RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 46 (2), II Sem. 2008, pp. 53-79.
CL ISSN 0033 - 698X

INTERFAZ COMPUTACIONAL DE APOYO AL ANÁLISIS TEXTUAL: "EL MANCHADOR DE TEXTOS"*

INFORMATIONAL DENSITY IN TEXTS: COMPUTATIONAL APPROACH USING "EL MANCHADOR DE TEXTOS"

RENÉ VENEGAS
Pontificia Universidad Católica de Valparaíso.
Valparaíso. Chile
rene.venegas@ucv.cl

RESUMEN

En este artículo se describe y ejemplifica el funcionamiento de la herramienta computacional denominada "El Manchador de Textos". Esta herramienta permite localizar y destacar rasgos lingüísticos co-ocurrentes en corpus digitalizados. De este modo, tales rasgos pueden ser visualizados en el o los textos por medio de colores asociados a cada uno de ellos. Junto a esto último la herramienta permite cuantificar la co-ocurrencia de los rasgos, a través de un índice que mide la densidad de aparición conjunta de los rasgos lingüísticos investigados. Como demostración del funcionamiento de la herramienta, se investigó la aparición conjunta de una serie de rasgos lingüísticos asociados a la informatividad en una muestra diversificada de textos (científicos, de divulgación de la ciencia y periodísticos). Los resultados muestran diferencias significativas entre los índices de densidad informativa, según los distintos tipos de registros a los cuales pertenecen los textos estudiados.

Palabras clave: Lingüística de corpus, herramienta computacional, análisis de textos, rasgos lingüísticos.

ABSTRACT

In this paper the computational tool called "El Manchador de Textos" is described and exemplified. This tool helps find and highlight co-occurring linguistic features in digital texts and corpus. These features can be visualized in texts through colors, associated to each of these features. The computational tool also helps quantify co-occurring linguistic features producing an index that represents the density of the text prose. As an example of the use of the tool, a research focused on co-occurring informativity features in a diversified

*Este artículo ha sido escrito en el marco del Proyecto FONDECYT 1060440.

sample of texts is conducted (scientific, science popularization and journalistic). The results show statistical differences between the registers, considering the informativity indexes.

Keywords: Corpus linguistics, computational tool, text analysis, linguistic features.

Recibido: 17-10-2007. Aceptado: 21-10-2008

1. INTRODUCCIÓN

Se ha vuelto un lugar común, hoy en día, afirmar que el estudio del lenguaje y de las lenguas debe llevarse a cabo de modo interdisciplinario. En el caso de la descripción de las lenguas esto se hace cada vez más válido, dada la mayor participación de lingüistas en el desarrollo de herramientas para la entrega de servicios de información basados en tecnologías de uso masivo como la Internet y la telefonía celular. A pesar de estos avances altamente interdisciplinarios, aún es habitual percibir cierto escepticismo entre algunos especialistas cuando se proponen resultados lingüísticos obtenidos o acreditados por medio de herramientas informáticas. Este escepticismo puede ser superado promoviendo el desarrollo de herramientas cuyo uso sea "amigable". En este sentido, las interfaces computacionales que actúan entre los programas que realizan complejos análisis algorítmicos y la visualización en web de los sistemas de consulta y los correspondientes resultados, facilitan, sin duda, la tarea de los usuarios al momento de utilizar estas herramientas en tareas de investigación. De este modo, la aparición de mayor cantidad de este tipo de interfaces permite prever un mayor desarrollo de estudios, cada vez más y mejor sustentados en datos empíricos robustos, permitiéndoles a los analistas defender con mayor fuerza sus hipótesis y avanzar así en la explicación del fenómeno lingüístico e incluso discursivo al que se encuentran abocados.

A partir de los desarrollos generados por la lingüística computacional y la lingüística de corpus ha surgido un amplio número de herramientas que permiten el análisis de textos tanto orales como escritos. Sin duda, las más relevantes para la descripción y análisis lingüístico han sido aquéllas centradas en la construcción de sistemas de etiquetado y análisis morfosintáctico. Estas últimas, basadas en gramáticas artificiales creadas para este fin o en gramáticas de lengua natural adaptadas a los sistemas informáticos. Otro tipo de herramientas son aquellas que no consideran una gramática en su procesamiento y que se caracterizan por permitir el análisis de los textos desde la perspectiva de las frecuencias, agrupamientos y concordancias de unidades léxicas. Entre algunas de las más conocidas se encuentran: Wordsmith tools (http://www.lexically.net/wordsmith/), Concordancer (http://www.concordancesoftware.co.uk/) y AntConc (http://www.antlab.sci.waseda.ac.jp/).

Estos dos grandes tipos de herramientas permiten a los analistas dar cuenta

de múltiples fenómenos lingüísticos, tanto a nivel del léxico, de la morfología, de la sintaxis, de la semántica (asociada a la sintaxis y al estudio de colocaciones) e incluso a nivel del texto, a través del estudio de patrones de co-ocurrencia sistemática de rasgos lingüísticos con proyección textual (Parodi, 2005).

Este último nivel es particularmente interesante, pues en general cada vez que los analistas del texto utilizan estas herramientas se ven forzados a estudiar el texto de modo muy poco natural, fundamentalmente porque todas estas herramientas están diseñadas para presentar los resultados como partes aisladas (en listas de frecuencias o de agrupamientos de lexemas o de partes de la oración) y en el mejor de los casos en cotextos lingüísticos reducidos (listado de colocaciones o coligaciones), presentándose, en todos los casos, los resultados como textos "verticalizados", esto es como listas de datos y no en la forma en que los textos han sido producidos originalmente.

Desde el marco anteriormente descrito, el equipo de investigadores de la Escuela Lingüística de Valparaíso[1], en particular por quienes han focalizado sus investigaciones desde ya hace algunos años en el área de la lingüística de corpus (Parodi y Venegas, 2004; Venegas, 2005, 2006 y 2007; Cademartori, Parodi y Venegas, 2006; Parodi, 2006, 2007a, 2007b, 2007c y 2007d; Gutiérrez, 2007; Sabaj, 2007), se ha propuesto, en el seno de sus investigaciones lingüísticas, crear herramientas para el análisis de textos en español y ponerlas a disposición de forma gratuita a la comunidad científica, en este caso particular "El Manchador de Textos".

En lo que sigue de este artículo nos proponemos describir la interfaz computacional denominada "El Manchador de Textos" y presentar una aproximación a las potencialidades de esta herramienta, a través de la indagación de rasgos lingüísticos asociados a la densidad informacional en cuatro corpus textuales.

2. ¿QUÉ ES EL MANCHADOR DE TEXTOS?

"El Manchador de Textos" (EMT) es una interfaz computacional. Su propósito es identificar automáticamente rasgos lingüísticos en corpus digitalizados. Los resultados obtenidos en esta búsqueda se presentan a través del coloreado de las palabras o estructuras lingüísticas que han sido buscadas, esto es el "manchado". Además, el programa permite calcular un índice que da cuenta de la aparición conjunta de los rasgos en el texto. En la Figura 1 se presenta, a modo de ejemplo, un extracto de un texto "manchado":

[1] Ver www.linguistica.cl

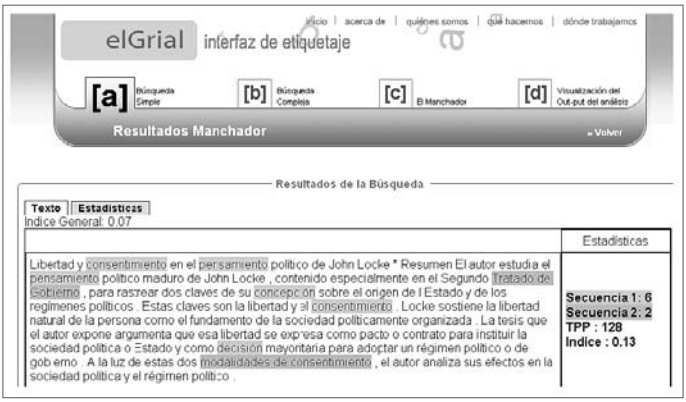


Figura1. Ejemplo de "manchado" de un párrafo de texto.

Como es posible observar, el párrafo ha sido "manchado", esto es, se han coloreado los rasgos lingüísticos correspondientes a algunos tipos de nominalizaciones y a la estructura nominal compuesta por sustantivo + de + sustantivo.

De este modo, a partir del "manchado", el investigador puede visualizar en qué sectores del texto aparecen tanto las agrupaciones como la forma en que estos rasgos se interrelacionan al interior de los párrafos de un texto.

La herramienta computacional que presentamos, además, permite calcular y mostrar la frecuencia de aparición de cada secuencia de rasgos. Esta información es relevante para calcular la co-ocurrencia sistemática (aparición conjunta) de los rasgos lingüísticos seleccionados según el total de palabras de cada párrafo del texto y, de este modo, del texto completo. A este proceso cuantitativo lo hemos denominado Índice de Densidad Lingüística y será explicado más adelante.

En suma, son productos de El Manchador de Textos tanto el "manchado" –con la consiguiente localización de los rasgos lingüísticos y la visualización de sus interacciones– como la determinación de un índice que expresa la densidad (co-ocurrencia de rasgos según el total de palabras del párrafo) con que se presentan los rasgos en cada uno de los párrafos del texto y del texto completo.

3. ¿CÓMO FUNCIONA EL PROGRAMA?

El programa El Manchador de Textos funciona sobre la base de la interfaz de interrogación lingüística El Grial, ya que ésta permite etiquetar morfo-sintácticamente los textos que se desean estudiar y, además, ejecutar la búsqueda de los rasgos lingüísticos. En este contexto, lo primero que se debe considerar en el funcionamiento de la herramienta es disponer de un texto o corpus de textos en formato *.txt que sean cargados en la interfaz de El Grial, con el fin de que tales textos

sean etiquetados de forma automática. Este proceso se puede realizar a través de la denominada Carga Temporal o solicitando autorización para incluir los textos en la base general de corpus de El Grial. También es posible utilizar los textos correspondientes a cualquiera de los corpus disponibles en la interfaz[2]. Además, es importante destacar que los textos que se carguen en El Grial deben estar debidamente segmentados, en relación con sus párrafos. En términos más concretos, se sugiere que luego de cada párrafo se incluyan dos líneas en blanco, de modo que el proceso automático considere los párrafos de manera independiente y pueda realizar adecuadamente tanto el manchado como el cálculo de la densidad lingüística asociada a los rasgos.

Una vez realizado lo anterior, el investigador podrá comenzar a realizar las búsquedas que considere pertinentes para sus propósitos de investigación.

En lo que sigue describiremos tres aspectos relevantes respecto de El Manchador de Textos, éstos son: a) indicaciones para hacer búsquedas con la herramienta b) el "manchado" y c) el índice de densidad lingüística.

3.1. ¿Cómo hacer consultas en el programa?

Para hacer consultas en El Manchador de Textos, primero se debe ingresar a la página www.elgrial.cl, una vez en ella se debe seleccionar la opción "Consulta de Corpus El Grial", tal como se presenta en la Figura 2.

Figura 2. Página de inicio de El Grial.

[2] Para mayores detalles sobre El Grial, el sistema de etiquetado y los corpus, ver Parodi, 2006 y 2007b.

Luego de esta selección se desplegará una pantalla con cuatro opciones, El Manchador de Textos se encuentra bajo la opción **[c]**. Una vez hecha la selección se presentará la siguiente pantalla (ver Figura 3).

Figura 3. El Manchador.

Para iniciar una consulta se debe seleccionar el botón "Seleccionar Corpus". Esta opción permite elegir el o los textos, incluidos en algún corpus, con los que se va a realizar la investigación. Como mencionamos anteriormente, la investigación se puede realizar utilizando uno de los corpus actualmente existentes en El Grial o también se puede cargar temporalmente un texto para realizar las consultas que se requieran.

De este modo, para elegir uno o varios textos a ser analizados, tanto en El Grial como en El Manchador de Textos, existen diversos criterios de selección. Estos criterios son: Modo, Registro, Textos, Corpus y/o Temas. Cabe señalar que estos criterios pueden ser utilizados en conjunto o aisladamente. Esto quiere decir que se puede acceder a los textos a partir de la combinación de criterios para una búsqueda más precisa, o a partir de la elección de un solo criterio para una búsqueda más general.

En la Figura 4, a modo de ejemplo, se ha elegido el Corpus ARTICOS (Artículos de Investigación Científica Originales), recuperándose automáticamente en la sección Documentos todos los textos asociados a este corpus. Como se observa, los textos aparecen codificados; por ejemplo, el primer texto de la lista aparece con el código BIO_194. Esto significa que es el artículo de investigación número 194 correspondiente al área de Biología[3]. Ahora bien, si no se desean consultar

[3] Mayor información sobre este y el resto de los corpus se encuentra en "Descripción de Corpus El Grial" en la página de inicio de la interfaz.

todos los textos a la vez, se selecciona el o los textos de interés para el investigador haciendo clic (o control + clic si son más de uno) sobre el código o códigos correspondientes.

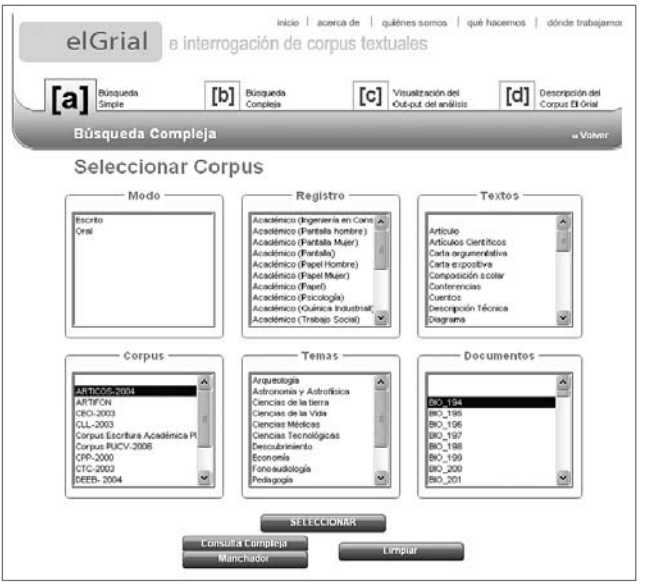


Figura 4. Selección de corpus y texto.

Una vez seleccionado el texto BIO_194, en nuestro ejemplo, se selecciona el botón "Manchador". Una vez realizado esto se despliega la pantalla para consultas (ver Figura 5).

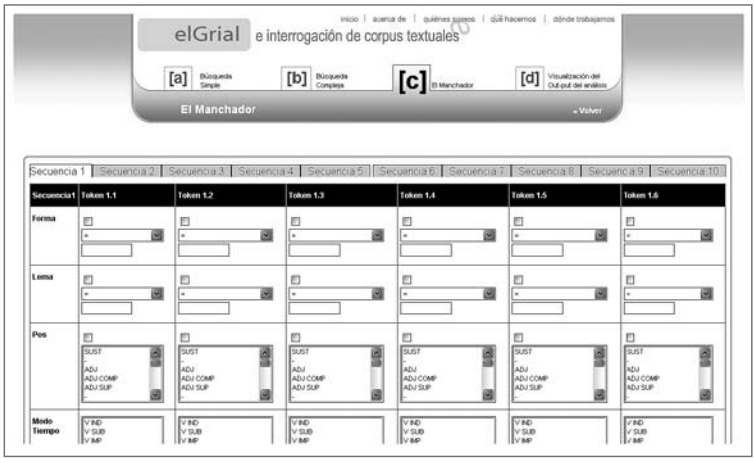


Figura 5. Página de consultas para El Manchador de Textos.

Como se observa en la Figura 5, la herramienta dispone de hasta 10 secuencias de búsqueda, siendo posible organizar la secuencia hasta en 6 unidades (o *tokens*). Por secuencia se entiende aquí el rasgo lingüístico a ser buscado, por ejemplo: *nominalización*. En tanto que por *token* se entiende la unidad que permite conformar (por sí misma o junto a otros *tokens*) el rasgo para la búsqueda en El Manchador. Ahora bien tales unidades pueden corresponder a distintos niveles de abstracción en la búsqueda (desde unidades más simples o superficiales a unidades más complejas o profundas de la lengua). Los niveles más importantes de búsqueda en El Manchador de Textos son:

a) Forma: En este nivel, la búsqueda se realiza a partir de la forma exacta de las palabras que se buscan. Para realizar esta búsqueda se ofrecen cuatro opciones, simbolizadas de la siguiente manera:
   1) = Busca la palabra exactamente como ha sido escrita, por ejemplo, "investigación".
   2) { Busca un conjunto de palabras, por ejemplo, "investigación", "investigaciones", "investigados", etc.
   3) * Busca partes de una palabra, por ejemplo, el sufijo "ción".
   4) {* Busca un conjunto de partes de palabras, por ejemplo, los sufijos "ción", "miento", "dor", etc.

   Las palabras de búsqueda o terminaciones deben ser ingresadas en el tercer casillero. Un ejemplo de búsqueda usando este nivel es el que se presenta en la Figura 6.

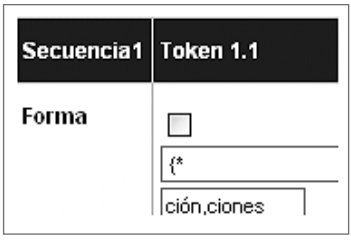


Figura 6. Ejemplo de búsqueda utilizando Forma.

Como vemos, la búsqueda está especificada para encontrar en el texto todas las palabras terminadas con los sufijos "ción" y "ciones". Un ejemplo del resultado se presenta en la Figura 7. Se identifican así 6 sustantivos terminados en "ción" y "ciones" en el párrafo.

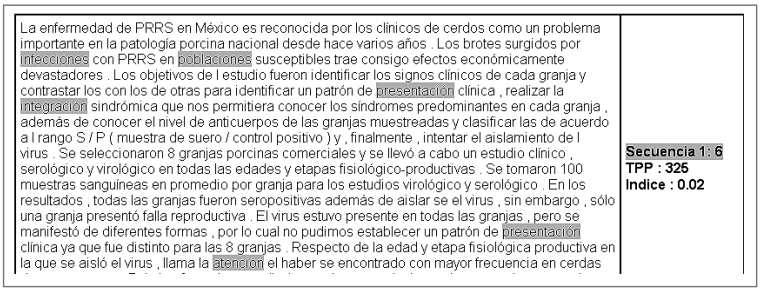


Figura 7. Resultado de búsqueda por Forma.

b) Lema: En este caso lo que se indaga es un conjunto de palabras paradigmáticamente asociadas a la palabra buscada. Las opciones de búsqueda corresponden a las tres primeras descritas anteriormente para Forma: [0] [*] y [{]. Así, por ejemplo, si se desea buscar el conjunto de palabras paradigmáticamente asociadas al verbo "ser", la búsqueda sería como se presenta en la Figura 8.

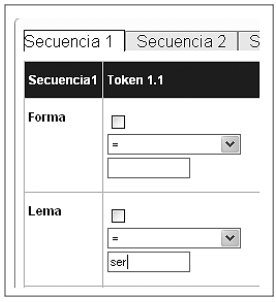


Figura 8. Búsqueda por Lema.

El resultado de esta búsqueda corresponderá a todas las flexiones del verbo "ser" en el texto. La Figura 9 muestra un ejemplo de los resultados obtenidos. En este caso se encuentran 5 formas flexionadas del verbo "ser" en el párrafo.

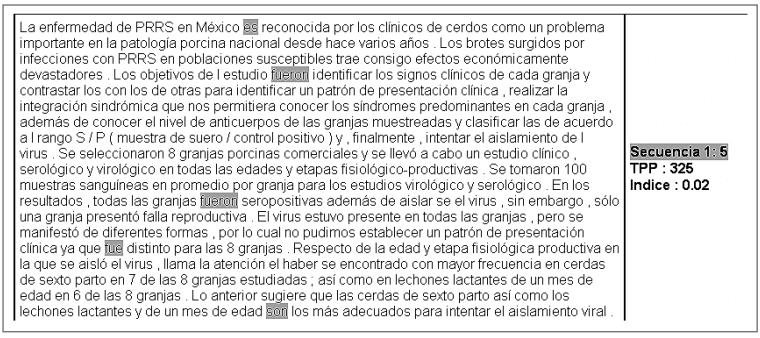


Figura 9. Resultado de la búsqueda por Lema.

c) Pos: En este nivel, lo que se busca corresponde a las partes de la oración (*Parts of Speech*). De este modo, las búsquedas se realizan consultando algún rasgo o combinación de rasgos morfológico de interés para el investigador. Cabe señalar que se ha separado el modo verbal y los tiempos para ampliar las posibilidades combinatorias de los rasgos. En este nivel se incluye un casillero con la abreviatura de cada rasgo morfológico posible de ser buscado (27 en total). Así, por ejemplo, si se desea identificar la presencia de adjetivos en el texto la búsqueda se realiza tal como lo presenta la Figura 10.

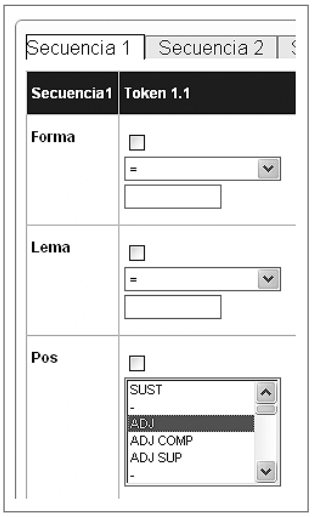


Figura 10. Búsqueda por Pos.

El resultado de esta indagación será la identificación de todos los adjetivos presentes en el texto, tal como se observa en la Figura 11. De esta manera se identifican 6 adjetivos en el párrafo.

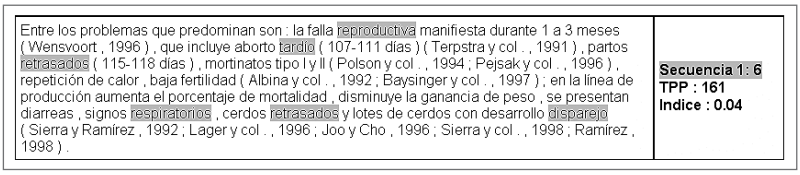


Figura 11. Resultados de la búsqueda por Pos.

Como hemos visto hasta ahora, para cada *token* es posible seleccionar un nivel de búsqueda de un rasgo determinado. Sin embargo, también existe la posibilidad de combinar *tokens* para identificar rasgos lingüísticamente más complejos. De este modo, por ejemplo, puede interesar buscar la secuencia una cadena de *tokens*

que permita identificar en el texto *Frases preposicionales como complemento de nombre*. La Figura 12 ejemplifica la búsqueda para este rasgo.

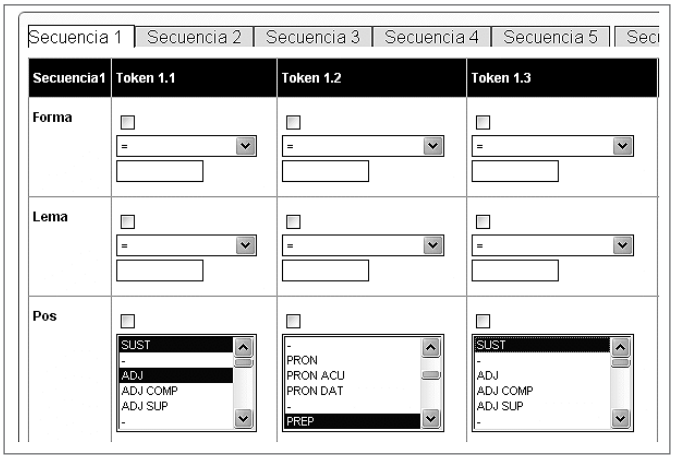


Figura 12. Búsqueda de *Frases preposicionales como complemento de nombre*.

Como se observa en el *token* 1.1 se ha seleccionado en Pos los sustantivos (SUST) y adjetivos (ADJ) y se ha combinado con el *token* 1.2 en el que se ha seleccionado en POS las preposiciones (PREP) y con el *token* 1.3 los sustantivos (SUST). La Figura 13 presenta un ejemplo del resultado de esta búsqueda.

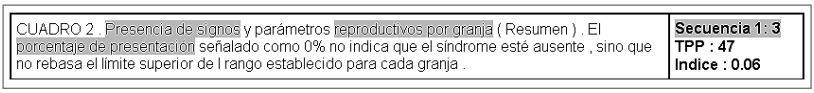


Figura 13. Resultados de la búsqueda de *Frases preposicionales como complemento de nombre*.

Como es posible notar en este párrafo, se han manchado 3 ocurrencias de la cadena buscada. La primera y la tercera corresponden a la cadena SUST+PREP+SUST, en tanto que la segunda corresponde a ADJ+PREP+SUST.

Pues bien, hemos visto hasta ahora las posibles búsquedas de rasgos utilizando una sola secuencia en sus distintos niveles y en combinación de *tokens*. A ello se debe que el manchado siempre corresponda al mismo color (verde en este caso). Sin embargo, como bien se sabe, la presencia de un solo rasgo en la descripción de la mayoría de los fenómenos lingüísticos o discursivos no es suficiente para dar cuenta de ellos. De este modo, se hace necesario combinar secuencias que den cuenta de múltiples rasgos, los que, presentándose conjuntamente, evidencian de forma más precisa un fenónemo lingüístico o discursivo determinado. A modo

de ejemplo, si se está interesado en buscar la presencia de modalización en algún texto o textos, se pueden indagar los cinco primeros rasgos de la dimensión Foco Modalizador, identificados por Parodi (2005):
a) Forma activa "ser".
b) Verbos Atenuadores.
c) Verbos modales de posibilidad.
d) Adverbio de modo.
e) Adjetivos predicativos.

De este modo, una forma posible de búsqueda en El Manchador de Texto sería la que se presenta en la Tabla I:

Tabla I. Forma de búsqueda del Foco Modalizador

RASGO LINGÜÍSTICO	**FORMA DE BÚSQUEDA**
Formas activas "ser"	**Secuencia 1: Token 1.1**: FORMA { soy, somos, eres, es, son, seré, seremos, serás, será, serán, fui, fuimos, fuiste, fue, fueron, iba, íbamos, ibas, iban, iría, iríamos, irías, irían **Token 1.2:** POS = SUST ADJ PRON DET
Verbos Atenuadores	**Secuencia 2: Token 2.1** LEMA { parecer, creer, estimar, pensar **Token 2.2** FORMA = que
Verbos modales de posibilidad	**Secuencia 3: Token 3.1**: LEMA = poder **Token 3.2**: POS = V VBD INF
Adverbios de modo	**Secuencia 4: Token 4.1**: FORMA * mente
Adjetivos predicativos	**Secuencia 5: Token 5.1**: LEMA { ser, estar **Token 5.2**: POS = ADV ADJ

El resultado de esta indagación, en una noticia del diario *El Mercurio* que trata sobre temas de arqueología (EME-ARQ 12) del corpus DICIPE (Divulgación de la Ciencia en la Prensa Escrita), se presenta en la Figura 14.

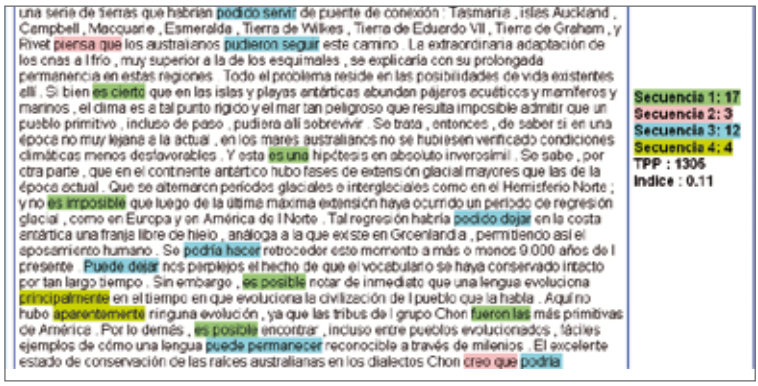


Figura 14. Resultado de la búsqueda del Foco Modalizador.

Como observamos, en el texto de la Figura 14 se aprecia una co-ocurrencia de los cinco rasgos de la modalización indagados. Además, notamos que la herramienta "mancha" con distintos colores cada secuencia de búsqueda, permitiendo su rápida identificación y la relación existente en el texto entre cada una de las secuencias. De este modo, utilizando este conjunto de rasgos sería posible describir y comparar textos de diversos tipos en relación a la mayor o menor presencia de modalización.

3.2. ¿Qué es el "manchado" del texto?

Como ya se ha planteado y ejemplificado anteriormente, se le llama "manchado" al proceso de búsqueda, localización y señalización de determinados rasgos lingüísticos indagados, asociándole a cada rasgo un color que lo identifica en el texto. El "manchado" posee dos aspectos importantes que lo justifican: Por una parte, facilita una rápida detección de uno o varios rasgos lingüísticos buscados en el texto, de modo contextualizado y sin "verticalizar" el texto, es decir, sin crear listas de palabras que desarticulan el texto. Por verticalizar se entiende aquí el proceso de entrega de resultados disponibles en los programas de concordancia. En estos programas, a partir de un nodo (palabra de búsqueda) se entrega un cotexto restringido (y variable en tamaño) de todas las apariciones de la palabra buscada. Por ejemplo, si buscamos la palabra "datos" en un programa de concordancia cualquiera en un texto de química, el resultado sería, aproximadamente, el que se presenta en la Figura 15.

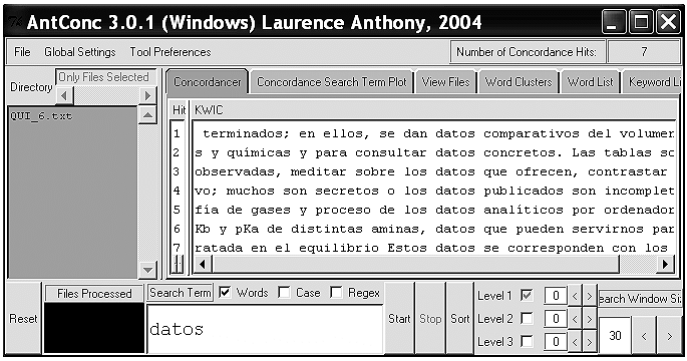


Figura 15. Búsqueda de "datos" en AntConc 3.0.1.

Como se observa, la diferencia radica en que el resultado de la búsqueda presenta el texto en su estado original. Esta horizontalización, sin duda, es una propiedad distintiva de esta herramienta y favorece una aproximación más natural al texto mismo, dado que permite identificar no sólo el cotexto inmediato sino que todo el párrafo o párrafos en los que la palabra buscada aparece. Cabe señalar que en el proceso de entrega de los datos, la herramienta considera toda la información correspondiente al etiquetado morfosintáctico, sin embargo, esta información no se muestra al usuario, precisamente para no "destruir" la conformación original del texto.

Por otra parte, el "manchado" permite la visualización de varios rasgos lingüísticos simultáneamente, como se presentó en la Figura 14. Esto se realiza otorgándole a cada rasgo lingüístico un color, permitiendo al investigador identificar rápidamente dónde aparecen, de qué manera está relacionado con otros y dónde se concentran los distintos rasgos lingüísticos investigados.

3.3. ¿Cómo se cuantifican los datos?

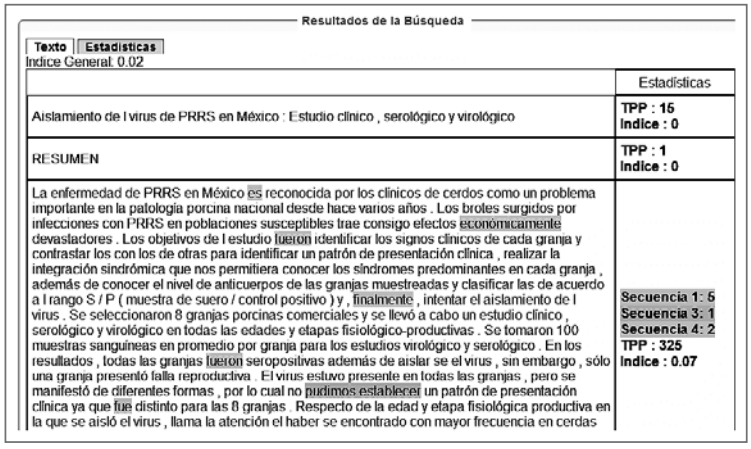
En el programa se incorporó una fórmula matemática que calcula tanto la frecuencia de aparición de cada rasgo como la co-ocurrencia sistemática de estos rasgos lingüísticos en el párrafo y el texto. A esta fórmula matemática se le ha dado el nombre de Índice de Densidad Lingüística del Texto (IDLT). Lo interesante de este dato numérico es que permite dar cuenta del modo en que un texto se caracteriza y se puede distinguir de otros según la presencia conjunta de varios rasgos lingüísticos asociados a algún fenómeno identificable en los textos[4]. De este modo, la caracterización del texto depende de los rasgos que el investigador elija, así es posible calcular un índice relacionado con algún fenómeno lingüístico

[4] Para una descripción detallada de la fórmula, ver Anexo 1.

o discursivo determinado, por ejemplo, la informatividad, la modalización, la argumentatividad, etc.

En relación con lo anterior, es importante recalcar que la densidad lingüística es una representación numérica de la relación existente entre los rasgos. Por ello, es necesario que el investigador distinga claramente que el fenómeno que investiga puede ser examinado a partir de rasgos lingüísticos identificables en la materialidad del texto. En este sentido, la herramienta apoya la localización y la cuantificación del fenómeno, pero es el investigador quien debe determinar qué rasgos lingüísticos lo evidencian.

El Índice de Densidad Lingüística se consigna de dos maneras en la herramienta. La primera es en la hoja "Texto" de la página "Resultados de la Búsqueda". En esta hoja, junto al resultado del "manchado" se consigna, en la columna "Estadística", la frecuencia de aparición de cada secuencia, el total de palabras y el índice de densidad del párrafo. Además, en esta hoja en el extremo superior izquierdo se presenta el dato de densidad total del texto. La Figura 16 presenta el "manchado" y el cálculo de los valores de frecuencia y de densidad para el primer párrafo del texto BIO_194, en el cual se han indagado los rasgos de modalización descritos anteriormente.


Resultados de la Búsqueda

Texto | Estadisticas

Indice General: 0.02

	Estadísticas
Aislamiento de l virus de PRRS en México : Estudio clínico , serológico y virológico	TPP : 15 Indice : 0
RESUMEN	TPP : 1 Indice : 0
La enfermedad de PRRS en México es reconocida por los clínicos de cerdos como un problema importante en la patología porcina nacional desde hace varios años . Los brotes surgidos por infecciones con PRRS en poblaciones susceptibles trae consigo efectos económicamente devastadores . Los objetivos de l estudio fueron identificar los signos clínicos de cada granja y contrastar los con los de otras para identificar un patrón de presentación clínica , realizar la integración sindrómica que nos permitiera conocer los síndromes predominantes en cada granja , además de conocer el nivel de anticuerpos de las granjas muestreadas y clasificar las de acuerdo a l rango S / P (muestra de suero / control positivo) y , finalmente , intentar el aislamiento de l virus . Se seleccionaron 8 granjas porcinas comerciales y se llevó a cabo un estudio clínico , serológico y virológico en todas las edades y etapas fisiológico-productivas . Se tomaron 100 muestras sanguíneas en promedio por granja para los estudios virológico y serológico . En los resultados , todas las granjas fueron seropositivas además de aislar se el virus , sin embargo , sólo una granja presentó falla reproductiva . El virus estuvo presente en todas las granjas , pero se manifestó de diferentes formas , por lo cual no pudimos establecer un patrón de presentación clínica ya que fue distinto para las 8 granjas . Respecto de la edad y etapa fisiológica productiva en la que se aisló el virus , llama la atención el haber se encontrado con mayor frecuencia en cerdas	Secuencia 1: 5 Secuencia 3: 1 Secuencia 4: 2 TPP : 325 Indice : 0.07

Figura 16. Ejemplo del resultado del manchado y del cálculo de frecuencia y densidad.

Como se observa, para la presencia de cada secuencia se calcula la frecuencia de cada ocurrencia de los rasgos (5, 1 y 2 respectivamente); así también, se consigna el total de palabras (325) y el índice de densidad (0,07) de los rasgos que aparecen en cada párrafo del texto. Se observa, además, en el extremo superior de esta página, el "Índice General" de densidad, correspondiente al promedio de densidad de todos los párrafos del texto.

Una segunda forma de ver los datos cuantitativos asociados a la fórmula IDLT es observando la hoja de "Estadísticas" de esta misma Página. La Figura 15 ejemplifica los resultados obtenidos para la misma búsqueda en el texto. De este modo, se distinguen las frecuencias de las secuencias, el total de palabras y la densidad de cada párrafo, tal como se muestra en la Figura 17.

Resultados de la Búsqueda

Texto | Estadisticas

Parrafo	Secuencia1	Secuencia2	Secuencia3	Secuencia4	TPP	Densidad por Parrafo
1	0	0	0	0	15	0
2	0	0	0	0	1	0
3	5	0	1	2	325	0.07
4	0	0	0	0	11	0
5	0	0	0	0	1	0
6	2	0	0	1	102	0.06
7	1	0	0	0	161	0.01
8	1	0	1	3	242	0.06
9	1	0	0	3	143	0.06
10	1	0	0	1	35	0.11
11	0	0	0	0	3	0
12	0	0	0	0	105	0
13	0	0	0	0	17	0

Figura 17. Ejemplo de visualización de los datos estadísticos de los primeros 13 párrafos del texto BIO 194.

Como se puede observar, la mayor densidad de los rasgos lingüísticos asociados a la modalización se encuentra en el párrafo 10 (0,11). Sin embargo, en general presenta una baja densidad de esta función comunicativa en el texto. Al final de esta misma página se consigna la Densidad Total del Texto, la que alcanza a 0,02. Cabe recordar que este valor en sí mismo no significa nada, a menos que se compare con otros textos o con otras características o dimensiones lingüísticas o discursivas.

En lo que sigue presentamos una aproximación al uso de la herramienta en un estudio comparativo de textos acorde a un conjunto de rasgos que expresan densidad informativa.

4. LA DENSIDAD INFORMACIONAL EN REGISTROS DIVERSIFICADOS: UNA INVESTIGACIÓN

En este apartado nos proponemos presentar una aproximación a las potencialidades de El Manchador de Textos. Para cumplir con este objetivo, detallamos un estudio en el cual se comparan los rasgos lingüísticos asociados a la densidad informacional, tal como son presentados por Parodi (2005) y por Parodi y Venegas (2004), en cuatro corpus textuales pertenecientes a registros diversos de la lengua española.

4.1. El corpus

Para llevar a cabo esta aproximación se utilizaron los siguientes corpus: ARTICOS, DICIPE y NEMOL. A continuación describiremos brevemente cada uno de ellos:

ARTICOS (Artículos de Investigación Científica Originales Scielo): Este corpus está conformado por 642 artículos de investigación científica en español, recolectados del indexador Scielo (Scientific Electronic Library Online). Su fuente está compuesta por artículos científicos pertenecientes a las áreas de ciencias biológicas, exactas y sociales. Su registro es el científico y consta de 2.471.389 palabras.

DICIPE (Divulgación de la Ciencia en la Prensa Escrita): Este corpus contiene 412 textos de divulgación de la ciencia y la tecnología en cinco periódicos chilenos de circulación nacional, su registro es el periodístico de divulgación científica y sus fuentes son: El Mercurio, Las Últimas Noticias, La Nación, La Tercera y La Cuarta. Fue recolectado desde el 1 de marzo hasta el 31 de mayo de 2004 y consta de 204.598 palabras.

NEMOL (Noticia de El Mercurio Online): Este corpus está constituido por 152 noticias de El Mercurio Online. El modo de recolección de estos textos fue vía Internet y fueron recolectados entre el 1 y el 30 de mayo de 2006. Su registro es el periodístico y consta de 60.800 palabras.

4.2. Objetivo

El objetivo de la presente investigación es comparar el grado de densidad informacional de una serie de textos de registro diversificado, a partir de 5 rasgos lingüísticos cuya función comunicativa evidencia la presencia de densidad informacional (Parodi, 2005). Estos rasgos son: verbo modal de obligación, verbos en modo subjuntivo, nominalizaciones, participios en función adjetiva y frases preposicionales como complemento del nombre.

4.3. Procedimientos

En primer lugar, se seleccionaron por azar simple cinco textos digitalizados (*.txt) de cada corpus fuente, los que constituyen una muestra de trabajo correspondiente a 15 textos (23.967 palabras). En la Tabla II se presenta la muestra de textos según el corpus fuente y el total de palabras de cada texto.

Tabla II. Descripción de la muestra de trabajo.

CORPUS FUENTE	TEXTOS DE LA MUESTRA	PALABRAS
ARTICOS	Efectos del tiempo de transporte de novillos previo al faenamiento sobre el comportamiento, las pérdidas de peso y algunas características de la canal.	6226
	Efecto de melatonina sobre la secreción pulsátil de hormona luteinizante y de hormona del crecimiento en borregas con restricción alimenticia.	2524
	Sensibilidad tisular a la insulina antes, durante y después de un ayuno en ovejas prepúberes.	3328
	Escherichia coli aislada de cerditos diarreicos. Presunción de cepas productoras del factor citotóxico necrosante (cnf).	2080
	Prevalencia de lesiones podales en ovinos de 25 explotaciones familiares de la Provincia de Valdivia, Chile.	3940
NEMOLDICIPE	Cobre dispara superávit fiscal en primer trimestre	463
	Pearl Jam lanzó nuevo disco que retoma sus raíces	336
	Llegó a Chile el Dalai Lama	330
	Pellegrini anuncia que gira del Villarreal a Chile quedó postergada	218
	Ministro: Operadores deberán dar explicaciones por atraso del Transantiago	400
	Consultorios elevaron ayer la cantidad de consultas hasta en un 50%	333
	"La democracia no se exporta"	1982
	¿Qué es en estos días ser un chileno (a)?	517
	¡Somos 10!: Confirmada existencia de nuevo planeta	248
	Aceite de emú: fórmula chilena rinde examen	1042
	TOTAL DE PALABRAS	**23967**

En segundo lugar, utilizando el programa El Manchador de Textos, se indagaron en los 20 textos las secuencias correspondientes a cada uno de los cinco rasgos lingüísticos asociados a la densidad informacional, tal como se describe en la Tabla III. Cabe mencionar que la búsqueda en el programa de cada rasgo lingüístico se realiza por secuencias. Esto, debido a que existen rasgos que se conforman con más de un *token*, ya sea como forma léxica o como parte de la oración (POS), y en ocasiones como la combinación de ambos. La Tabla III presenta las secuencias correspondientes a los 5 rasgos mencionados más arriba.

Tabla III. Ejemplo de rasgos lingüísticos asociados a la densidad informacional y su búsqueda en El Manchador de Textos.

RASGO LINGÜÍSTICO	FORMA DE BÚSQUEDA
Verbo modal de obligación	Secuencia 1: Token 1.1: Forma { debo, debemos, debes, debe, deben, deberé, deberemos, deberás, deberá, deberán, debería, deberíamos, deberías, deberían, debiese, debiésemos, debieses, debiesen
Verbos en modo subjuntivo	Secuencia 2: Token 2.1: Pos = V SUB
Nominalizaciones	Secuencia 3: Token 3.1: Forma {* ción, ciones, sión, siones, miento, mientos
Participios en función adjetiva	Secuencia 4: Token 4.1: Pos = SUST Token 4.2: Pos = VBD PART
Frases preposicionales complemento del nombre	Secuencia 5: Token 5.1: Pos = SUST ADJ Token 5.2: Pos = PREP Token 5.3: Pos = SUST

Para mayor claridad, en la Figura 18 se presenta un ejemplo del "manchado" obtenido en uno de los párrafos del primer texto de la Tabla II. Como se observa, en este párrafo se han coloreado automáticamente los rasgos de las distintas secuencias presentadas en la Tabla 3 con un color particular y se ha calculado la densidad informacional (aparición conjunta de las secuencias) por párrafo, cuyo resultado es el índice presentado en la columna de la derecha.

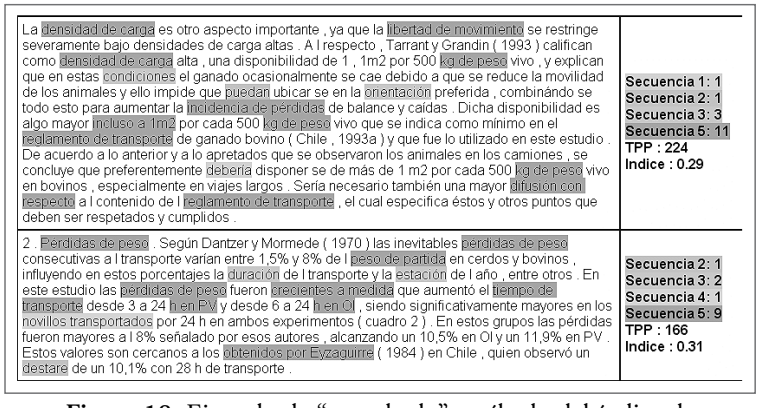


Figura 18. Ejemplo de "manchado" y cálculo del índice de densidad informacional de un párrafo de texto. (Corpus ARTICOS; Archivo BIO_205).

En este ejemplo, correspondiente a los párrafos 40 y 41 del artículo de investigación científica, puede verse claramente la distribución de cada una de las manifestaciones textuales de las secuencias y su relación con las demás en ambos párrafos. De este modo, podemos observar la densidad que estos párrafos tienen, no sólo en términos cualitativos (agrupación de distintos colores) sino también en términos cuantitativos (0,29 y 0,31, respectivamente).

Ahora bien, como mencionamos anteriormente –a partir de la información estadística– la herramienta calcula un índice general de densidad, en este caso informacional, del texto completo. Precisamente, este índice será el dato que nos permitirá comparar los textos de cada registro. Por último, como ya se sabe, junto a la página de resultado del manchado existe una página en la cual se presenta la estadística general por párrafos del texto, en ella se consigna la frecuencia por rasgos en cada párrafo, la cantidad de palabras totales de los párrafos y la densidad de cada párrafo, tal como se observa en el ejemplo de la Tabla IV.

Tabla IV. Datos estadísticos generales del texto BIO_205.

Párrafo	Secuencia1	Secuencia2	Secuencia3	Secuencia4	Secuencia5	TPP	Densidad por Párrafo
1	0	0	2	0	2	27	0.3
2	0	0	0	0	0	1	0
3	0	0	3	0	4	134	0.1
4	0	0	2	0	3	113	0.09
5	0	0	3	1	6	265	0.11
6	0	0	4	0	3	88	0.16
7	0	0	0	0	0	13	0
8	0	0	0	0	0	1	0
9	0	0	5	1	0	134	0.09
10	0	1	9	0	8	160	0.34

Todos estos datos son posibles de ser copiados y pegados en una tabla Excel o en algún programa estadístico, con lo que el análisis cuantitativo puede ser extendido a otros cálculos requeridos por el investigador.

Como mencionamos, la densidad por párrafo es calculada a través de la fórmula $IDL_x T$. A modo de ilustración presentamos la aplicación de la fórmula para el caso del párrafo presentado de la Figura 16.

Así:

$IDL_xT \rightarrow IDL_iT$ = Promedio IDL_iP's

Donde la equis es reemplazada por la i (de informacional), de este modo la fórmula expresada para el cálculo por párrafos se expresa como:

$$IDL_iP = \frac{\sum(fR_i) * \sum(TR_i)}{\sum pP}$$

Donde el índice de densidad informacional expresado lingüísticamente en el párrafo es el cuociente de la suma de las frecuencias de cada secuencia de rasgos asociados a la informatividad y el producto de esta sumatoria con la cantidad de tipos de rasgos presentes en el texto dividido por la sumatoria de palabras en el párrafo (lo que permite normalizar el dato para luego promediarlo con todos los párrafos del texto).

Ahora bien, si llenamos las variables de la fórmula obtenemos lo siguiente:

$$IDL_iP = \frac{(1 + 2 + 1 + 9) * 4}{166} = 0{,}31$$

El valor que presenta la densidad informacional del párrafo es promediado por el resultado obtenido para cada párrafo en el texto. De este modo, se establece un valor que representa la densidad informacional total del texto. Cabe señalar que, si bien la fórmula identifica la co-ocurrencia de los rasgos, ésta no da cuenta de la importancia o peso estadístico que cada secuencia de rasgos pueda tener en relación con las otras secuencias de rasgos.

4.4. Resultados

Acorde a los procedimientos previamente presentados, los resultados obtenidos en esta aproximación son los que se presentan en la Tabla V.

Tabla V. Índices de densidad informacional por cada uno de los textos investigados.

Corpus fuente	Código del texto	Textos de la muestra	Índice
ARTICOS	A1	Prevalencia de lesiones podales en ovinos de 25 explotaciones familiares de la provincia de Valdivia, Chile	0,46
	A2	Escherichia coli aislada de cerditos diarreicos. Presunción de cepas productoras del Factor Citotóxico Necrosante (CNF)*	0,49
	A3	Sensibilidad tisular a la insulina antes, durante y después de un ayuno en ovejas prepúberes*	0,65
	A4	Efecto de melatonina sobre la secreción pulsátil de hormona luteinizante y de hormona del crecimiento en borregas con restricción alimenticia	0,76
	A5	Efectos del tiempo de transporte de novillos previo al faenamiento sobre el comportamiento, las pérdidas de peso y algunas características de la canal*	0,70
NEMOL	N1	Ministro: Operadores deberán dar explicaciones por atraso del Transantiago	0,45
	N2	Cobre dispara superávit fiscal en primer trimestre	0,21
	N3	Pearl Jam lanzó nuevo disco que retoma sus raíces	0,22
	N4	Llegó a Chile el Dalai Lama	0,09
	N5	Pellegrini anuncia que gira del Villarreal a Chile quedó postergada	0,32
DICIPE	D1	"La democracia no se exporta"	0,13
	D2	¿Qué es en estos días ser un chileno (a)?	0,12
	D3	¡Somos 10!: Confirmada existencia de nuevo planeta	0,21
	D4	Aceite de emú: fórmula chilena rinde examen	0,3
	D5	Baja el smog, pero aumentan enfermedades respiratorias	0,27

Como se puede observar en la Tabla V, y se verifica en el Gráfico I, los textos que presentan el mayor índice de densidad informacional son aquéllos obtenidos del corpus ARTICOS, destacando los textos A4 y A5 con los puntajes más altos y presentando, en general como grupo, los puntajes más altos en comparación con los demás textos. Los textos correspondientes a NEMOL y DICIPE presentan en general índices menores y mayor variación interna que ARTICOS.

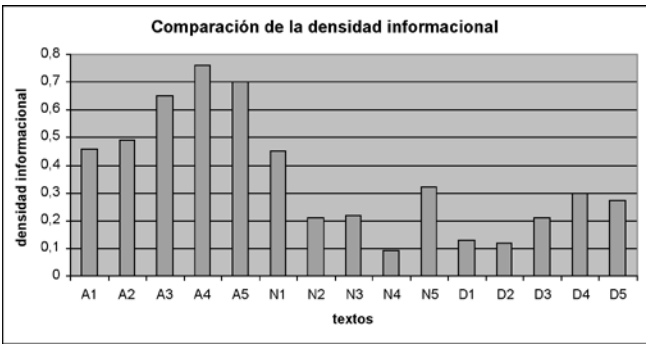


Gráfico I. Comparación de los índices de densidad informacional en los textos de la muestra.

Ahora bien, si observamos el índice promedio de densidad informacional en cada uno de los registros, tal como se expresa en la Tabla VI, nos percatamos de la existencia de un continuum que va desde un registro científico altamente denso informacionalmente a uno periodístico en donde se evidencia un menor índice de densidad informacional.

Tabla VI. Índice promedio de densidad informacional.

CORPUS	REGISTRO	ÍNDICE PROMEDIO DE DENSIDAD INFORMACIONAL
ARTICOS	Científico	0,615
DICIPE	Divulgación científica	0,402
NEMOL	Periodístico	0,258

El Gráfico II evidencia con mayor claridad la presencia del *continuum* que existe en cuanto a la densidad informacional entre los registros estudiados. Por otra parte, se observan diferencias cuantitativas significativas (p= 0,05) entre el registro científico y el registro periodístico.

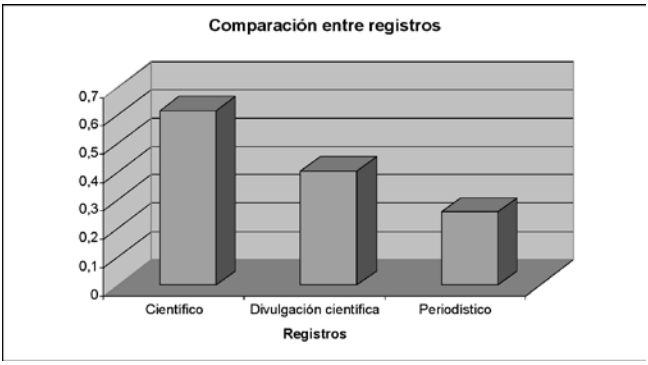


Gráfico II. Índice de densidad informacional entre registros.

Es interesante observar que entre el registro científico y el de divulgación científica no existe diferencia significativa, lo que podría ser atribuible a que comparten temáticas científicas y también a que la co-ocurrencia de rasgos lingüísticos compartidos condensa esta información. En cuanto al registro de divulgación científica y el periodístico, la no distinción estadística es atribuible a que ambos registros comparten, seguramente, una co-ocurrencia similar de otros rasgos funcionales diferentes al de la densidad informacional y propios del registro periodístico (por ejemplo: una modalidad narrativa similar).

5. COMENTARIOS FINALES

La herramienta "El Manchador de Textos" reúne características que conjugan aspectos que pueden ser de interés para los investigadores cuyo objeto de estudio son los textos, esto porque la herramienta permite: a) identificar, destacando por medio de colores o "manchando", la posición, en un texto cualquiera, de un conjunto de rasgos lingüísticos y b) cuantificar la co-ocurrencia de estos rasgos, seleccionados para la descripción de algún fenómeno lingüístico textual e incluso discursivo. Por lo mismo, esta herramienta permite a los investigadores realizar sus estudios desde una perspectiva tanto cualitativa como cuantitativa, a través de una interfaz fácil de usar.

Desde el punto de vista cuantitativo la herramienta ofrece dos opciones a tener en cuenta: 1) el dato de frecuencia de cada rasgo buscado y 2) un índice de densidad de co-ocurrencia sistemática. Este último dato es particularmente interesante, puesto que permite comparar una serie de textos en función de la selección de rasgos. Además, se presentan dos opciones de visualización de los datos cuantitativos. Una en correspondencia con el manchado y otra en la página de estadísticas, la que a su vez tiene la posibilidad de ser copiada en una página Excel o de algún programa estadístico para hacer nuevos cálculos estadísticos.

Un aspecto diferenciador de esta herramienta de análisis textual respecto de otras existentes es que no sólo busca rasgos lingüísticos por forma, sino también por partes de la oración y por la combinación de estos dos aspectos. Esto la convierte en una herramienta potente en cuanto a su poder descriptivo. Además, la presentación final del texto analizado mantiene su forma original, "horizontalizada". Sin duda, esta característica es una de las más innovadoras de la herramienta y permite al investigador tener un acercamiento "ecológico" al texto y observar las relaciones de los rasgos lingüísticos seleccionados para el análisis.

En cuanto a la investigación presentada, podemos establecer que la herramienta permite distinguir un continuum entre los registros estudiados, acorde con la densidad informacional que éstos presentan. De este modo, el registro científico es el que mayor densidad informacional presenta, diferenciándose significativamente del registro periodístico. Esto último está en concordancia con los hallazgos presentados en Parodi (2005):

> " […] es esclarecedor que la Dimensión 5 "Foco Informacional" se presente con un puntaje promedio positivo (0,9), lo que permite identificar y distinguir significativamente el CTC (Corpus Técnico Científico) del CLL (Corpus de Literatura Latinoamericana) y CEO (Corpus de Entrevistas Orales). Las nominalizaciones, frases preposicionales como complemento del nombre y participios pasivos adjetivos, entre otros, son rasgos de alta relevancia en estos textos escritos. Estos rasgos contribuyen al cumplimiento del dispositivo informacional, el cual se aleja claramente de contenidos interpersonales y afectivos". (Parodi, 2005:102)

Esta semejanza en los resultados de ambas investigaciones permite comprobar de forma empírica funciones expresadas lingüísticamente en los textos, en nuestro caso densidad informacional, a pesar de la diferencia metodológica entre ambas investigaciones. Y, de este modo, corroborar en alguna medida las investigaciones que consideran el discurso científico como más abstracto y denso, en cuanto a la presentación de la información en él contenida (Ciapuscio, 2000 y 2003; Cassany, López y Martí, 2000; Moyano, 2000; López, 2002; Mogollón, 2003, Parodi, 2007b y 2007c).

Por último, proyectamos realizar mejoras en la herramienta que permitan una mejor interacción entre el investigador y "el manchado" que hace la herramienta, de modo que puedan modificarse algunos rasgos de dudoso etiquetado o incluir algunos no detectados adecuadamente, aspectos que a su vez influirían en el cálculo final del índice de co-ocurrencia de los rasgos investigados. Esto haría que la herramienta fuera mucho más flexible y abierta a las necesidades de los investigadores.

REFERENCIAS

Cademartori, Y., Parodi, G. y Venegas, R. 2006. "El discurso escrito y especializado: Caracterización y funciones de las nominalizaciones en los manuales técnicos". En *Literatura y Lingüística*, 17, 243-265.

Cassany, D.; López, C.; Martí, J. 2000. "Divulgación del discurso científico: La transformación de redes conceptuales. Hipótesis, modelo y estrategias". En *Discurso y sociedad, 2* (2), 73-103.

Ciapuscio, G. 2000. "Hacia una tipología del discurso especializado". En *Discurso y Sociedad, 2* (2), 39-71.

Ciapuscio, G. 2003. *Textos especializados y terminología*. Barcelona: IULA.

Gutiérrez, R. M. 2007. "Oralidad, escritura y especialización: Una caracterización desde el sistema de la modulación". En Giovanni Parodi (Ed.). *Lingüística de corpus y discursos especializados: Puntos de mira*. Valparaíso: Ediciones Universitarias de Valparaíso, pp. 149-178.

López, C. 2002. "Aproximaciones al análisis de los discursos profesionales". En *Revista Signos, 35* (51-52), 195-215.

Mogollón, I. 2003. "Paradigma científico y lenguaje especializado". En *Revista de la Facultad de Ingeniería de la Universidad Central de Venezuela*, 18 (3), pp: 5-14.

Moyano, E. 2000. *Comunicar ciencia*. Buenos Aires: Secretaría de Investigaciones. Universidad Nacional de Lomas de Zamora.

Parodi, G. 2005. "Lingüística de corpus y análisis multidimensional: Exploración de la variación en el corpus PUCV-2003: Una aproximación multiniveles". En G. Parodi (Ed.), *Discurso especializado e instituciones formadoras*. Valparaíso: Ediciones Universitarias de Valparaíso PUCV, 83-125.

Parodi, G. 2006. "El Grial: interfaz computacional para anotación e interrogación de corpus en español". RLA, 44 (2), 91-116.

Parodi, G. 2007a. El discurso especializado escrito en el ámbito universitario y profesional: Constitución de un corpus de estudio. *Revista Signos*, 2007, 40 (63), 147-178.

Parodi, G. 2007b (editor). *Lingüística de corpus y discursos especializados: Puntos de mira*. Valparaíso: Ediciones Universitarias de Valparaíso PUCV.

Parodi, G. 2007c. *Working with Spanish corpora*. Londres: Continuum.

Parodi, G. 2007d. *Lingüística de Corpus*. Buenos Aires: Eudeba.

Parodi, G. y Venegas, R. 2004. *"Bucólico*: Aplicación computacional para el análisis de textos: Hacia un análisis de rasgos de la informatividad". En *Literatura y Lingüística*, 15, 2004, 223-257.

Sabaj, O. 2007. "Hacia una matriz de rasgos lingüísticos con impacto textual: Un estudio exploratorio". En *Revista Signos*, 2007, 40 (63), 197-218.

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

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

Venegas, R. 2007. "Clasificación de textos académicos en función de su contenido léxico-semántico". *Revista Signos*, 2007, 40 (63), 239-271.

ANEXO 1

Descripción de la fórmula IDLT

En términos formales la fórmula IDLT se define de la siguiente manera:

$$IDL_xT = \overline{X}\, IDL_xP's$$

Así, el Índice de Densidad Lingüística del Texto calculado a partir de un conjunto de rasgos lingüísticos co-ocurrentes sistemáticamente, es igual al promedio del Índice de Densidad Lingüística de los Párrafos del texto (IDL_xP's).

A su vez, el IDL_xP's se define de la siguiente manera:

$$IDL_xP = \frac{\sum(fR_x) * \sum(TR_x)}{\sum pP}$$

Donde:

IDL_xP = cuociente de densidad lingüística por párrafo

x = función comunicativa determinada

$\sum(fR_x)$ = sumatoria de las frecuencias (ocurrencias) de cada rasgo lingüístico buscado

$\sum(TR_x)$ = sumatoria de la cantidad de tipos de rasgos que aparecen en el párrafo (co-ocurrencia)

$\sum pP$ = total de palabras del párrafo

Como se puede deducir, el Índice de Densidad Lingüística del Texto está determinado por el promedio de los índices de Densidad Lingüística de cada Párrafo del texto. A su vez, la división por el total de palabras de cada párrafo asegura una normalización que permite la comparación del índice final con los índices obtenidos de otros textos, independiente de la cantidad de palabras que éstos posean.