por [PARODI Y GRAMAJO \(2003\)](#), ya que estos rasgos lingüísticos nos permiten, desde un análisis más cuantitativo, identificar la función referencial y la orientación informativa predominante en textos de este tipo.

El análisis más fino llevado a cabo sobre las nominalizaciones y los sufijos derivacionales constitutivos de las mismas, permite concluir que estas organizaciones lingüísticas que recategorizan elementos gramaticales y compactan, muchas veces, altas cantidades de información son un recurso importante en los Manuales Técnicos de los tres ámbitos de especialización en indagación. El sufijo derivacional *-ción* se posiciona como el mayoritariamente empleado para efectuar este proceso. Sin lugar a dudas, estas construcciones que aglutinan información se transforman en unidades de importante abstracción y se vuelven cruciales para el adecuado procesamiento del discurso y su correspondiente comprensión lingüística.

Por último, los datos aportados en esta investigación constituyen antecedentes vitales para pensar en la necesidad de una didáctica del discurso especializado escrito. Si bien es cierto en los últimos años se han hecho esfuerzos ingentes para apoyar la educación lingüística en la educación escolar básica y secundaria, es evidente que también se hace urgente alfabetizar a las comunidades técnico-científicas. De modo particular, se debe proyectar una didáctica especializada en la comprensión y producción del discurso escrito de los textos que efectivamente circulan en los diferentes medios especializados. Si como bien sabemos los textos escritos vehiculan gran parte del saber comunitario particular se constituyen en un medio central para la paulatina transformación de los actuales legos en los futuros expertos del conocimiento especializado. Leer y escribir, por lo tanto, son herramientas fundamentales en el proceso de incorporación a comunidades discursivas especializadas. Esta es una de las sendas que espera que la investigación lingüística le brinde luces; si la lingüística se vale de la alianza con la informática evidentemente lo hará mejor y en plazos más breves.

LITERATURA Y LINGÜÍSTICA

Referencias bibliográficas

- Arnoux, E., Nogueira, S. y Silvestri, A. (2002). "La construcción de representaciones enunciativas: el reconocimiento de voces en la comprensión de textos polifónicos" . *Signos*, vol. 35, N° 51y 52: 129-48.
- Biber, D. (1986). "Spoken and written textual dimensions in English: resolving that contradictory findings" . *Language*, N° 62: 384-414.
- Biber, D. (1988). "Variation across Speech and Writing" , Cambridge: CUP.
- Bod, R. (2003). "Introduction to elementary probability theory and formal stochastic language theory" . En R. Bod, J. Hay y S. Jannedy (Eds.) *Probabilistic linguistics* (pp. 11-37). Cambridge: MIT Press.
- Burdach, A.M. (2000). "El lexico científico y tecnico: un recurso publicitario persuasivo" . *Onomazein*, N° 5: 189-208.
- Cabot, C. (2000). "Computer applications to reading and writing abilities and the culture of the Spanish speaking world" . *Eurocall Journal*, May: 184-208.
- Chafe, W. (1982). "Integration and involvement in speaking, writing and oral literature" . En D. Tannen (Ed.). *Spoken and written language: exploring orality and literacy*. (pp. 35-53). Norwood, N.J.: Ablex.
- Chafe, W. (1985). "Linguistic differences produced by differences between speaking and writing" . En D. R. Olson, N. Torrence y A. Hidiyard (Eds.). *Literature, language and learning: the nature and consequences of reading and writing*. (pp. 105-123). Cambridge: Cambridge University Press.
- Chapelle, C. (2001). "Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing and Research" . Cambridge: Cambridge University
- Ciapuscio, G. (1992). "Impersonalidad y desagentivación en la divulgación científica" . *Lingüística Española Actual*, vol. 14, N° 2: 183-205.
- Echeverría, M. y Ramos, (2002). "Antes'98: Un tutorial interactivo para el análisis de textos" . En G. Parodi (Ed.). *Lingüística a interdisciplinariedad: desafíos del nuevo Milenio. Ensayos en Honor a Marianne Peronard*. (pp. 375- 385). Valparaíso: Ediciones Universitarias de Valparaíso de la Universidad Católica de Valparaíso.
- Echeverría, M. (2002). "Programas computacionales para el español como lengua materna" . *Signos*, vol.35, N° 51-52: 163-193.
- Ferreira, A., Campos, D. y Ruggeri, E. (1998). "VERBUM: Una aplicación multimedial para la enseñanza del Latín" . *Estudios Clásicos*, N° 114: 121-134.
- Ferreira-Cabrera, A. y Atkinson-Abudity, J. (2002). "A model for generating explanatory web-based natural-language dialogue interactions for document filtering" . *Journal of Research and Practice in Information Technology*, vol. 43, N° 1: 2-19.
- Graesser, A. C., VanLehn, K., Rose, C., Jordan, P., y Harter, D. (2001). "Intelligent tutoring systems with conversational dialogue" . *AI Magazine*, N° 22: 39-51.

LITERATURA Y LINGÜÍSTICA

Halliday, M. (1993). "On language and physical science" . En M. Halliday y J. Martin (Eds.). *Writing science. Literacy and discursive power*. (pp.54-68). Pittsburgh: University of Pittsburgh Press.

Halliday, M y Martin, J. (1993). (Eds.). "Writing science Literacy and discursive power" . Pittsburgh: University of Pittsburgh Press.

Jackson, P. y Moulinier, I. (2002). "Natural language processing for online applications Text retrieval, extraction and categorization": Amsterdam: John Benjamins.

Janda, R. (1985). "Note-taking as simplified register" . *Discourse Processes*, N° 84: 437454.

Jurafsky, D. y Martin, J. (2000). "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition" . New Jersey: Prentice-Hall.

Jurafsky, D. (2003). "Probabilistic modeling in psycholinguistics: linguistic comprehension and production" . En R. Bod, J. Hay y S. Jannedy (Eds.) *Probabilistic linguistics* (pp. 39-95). Cambridge: MIT Press.

Manning, C. y Schütze, H. (1999). "*Foundations of statistical natural language processing*" . Cambridge, Massachusetts: The MIT Press.

Marinkovich, J. y Cademartori, Y. (2004). "Foco Narrativo y Foco Informativo: dos dimensiones para una descripción de los manuales en la formación técnico-profesional" . *Revista Signos*, vol 37, N° 55: (en prensa).

[[SciELO Chile](#)]

Martin, J. R. (1993). "Genre and literacy - modelling context in educational linguistics" . *Annual Review of Applied Linguistics*, N° 13: 141-172.

Moreno, A. (1998). "*Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos*" . Madrid: Síntesis.

Parodi, G. (2002). "Comprensión lingüística: ¿Hacia dónde vamos desde donde estamos?" . En G. Parodi (Ed.). *Lingüística a interdisciplinariedad: desafíos del nuevo Milenio. Ensayos en Honor a Marianne Peronard*. (pp.47- 67). Valparaíso: Ediciones Universitarias de Valparaíso de la Universidad Católica de Valparaíso.

Parodi, G. (2003). "Reading-writing connections: Discourse-oriented research." *Reading and Writing. Interdisciplinary Journal*, (en prensa).

Parodi, G. (2004a) "Textos de especialidad y comunidades discursivas técnico-profesionales: una aproximación basada en corpus computarizado" . *Revista Estudios Filológicos*, (en prensa).

Parodi, G. (2004b) "Lingüística de corpus y análisis multidimensional: exploración de la variación en el corpus PUCV- 2003" . *Revista Española de Lingüística*, (en prensa).

Parodi, G. y Gramajo, A. (2003): "Los tipos textuales del corpus PUCV-2003: una aproximación multiniveles" , *Revista Signos*, vol. 36, N° 54: 207-223.

[[SciELO Chile](#)]

LITERATURA Y LINGÜÍSTICA

Núñez, P., Gramajo, A. y Parodi, G., (2003). "LECTES, Programa para mejorar las competencias de lectura y escritura a través de la Web" . Ponencia presentada en el II Congreso Internacional Cátedra UNESCO, Lectura y Escritura. "Comprensión y producción de Textos Escritos: de la Reflexión a la Práctica en el Aula" . Valparaíso, Chile.

Peronard, M., Gómez, L., Parodi, G. y Núñez, P. (1998). "*Comprensión de textos escritos: de la teoría a la sala de clases*" . Santiago de Chile: Editorial Andrés Bello.

Picallo, M. (1999). "La estructura del sintagma nominal: la nominalización y otros sustantivos con complementos argumentales" . En I. Bosque y V. Demonte (Coords.). *Gramática Descriptiva de la Lengua Española*. (pp. 363-394). Madrid: Espasa Calpe.

Warschauer, M. y Kern, R. (2000). *Network-based language teaching: Concepts and practice*. New York: Cambridge University Press.