Káñina, Rev. Artes y Letras, Univ. Costa Rica. XXXVI (Especial): 135-143, 2012 / ISSN: 0378-0473

CALCULADOR LÉXICO-ESTADÍSTICO DE FRECUENCIA DISPERSIÓN Y USO (CaLeFDU)

*Jorge Antonio Leoni de León**

RESUMEN

CaLeFDU (por *Calculador Léxico-estadístico de Frecuencia Dispersión y Uso*) es un programa que, dado un corpus en formato texto con marcadores de exclusión y desambiguación, lematiza automáticamente los vocablos que lo conforman a través de series de consultas al etiquetador morfológico del analizador sintáctico Fips (LATL 2008). Una vez obtenidos los resultados de Fips, CaLeFDU efectúa los cálculos de frecuencia, dispersión y uso para todos los vocablos asociados con un lema tomando en consideración su distribución por campos temáticos. Al final, CaLeFDU produce un diccionario léxico–estadístico con el inventario de todos los lemas recuperados y vocablos asociados.

Palabras clave: Análisis sintáctico profundo, lematización, frecuencia léxica, lexicografía, análisis morfológico, procesamiento del lenguaje natural.

ABSTRACT

CaLeFDU is an application, written in Perl, that, given a corpus, calculates the values for frequency, dispersion and usage. CaLeFDU is based in the deep syntactic parser Fips developed at the Language Technology Laboratory at the University of Geneva (LATL 2008). Fips helps lemmatize word occurrences in order to produce a statitistical dictionary.

Key Words: Deep syntactic parsing, lemmatization, lexical frequency, lexicography, morphological analysis, lexicography, Natural Language Processing.

1. Introducción

La Lexicografía, como ciencia descriptiva, pretende dar cuenta del léxico empleado por todos los hablantes de una lengua determinada, en nuestro caso, el español. Sin embargo, a fin de que los resultados sean representativos, no basta con disponer únicamente de un listado de términos y sus definiciones: también necesitamos tener una idea del "peso" de cada término, de sus sentidos y los contextos en que estos se producen. De esta manera, es posible desligarse de la subjetividad establecida a través de los gustos y la memoria (siempre imperfecta) de los redactores. Como consecuencia, se obtienen descripciones más ajustadas a la realidad de la lengua y, por extensión, a las necesidades de los usuarios de obras lexicográficas.

Aunque las técnicas que permiten resultados más objetivos son diversas, todas tienen en común el uso de corpus como fuente primaria de información. Los métodos empleados para extraer información del corpus, y como ésta se emplea finalmente, marca, junto con los procedimientos de selección y recolección, la diferencia en cuanto a la perspectiva que caracteriza a cada obra lexicográfica contemporánea. El trabajo lexicográfico pasa, entonces, por el levantamiento de un inventario de términos a partir de sus realizaciones en una colección de textos que pueden ser transcripciones de entrevistas con informantes. Para ilustrar una parte de la labor de extracción, tomemos una muestra del capítulo XXII de la segunda parte

* Profesor de la Escuela de Filología, Lingüística y Literatura. Universidad de Costa Rica.
Recepción: 16/05/11. Aceptación: 04/07/11.

del Quijote[1]; el propósito es establecer una lista parcial de los lexemas presentes:

> (1) Pero nunca **pensé** que hacía mal en ello: que toda mi intención era que todo el mundo **se holgase** y **viviese** en paz y quietud, sin **pendencias** ni **penas**; pero no me **aprovechó** nada este buen **deseo** para **dejar** de **ir adonde** no **espero volver**,[...]

De la muestra (1) podemos excluir las conjunciones, las preposiciones y la mayor parte de los determinantes. Si seguimos la tradición en cuanto a las formas canónicas, las formas en singular de los sustantivos y las de infinitivo para los verbos son fáciles de identificar. Sin embargo, los verbos conjugados y los sustantivos en plural (para no hablar que de dos categorías gramaticales), algunos de los aparecen marcados en negritas, deben ser identificados y asociados a una forma canónica. En el cuadro 1 tenemos un ejemplo de resultados parciales:

CUADRO 1
Resultados de frecuencias simples en una muestra pequeña

		Frecuencia
Lema	excluir *v.*	
Realización	excluir	1
	Total	1
	...	
	ser *v.*	
	era	1
	Total	1

Dada una muestra mayor, la dificultad se multiplica si tomamos en cuenta los tiempos compuestos y toda la fenomenología asociada a los pronombres clíticos (como *darme* o *dándoselo*), además de las unidades fraseológicas, siempre muy frecuentes y de formas canónicas no necesariamente consensuadas. De esta manera, vemos no sólo como el trabajo es enorme (en parcitular en el caso de corpus consistentes en millones de vocablos), sino que es fácil comprender la necesidad de encontrar métodos de procesamiento automático que faciliten la tarea.

Ahora bien, tal y como ya lo señalamos, la vida de un lexema no se limita a su significado, sus rasgos gramaticales y la cantidad de veces que aparece en un corpus (frecuencia), sino que su valor también está asociado a los contextos en que se utiliza. Dichos contextos o universos pueden ser establecidos según criterios operacionales. De esta manera, en Leoni de León (1997), los universos fueron asociados con las secciones de los periódicos, de manera similar a la metodología empleada por Barahona Novoa (1996) para la radio; en el caso de entrevistas, los universos pueden darse en el continuum de la expresión y no estar tan bien delimitados, sino que depende de la pericia del entrevistador para elicitar vocabulario (Murillo Rojas 1996; Murillo Rojas y Sánchez Corrales 2002).Sin importar cuál sea la concepción de tema o universo en la que nos basemos, lo que es necesario encontrar es la razón resultante de la relación entre un lexema y los universos en los que aparece con respecto a un total de contextos temáticos y vocablos empleados. Los índices obtenidos darán el peso de un lexema en el discurso.

En nuestro caso, nosotros nos propusimos establecer un método de extracción de datos morfosintácticos a partir de un corpus manualmente codificado. Nuestro objetivo fue crear una aplicación para la redacción de diccionarios lexicoestadísticos sobre la base de la frecuencia, la dispersión y el uso, a partir del procesamiento automático de un corpus. Dicha aplicación la bautizamos "Calculador Léxico- estadístico de Frecuencia Dispersión y Uso", CaLeFDU, según sus siglas[2]. Adicionalmente, también buscamos contar con un programa que recupere los valores gramaticales de los vocablos de un corpus a través de un párser (analizador sintáctico), estableciendo indicadores de resultados para las diferentes etapas del análisis lexicoestadístico, que faciliten la corrección manual de errores en la codificación del coprus o que incluso permitan recuperar la mayor cantidad de información posible en caso de fallas del sistema computacional o de la consulta de datos. El resultado final es un programa que no sólo calcula los valores para un diccionario lexicoestadístico, sino que además produce un archivo, en formato de diccionario, con los resultados de los cálculos. El corpus que empleamos consistió en redacciones de pruebas de bachillerato de colegios públicos y privados

de las veinte regiones educativas de Costa Rica, debidamente codificadas y clasificadas según ciclo, universo, tipo de institución educativa y sexo de los participantes. Estos datos forman parte de los resultados del proyecto de investigación "Diccionario de la Lengua Española Secundaria"[3] a cargo del Prof. Víctor Ml. Sánchez Corrales, con el cual colaboramos para acelerar la lematización de una muestra de redacciones codificadas de las pruebas de bachillerato de colegios públicos y privados de las veinte regiones educativas.

El corpus del proyecto "Diccionario de la Lengua Española Secundaria" consistía en una serie de archivos de texto, uno por cada redacción. La figura 1. ilustra la estructura de cada archivo, donde las primeras cuatro líneas corresponden a marcadores predeterminados para el cruce de variables; estas consisten en un símbolo seguido de un número: el signo de cierre de interrogación "?" indica el ciclo educativo al cual pertenece el informante, la almohadilla "#" se refiere al colegio donde se recolectó la muestra, el símbolo de dólar "$" brinda información sobre el género, mientras que la arroba "@" apunta a los datos del "universo" al cual pertenece la redacción. Es de notar que CaLeFDU no tiene ningún límite establecito para la cantidad de universos posibles, sino que simplemente utilizará el número mayor que encuentre para determinar el total. En cuanto a la codificación, los signos de puntuación y las palabras entre barras, "/. . . /", son ignorados al momento del cálculo. Adicionalmente, los sufijos alfanuméricos de los vocablos distinguen significados. Los textos fueron codificados manualmente; actualmente una opción para codificar automáticamente o incluso obviar todo manipulación del texto está en estudio.

En las líneas que siguen, pasamos a describir el funcionamiento de CaLeFDU y los principios que los sustentan. La sección 2 está dedicada a las definiciones de los conceptos fundamentales y la fórmula empleada. Los recursos con que contamos y la implementación están explicados en la sección 3. La aplicación es descrita, de manera general, en la sección 4, donde también se explican los archivos auxiliares y la creación del diccionario léxico–estadístico. Las conclusiones están contenidas en la sección 5.

```
?2
#2
$2
@2

/El/ "piercing" ¿moda peligrosa?

¿/Será/ alarmante /y/ peligroso /un/ artefacto
/de/ este6 calibre1? ¿Merece /su/ debida
atención, /un/ tema como5 éste6? ¿/O/ /es/
simplemente una6 moda, un6 grito1 /de/
superficialidad /de/ /los/ jóvenes2 /por/
/demostrarse/ demostrar /se/ vanidad?

/Son/ inexorables todas6 estas6 incógnitas
/que/ habitan /mi/ mente /y/ divagan /en/
busca1 /de/ respuestas, provocando3d un6
torbellino /de/ desesperación, /en/ busca1
/de/ algo1 más5 /que/ no /sea/ duda1 /ni/
pesadumbre. /Por/ esta6 razón1b, /me/ preparo
/para/ evacuar todo6 pensamiento1b /que/ no
/se/ encuentre3c claro4b /y/ engrandezca
/mi/ ignorancia /hacia/ un6 tema /de/ esta6
relevancia.
```

FIGURA 1
Codificación del corpus

2. Definiciones y fórmula empleada

Tomando como referencia la 1, la frecuencia (F) consistirá en el número de realizaciones de los vocablos[4] asociados a un lexema. El sentido de establecer universos es, entonces, establecer la distribución de los lexemas en el conjunto de universos; este dato, conocido como dispersión (D), podemos definirlo como la distribución de las realizaciones de un lema en 'contextos' temáticos diferentes. La relación entre la frecuencia y dispersión nos da el uso (U), dicho en otros términos, el uso es la frecuencia matizada por la dispersión.

La expresión matemática de estas relaciones ya nos estaba dada por la metodología empleada por el programa ELEXHICÓS, arriba citada, que recurre a la forma de Juilland empleada por Morales (1986), de manera que nosotros simplemente debíamos implementarla en un marco computacional. Morales (1986), entendiendo la frecuencia (F) como la cantidad absoluta de las realizaciones de un lexema, utilizó tal y como la reproducimos en la figura 2, donde:

- *n* es el número de mundos, universos o temas.
- *x* es la frecuencia de cada vocablo en cada universo.
- $\Sigma(x_i) = T$ es la frecuencia simple.
- $\Sigma(x_i)^2 = T$ es la frecuencia simple elevada al cuadrado.

Consecuentemente, el uso (U) es igual a la frecuencia (F) por la dispersión (D):

(2) Uso

U = F (D).

$$D = 1 - \frac{\sqrt{nx_i^2 - T^2}}{2T}$$

FIGURA 2
Fórmula de la dispersión

Estas fórmulas y la estructura general del diccionario (cuadro 1) son los requisitos fundamentales de CaLeFDU para obtener los valores distributivos los vocablos lematizados, donde haya un cálculo de las frecuencias por universos más las sumatorias y valores derivados según la fórmula de la figura 2. Estos resultados tendrán una presentación similar a los del cuadro 2, donde se muestra la distribución del verbo "venir" (sentido 3, definido por los lexicógrafos) para los seis universos del corpus, así como los valores finales correspondientes derivados del proceso de cálculo.

CUADRO 2
Estructura de un artículo en el diccionario estadístico

venir–3						
venía3	*0*	*0*	*0*	*1*	*0*	*0*
venga3	*0*	*0*	*0*	*1*	*0*	*0*
vengo3	*1*	*0*	*0*	*0*	*0*	*0*
vienen3	*0*	*0*	*0*	*3*	*0*	*1*
viene3	*0*	*1*	*0*	*2*	*0*	*0*
vinimos3	*0*	*0*	*0*	*1*	*0*	*0*
Totales:	**1**	**1**	**0**	**8**	**0**	**1**
$\Sigma(x_i) = T$ [frecuencia simple]: **11**						
TOTALES qd:	*1*	*1*	*0*	*64*	*0*	*1*
$\Sigma(x_i)^2$: 67						
Universos: **6**						
DISPERSIÓN: **0.23804297207999**						
USO: **2.61847269287989**						

3. Recursos e implementación

El proceso de lematización automatizado exige la utilización de un sistema que establezca los valores morfosintácticos, de cada vocablo, asociándos con una forma canónica correspondiente. Ahora bien, las lenguas naturales se caracterizan por su alto nivel de ambigüedad. Así, por ejemplo, a nivel léxico, una secuencia de caracteres puede representar más de una categoría gramatical, tal y como lo ilustra el término "recibo" en (3) y (4):

3) El recibo fue firmado por el director.
4) Recibo este producto en consignación.

En los ejemplos (3) y (4), la desambiguación de "recibo" no presenta ninguna dificultad para un hablante del español, puesto que para toda persona que se expresamente correctamente en dicha lengua es bastante obvio que en (3) "recibo" es un sustantivo, mientras que en (4) se trata de un verbo. Sin embargo esto no se cumple para una computadora, para lo cual el desarrollo de un sistema computacional capaz de efectuar este tipo de distinciones puede constituir todo un reto a largo plazo. Debido que por el momento carecemos de un procesador de lenguaje natural capaz de hacer ese tipo de distinciones, recurrimos al sistema de análisis sintáctico[5] Fips (Wehrli 2007; Leoni de León, Schwab y Wehrli 2008) del *Laboratoire d'Analyse et de Technologie du Langage* (LATL 2008) de la Universidad de Ginebra, el cual no sólo conocemos bien por otros proyectos, sino que además es posible interactuar con él por medio de un *web service*.

Ahora bien, Fips cuenta tanto con un analizador sintáctico, como con un etiquetador morfológico ("tagger" en inglés). Esto significa que dada una entrada como la que tenemos en (5), Fips retorna esa misma frase enriquecida con un etiquetado sintáctico, libremente inspirado en la gramática generativa chomskyana (N. Chomsky 1995; Noam Chomsky 2004), entre otras fuentes (Wehrli 2007). El resultado correspondiente a (5) lo tenemos en (6):

5) Input: Veo las nubes en el cielo.
6) Ouput: [TP[DP] Veo [VP [DP las [NP nubes]][PP en [DP el [NP cielo]]]]]

En (6) podemos apreciar cómo Fips marca el sujeto desinencial con un "[DP]" vacío que, en otras circunstancias, estaría ocupado con un sujeto léxicamente realizado. Además, los complementos directo y circunstancial no sólo fueron reconocidos, sino que están correctamente asociados con el sintagma verbal. El hecho de que el verbo conjugado no esté dentro del VP se debe a que Fips modeliza el movimiento del verbo hacia la posición de la inflexión donde se chequea la concordancia entre el sujeto y el verbo.

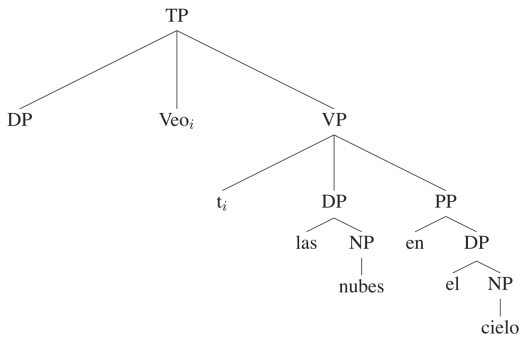

FIGURA 3
Análisis sintáctico de Fips

Tal y como se desprende de (6), Fips produce estructuras ternarias y no binarias como cabría esperar por el modelo teórico del cual se inspira. La figura 3 muestra la estructura arborescente correspondiente; en ella se aprecian más fácilmente la relaciones establecidas por medio de los subíndices que señalan tanto el origen[6] ("t_i"), como el destino del elemento desplazado (el verbo "veo$_i$"). Estas capacidades de Fips para el análisis sintáctico profundo pueden ser útiles, en otro momento, para proponer una lematización directa partir de un corpus no anotado. Sin embargo, en esta ocasión, sólo explotamos el etiquetador morfológico, el cual produce un desglose de los rasgos morfosintácticos de los vocablos que complementa los resultados del análisis sintáctico (figura 3). De esta manera, obtenemos los lemas correspondientes a los vocablos del *input* (columna *lema*), además del conjunto de rasgos asociados (columna *rasgos*). Adicionalmente, en la columna *función* están desplegadas las informaciones relativas al caso (objeto directo, OBJ) o a la relación de los complementos con el sujeto (FPO).

CUADRO 3
Resultados del etiquetador morfológico de Fips

Vocablos	*Rasgos*	*Secuencia*	*Lema*	*Función*
Veo	VER-IND-PRE-1-SIN	511017884	ver	
las	DET-PLU-FEM	511007887	el	OBJ
nubes	NOM-PLU-FEM	511013770	nube	
en	PRE	511007093 15	en	FPO
el	DET-SIN-MAS	511007887	el	
cielo	NOM-SIN-MAS	511002378	cielo	
.	PONC-point	0	.	

4. La Aplicación

CaLeFDU es un programa orientado a objetos escrito en Perl[7] que funciona desde la línea de comando (figura 4). CaLeFDU consiste en dos módulos: Dispuso.pm y Fips.pm. El programa, tal y como nosotros lo utilizamos, está definido en un programa llamado Herdacuso.pl, que se limita a emplear las funciones exportadas por Dispuso.pm.

Como lo hemos señalado anteriormente, nosotros recibimos un corpus ya codificado, del cual había que excluir los vocablos entre barras (figura 1). El corpus estaba conformado por un conjunto de documentos de texto, almacenados en una carpeta específica en la cual CaLeFDU ingresa para leer cada uno y extraer los vocablos válidos (según los parámetros de la codificación). Este procedimiento es repetido para cada archivo de texto en vez de ser aplicado a la totalidad. Esta estrategia permite la creación de bitácoras precisas para el registro de los procedimientos en cada etapa del cálculo; esto luego facilita la detección de errores por nombre de archivo (figura 5) o secuencias de códigos (figura 6), en las que también se incluyeron las frecuencias simples.

Luego de la creación de las bitácoras, CaLeFDU consulta cada vocablo con el

etiquetador morfológico de Fips; en el caso de vocablos con frecuencias simples mayores a 1 (F > 1), no es necesario efectuar la consulta más de una vez, puesto que para este fin el trabajo fue realizado sobre sumatorias de las frecuencais de los vocablos. De los datos obtenidos para cada uno, únicamente retuvimos las informaciones de rasgo y lema. El módulo a cargo de estas labores fue Fips.pm (Leoni de León 2008).

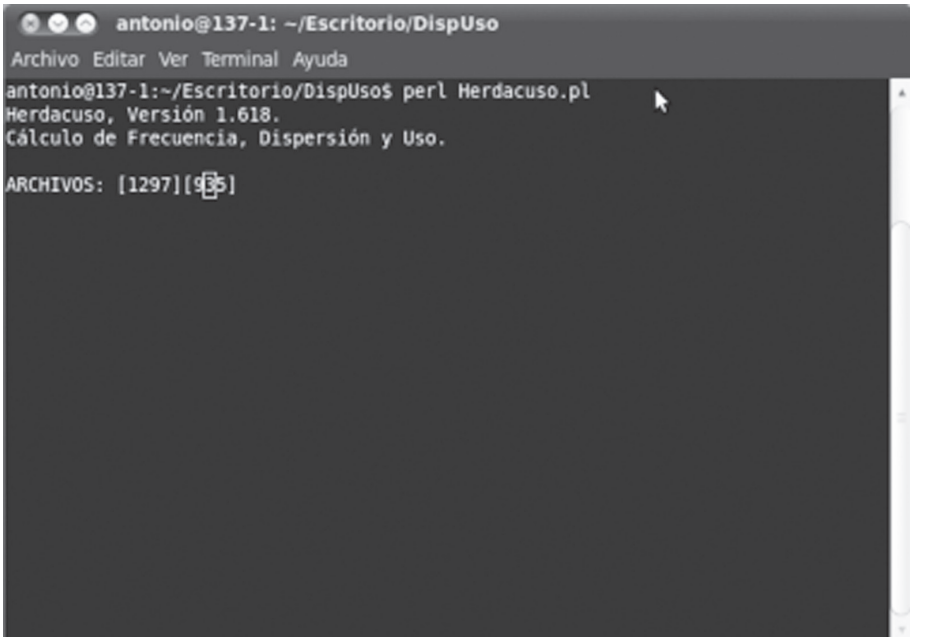

FIGURA 4
Inicio del programa

Las bitácoras resultaron ser instrumentos muy útiles. Ellas permitieron identificar y localizar muy fácilmente varios tipos de errores, con lo que las modificaciones en el corpus podían ser realizadas rápidamente. Sin embargo, cada vez que uno o varios errores eran corregidos y una nueva revisión era hecha, debíamos ejecutar todo el programa de nuevo. Como es sabido, la consulta de miles de ítemes léxicos vía web puede tomar mucho tiempo debido a los tiempos de conexión o a la disponibilidad del ancho de banda según el horario en el que se efectúe el cálculo. De ahí que CaLeFDU utiliza la técnica de programación conocida como *marshalling*, con la cual se transforma un conjunto de datos en un formato que permite almacenarlos para ser explotados posteriormente[8]. Por medio del *marshalling* guardamos los resultados previos de Fips, de manera que CaLeFDU sólo solicitaba los vocablos faltantes. Esto fue particularmente útil para procesos incompletos o interrumpidos.

Durante la ejecución del programa, la ventana de la línea de comando proveía una serie de informaciones útiles que permitían saber si el vocablo analizado tenía alguna acepción marcada en el corpus, cuáles rasgos le correspondían, qué forma canónica le iba a ser asignada, si los datos estaban siendo recuperados directamente de Fips o si eran extraídos por medio de *marshalling*. Además, hay un indicador del avance con respecto al total de vocablos encontrados. La figura 7 ilustra estas informaciones para los números, los cuales, por cierto, no fueron eliminados del corpus en la codificación manual.

Antes de crear el diccionario estadístico propiamente dicho, CaLeFDU produce dos índices: DispUso-indice.txt (figura 8) y DispUso-tabla_borrador.txt (figura 9). El primero

consiste una lista de los lemas encontrados con los vocablos que le fueron asociados y las frecuencias simples correspondientes; el segundo reproduce una lista con los datos asociados a cada vocablo que indica las frecuencias por variables y universos, los rasgos, la forma canónica y las acepciones. El índice DispUso-indice.txt puede ser útil en caso de querer realizar modificaciones manuales en cuanto a la clasificación de los vocablos. El archivo DispUso-indice.txt permite comprender mejor los datos utilizados en el cálculo final de frecuencia, dispersión y uso.

```
ARCHIVO: texto/CBIJ_0622_cod.txt

amiga -> 2
anemia -> 1
angustiado -> 1
años -> 1
a+pesar+de10 -> 1
[...]
```

FIGURA 5
Archivo INDEX-vocarchfg.txt producido automáticamente

```
DATOS POR SECUENCIA

[...]

TOTAL DE LA SECUENCIA: 14415
18. SECUENCIA DE CÓDIGOS por archivo: =CICLO=2=COLEGIO=2=SEXO=1=UNIVERSO=3
     abastecen 1
     absoluta 1
     accesibles 1
     acceso1 1
     aceptar 1
     acostumbraba 1
[...]
```

FIGURA 6
Archivo DispUso-data.txt producido automáticamente

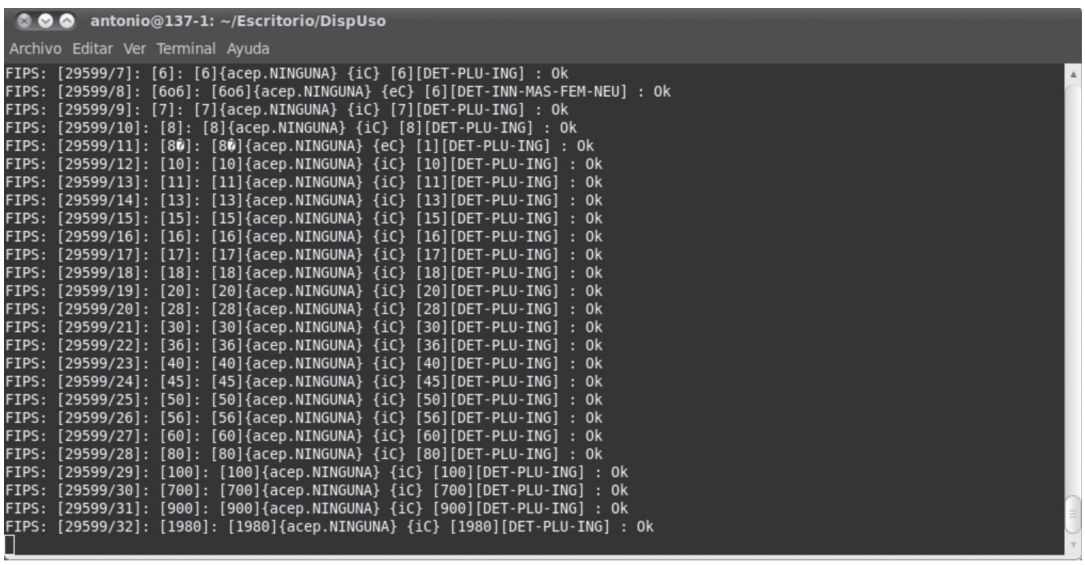

FIGURA 7
Informaciones en la ventana de terminal

```
INDEX: ENTRADA -> VOCABLOS -> FRECUENCIA

[...]

faltar
     faltó      1
     falté      1
     faltaba    1
     faltaban   1
     faltan     1
     faltar     1
     faltaron   1
     falte      1
     falten     1
falta-1a
     faltas1a   1
falto
     falta      1
     faltos     1
[...]
```

FIGURA 8
Archivo DispUso-indice.txt

```
  logra

       CICLO=1 -> 10
       CICLO=2 -> 5
       COLEGIO=1 -> 7
       COLEGIO=2 -> 8
       DIC_ACEPCION -> NINGUNA
       DIC_FORMA_CANONICA -> lograr
       DIC_RASGOS -> VER-IND-PRE-3-SIN
       DIC_VOCABLO -> logra
       SEXO=1 -> 12
       SEXO=2 -> 3
       UNIVERSO=1 -> 1
       UNIVERSO=2 -> 1
       UNIVERSO=4 -> 6
       UNIVERSO=6 -> 7
```

FIGURA 9
Archivo DispUso-tabla_borrador.txt

La última etapa del proceso es la creación del diccionario, en formato texto, en el archivo Dispuso-DICO.txt según el esquema mostrado en el cuadro 2. Aunque en esta versión contiene algunas informaciones superfluas, como resultados de cálculos intermedios, que no son necesarias ni para la versión final publicable, ni para los lexicógrafos, decidimos mantenerlas para estimar mejor la eficacia de CaLeFDU y de la fórmula de cálculo empleada; siempre será sencillo eliminarlas automáticamente en un proceso de edición posterior, si se quiere publicar el diccionario.

5. Conclusiones

Aunque el programa descrito aquí es apenas una primera versión, CaLeFDU mostró ser de gran valor en la lematización automática de corpus codificados manualmente. Sin embargo, por diversos motivos, los resultados no están exentos de errores; la principal razón es la ambigüedad del lenguaje humano y la imperfección de la codificación manual, a lo que también hay que agregarle los límites del léxico de Fips frente a la riqueza del lenguaje en el mundo real. En esta investigación propusimos varias alternativas para compensar cualquier tipo de error, sin embargo aún se requiere una implementación más profunda de dichas técnicas. Por ejemplo, el *marshalling* puede ser perfeccionado para recuperar estados incompletos de cálculo; también el índice del archivo DispUso-indice.txt podría ser mejor utilizado para las correcciones manuales. Sin duda alguna, también se puede hacer mucho para ofrecer mejores formatos de salida, además del texto simple.

En algún momento señalamos la posibilidad de prescindir de la codificación manual. Esta es una posibilidad que queremos explorar en el futuro próximo en una versión que ofrezca la opción entre un tratamiento a partir de un texto no anotado y uno que tome en consideración la codificación manual. Todavía queda pendiente sopesar las ventajas y desventajas de cada uno de esos métodos. Esto apunta también a la necesidad de contar con nuestro propio analizador sintáctico, al que le podamos realizar mejorías directamente.

Finalmente, aunque no era el objeto de esta investigación, era algo que ya nos estaba dado, recomendamos la actualización de la fórmula. La que se utilizó ahora funciona bien para un máximo de cuatro universos. Desde nuestra perspectiva, lo mejor sería que en una versión futura de CaLeFDU se incluyan en opción varias fórmulas de cálculo de dispersión o, ¿por qué no?, la posibilidad de realizar otros cálculos sobre el corpus.

Notas

1. http://www.cervantesvirtual.com/obra/el-ingenioso-hidalgo-don-quijote-de-la-mancha--0/.
2. CaLeFDU fue inscrito en el Instituto de Investigaciones Lingüísticas (proyecto No 745-B0-183 de la Vicerrectoría de Investigación de la Universidad de Costa Rica).
3. Proyecto No 745–A4–122 de la Vicerrectoría de Investigación de la Universidad de Costa Rica.
4. Donde vocablo se define como una secuencia de caracteres entre espacios en blanco.
5. En inglés, "parsing".
6. Técnicamente denominado "huella".
7. Información en línea en http://www.perl.org
8. Otra alternativa, habría sido guardarlos en una base de datos directamente; sin embargo, para las necesidades del proyecto, el *marshalling* se ajustó perfectamente y dio muy buenos resultados.

Referencias

Barahona Novoa, José Alberto. 1996. «Léxico básico de la radio costarricense». Tesis de lic. San José, Costa Rica: Posgrado en Lingüística, Universidad de Costa Rica.

Chomsky, N. 1995. The minimalist program. Current studies in linguistics series. The MIT Press. ISBN: 9780262531283. URL: http://books.google.com/books?id=vtPQiYCNpjgC.

Chomsky, Noam. 2004. «Structures and Beyond: The Cartography of Syntactic Structures». En: ed. por A. Belletti e Italy) Certosa di Pontignano (Pontignano. Vol. 3. Structures and Beyond. Oxford University Press. Cap. Beyond Explanatory Adequacy, págs. 104-131. ISBN: 9780195171976. URL: http://books.google.com/books?id=e5qwmk4uQcC.

ELEXHICÓS. 2008. Estudios de Lexicografía Hispano-Costarricense (ELEXHICÓS). Página web. URL: http://www.lexicografia.ucr.ac.cr.

LATL. 2008. Laboratoire d'Analyse et de Technologie du Langage. Página web. [Dirección electrónica : http://www.latl.unige.ch/ ; Visitada el: 28 de abril de 2008]. Université de Genève. Ginebra, Suiza.

Leoni de León, Jorge Antonio. 1997. «Léxico Básico de la Prensa Escrita Costarricense». Tesis de Licenciatura. Escuela de Filología, Lingüística y Literatura de la Universidad de Costa Rica.

Leoni de León, Jorge Antonio. 2008. «Modèle d'analyse lexico-syntaxique des locutions espagnoles». [Publicado en línea: http://www.unige.ch/cyberdocuments/theses2008/LeonideLeonJA/meta.html]. Tesis doct. Ginebra, Suiza: Université de Genève. URL: http://www.unige.ch/cyberdocuments/theses2008/LeonideLeonJA/meta.html.

Leoni de León, Jorge Antonio, Sandra Schwab y Éric Wehrli. 2008. «Análisis sintáctico profundo del español: un ejemplo del procesamiento de secuencias idiomáticas». En: Procesamiento del Lenguaje Natural. Ed. por Paloma Martínez Fernández, Dolores Cuadra Fernández y F. Javier Calle Gómez. 41. Sociedad Española para el Procesamiento del Lenguaje Natural, Departamento de Informática, Universidad de Jaén. Jaén, págs. 37-44. URL: http://www.sepln.org/revistaSEPLN/revista/41/sec1-art5.pdf .

Morales, Amparo. 1986. Léxico básico del español de Puerto Rico. Academia Puertorriqueña de la Lengua Española.

Murillo Rojas, Marielos (1996). «Léxico básico de los niños preescolares de la subregión educativa San José, Costa Rica». Tesis de lic. San José, Costa Rica: Posgrado en Lingüística, Universidad de Costa Rica.

Murillo Rojas, Marielos y Víctor Manuel Sánchez Corrales. 2002. Léxico básico de los niños preescolares costarricenses. San José, Costa Rica: Editorial de la Universidad de Costa Rica.

Wehrli, Éric. 2007. «Fips, A "Deep" Linguistic Multilingual Parser». En: ACL 2007 Workshop on Deep Linguistic Processing. Prague, Czech Republic: Association for Computational Linguistics, págs. 120-127. URL: http://www.aclweb.org/anthology/W/W07/W07-1216.