Káñina, Rev. Artes y Letras, Univ. Costa Rica XXXVI (2): 171-179, 2011 / ISSN:0378-0473

"DÍGALO": HERRAMIENTA DE APOYO BÁSICO PARA ESTUDIANTES DE L2**

*Jorge Antonio Leoni de Léon**

RESUMEN

Se describe aquí una herramienta de apoyo al aprendizaje básico del léxico y la escritura en una lengua extranjera, basada en el Procesamiento del Lenguaje Natural y el conocimiento lingüístico. Esta herramienta, llamada "Dígalo", se ha sido desarrollado para el español, pero nuestro objetivo es ampliar su aplicación a otros idiomas de Costa Rica. "Digalo" es una aplicación web que consiste en cuatro módulos especializados: el primero, Hable.pm, gestiona la coherencia de datos, el segundo, Hamin.pm, muestra la aplicación en sí, el tercero, Lexadal.pm organiza la estructura de datos, y el cuarto, Fips.pm, se comunica con el analizador Fips del Laboratorio de Idiomas de la Tecnología de la Universidad de Ginebra (http://www.latl.unige.ch) para comprobar las producciones de los usuarios de la frase. En la pantalla "Digalo" tiene dos partes, un léxico y una forma de escribir frases en un idioma extranjero. Si la flecha del mouse apunta sobre una palabra, se muestra una imagen que lleva a un vínculo entre la palabra y su significado (como una imagen de la casa para la palabra "casa"). Cuando el usuario/a hace clic sobre el formulario, éste/a puede comenzar a escribir una frase incluyendo una o varias palabras del léxico que se ofrece (dividido en cuatro categorías: sustantivos, adjetivos, verbos y adverbios). Una vez que el usuario ha introducido la frase, su validez se comprueba a través de Internet, por FIPS (Wehrli 2007). Si la frase es válida, el usuario puede escribir una nueva frase. Si la frase es incorrecta, se muestra un mensaje de error y se invita al usuario a reintentarlo. Esta metodología busca mejorar la competencia del usuario en el idioma extranjero a un nivel de principiante. "Digalo" ofrece una solución útil en el aprendizaje supervisado.
Palabras clave: herramientas de aprendizaje, L2, Lingüística Aplicada, análisis profundo, español como segunda lengua.

ABSTRACT

We describe a tool supporting the basic learning of the lexicon and writing in a foreign language, based on Natural Language Processing and Linguistic knowledge. This tool, called "Dígalo" (for the Spanish "say it!"), has been developed for Spanish, but we aim to expand its application to other languages of Costa Rica. "Dígalo" is a web application consisting in four specialized modules: First, Hable.pm manages data consistency; second, Hamin.pm displays the application itself; third, Lexadal.pm organizes the data structure; and Fips.pm communicates with the parser Fips at the University of Geneva's Language Technology Laboratory (http://www.latl.unige.ch) to check user's phrase productions. On the screen "Dígalo" has two parts, a lexicon and a form to write phrases in a foreign language. If the mouse's arrow points over a word, a picture is displayed bringing a link between the word and its meaning (so, a house picture for "casa"). When the user clicks over the form, she can start writing a sentence including one or several words from the lexicon offered (which is divided in four categories: nouns, adjectives, verbs and adverbs). Once the user has entered the phrase, its correctness is checked, via web, by Fips (Wehrli 2007). If the phrase is valid, the user can write a new phrase. If the phrase is incorrect, an error message is shown up and the user is invited to try again. This methodology searches to improve user's competence in foreign language at a beginner's level. "Dígalo" provides a solution useful in supervised learning.
Key Words: learning tools, L2, Applied Linguistics, NLP applications, deep parsing, Spanish L2.

---

* Profesor de la Escuela de Filología, Lingüística y Literatura, Universidad de Costa Rica.

** Este artículo corresponde a las I[as] Jornadas de Lexicografía, además esta Ponencia fue presentada en la 18[a] Conferencia Bienal de la Sociedad de Lingüística del Caribe, Universidad de las Antillas Occidentales, recinto de Cave Hill, Barbados, agosto de 2010. *Recepción: 29/04/11. Aceptación: 09/07/12.*

1. Introducción

La Lingüística Aplicada es un área que en los próximos años se verá particularmente beneficiada con los avances en el Procesamiento del Lenguaje Natural (PLN), en particular en lo que respecta a la enseñanza de segundas lenguas. En este artículo, presentamos una herramienta computacional destinada a apoyar el aprendizaje del léxico y la escritura en L2 a nivel de principiantes. Dicha herramienta, llamada "Dígalo", aunque ha sido desarrollada para el español, también está pensada para extender su uso a otras lenguas.

La aplicación "Dígalo" consiste en varios módulos, uno de los cuales está encargado de la detección de errores por medio del análisis sintáctico de los valores de entrada, es lo que denominamos "estimación de gramaticalidad". En los últimos años, la literatura lingüística se ha visto enriquecida con diversas propuestas en este sentido. La mayoría de las investigaciones se focalizan en la fonología. Así, Chul-Ho y col. (1998) proponen un sistema automático detección de errores de pronunciación en japonés como L2, con envío de diagnósticos a los usuarios; mientras que Abhinav Sethy y Johnson (2005) presentan un sistema de diálogo hombre–máquina, con reconocimiento de habla, basado en métodos probabilísticos, para detectar errores de pronunciación en dialectos árabes y en pashtu en el marco del "Tactical Language Training System" (TLTS). Por otra parte, Lee y Seneff (2006), según un modelo de generación basado en n-gramas, también describe un sistema de diálogo hombre–máquina, pero para estudiantes de inglés, que permite corregir errores fonéticos y conversacionales. Nuestra orientación es morfo–sintáctica, por lo que estas investigaciones no tienen mayores repercusiones en este trabajo.

En lo que concierne a la sintaxis encontramos sistemas especializados en la detección de errores puntuales, como las reglas de las preposiciones para el sueco (Eeg-Olofsson y Knutsson 2003) o el inglés (Chodorow, Tetreault y Han 2007), este último elaborado a partir de un corpus de textos de hablantes no nativos. Brück y Stenzhorn (2008) utilizan una estrategia dinámica para la detección automática de errores en las gramáticas de generación, lo que se aprovecha para deducir una regla gramatical incorrecta por medio de minería de datos; Tschichold (2003), por su lado, enfoca los aspectos léxicos del aprendizaje de lenguas asistido por computadora. No está de más mencionar que el aprendizaje de la escritura también ha sido objeto de incursiones, como lo ejemplifican Hu y col. (2009) con su sistema de corrección de trazos de los caracteres chinos. Las capacidades de Fips en la detección de errores no necesitan ser probadas: no es la primera vez que Fips es utilizado con este propósito. Finalmente, aprovechando las características de un analizador sintáctico profundo, L'Haire y Vandeventer Faltin (2003) presentan el proyecto FreeText, el cual es un sistema automatizado de detección de errores para estudiantes del francés como lengua extranjera, cuya metodología fue abordada con profundidad en Vandeventer Faltin (2003). Nosotros nos ubicamos en esta última tendencia, por las razones que detallamos a continuación.

Aunque es cierto que la mayoría de los esfuerzos están basados en métodos estocásticos, también es necesario decir que estos encuentran ciertos límites que nosotros queremos evitar. El aspecto más importante para nosotros, en este sentido, es que los métodos probabilísticos a pesar de que permiten sentar rápidamente las bases para analizar las formas más frecuentes de la oración simple (1a), carecen de la fineza suficiente para el análisis de relaciones estructurales profundas (1b) (Leoni de León, Schwab y Wehrli 2008):

(1) a. Ana rompió el récord.
b. El récord de Luis fue roto por Ana.

Así, por ejemplo, en (1a) tenemos una secuencia estándar Sujeto–Verbo–Objeto replanteada en (1b) como una pasiva, cuyo sujeto es modificado por un sintagma preposicional ("de Luis"). La distancia entre el núcleo de la frase ("fue roto") y el sujeto de la pasiva es, en términos generales, demasiado grande para analizadores superficiales ("shallow parsers") que son los empleados en los métodos estocásticos. Por este motivo, las relaciones complejas en frases similares a (1b) son

identificadas más eficazmente por medio de analizadores sintácticos profundos (Wehrli 2007; Leoni de León, Schwab y Wehrli 2008). Esto y el hecho de que los recursos necesarios para desarrollar analizadores superficiales no son abundantes en español, nos hacen optar por un analizador sintáctico profundo, concretamente por el párser Fips (Leoni de León, Schwab y Wehrli 2008; Wehrli 2007; Wehrli 2004) del Laboratorio de Análisis y de Tecnología del Lenguaje (LATL) de la Universidad de Ginebra, el cual está disponible en línea para varios idiomas (Laboratoire d'Analyse et de Technologie du Langage 2010).

Fips es un analizador sintáctico profundo multilingüe (disponible para inglés, francés, alemán, italiano, español y griego), cuya concepción teórica es una adaptación libre de la gramática generativa chomskyana, con influencias de los modelos minimalista (Chomsky 2004; Chomsky 1995; Chomsky 1993) y "Simpler Syntax" (Culicover y Jackendoff 2005), así como de la Gramática Léxico–Funcional (Bresnan 2001). Dada una frase de entrada (2a), Fips brinda, como salida, el etiquetado sintáctico correspondiente (2b), así como las funciones y los ragos de los elementos de la oración (cuadro 1):

(2) a. Anoche observamos la luna.

b. [TP[AdvP Anoche][DP] observamos [VP [DP la [NP luna]]]]

No entraremos en los detalles del análisis sintáctico efectuado por Fips, en las referencias de este artículo hay abundantes referencias en ese sentido. Sin embargo, sí es necesario poner de relieve que el etiquetado sintáctico de Fips reconoce el sujeto tácito, como es posible observarlo en el sintagma determinante vacío, [DP], en la oración (2b); de estar ocupada la posición de sujeto, en la columna vocablo aparecería el valor de entrada, y en la columna función estaría la indicación SUBJ, por sujeto. Dos características importantes apreciables en (2a) son que el análisis no es binario, sino trinario y que, en este caso, el sintagma principal está marcado como TP (sintagma de tiempo, correspondiente al sintagma de la inflexión); esto se debe a que el símbolo inicial, CP (sintagma complementante), está obviado por estar vacío, lo que no ocurriría, por ejemplo, si se tratra de una oración interrogativa. En el etiquetado morfológico, es sobre el sintagma determinante "la" que recae el valor de objeto (columna función del cuadro 1), esto por cuanto en Fips se modeliza la Hipótesis DP, según la cual, los sintagmas nominales son argumentos de los sintagmas determinantes.

En el caso de una oración mal formada (3a), Fips reenvía un análisis incompleto encabezado por la advertencia que tenemos en (3b):

(3) a. Anoche ella observamos la luna.
b. *** no analysis

"Dígalo" recupera los datos del etiquetador (cuadro 1), sólo si el análisis sintáctico es positivo.

Cuadro 1: Etiquetado morfológico

Vocablo	Rasgos	ID	Lema	Función
anoche	ADV	511016629	anoche	
observamos	VER-IND-PRE-1-PLU	511005165	observar	
la	DET-SIN-FEM	511007887	el	OBJ
luna	NOM-SIN-FEM	511013755	luna	

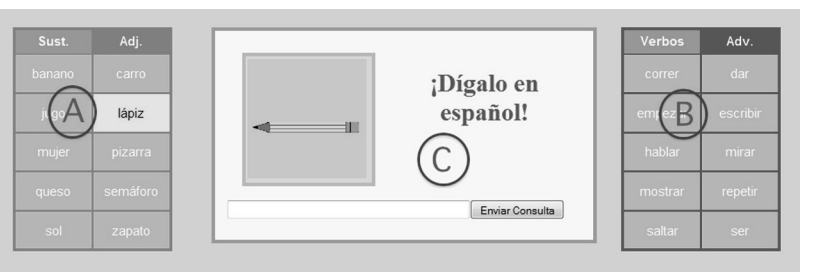
Este artículo está presentado en la forma de un "demo"; es decir, que nuestro objetivo primordial es mostrar el funcionamiento general de un software. Así, en la sección 2 elaboramos una descripción general de la aplicación "Dígalo" que incluye detalles de la interacción entre el usuario y el sistema. La sección 3 brinda datos muy generales sobre la arquitectura del programa, antes de llegar a las conclusiones en la sección 4.

2. Descripción general

"Dígalo" es una aplicación web; es decir, ha sido desarrollada usando la tecnología web como soporte principal. Esto implica que el programa está almacenado en un servidor, por lo que los usuarios deben interactuar con el sistema a través de un navegador. Es importante señalar que "Dígalo" se encuentra en desarrollo, por lo que está hospedado en un servidor de pruebas y no es libremente accesible por el momento.

Nuestra aplicación consiste en una sola ventana con dos columnas (que denominadas A y B en la figura 1) y un espacio intermedio (indicado como C). Las columnas A y B consisten en listas de palabras clasificadas según su categoría gramatical. En A tenemos dos pestañas, que permiten alternar entre sustantivos y adjetivos, y una lista de sustantivos; la columna B tiene una estructura similar para los verbos y los adverbios. El espacio intermedio C consiste en un cuadrado en el que se desplegarán las imágenes, la frase "Dígalo en español" , un campo de texto para introducir las frases y un botón para enviar las informaciones. El área de texto para los resultados no aparece en la pantalla hasta que se comience a enviar datos. En síntesis, el usuario dispone de dos columnas con el léxico, un área de imágenes dinámicas y un formulario para el envío de las frases por evaluar, tal y como aparece en la figura 1.

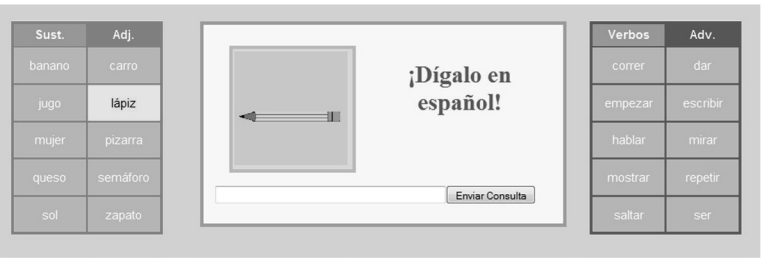
Figura 1: Aspecto general de "Dígalo"

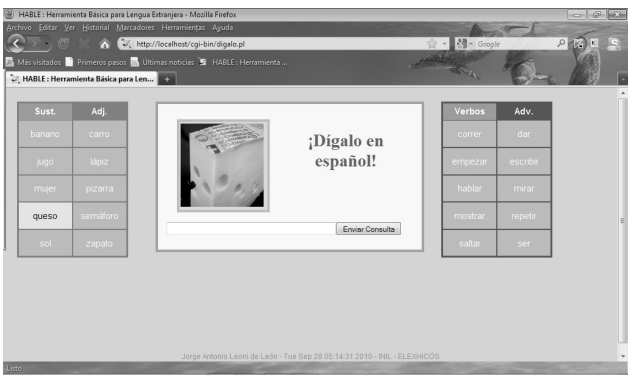


2.1. Interacción con el usuario

Cada entrada léxica de "Dígalo" está asociada con una imagen que la evoca. Para acceder a esta información, el usuario debe colocar el puntero del ratón sobre una palabra. El resultado de esta acción será una imagen en el espacio reservado. Así, si el puntero pasa sobre la palabra "casa", la celda del lema cambia de color anaranjado a amarillo y el usuario podrá ver la imagen de una casa en el cuadro correspondiente; en la figura 2 vemos los resultados de esta acción para "lápiz" y "queso".

Figura 2: Evocación de sentidos por medio de imágenes





En el campo de texto, el usuario debe introducir la frase por evaluar. Una vez presionado el botón de "enviar la consulta", los datos son enviados al servidor del LATL donde se hospeda la versión web de Fips, la cual devuelve como resultado el análisis sintáctico. "Dígalo" recibe un análisis completo en caso de una evaluación exitosa; si la frase no pudo ser analizada (por agramaticalidad, por ejemplo), lo que se recibe es una nota de análisis incompleto, como lo indicamos en el ejemplo (3b), la cual es interpretada como una inadecuación. En el primer caso, sobre el espacio del formulario, "Dígalo" despliega la exclamación "¡Correcto!"; en el segundo, el sistema invita a tratar una nueva versión de la frase indicando "¡Otra vez!". En la figura 3 vemos un ejemplo exitoso con la frase "La mujer escribe en la pizarra"; mientras que la figura 4 ilustra el resultado de una frase errónea (un error de concordancia entre el sujeto y el verbo en la frase "La mujer saltamos en la pizarra").

Figura 3: Resultados correctos

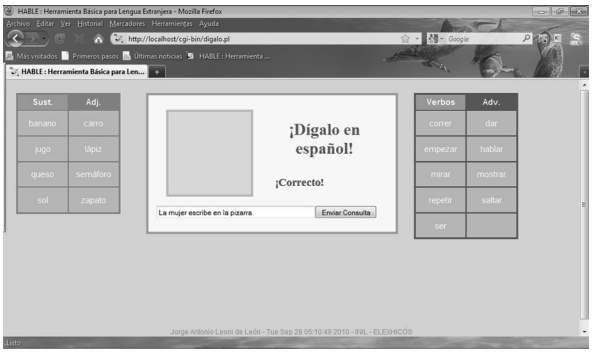
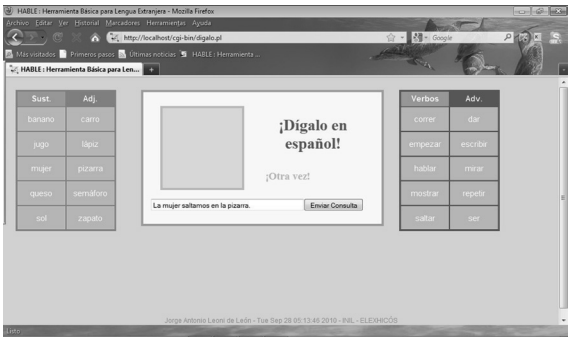


Figura 4: Resultados incorrectos



En el caso de una frase correcta, si se ha utilizado el léxico de alguna de las columnas, los ítemes léxicos empleados en la frase desaparecen de las columnas correspondientes: esto es lo que ocurre con "mujer", "escribir" y "pizarra" en la figura 3, los cuales son reconocidos en la frase introducida, sin importar qué modo o tiempo aparezcan, para ser asociados con una forma canónica que es bajo la que aparecen en las columnas del léxico. La asociación vocablo-lema es efectuada por el etiquedator de Fips (cuadro 1). Esta metodología busca mejorar la competencia léxica y las habilidades escritura de estudiantes principiantes de L2, de preferencia bajo supervisión de un profesor.

3. Arquitectura del sistema

"Dígalo" es una aplicación web completamente escrita en Perl (http://www.perl.org) consistente en cuatro módulos especializados. En primer lugar, tenemos Hable.pm, el cual administra la coherencia de los datos. En segundo lugar, el módulo Hamin.pm se encarga de desplegar la aplicación propiamente dicha; es decir, el conjunto de páginas en HTML, con los formularios y los datos que median entre varios estados. El inventario léxico, consistente en un lema asociado a una imagen y una categoría gramatical es administrado por el módulo Lexadal.pm, que administra la estructura jerárquica de datos. Finalmente, el módulo Fips.pm es el que establece la comunicación entre Fips, en los servidores del Laboratorio de Análisis y Tecnología del Lenguaje en la universidad de Ginebra, y "Dígalo", desarrollado por Leoni de León (2008) como parte de un modelo computacional léxico-sintáctico de las locuciones del español.

4. Conclusiones

"Dígalo" conjuga nuevas tecnologías, enseñanza de lenguas extranjeras y conocimiento teórico de la Lingüística en una original propuesta fácilmente adaptable para su utilización en el aula o, incluso, para el estudio en casa, bajo condiciones mínimas de supervisión. La expansión de este proyecto a otras lenguas está prevista; sin embargo primero es necesario afinar la versión en español. La dependencia hacia analizador sintáctico Fips del LATL, no nos permite aportar directamente nuestras propias modificaciones en el analizador sintáctico, por lo que es necesario emprender iniciativas que nos dirijan al desarrollo de nuestra propia tecnología de "parsing" (análisis sintáctico automatizado); el hecho de que "Dígalo" interactúe con Fips por medio de un servicio web, muestra de qué manera se pueden incluir otros sistemas similares. "Dígalo" busca ante todo mejorar la competencia de estudiantes de L2. Las posibilidades son inmensas, por lo que esperamos mejorar sensiblemente esta aplicación en el futuro próximo.

Reconocimientos

Quiero agradecer a la Rectoría, la Vicerrectoría de Investigación, el Instituto de Investigaciones Lingüísticas (INIL) y la Escuela de Filología, Lingüística y Literatura de la Universidad de Costa Rica por su apoyo al proyecto No 745–A8–188, del cual "Dígalo" forma parte, y que me permitió presentar esta propuesta en la Conferencia Bienal de la Sociedad Caribeña de Lingüística.

Referencias

Abhinav Sethy Nicolaus Mote, Shrikanth S. Narayanan y Lewis Johnson. 2005. Modeling and automating detection of errors in Arabic language learner speech. *InterSpeech ISCA*: 177-180.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Brück, Tim vor der y Holger Stenzhorn. 2008. *A Dynamic Approach for Automatic Error Detection in Generation Grammars*. ECAI: 837-838.

Chodorow, Martin, Joel Tetreault y Na-Rae Han. 2007. *Detection of Grammatical Errors Involving Prepositions*. *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*. Online documents at URL < http://www.aclweb.org/anthology/W/W07/W07-1600.pdf>.

Chomsky, Noam. 1993. *A Minimalist Program for Linguistic Theory*. *The View from Building 20*. Cambridge, Massachusetts: 3-52.

_____.1995. The Minimalist Program. Cambridge, Massachusetts: MIT Press.

_____. 2004. Beyond Explanatory Adequacy. Structures and Beyond: The Cartography of Syntactic Structures. Ed. por Adriana Belletti. (3). Oxford: Oxford University Press: 104-131.

Chul-Ho, Jo y col. 1998. Automatic pronunciation error detection and guidance for foreign language learning. ICSLP (Paper 0741). Online documents at URL < http://www.shlrc.mq.edu.au/proceedings/icslp98/PDF/AUTHOR/SL980741.PDF>.

Culicover, Peter W. y Ray Jackendoff. 2005. Simpler Syntax. Oxford Linguistics. Oxford University Press.

Eeg-Olofsson, Jens y Ola Knutsson. 2003. Automatic grammar checking for second language learners - the use of prepositions. Nodalida. Online documents at URL <http://www.nada.kth.se/~knutsson/eegolofsson_knutsson.pdf>.

Hu, Zhihui y col. 2009. A Chinese Handwriting Education System with Automatic Error Detection. JSW 4.2: 101-107.

Laboratoire d'Analyse et de Technologie du Langage. 2010. Online documents at URL < http://www.latl.unige.ch>.

Lee, John y Stephanie Seneff. 2006. Automatic Grammar Correction for Second–Language Learners. INTERSPEECH–2006 ICSLP, Ninth International Conference on Spoken Language Processing. Pittsburgh, PA, USA.

Leoni de León, Jorge Antonio. 2008. Modèle d'analyse lexico-syntaxique des locutions espagnoles. Ph.D. tesis, Université de Genève, Ginebra, Suiza, 24 de mayo. Online documents at 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. 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: 37-44. Online documents at URL < http://www.sepln.org/ revistaSEPLN/revista/41/sec1-art5.pdf>.

L'Haire, Sébastien y Anne Vandeventer Faltin. 2003. Error Diagnosis in the FreeText Project. CALICO Journal 20.3: 481-495. Online documents at URL < http://sebastien.lhaire.org/publis/06 L'haireVandeventer.pdf >.

Tschichold, Cornelia. 2003. Lexically Driven Error Detection and Correction. CALICO Journal 20.3: 549-559.

Vandeventer Faltin, Anne. 2003. Syntactic error diagnosis in the context of computer assisted language learning. Tesis doct. Geneva: Université de Genève.

Wehrli, Éric. 2004. Un modèle multilingue d'analyse syntaxique.Structures et discourse:mélanges offferts à Eddy Roulet. Ed. por Antoine Auchlin y col. Langue et pratiques discursives. Montréal: Éditions Nota Bene: 311-329.

Wehrli, Éric. 2007. Fips, A "Deep" Linguistic Multilingual Parser. ACL 2007 Workshop on Deep Linguistic Processing. Prague,

Czech Republic: Association for Computational Linguistics: 120-127. Online documents at URL < http://www.aclweb.org/anthology/W/W07/W07-1216 >.

Wikipedia (2010). Wikimedia Commons — Wikipedia, La enciclopedia libre. Online documents at URL < http://es.wikipedia.org/w/index.php?title=Wikimedia Commons& oldid=40474392 >.