RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 44 (2), II Sem. 2006, pp. 91-115.

CL ISSN 0033 - 698X

EL GRIAL: INTERFAZ COMPUTACIONAL PARA ANOTACION E INTERROGACION DE CORPUS EN ESPAñOL*

EL GRIAL: COMPUTATIONAL INTERFACE FOR THE ANNOTATION AND INTERROGATION OF CORPORA IN SPANISH

GIOVANNI PARODI
Pontificia Universidad Católica de Valparaíso. Valparaíso, Chile
gparodi@ucv.cl

RESUMEN

El Grial (www.elgrial.cl) es una interfaz computacional que permite tanto la realización de anotaciones morfosintácticas en textos planos en lengua española como la interrogación o consulta en forma de base de datos de los corpora allí reunidos. En este artículo se da cuenta de los objetivos para la creación de este sitio web dinámico junto a sus características y funcionalidades centrales. También se describen los corpora de base que dan sustento al sitio, los cuales son parte de las investigaciones realizadas en la PUCV y que están disponibles para indagación. Otro aspecto destacable lo constituye la opción de acceso temporal que ofrece esta herramienta informática al investigador para levantar, etiquetar y consultar otros corpora de manera gratuita y en línea durante un periodo de tiempo determinado. Como una forma de mostrar la utilidad de este recurso informático, se ejemplifican y muestran pantallas de algunas de sus herramientas.

Palabras claves: Interfaz computacional, El Grial, anotación lingüística, corpora textuales.

ABSTRACT

El Grial (www.elgrial.cl) is a computational interface that allows not only tagging and parsing of texts in Spanish language but also interrogation of the corpora collected and stored in the web site. In this article we give a general overview of the objectives for its design and implementation; also, a description of its main functions and components is given. Also, the corpora stored are described, all of them produced as result of research conducted at the PUCV. Another relevant feature of the interface is the temporal loading and access system that allows researchers to add new texts to the system, have them tagged

* El Grial es una herramienta desarrollada en el marco del Proyecto FONDECYT 1060440.

and then interrogated. As a way to show all the potentialities of the web site and tools associated some of the resources and exemplified and commented.

Keywords: Computational interface, El Grial, linguistic annotation, text corpora.

Recibido: 21-09-2006. Aceptado: 04-12-2006.

INTRODUCCION

EL DESARROLLO de la investigación en la lingüística ha venido enfrentando cambios radicales en los últimos 10 a 20 años. No tanto por los cambios paradigmáticos o los nuevos desafíos teóricos, sino por la vertiginosa tecnologización que ha avanzado de manera decidida como apoyo a la labor científica y académica. Es muy cierto que en Latinoamérica, en parte por dificultades presupuestarias, no hemos logrado implementar los recursos necesarios en términos comparativos con otros países. Esto ha dejado, de algún modo, a la indagación del español disminuida o está produciendo que su proyección se realice desde otros hemisferios. Pero también esto ha sido producto de cierta falta de decisión y vanguardia en los equipos de investigación en nuestro continente. Afortunadamente, en los años recientes se han llevado a cabo una serie de implementaciones digitales y computacionales para el español, algunas desarrolladas en España y otras en Latinoamérica (Ruiz Miyares, 2001; Rojo, 2001; 2002; Parodi y Venegas, 2004; Castel y Miret, 2004; Venegas, 2006; Parodi, 2007; Torner y Battaner, 2006).

Dentro de este escenario, y como un modo de progresar hacia una lingüística interdisciplinaria y contar con recursos tecnológicos de punta, en la Pontificia Universidad Católica de Valparaíso, Chile, hemos desarrollado una herramienta computacional denominada El Grial, disponible en un sitio web gratuito, que apoya la investigación de la lengua española en el marco de la hoy llamada Lingüística de Corpus (Leech, 1992, 2002; Tognini-Bonelli, 2001; Teubert, 2005; Parodi, 2005a, 2007). En este artículo nos proponemos describir los fundamentos para la construcción de esta interfaz computacional, explicar su funcionamiento y ejemplificar sus funciones básicas (cuando corresponde y es posible). En la parte final del artículo se comentan proyecciones y desafíos para esta interfaz y la investigación asociada.

OBJETIVOS Y FUNCIONALIDADES DEL PROGRAMA

La herramienta informática El Grial (www.elgrial.cl) fue creada inicialmente con el objetivo de apoyar la investigación y la docencia en los Programas de Postgrado en Lingüística de la Pontificia Universidad Católica de Valparaíso, Chile. Paralela-

mente, se decidió otorgarle un carácter más versátil como un sitio web abierto que acogiera tanto la herramienta de etiquetaje morfosintáctico como una base de almacenamiento y una interfaz de consulta de corpus electrónicos. Este servicio de recursos y datos se ofrece ahora a investigadores en el ámbito de la lingüística tanto a nivel nacional como internacional, dado que permite la incorporación de nuevos conjuntos de textos para su exploración como también el análisis de los textos ya existentes en el sistema.

De este modo, El Grial es un sistema computacional que cumple cuatro funciones básicas:

1) Anotar morfosintácticamente textos en archivos digitales planos en lengua española.
2) Recuperar esta información en forma de consultas de bases de datos.
3) Organizar y administrar los corpus recopilados por los equipos de investigación de la *Escuela Lingüística de Valparaíso* (www.linguistica.cl) de la Pontificia Universidad Católica de Valparaíso, Chile, y
4) Ofrecer la posibilidad de cargar y anotar corpus de modo temporal y consultarlos de modo gratuito, ya sea a través de una consulta en línea o con un permiso de carga temporal por un período de tiempo determinado.

Algunas otras motivaciones para la construcción del sitio El Grial han sido:

a) Poner a disposición una interfaz de interacción amigable que apoye a los lingüistas e investigadores no necesariamente especialistas en informática.
b) Ofrecer una herramienta computacional de uso gratuito y en línea a la comunidad de investigadores con textos etiquetados en lengua española.
c) Aportar a la tecnologización de la investigación.
d) Estimular el uso de corpus progresivamente crecientes en las investigaciones en lengua española.
e) Impulsar líneas de investigación en torno a la Lingüística de Corpus.

El sitio El Grial cumple, al mismo tiempo, la utilidad de hacer visible una línea de investigación en desarrollo por parte de los académicos de la denominada *Escuela Lingüística de Valparaíso,* Chile (www.linguistica.cl). Muchas de estas investigaciones se realizan con fondos gubernamentales esencialmente al servicio de propósitos exclusivamente científicos. En este contexto, es relevante destacar que el sitio web www.elgrial.cl busca fines netamente académicos y no se contempla, de ningún modo, acciones comerciales o propósitos de lucro. Por ello, es un sitio patrocinado íntegramente por la *Pontificia Universidad Católica de Valparaíso* y no se considera la participación de auspiciadores con ningún tipo de propósito comercial.

Vale la pena señalar que hemos escogido el nombre de *El Grial* para identificar

la interfaz de etiquetaje y consulta computacional de corpus textuales construida por el equipo que coordinó de manera muy intencionada para mostrar la relación existente entre la mítica leyenda de raíz céltico-cristiana y el espíritu que inspira al grupo de investigación en esta búsqueda de conocimiento y ojalá de sabiduría, esperando aportar a un proyecto académico internacional, no sólo con información bruta sino impulsando nuevos desafíos académicos mancomunados.

MULTINIVELES DE ANOTACION

Las herramientas que componen el sistema computacional permiten etiquetar (clasificar gramaticalmente las palabras de un texto) y analizar los tipos de estructuras lingüísticas que aparecen en distintos tipos de textos escritos (periodísticos, académicos, científicos, legales, literarios, etc.). Las máquinas computacionales que subyacen a El Grial y que posibilitan la anotación gramatical provienen de un programa llamado Connexor y que cuenta para el español con dos desarrollos: un etiquetador morfológico (*tagger*) y un analizador sintáctico (*parser*). El etiquetador morfológico es un analizador rápido que enriquece las formas textuales y etiqueta los textos de acuerdo a las clásicas partes de la oración (POS, por su sigla en inglés), morfología y entidades de significado básico. Produce lo que se denomina etiquetas morfológicas superficiales.

Por su parte, el analizador sintáctico (basado en una gramática funcional de dependencias: GFD) entrega, a la vez, información morfológica básica y también de la dependencia funcional que representan las relaciones de información al interior de la oración. Codifica información acerca de objetos y hechos (nombres, organizaciones y lugares); acciones (quién hizo qué a quién) y circunstancias (dónde, cuándo, cómo, por qué). Su output contiene 5 campos: posición de la palabra, palabra; lema, dependencia funcional; etiqueta funcional (etiqueta sintáctica de superficie y etiquetas morfológicas).

Dado que estas máquinas computacionales del Connexor se encuentran únicamente disponibles para equipos con sistema Linux, decidimos construir una interfaz amigable y versátil en Windows que permitiera un trabajo más expedito para no expertos en este sistema. También, como se ha dicho, El Grial supera largamente a un programa de marcaje morfosintáctico, ya que su construcción de almacenaje y consulta de corpus brinda potencialidades infinitas de investigación; al mismo tiempo, también es lícito señalar que todo ello ha supuesto una inversión de recursos y tiempo considerable que van mucho más allá de lo ya complejo que supone una herramienta de anotación lingüística. Veamos pues, a continuación, de qué se trata todo esto.

Tres tipos de etiquetas: multiniveles de anotación

Los dos tipos de anotaciones más arriba comentados se operacionalizan en tres tipos de etiquetas, las cuales alcanzan un total de 41 etiquetas básicas, pero cuya productividad específica supera las 70. Ellas son las siguientes:

1. Etiquetas morfológicas [POS]. Son 11 (de base)
2. Etiquetas de dependencia funcional [EDF]. Son 27
3. Etiquetas sintácticas de superficie [ESS]. Son 13

Ahora bien, dado que se detectaron diversos tipos de problemas tanto con el nombre de algunas de las etiquetas provistas por Connexor (sólo disponibles inicialmente en idioma inglés), como con el grado de precisión del análisis ejecutado (porcentaje de error cercano al 4% en el nivel morfológico y rondando el 13% en el nivel sintáctico), se procedió paralelamente en tres frentes de acción. Se buscaba, por una parte, incrementar el nivel de fiabilidad de las anotaciones y, por otra, producir etiquetas en español que fueran transparentes y acertadas en su nominación.

Paso 1. Como se dijo, debido a que se comprobó que ciertas etiquetas que inicialmente la maquina morfosintáctica del Connexor anotaba resultaban en algunos casos ambiguas y no siempre acertadas, se optó por llevar a cabo una comprobación del grado de precisión de cada una de ellas a partir de un corpus de prueba y contraste. Esta indagación empírica nos condujo a eliminar 3 etiquetas que no parecían discriminar en su anotación, no siendo consistente el criterio que subyacía en el etiquetaje. Así, se llegó a las 41 etiquetas de base (sin sumar las subcategorías), agrupadas en tres tipos de anotaciones.

Paso 2. El procedimiento de indagación y comprobación del grado de fiabilidad de cada anotación también probó ser una estrategia muy útil para revisar y determinar el nombre correspondiente, según la gramática del español, para cada una de las etiquetas. Ello debido a que, por un lado, las etiquetas y las abreviaturas de estos nombres no resultaban siempre de alta transparencia para el investigador y, por otro, a que el programa original (aunque sigue una gramática del español) proporcionaba etiquetas y abreviaturas sólo en lengua inglesa. De este modo, fue necesario llevar a cabo una exploración basada en corpus y realizar una interpretación gramatical de los ejemplos marcados bajo determinadas etiquetas, buscando nombres adecuados y simples (pero certeros) según la gramática del español.

Paso 3. En el tercer eje de acción, con el fin de alcanzar el mayor porcentaje posible de certeza en el etiquetaje, se diseñó y dotó a El Grial de una plataforma de revisión

y corrección manual de las anotaciones automáticas iniciales. Esta herramienta tecnológica adicional brinda la posibilidad de contar con textos anotados con un alto porcentaje de fiabilidad a través del cual se puede alcanzar el 100%. Esto quiere decir que una vez aplicado automáticamente el proceso de anotación, se revisa cada texto a través de una interfaz de manera manual y se corrigen las etiquetas que pudieran estar asignadas erróneamente. Por supuesto que para ello se ha debido especializar a personal idóneo tanto en el manejo del sistema como en la competencia gramatical pertinente; además, se debe contemplar el tiempo requerido para esta fase de revisión que obviamente es lenta y compleja. No obstante ello, tanto la posibilidad de contar con la creación de una interfaz de corrección manual como la alternativa de llegar a disponer de un corpus etiquetado fiablemente en un 100%, son logros que robustecen indudablemente a nuestro sistema El Grial.

A continuación, en la Tabla 1 se presentan las once etiquetas morfológicas de base o clásicamente conocidas bajo la sigla en inglés POS (*Part of Speech*). Ellas se diferencian de los otros dos grupos de etiquetas porque poseen categorías y subcategorías.

Tabla 1. Etiquetas morfológicas de El Grial.

Etiquetas Morfológicas		
Categoría gramatical	Subcategorías	Explicación
SUST		Nombre
Género	FEM	Femenino
	MSC	Masculino
	AMB	común, no indicado
Número	SG	Singular
	PL	Plural
	ABR	Abreviación
	PROP	nombre propio
ADJ		Adjetivo
	COMP	Comparativo
	SUP	Superlativo
	Género y número si se aplica	
Numeral		Numeral
	CARD	Cardinal
	ORD	Ordinal
	Género y número si se aplica	
PRON		Pronombre
Caso	ACU	Acusativo
	DAT	Dativo
Género	MSC	Masculino

	FEM	Femenino
	AMB	Ambiguo
Número	SG	Singular
	SG1	singular, primera persona
	SG2	singular, segunda persona
	SG3	singular, tercera persona
	PL	Plural
	PL1	plural, primera persona
	PL2	plural, segunda persona
	PL3	plural, tercera persona
Subtipos	INT	adverbio interrogativo o pronombre
	PER	Pronombre personal
	POS	Pronombre posesivo
	DEM	Pronombre demostrativo
	REFL	Pronombre reflexivo
	REL	Pronombre relativo
PREP		Preposición
DET		Determinante
	Género y número si se aplica	
ADV		adverbio
V		Verbo
Modo	IND	Indicativo
	SUB	Subjuntivo
	IMP	Imperativo
Tiempo	PRES	Presente
	IMPF	Imperfecto
	PRET	Pretérito
	CND	Condicional
	FUT	Futuro
Número	SG1	singular, primera persona
	SG2	singular, segunda persona
	SG3	singular, tercera persona
	PL1	plural, primera persona
	PL2	plural, segunda persona
	PL3	plural, tercera persona
VBD	INF	Infinitivo
	PART	Participio
	GER	Gerundio
CS		conjunción subordinada
CC		conjunción coordinante
INTERJ		Interjección

Tal como ya se adelantó, estas once etiquetas pueden alcanzar una alta productividad y cubrir más de setenta anotaciones efectivas. Por ejemplo, en el caso de la etiqueta *Pronombre*, ésta se ha contabilizado como una sola pero ella cubre toda la gama de pronombres existentes en español y sus respectivas combinatorias de género y número. Así, existen anotaciones particulares para cada una de ellas y, por ende, en la práctica se cuenta con más de 30 posibilidades de anotación altamente subespecificada. Algo similar ocurre con la etiqueta de *Sustantivo* y *Adjetivo*. En ambos casos su riqueza también es mucho mayor de la que se cuenta en una sola anotación.

La Tabla 2 muestra las etiquetas de dependencia funcional [EDF], cuyo número alcanza a 27. Cabe destacar que algunas de estas etiquetas se superponen con ciertas de las incluidas en la Tabla 1, pero en ningún caso esto genera dificultades de procesamiento. Es sólo una cuestión de aproximación gramatical al texto y no afecta el análisis propiamente tal.

Tabla 2. Etiqueta de dependencia funcional de El Grial.

Etiqueta de Dependencia Funcional	EDF
Nombre de la etiqueta	Etiqueta abreviada
Sintagma verbal	SV
Auxiliar de verbo compuesto	AUX
Preposición	PREP
Pronombre enclítico	ENC
Complemento de régimen preposicional	CRPRE
Sujeto	SUJ
Objeto directo	OD
Atributo subjetivo	AS
Objeto indirecto	OI
Frase adverbial de participio	FRAP
Vocativo	VOC
Frase adverbial de tiempo	FRAT
Frase adverbial de duración	FRAD
Frase adverbial de frecuencia	FRAF
Frase adverbial de cantidad	FRAC
Frase adverbial modo	FRAM
Frase adverbial de lugar (CCL o adv. de lugar)	FRAL
Frase preposicional de dirección	FRPD
Frase preposicional de finalidad	FRPF
Cláusula de causa/efecto o finalidad	CLCE/F
Cláusula de condicionalidad	CLC
Adjetivos numerales cardinales	ADJNC
Determinantes	DET
Adverbio de negación	ADVN
Modificador prenominal	MPREN
Adjetivo postpuesto	ADJPOST
Modificador nominal (frases preposicionales y cláusulas relativas)	MN

Por último, se entrega la tercera tabla con las denominadas etiquetas sintácticas de superficie. Nuevamente cabe señalar que existe entrecruzamiento entre algunas etiquetas con las listadas en las Tablas 1 y 2, pero esto no afecta el procesamiento de la información en el sistema. Estas son 13 anotaciones:

Tabla 3. Etiquetas sintácticas de superficie de El Grial.

Nombre de la etiqueta	Etiqueta abreviada
Frase verbal simple conjugada	FRVconj
Verboides	Vbd
Verbo auxiliar	Vaux
Adverbio de frase adjetiva	ADVfradj
Adverbios	Adv
Núcleo de frase nominal	Nn
modificador pronominal	Mpre
Adjetivo especificativo	ADJes
Artículos	Art
Adjetivos numerales cardinales	ADJnc
Conjunciones	Conj
Preposiciones	Prep
Interjecciones	Interj

DESCRIPCION OPERATIVA DE LA INTERFAZ EL GRIAL

Una vez descrito el sistema de etiquetas que conforma el sistema de anotaciones morfosintácticas de El Grial, procedemos a describir y ejemplificar el funcionamiento del programa tanto en su modo de consulta como de carga de corpus. Para ello mostraremos las principales pantallas de la interfaz gráfica computacional.

En la Figura 1 se muestra la pantalla de inicio, luego de un flash de entrada (el que es posible saltar). En ella se entrega información acerca del Programa o, alternativamente, se permite acceder directamente a la siguiente pantalla para iniciar una sesión de trabajo. En las pantallas siguientes siempre habrá posibilidad (a través de algunos links permanentes o de botones de ayuda) de encontrar información variada acerca del programa o del equipo académico y su actividad científica.

Figura 1. Pantalla de inicio El Grial.

Como se aprecia en la Figura 2, una vez iniciada la sesión de trabajo se ofrecen dos alternativas: trabajar con los corpus existentes en el sistema y denominados de manera genérica El Grial o, si se desea anotar un corpus nuevo y consultarlo, se accede a través de la opción *Carga y Consulta de Corpus Temporal*.

Figura 2. Acceso a Consulta de corpus El Grial y Carga y consulta de corpus temporal.

Para explorar los corpus recolectados por el equipo PUCV es posible encontrar un detallado anexo con su nombre, sigla, conformación y tamaño. Estimamos que esta posibilidad de acceso a información pormenorizada de los corpora es una fortaleza que exhibe El Grial en comparación con otros sistemas similares en los cuales no es posible encontrar descripción de los textos que componen cada corpus. Sin lugar a dudas, la explotación para cualquier investigación a partir de los corpus de El Grial se ve apoyada y cuenta con sustento descriptivo valioso.

Cabe destacar otra fortaleza de El Grial que nos resulta relevante en términos comparativos, tal como se señaló más arriba. Esta es la que se muestra en la Figura 2 y dice relación con nuestra decisión de crear una opción de carga y anotación temporal así como de consulta para los investigadores que deseen trabajar con su propio corpus en el ambiente de El Grial. Tal como ya se ha indicado más arriba, se diseñó este acceso con el propósito específico de brindar a la comunidad científica una herramienta de trabajo en línea y sin costo, buscando apoyar las investigaciones desde el marco de la Lingüística de Corpus (LC).

Ahora bien, si se procede a través de esta opción (*Carga y Consulta de Corpus Temporal*), habrá que seguir ciertos pasos de manera secuencial, tal como contar previamente con un texto plano, ser consciente que la carga estará en el sistema por un tiempo disponible limitado para todo tipo de consultas, conocer los descriptores mínimos requeridos para este proceso de acceso limitado en el tiempo.

Figura 3. Ingreso al proceso de documento temporal.

Si no se trabaja con un corpus nuevo que se desee etiquetar, se procede a través de la consulta de los corpus existentes en la base de datos. En la siguiente figura, se muestra la pantalla de *Consulta de Corpus El Grial*. Una vez desplegada, se debe –inicialmente– seleccionar el tipo de consulta que se desea realizar y posteriormente elegir el corpus de trabajo o el texto en cuestión (ya sea, *Búsqueda Simple* o *Búsqueda Compleja*).

Figura 4. Seleccionar tipo de consulta.

Ahora bien, sea que se ha optado por consultar el corpus El Grial o incluso si ya se ha cargado un nuevo corpus a través de la opción Carga Temporal, tal como la muestra lo Figura 4, se procede a la siguiente pantalla en que se nos ofrecen tres alternativas de trabajo y una opción de información. A través de esta pantalla, podemos seleccionar el tipo de consulta que se quiere realizar: simple, compleja o visualización del output, o sea, obtener el corpus seleccionado con las anotaciones realizadas por el programa.

La decisión del tipo de consulta obliga a seleccionar un texto, un corpus o varios de ellos, según el propósito de la indagación. Dicho de otro modo, al optar por una consulta de tipo simple o compleja el sistema lleva a otra pantalla (ver Figura 5) en la que se ofrece el menú de acceso al corpus con una serie de variables posibles de estudiar. De ellas, se deberá realizar una selección, la cual se constituirá en el corpus seleccionado de trabajo con el cual se realizarán las consultas posteriores.

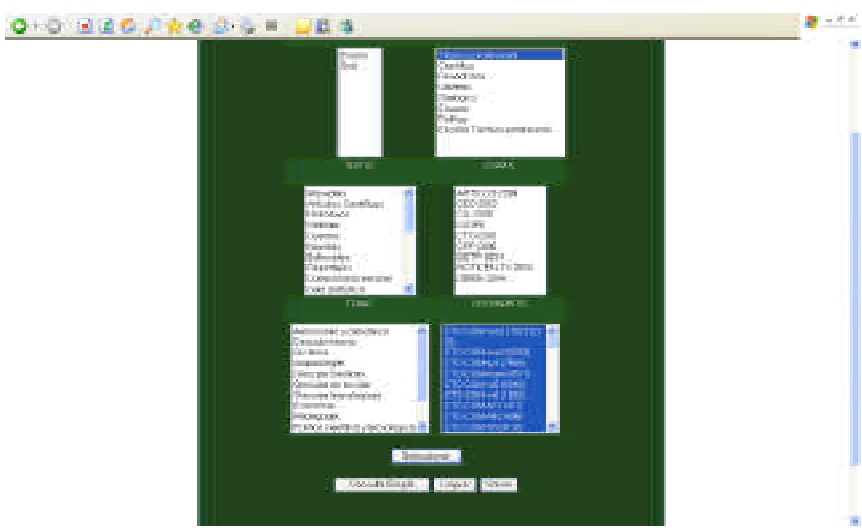

Figura 5. Seleccionar corpus El Grial.

Tal como lo muestra la Figura 5, es factible realizar la selección del corpus por medio de una serie de alternativas disponibles. Algunas pueden ser muy focalizadas en razón a un texto determinado u otras pueden ser más generales e interesadas en un registro, modo o tema específico.

Esta pantalla entrega información del Corpus El Grial a través de seis descriptores, que a su vez se constituyen en seis opciones de búsqueda e interrogación. Cada uno de ellos, de acuerdo a su naturaleza y decisión del equipo PUCV, cuenta con algunas subcategorizaciones o especificaciones, que hacen más rica y profunda la información disponible. A la vez, entregan mayores alternativas de indagación y comparación en las consultas, ya que es factible combinar más de un descriptor. Por ejemplo, si se seleccionó el modo escrito y, además, se selecciona el registro científico, sólo se procesarán aquellos textos que cumplan con estas dos condiciones.

Cabe destacar que, de cierto modo, la presentación de los seis niveles tiende a seguir un orden desde lo más general a lo más específico, particular e individual. Esto quiere decir que se parte con opciones tales como Modalidad de Lengua (oral o escrita), una categoría dicotómica y se llega a otra como Documento en que es posible seleccionar un solo texto de un subcorpus.

Los seis niveles en cuestión son:

1. Modo
2. Registro
3. Textos
4. Corpus
5. Tema
6. Documento

En cuanto al Modo, se cuenta con un acceso a textos escritos y otro a textos orales. Si al efectuar la selección sólo se marca una de estas opciones, el sistema incluirá en la consulta todos los textos de El Grial que caben bajo esta clasificación, incluyendo textos y corpus de diversa índole

La etiqueta Registro dice relación con los tipos de corpus que componen El Grial. En la actualidad son ocho: Técnico-profesional, Científico, Periodístico, Literario, Dialógico, Escolar, Político, Escolar Técnico-profesional.

Textos ha sido elegido para dar cuenta de las clases textuales que integran los ocho registros ya comentados. Ellas alcanzan en la actualidad una variedad que llega a veinte y siete, tales como ley, reglamento, manual, entrevista oral, glosario, instructivo, etc.

Por su parte, la etiqueta Corpus, a través de una sigla, permite englobar todo un subcorpus de El Grial a la vez, el cual obviamente queda cruzado por muchas de las variables ya descritas. Esto quiere decir que si se selecciona uno de los corpus a través de esta opción se realiza, paralelamente, una opción por un conjunto de las otras categorías descritas. Por ejemplo, si se elige el subcorpus DICIPE, se opta por una modalidad escrita, periodística, monológica y que, a la vez, incluye una variedad de clases textuales (noticias, reportajes, editoriales, etc.).

Por último, Tema es una categoría que se aplica a los Documentos. Cabe señalar que no todos los documentos están clasificados temáticamente, por ello algunos son etiquetados como “sin tema”. En todo caso, la mayoría sí lo está y con esto se intenta aportar mayores detalles de cada uno de los subcorpus. En Temas se indican los tópicos que se abordan en los diferentes textos o corpus (arqueología, ciencias médicas, ciencias de la vida, pedagogía, etc.).

En Documentos se entrega un detalle de cada texto que compone cada subcorpus; así, encontramos información del subcorpus al que pertenece el documento, la numeración del texto dentro del corpus, su clase textual y el número de palabras que lo componen. A modo de ejemplo, si tomamos de la ventana Documento el texto CTC-COM-ma1 (102.312), tenemos que se ha incorporado una gran cantidad de información descriptiva de alto poder en esta etiqueta. En primer lugar, cabe señalar que se llega a ella una vez que se selecciona el Registro Técnico-Profesional. Ahora bien, un subcorpus de este registro es el denominado CTC

(Corpus Técnico-Científico), el que se compone de textos que leen obligatoriamente alumnos de liceos técnico-profesionales de educación secundaria diferenciada en tres especialidades: área comercial, marítima e industrial. Como se aprecia en la etiqueta del ejemplo, el texto en cuestión pertenece al CTC del área Comercial (COM) y en la clase textual Manual (ma). En la misma etiqueta también se consigna que es el primer texto de este subcorpus (1) y que cuenta con un total de 102.312 palabras. Toda esta información se encuentra también disponible y con mayores detalles en el sistema en el menú derecho a través del botón *Descripción del Corpus El Grial*.

Ahora bien, una vez que se ha seleccionado un texto o un corpus a través de alguno o varios de los mecanismos más arriba descritos, se procede a ejecutar el tipo de consulta que previamente se había seleccionado: simple o compleja. Este botón nos llevará a otra pantalla en la que se desplegarán opciones más detalladas de la búsqueda misma. La siguiente Figura 5 nos permite visualizar el caso de la *Búsqueda Simple*.

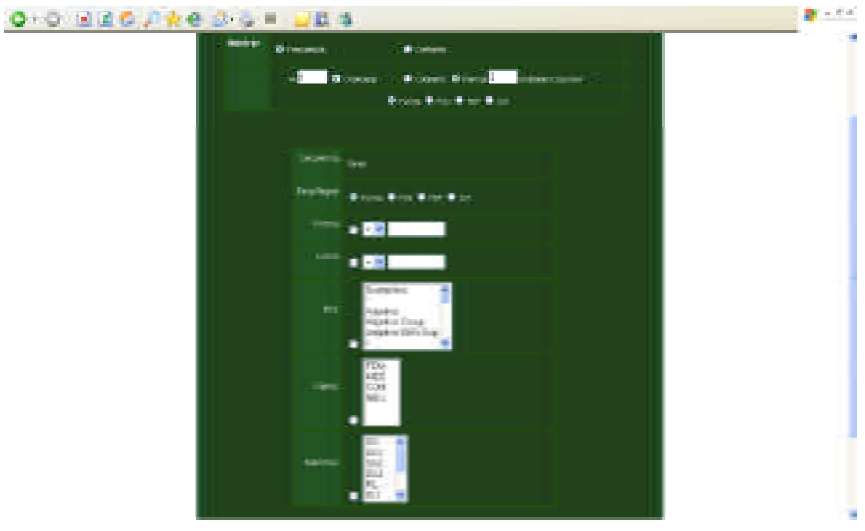

Figura 6. Búsqueda simple.

La denominada *Búsqueda Simple* es la primera y más elemental función de interrogación con que cuenta El Grial. A través de ella se permite realizar consultas básicas acerca de un corpus previamente anotado, morfológica y sintácticamente.

Esta búsqueda posibilita acceso a información de un corpus según dos modalidades: por frecuencia y en contexto. Estas dos opciones de despliegue de datos se aplican sobre tres categorías: formas, lemas y partes de la oración (POS), a las que se puede agregar información de género y número.

La *Búsqueda Simple* permite consultar por una palabra específica dentro de un corpus o bien indagar un texto o corpus de modo general (para conocer, por ejemplo, las más altas frecuencias de ocurrencia por categoría gramatical o por forma o lema). Una opción disponible es escribir en la ventana de la página la palabra que se desea indagar, elegir el modo de consulta (por frecuencia o en contexto) y la categoría que queremos obtener como resultado (forma, lema o parte de la oración). Si, por el contrario, se busca visualizar toda la información del texto o corpus desplegada por frecuencia, sólo debemos elegir el modo de consulta y las categorías que se quieren obtener como resultado sin necesidad de escribir nada en la pantalla. En el capítulo anterior se ejemplificó esta función abordándolo a modo de conteo de frecuencias, que es otra opción disponible.

Si volvemos a observar la Figura 5, se comprueba que, a través de esta pantalla de trabajo, se cuenta no sólo la opción de seleccionar una categoría sino que también se puede subespecificar una o varias subcategorías tales como género y número en el caso de los sustantivos o adjetivos. Todo ello revela la riqueza y profundidad con que una interrogación puede ser explorada y de las infinitas posibilidades de consulta disponibles en virtud de las preguntas de investigación o las hipótesis por contrastar.

Cualquiera de las búsquedas de que dispone la interfaz El Grial ha sido diseñada para llevar a cabo diversos tipos de análisis cuantitativo de los textos y de los corpus aquí registrados o de otros por ingresar; de esta manera, es posible conocer la frecuencia de una *palabra objetivo* en un texto dado o en todo un 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 está siendo buscada. Estas funciones se aplican tanto a la *Búsqueda Simple* como a las otras disponibles.

A continuación, en la Figura 7 se presenta un tipo de herramienta de mayor complejidad, que brinda alternativas más intrincadas de indagación y con potencialidades mucho mayores.

Tal como se decía más arriba, El Grial también ha sido diseñado para posibilitar interrogaciones de mayor versatilidad en el análisis de un corpus. A través de esta opción es factible llevar a cabo búsquedas avanzadas a partir de más de una categoría, es decir, una serie combinada de categorías. En este tipo de consultas se incorporan todas las funciones de la *Búsqueda Simple*, pero además se puede recuperar información sintáctica e información de dependencia gramatical. La principal di-

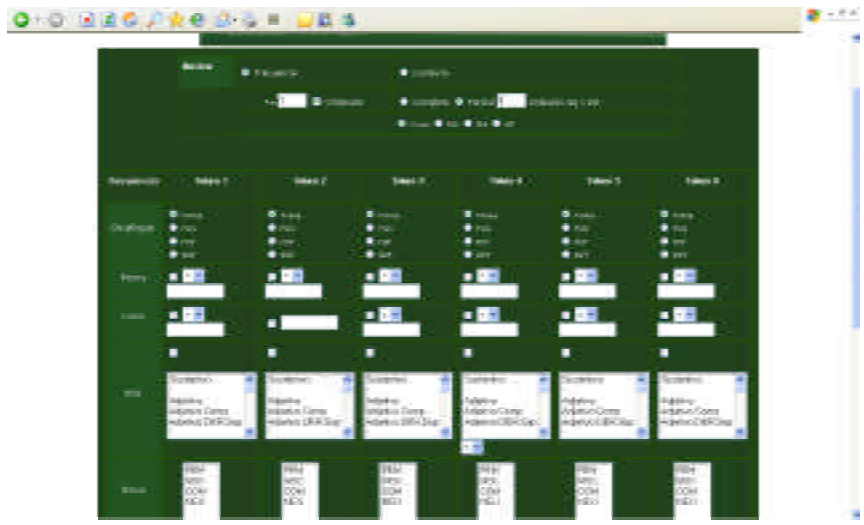

Figura 7. Búsqueda compleja.

ferencia entre este tipo de *Búsqueda Compleja* y la simple es la posibilidad de interrogar un corpus por cadenas sintácticas específicas, combinando incluso formas, categorías y subcategorías. En efecto, por ejemplo, es posible seleccionar una forma, seguido de una etiqueta sintáctica, tal como se muestra en la siguiente secuencia:

[Forma = para; lema = ser; Etiqueta sintáctica = Participio]

A partir de una consulta así, obtendremos secuencias como:

1 Para ser entregados
2 Para ser amados
3 Para ser investigados

Ahora bien, debido a que el programa cuenta además con etiquetas de dependencia funcional, es posible obtener una secuencia a partir de la selección de solamente una etiqueta. De este modo, si, por ejemplo, se selecciona la etiqueta *Sujeto*, se obtiene como resultado todas las secuencias de superficie que cumplen esa función gramatical.

Otras opciones para explotar la *Búsqueda Compleja* se ejemplificaron en el capítulo anterior en el marco del estudio de las colocaciones.

Visualización del output

La interfaz El Grial permite además la posibilidad de ver los resultados completos del análisis de anotación morfosintáctica y de dependencia funcional que realiza el programa. Estos se presentan en formato de tablas con seis columnas. En la primera columna se entrega la identificación del documento a través del número de clasificación en el subcorpus en que se incrusta. En la segunda, se presenta la numeración correlativa de los elementos de la oración en análisis (considerando la separación de punto a punto). La tercera columna consigna las formas textuales o superficiales, es decir, la palabra tal como aparece en el texto. Cada vez que comienza una nueva oración, la enumeración se inicia nuevamente. En la cuarta columna aparece la lematización de la forma textual de la tercera columna. Como se sabe, el lema corresponde a:

a) el infinitivo para el caso de los verbos
b) el masculino singular para el caso de los sustantivos, adjetivos y pronombres.

En la quinta columna se muestra la relación sintáctica asociada al número de la primera columna. De este modo, si en la tercera columna aparece det:>3 significa que esa palabra es el determinante de la palabra analizada con el número 3. En la última columna se entrega información sintáctica y morfológica. La información sintáctica es la primera etiqueta que aparece y es antecedida por el símbolo &. La información morfológica se presenta desde la segunda parte de la etiqueta en adelante y corresponde a la categoría gramatical y las marcas de género y número.

Un ejemplo de los resultados de este programa se presenta a continuación y se detalla posteriormente en la Tabla 4:

Registro: Técnico-Profesional
Modo: Escrito
Corpus: CTC-Com-ma
Clase textual: Manual
Area: Marítimo
Identificación Documento: 21

Tabla 4. Visualización del output: texto anotado.

id_docu.	Posición	Forma	Lema	Dep Func	POS
21	1	CARACTERiSTICAS	característica	Main:>0 |	&NH N FEM PL |
21	2	DE	De	pm:>4 |	&PM> PREP |
21	3	LA	La	det:>4 |	&DN> DET FEM SG |
21	4	CONTABILIDAD	contabilidad	Mod:>1 |	&NH N FEM SG |
21	5				

(continuación Tabla 4)

id_docu.	Posición	Forma	Lema	Dep Func	POS
21	1	Las	Las	det:>2 |	&DN> DET FEM PL |
21	2	Características	característica	null |	&NH N FEM PL |
21	3	De	De	pm:>5 |	&PM> PREP |
21	4	La	La	det:>5 |	&DN> DET FEM SG |
21	5	Información	información	mod:>2 |	&NH N FEM SG |
21	6	Contable	contable	ads:>5 |	&
21	7	Se	Se	obj:>9 |	&NH PRON |
21	8	Puede	Poder	null |	&+FM V IND PRES SG3 |
21	9	Resumir	resumir	obj:>8 |	&-FM V INF |
21	10	En	En	null |	&PM> PREP |
21	11	:			

Corpus disponibles en línea

La herramienta y base de datos El Grial también cumple con el objetivo de administrar todos los corpus disponibles (actualmente seis), recolectados a partir de diversos proyectos de investigación y de otras tantas tesis de postgrado. A través del botón *Descripción del Corpus El Grial* se despliega una tabla de cuatro columnas que especifica información pormenorizada de los corpus, sus características, el número de documentos que lo componen y el número de palabras de cada texto y de cada corpus. La Tabla 5 muestra estos datos:

Tabla 5. Descripción del Corpus El Grial.

Configuración de los corpus que componen El Grial			
Nombre	Características generales	Nº doctos.	Nº palabras
ARTICOS	Artículos de investigación científica en español, recolectados del indexador Scielo	642	2.471.389
NOTICENTV-2000	Noticiarios centrales de cuatro canales de televisión abierta de Chile	270	84.809
DETP- 2004	Resúmenes obtenidos como parte de pruebas de comprensión aplicadas a alumnos de especialidades de la formación técnico-profesional diferenciada de enseñanza	27	40.449
DICIPE-2004	Textos de divulgación de la ciencia y la tecnología en cinco periódicos chilenos de circulación nacional	411	204.598
CPP-2000	Artículos sobre políticas públicas acerca de la pobreza	20	234.818
PUCV-2003	Corpus constituido por tres subcorpora, a) Textos especializados de la formación técnico-profesional. b) Textos de literatura hispanoamericana. c) Entrevistas orales semidirigidas a estudiantes de cuarto año de enseñanza media	90	1.698.962

Como se comprende, éste es un corpus creciente y en desarrollo. En la actualidad, se están incorporando dos nuevos corpus de tamaño relativamente grande. Ellos pertenecen al Corpus PUCV-2006 que está siendo recolectado y que abarca, por un lado, los textos que se leen (en asignaturas del régimen obligatorio) como lectura obligatoria y complementaria en 4 carreras universitarias de la Pontificia Universidad Católica de Valparaíso: Trabajo Social, Psicología, Químico Industrial e Ingeniería en Construcción Civil. Por otro, el Corpus PUCV-2006 también contempla la recolección de textos de lectura fundamental en los cuatro escenarios laborales en que estos profesionales se desempeñen. Así, este corpus constituye (por sus características peculiares) una colección de discursos escritos única en Chile tanto por su naturaleza como por su tamaño, ya que no se tiene registro de otro corpus académico y profesional en los mismos cuatro ámbitos de indagación que pretenda llegar a los 50 millones de palabras.

¿Qué nos puede decir un corpus?: Conteo de palabras y algunos procesamientos básicos

Sin lugar a dudas la información que un corpus puede contener es infinita y cada investigador debe explorar y buscar respuestas a diversos tipos de preguntas que un determinado corpus puede estimular o para lo cual ha sido recolectado. Como es obvio, son múltiples los niveles y tipos de datos que pueden provenir de una colección de textos, dependiendo (entre otros) de si éste se encuentra etiquetado o no. Si lo está, la información variará según el tipo de etiquetas de que haya sido provisto, o sea, de la gramática que le subyace.

Las posibilidades de consulta de un corpus varían desde una simple lista de palabras para catalogar estructuras gramaticales o porcentaje de ocurrencia léxica que pueden revelar patrones de asociaciones lingüística y no lingüística hasta complejas búsquedas avanzadas a través de operadores booleanos y cadenas o secuencias léxicas o gramaticales, entre otros. En este caso, El Grial permite realizar análisis, explorando rasgos lexicales individuales o agrupamientos de rasgos co-ocurrentes a lo largo de un texto o de un grupo de textos o de todo un corpus o varios corpus.

Una de las herramientas más básicas y clásicas que extraen información de un corpus es la frecuencia de ocurrencia. A través de ella, lo que se obtiene es una lista de palabras, ya sea organizada alfabéticamente o por orden de frecuencia de ocurrencia (desde la más hasta la menos frecuente). Estas listas pueden –entre otros– resultar de alta utilidad lexicográfica, dado que ellas son de ayuda para decidir la lista de voces que, por ejemplo, pueden incluirse en un diccionario, considerando por supuesto su frecuencia de uso. También pueden ofrecernos índices de frecuencia en los que se muestre el ratio palabra/forma o tipo/caso (*type/token*), en otras

palabras, el número total de palabras de un texto frente al número de palabras diferentes que aparecen en el mismo texto. Este tipo de análisis se puede efectuar en El Grial a través de la interfaz denominada *Búsqueda Simple*, ya descrita en términos generales más arriba..

Con el fin de ejemplificar esta función, se ha realizado un conteo de frecuencia de un texto perteneciente a un corpus especializado escrito de la base El Grial. La Figura 8 muestra el resultado de esta búsqueda:

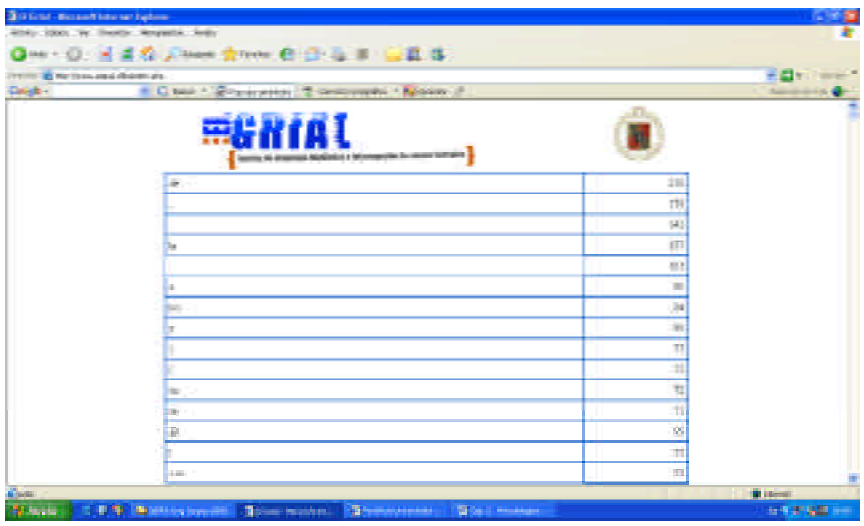

Figura 8. Uso del recurso Frecuencia de ocurrencia.

Como bien se sabe, preposiciones, artículos y conjunciones, o sea, palabras funcionales suelen ser las que mayor ocurrencia presentan en la mayoría de los textos. Así resulta ser el caso en este texto de muestra en que sólo observamos las primeras quince mayorías, incluidos los signos de puntuación y los espacios en blanco, y ninguna es una palabra de contenido. Resulta interesante comentar que este ejercicio está realizado a partir de un artículo de investigación científica del área de ciencias biológicas de una revista indexada. Por ello, es posible verificar el alto número de oraciones (seguramente breves) que se revelan a partir del número importante de puntos registrados y también el lugar relevante que el número "1" tiene en este listado con un total de 55 ocurrencias. Muy posiblemente también habrá una cantidad de información explicativa o aposicional, incrustada por medio de marcadores tales como los paréntesis redondos.

Otra de las herramientas de manejo de corpus, de gran utilidad y versatilidad

para el estudio lingüístico, lo constituyen las llamadas concordancias o también denominadas KWIC (del inglés, *Key Word in Context*). A través de este procedimiento se obtienen líneas de concordancia de una palabra objeto (en estudio) en su contexto lingüístico, en el que se consigna una colección de todas las apariciones de la palabra en búsqueda en un texto o conjunto de textos, junto con un número determinado (normalmente por el investigador) de palabras de cotexto anterior y posterior (la palabra en estudio o nodo normalmente se entrega en medio, resaltada en pantalla con un formato o color diferente del resto del cotexto). A través de este medio se puede visualizar a la vez una gran cantidad de ejemplos de uso de una palabra o un grupo de palabras. La mayoría de los programas computacionales para este procedimiento permiten obtener un número determinado de líneas (50 ó 100, o todas aquellas que contenga el texto o el corpus en estudio) y ordenarlas posteriormente de formas diversas: por ejemplo, alfabéticamente, de acuerdo con la palabra inmediatamente anterior o posterior a la palabra núcleo. Del mismo modo que en el caso anterior, la Figura 9, obtenida a través del programa El Grial, nos permite visualizar la búsqueda en contexto de la palabra *puente* en uno de los corpus del sitio del mismo nombre. Se ha definido esta búsqueda con un cotexto de 6 palabras a cada lado y se ha especificado buscar por *forma* y no por *lema* (opción también disponle en este programa).

Figura 9. Búsqueda de Concordancias.

Una característica provechosa del programa El Grial es que no limita la cantidad de líneas entregadas en la búsqueda y las provee todas. Por razones de espacio,

aquí sólo se registran las primeras doce, pero es posible acceder a visualizar las 108 apariciones en este corpus, tal como se indica en la esquina superior izquierda de la pantalla. En este caso, la palabra en búsqueda es entregada en una columna central y destacada en color rojo.

La última herramienta de búsqueda que comentamos es la noción de colocación. Si bien ella ha sido abordada de modo diferente en la literatura, entendemos sucintamente por ella la co-aparición, es decir, aparición simultánea de dos o más palabras en un segmento de texto en el que la distancia entre los elementos de la colocación no sobrepasa las cuatro o cinco palabras. En estos contextos, estas unidades fraseológicas presentan un alto interés de estudio y su productividad como indagación de combinaciones lingüísticas es ilimitada en los beneficios para –entre otros– la construcción de diccionarios y gramáticas. Del mismo modo, su utilidad para el diseño de materiales educativos y para el proceso de enseñanza/aprendizaje de las lenguas maternas y segundas o extranjeras y la traducción son altamente relevantes. Cabe destacar que las posibilidades de indagación a través de esta herramienta son tan diversas y versátiles como el programa con que se cuente lo permita, hecho que por lo demás se aplica a los dos otros procesamientos comentados más arriba. En el caso de El Grial, a través de su denominada *Búsqueda Compleja*, se entrega una variada y rica gama de alternativas de estudio. Con el fin de mostrar una opción diversa, ejemplificaremos la búsqueda colocacional de una cadena de tres categorías gramaticales que suelen constituir un grupo o frase nominal: Sustantivo/Adjetivo/Adjetivo. La Figura 10 entrega el resultado de una búsqueda compleja en un texto del corpus El Grial:

Figura 10. Resultado de una Búsqueda Compleja: colocaciones.

Como se aprecia en la esquina izquierda superior, se obtuvieron trece combinaciones de esta secuencia triádica en el texto objeto de análisis. La lista de búsqueda brinda una columna central con la cadena indagada, destacada en color rojo. La pantalla aquí copiada nos permite visualizar cinco de ellas. Los reproducimos nuevamente, dada la posible dificultad en su lectura:

1 catalizador bimetálico preparado.
2 soportes alternativos estudiados.
3 catalizadores bimetálicos empleados.
4 actividades catalíticas comparables.
5 arcillas pilareadas obtenidas.

Evidentemente son múltiples las conjeturas que pueden establecerse a partir de estos datos. Baste apuntar que tres de los cinco grupos actúan como sujetos gramaticales y que en cuatro de ellos se encuentran participios pasados en función adjetiva postmodificadora (sin duda una alta ocurrencia significativa en este tipo de texto: artículo de investigación científica del área de ciencias exactas).

PALABRAS FINALES

Con la ejemplificación de algunas de las herramientas de búsqueda disponibles en El Grial, cerramos este artículo. Estamos ciertos que sería posible realizar una pormenorizada explicación y mayor ejemplificación de las múltiples funciones y posibilidades que brinda el sitio y sus herramientas. En parte, no ahondamos en ello pues consideramos que éstas son prácticamente infinitas y preferimos insinuar algunas de las más relevantes y esperamos motivar así la curiosidad del lector para que por sí mismo explore el sitio e indague alternativas.

En la actualidad el sitio El Grial se encuentra en una fase de desarrollo de nuevas alternativas de apoyo a la investigación. Una de las opciones que pronto brindará será la comparación de frecuencias normalizadas a partir de todos los corpus disponibles. Esta función estará disponible a partir de todas las etiquetas de base del programa. Así, no será necesario que cada investigador realice las búsquedas básicas reiterada e innecesariamente, ya que éstas habrán sido ya efectuadas y almacenadas en bases de datos. A partir de ella se podrá establecer comparaciones multirregistros según los intereses de cada investigador. También se está implementando una herramienta denominada *Manchador de Textos*, la cual en base a la aplicación de ciertas fórmulas permitirá, por una parte, el "manchado" con colores diferentes ciertos rasgos seleccionados en el texto de análisis por el investigador, con el fin de observar su posible co-ocurrencia sistemática. Por otra parte, y más

complejamente, esta herramienta busca la determinación de un nivel de complejidad lingüística de cada texto a partir de rasgos seleccionados por el investigador y otros que estarán disponibles en el sistema. El objetivo que se persigue es la posible investigación de diferencias en principio de corte exclusivamente lingüísticas entre textos de complejidad diversa, pero con usos y proyecciones diversas (Parodi, 2005b).

Como se aprecia, el continuo desarrollo de la interfaz El Grial refleja en parte las preocupaciones de los miembros de la *Escuela Lingüística de Valparaíso* por el avance de las investigaciones y el apoyo a la tecnologización de la misma. En este sentido, nuestro fin último es abierto a la comunidad científica y de colaboración fraterna.

REFERENCIAS BIBLIOGRAFICAS

Castel, V. y Miret, A. 2004. "Generación de textos escritos en un marco sistémico funcional formal". En V. Castel, M. Aruani y V. Ceverino (eds.), *Investigaciones en ciencias humanas y sociales: del ABC disciplinar a la reflexión metodológica* (pp. 175-224). Mendoza: Editorial de la Facultad de Filosofía y Letras de la Universidad Nacional de Cuyo.

Leech, G. 1992. "Corpora theories of linguistic performance". En J. Svartvik (ed.), *Directions in Corpus Linguistics.* New York: Mouton de Gruyter, pp. 105-122.

————. 2002. "Sobre la importancia de los corpus de referencia". *Donosit* 24-25: 1-3.

Parodi, G. 2005a. *Discurso especializado e instituciones formadoras.* Valparaíso: EUVSA.

————. 2005b. "La comprensión del discurso especializado escrito en ámbitos técnico-profesionales: ¿Aprendiendo a partir del texto?", *Signos* 38(58): 221-267.

————. 2007. *Lingüística de corpus.* Buenos Aires: Eudeba.

Parodi, G. y Venegas, R. 2004. "BUCOLICO: Aplicación computacional para el análisis de textos. Hacia un análisis de rasgos de la informatividad". *Revista Lingüística y Literatura* 15: 223-251.

Rojo, G. 2001. "La explotación de la base de datos sintácticos del español actual" (BDS). En J. Kock (Ed.), *Lingüística con corpus. Catorce aplicaciones sobre el español*. Salamanca: Universidad de Salamanca, pp. 255-286.

———. 2002. "Sobre la lingüística basada en el análisis de corpus". Hizkunza-corpusak. *Oraria eta geroa*, 1-17.

Ruiz Miyares, L. 2001. "Desarrollo de un modelo computacional para el procesamiento de corpus textuales basado en la etiquetación automática". Tesis doctoral. Universidad de Oriente, Santiago de Cuba.

Teubert, W. 2005. "My version of corpus linguistics". *International Journal of Corpus Linguistics* 10(1): 1-13.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.

Torner, S. y Battaner, P. 2006. *El Corpus PAAU 1996. Estudios descriptivos, textos y vocabulario*. Barcelona: Instituto Universitario de Lingüística Aplicada.

Venegas, R. 2006. "La similitud léxico-semántica en artículos de investigación científica en español: Una aproximación desde el Análisis Semántico Latente". *Signos* 39(60), 75-106.