MEMORIA RESUMIDA DEL TRABAJO

1. TÍTULO DEL TRABAJO

AZRAEL: A-Z Reconocedor Automático de Español

2. AUTORES DEL TRABAJO:

Apellidos, nombre (coordinador)	Universidad:
Álvarez Mellado, Elena	Universidad Complutense de Madrid
Apellidos, nombre	Universidad:
Apellidos, nombre	Universidad:

3. TUTOR O TUTORES DEL TRABAJO (si los hubiera)

Apellidos, nombre	Organismo:
Basterrechea Molina, Eduardo	Universidad Complutense de Madrid
Apellidos, nombre	Organismo:
Apellidos, nombre	Organismo:

4. BREVE RESUMEN DE LA INVESTIGACIÓN (máx. 500 palabras)
(Debe indicarse la aportación personal del candidato al trabajo y el grado de intervención del tutor/es en el mismo)

La motivación principal al realizar este trabajo ha sido reflexionar sobre la siguiente cuestión: ¿qué hace que una palabra sea española? De esta pregunta nació AZRAEL, un programa informático capaz de detectar textos en castellano. El aspecto básico sobre el que se ha realizado la detección del idioma es la forma de las palabras, ya que es este (y no la existencia de significado) el mecanismo fundamental por el que un hablante reconoce el idioma de una palabra. La aspiración, por lo tanto, no ha sido que AZRAEL detecte exclusivamente lo que un diccionario consideraría español, sino que hemos querido que nuestro detector admita como español todo aquello que un hablante admitiría, es decir, todo aquello que tenga apariencia de español, al margen de que tenga significado o no. Un detector de estas características aceptaría lo que normativamente se considera español, pero también otras producciones menos convencionales, como neologismos, términos propios de la jerga callejera, palabras sin sentido propias de juegos y canciones infantiles o palabras inventadas.

Para la creación de AZRAEL, necesitábamos partir de un patrón de la estructura silábica de las palabras del español. A falta de estudios en profundidad sobre el tema, la investigación ha consistido también en realizar un estudio exhaustivo de la estructura de la sílaba en español a partir de un corpus del español de más de 600.000 palabras. A partir de este corpus, hemos analizado qué combinaciones de letras son posibles dentro de las sílabas del español, y qué combinaciones están prohibidas. Este estudio ha supuesto la clave fundamental para la detección del idioma. La estructura silábica del español apenas sí ha sido abordada previamente con la profundidad que el tema merece, aun cuando resulta una fuente inagotable de información para caracterizar una lengua y para extraer datos sobre la naturaleza de la palabra. Por tanto, con el estudio de la sílaba no sólo hemos buscado llegar a un patrón que se pueda implementar en un programa informático, sino también comprobar cuánta información podemos sacar de la palabra a partir de su estructura silábica.

El resultado del proyecto ha consistido en la creación de AZRAEL (A-Z Reconocedor Automático de Español), programa de detección automática de español, y de dos programas auxiliares: el Silabeador, encargado de la división silábica de las palabras a detectar, y ESDRA (Etimólogo Selectivo Del Reconocedor Automático). Asimismo, se ha realizado un estudio exhaustivo de la estructura silábica del español que ha revelado la estrecha relación entre la sílaba, los procesos de prefijación y el origen de las palabras. La concepción de la idea, el estudio teórico y el desarrollo del programa han sido realizados por la candidata, con el apoyo y la ayuda del tutor en el manejo del corpus.

5. ANTECEDENTES DEL TEMA TRATADO (Máx. 2000 palabras)

El interés por la naturaleza del lenguaje es casi tan antiguo como el lenguaje mismo. La primera gramática de la que se tiene testimonio es la gramática del sánscrito de Panini, que se remonta al siglo V a.C. aproximadamente. En occidente, es en Grecia donde nacen los primeros estudios gramaticales y donde aparece por primera la diatriba sobre la relación entre significante y significado, que retomaría Saussure a principios del siglo XX.

Pero más allá de cuestiones filosóficas sobre la naturaleza del lenguaje, el problema de las diferencias entre lenguas ha surgido de la propia convivencia e interacción entre poblaciones lingüísticamente distintas. Los orígenes de la comparación lingüística se remontan a la Edad Media. Uno de estos primeros trabajos comparativos lo realizó Dante Alighieri, clasificando las lenguas en base a la forma de la palabra *sí*, distinguiendo las que procedían de la forma latina *sic*, de las demás. A finales del siglo XVI, en Diatriba de Europaeorum Linguis, Joseph Justus Scaliger hace una nueva clasificación de las lenguas, esta vez caracterizándolas por la palabra "dios". Estos primeros trabajos de comparación de lenguas eran, por tanto, clasificaciones léxicas y constituyen el precedente de la Tipología Lingüística.

A lo largo del siglo XVII aparecen los primeros intentos de superar las limitaciones que suponían las diferencias lingüísticas mediante artefactos mecánicos. Este hecho se debe a dos causas: por un lado, la desaparición progresiva del latín como lengua universal para la comunicación científica, y por otro, el interés de racionalistas como Leibniz y Descartes en crear un lenguaje artificial, lógico y universal que estuviese libre de la ambigüedad de los lenguajes naturales. La idea era crear lenguajes con códigos numéricos en vez de palabras, para poder crear textos universales que cualquiera pudiera leer disponiendo de un diccionario en el que cada código remitiese al término equivalente.

Las propuestas de posibles lenguajes artificiales numéricos se sucedieron hasta el siglo XX, y algunos de estos diccionarios mecánicos se llevaron a cabo durante 1920 y 1930. Sin embargo, no es hasta los años cuarenta cuando se desarrollaron las primeras máquinas para el desciframiento de código, traducción y detección automática del lenguaje. El matemático inglés Alan Turing (que había trabajado durante la Segunda Guerra Mundial en el desciframiento de los mensajes secretos

alemanes) defendía que si un ordenador podía descifrar un código, también podría traducir un idioma extranjero. Tras esta afirmación se encuentra la noción de que un texto escrito en un idioma extranjero no es más un mensaje en un código desconocido pero susceptible de ser descifrado. La misma aproximación se encontraba en las cartas del matemático americano Warren Weaver del año 1947:

(…) Uno se pregunta si el problema de la traducción podría ser tratado como un problema de criptografía. Cuando miro un artículo en ruso, me digo: "Esto es en realidad inglés, pero está codificado en símbolos extraños. Ahora debo descodificarlos." (…)

El desarrollo de la informática y las propuestas de la Gramática Generativa de los años posteriores supusieron una revolución en la Lingüística. Los ordenadores ofrecían un filón de posibilidades en el tratamiento del texto, la traducción, y la extracción de información, mientras que Internet, a su vez, suponía una fuente inagotable de textos y material lingüístico. La Lingüística Computacional y el Procesamiento de Lenguaje Natural (PLN) nacieron como ramas mixtas en las que confluían el estudio lingüístico y el uso de herramientas informáticas. A partir de estas nuevas aproximaciones se han ido desarrollando diversos modelos de traducción y detección de idioma, si bien algunos problemas como la polisemia y la desambiguación no han podido ser resueltos todavía, y los resultados de los traductores automáticos no han logrado acercarse a la labor de traducción humana.

Actualmente existen múltiples detectores y traductores automáticos a libre disposición en Internet. Estos detectores siguen distintos métodos en la detección automática del idioma, como son a través de la consulta de un lexicón, la presencia de conectores, la detección de caracteres tipográficos particulares a una lengua, por frecuencia de aparición de letras y a través del modelo de de n-gramas.

AZRAEL comparte algunos rasgos con estos métodos, si bien se trata de una aproximación distinta y novedosa al problema de la detección automática del lenguaje. Esta nueva aproximación supone algunas mejoras respecto a otros enfoques, y al mismo tiempo plantea nuevos retos a resolver. En cualquier caso, ninguno de estos planteamientos es totalmente satisfactorio por sí

solo, ya que todos resuelven algunos aspectos, pero acarrean problemas en otros. Quizá el mejor enfoque para la detección de idioma sea desde una perspectiva sinérgica, en el que varios mecanismos analicen distintos aspectos del texto, aumentando la fiabilidad de la detección.

6. OBJETIVOS DE LA INVESTIGACIÓN (Máx. 1000 palabras)

La creación de AZRAEL responde a la necesidad actual de desarrollar instrumentos informáticos que nos ayuden en el tratamiento automático del lenguaje. La detección del idioma es fundamental para el uso de herramientas tan extendidas como los procesadores de texto y los correctores automáticos. Asimismo, la Red pone a nuestra disposición una inmensa cantidad de información en múltiples idiomas que necesita ser procesada lingüísticamente para poder sacarle partido.

Nuestro propósito es, por lo tanto, crear un detector de idioma eficaz que suponga una mejora respecto a los ya existentes. Esto quiere decir que la aproximación con la que nos acercaremos al tema será distinta de las que se han planteado hasta ahora, ya que nos centraremos fundamentalmente en la detección en base a la estructura silábica, lo que supone una novedad a este campo de investigación.

El aspecto básico sobre el que se pretende realizar la detección del idioma es la forma de las palabras, ya que, es este (y no la existencia de significado) el mecanismo fundamental por el que un hablante reconoce el idioma de una palabra. La aspiración, por lo tanto, no es que detecte exclusivamente lo que un diccionario consideraría español, sino que nuestro objetivo es mucho más ambicioso: queremos que nuestro detector admita como español todo aquello que un hablante admitiría, es decir, todo aquello que tenga apariencia de español, al margen de que tenga significado o no. Un detector de estas características aceptaría lo que normativamente se considera español, pero también otras producciones menos convencionales, como son los neologismos, términos propios de la jerga callejera, palabras propias de juegos y canciones infantiles, palabras inventadas, etc, , ya que, según hemos comprobado, están también regidas por las mismas leyes silábicas

No pretendemos hacer un programa que distinga español canónico del que no lo es, por lo tanto, no podemos restringir un idioma a lo que normativamente se considera correcto cuando el objetivo es la creación de una herramienta práctica que detecte desde textos literarios o académicos hasta producciones propias del lenguaje coloquial.

El segundo objetivo que perseguimos es la realización de un estudio en profundidad de la estructura de las palabras del español en base a la sílaba. La estructura de la sílaba resulta una herramienta muy útil para caracterizar una lengua, y es una fuente inagotable de información para caracterizar el idioma y para extraer datos sobre la naturaleza de la palabra. Por tanto, con el estudio de la sílaba no sólo buscamos llegar a un patrón que se pueda implementar en un programa informático, sino también extraer toda la información que sea posible a través de la forma de la palabra.

7. METODOLOGÍA EMPLEADA (Máx. 500 palabras)

AZRAEL propone un método novedoso en el campo de la detección: la sílaba como forma de reconocer un idioma. La detección del idioma a través de la sílaba conlleva múltiples ventajas.

En primer lugar, no dependemos de un inventario del léxico de una lengua, que resultaría muy costoso de manejar desde el punto de vista informático y que siempre resultaría incompleto. Por otro lado, el léxico de una lengua está en constante cambio, aparecen palabras nuevas, otras caen en desuso, pero la estructura de la sílaba se mantiene estable durante siglos. No nos es útil un diccionario del siglo XIX para leer un texto actual, pero las normas silábicas que regían el español en el siglo XIX siguen siendo válidas en el español de hoy.

En segundo lugar, la detección del español a través de la sílaba nos permite acercarnos a textos menos convencionales, ya que las reglas silábicas que rigen el español normativo y académico rigen también la estructura de los lenguajes inventados, la jerga callejera o las palabras propias de los juegos de niños.

Por último, el estudio de la sílaba nos permitirá sacar gran información de la naturaleza de las palabras, de su etimología y de los procesos morfológicos que ocurren en los préstamos léxicos desde una lengua a otra.

Puesto que AZRAEL es un detector de texto escrito, en la realización del programa se ha tomado el carácter tipográfico como elemento mínimo de estudio. Los caracteres se combinan siguiendo un patrón determinado, conformando una sílaba, que es la unidad en base a la cual se analiza el texto. Es decir, el carácter es la unidad estructural, mientras que la sílaba es la unidad funcional. Metafóricamente, podría verse como los átomos de un ser vivo: el conjunto de tipos de átomos que conforman un ser vivo es cerrado. Existe una inmensa cantidad de posibles moléculas que esos átomos podrían formar combinándose unos con otros, pero lo cierto es que sólo encontramos unas moléculas determinadas en el interior de los seres vivos; sólo son válidas algunas de las innumerables combinaciones que serían posibles. De igual modo, los caracteres que existen en una lengua podrían combinarse de múltiples formas distintas, pero la realidad es que sólo encontramos un subconjunto de todas las combinaciones que teóricamente serían posibles.

Por otro lado, hemos definido palabra como la secuencia de caracteres entre dos espacios en blanco. Esto significa que para nosotros una secuencia como *dáselo* es una sola palabra y *se lo da* son tres. Este enfoque deja al margen otras aproximaciones semánticas o morfológicas sobre la problemática del tema, pero resulta muy eficaz desde el punto de vista del procesamiento automático del texto, ya que es sencillo, útil y sobre todo completo.

El método seguido para la creación de AZRAEL ha consistido en, una vez realizado el estudio teórico de la estructura silábica, pasar el conocimiento de la estructura silábica del español a lenguaje de programación Prolog para lograr un programa que discrimine el español en base a la estructura de la sílaba. Una vez escrito el código del programa, se han realizado diversas pruebas con distintos tipos de textos para comprobar el buen funcionamiento de AZRAEL.

8. RESULTADOS OBTENIDOS (Máximo 3000 palabras)

AZRAEL supone múltiples aportaciones en el campo de la detección automática del idioma.

1. Aportaciones en el enfoque

AZRAEL es una aproximación nueva al problema de la detección automática del idioma. El modelo teórico en el que está basado el programa es la estructura silábica del español, lo que supone un método distinto al de otros detectores.

Además, aunque la versión de AZRAEL que presentamos está exclusivamente orientada a la detección de textos en español, este modelo de reconocimiento del idioma es aplicable a otras lenguas, lo que ofrece una inmensa cantidad de posibilidades para futuros detectores. De hecho, puesto que la estructura silábica permite distinguir unas lenguas de otras, se abre un prometedor campo de investigación en Tipología Lingüística, ya que sería posible llegar mediante comparación de lenguas a una clasificación tipológica de las lenguas en base a su estructura silábica. En este sentido, AZRAEL aúna dos ramas de la Lingüística: por un lado, como herramienta para el tratamiento automático del texto, la Lingüística Computacional y el Procesamiento de Lenguaje Natural; por otro lado, como estudio que busca elementos que caractericen a una lengua, la Tipología Lingüística. Ambas disciplinas tienen mucho que aportarse y esperamos que este trabajo impulse líneas de investigación comunes.

2. Estudio teórico de la sílaba en español.

2.1 Estructura silábica en español

Como no ha sido posible encontrar tratados exhaustivos de la sílaba es español, hemos realizado nuestro propio estudio del tema. Este estudio silábico está diseñado en base al texto escrito (que es el formato que nos interesa para la detección). De forma somera, podemos decir que el estudio de la sílaba nos ha llevado a la siguiente clasificación de las letras del español:

Consonantes intercaladas: R, L

Consonantes tipo 1: B, C, F, G, P.

Consonantes tipo 2: T, D

Vocales fuertes: A, E, O, Í, Ú, Á, É, Ó

Consonantes finales: R, S, L, N, D, Z.

Consonantes pseudofinales: ver tabla y contextos

CONSONANTE PSEUDOFINAL	CONTEXTO CONSONÁNTICO
C	t
	c
G	n
M	p
	b
	n
P	t
	c

Al margen de esta clasificación nos encontramos con letras que tienen comportamientos particulares:

- La S puede ir cerrando una sílaba detrás de una consonante final o a un prefijo, ya que es la única consonante que tiene la capacidad de unirse a una consonante final para acabar la sílaba.
- La Y la hemos considerado consonante y vocal, porque puede desempeñar un papel u otro según el contexto silábico en el que se encuentre.
- La Q sólo puede funcionar si va seguida de una U, lo que quiere decir que por sí sola no es una letra autónoma e independiente.
- La LL, la RR y la CH son combinaciones de dos letras que representan un único sonido, pero que a diferencia de la Q, pueden funcionar también solas. Son las *consonantes emparejadas*.

Estas son las reglas básicas de la estructura silábica del español. Sin embargo, existen dos casos en los que estas reglas no son aplicables: los prefijos y los cultismos.

2.2 El funcionamiento de los prefijos

Gracias al estudio silábico del español hemos descubierto que en el caso de los prefijos, las reglas de la derivación morfológica se superponen a las reglas de la estructuración silábica. Como los prefijos son partículas y aportan significado por sí mismos, los hablantes admiten que puedan ir seguidos por cualquier consonante, y por lo tanto, en estos casos dejan de ser aplicables las reglas de la estructuración silábica. Es decir, aunque los prefijos puedan ser sílabas, las reglas silábicas no se aplican en este caso, ya que los prefijos sin ser independientes, tienen significado autónomo. Este fenómeno se da con los prefijos *sub-* y *ex-*, pero también se aplican a las palabras con las partículas *ab-* y *ob-* al comienzo, ya que ambas eras prefijos en latín, y por lo tanto, en el momento de su creación estabas regidas por las mismas reglas de prefijación que actualmente rigen *ex-* y *sub-,* ya que de igual manera que los hispanohablantes no tienen problema en admitir cualquier consonante detrás de los prefijos *ex-* y *sub-*, los latinos podían construir palabras en las que *ab-* y *ob-* fueran seguidos de cualquier combinación de letras.

2.3 Cultismos y extranjerismos. Información etimológica.

Hasta aquí hemos descrito la estructura silábica del español. Sin embargo, es posible encontrar palabras españolas que se salgan de este patrón. Las combinaciones consonánticas más frecuentes que son ajenas a la estructura silábica del español son TL, TM, TN, PN, FT, GM, GD, CN y IT. Si analizamos de dónde provienen las palabras que contienen estas combinaciones, nos daremos cuenta de que todas son cultismos o extranjerismos, siendo la mayoría de los casos palabras provenientes del griego, aunque también hay términos heredados del náhuatl y del árabe.

En ningún caso es un inconveniente el hecho de que estas palabras se salgan de la estructura silábica del español, ya que este hecho nos proporciona valiosa información sobre el origen de estas palabras. De hecho, este fenómeno demuestra la solidez de las reglas de la sílaba: si una palabra no tiene la estructura silábica esperada es porque necesariamente viene de otra lengua.

No es solamente el idioma de procedencia el hecho que determina el aspecto de una palabra importada. Cuanto más reciente haya sido la incorporación, más se alejará de la estructura silábica del español; según vaya transcurriendo el tiempo, la forma de la palabra irá perdiendo las combinaciones anómalas y se irá adecuando a la estructura del español. De este modo, las palabras de origen griego que entraron al español a través del latín están ya adaptadas a la estructura silábica del español, en primer lugar porque el latín ya realizó una primera adaptación de la forma griega a la estructura silábica del latín, y en segundo lugar, porque estas palabras llevan siglos incorporadas al español. En contraposición, las palabras del griego que entraron como cultismos científicos a partir del siglo XVIII tienen una estructura anómala, porque han pasado directamente al español sin el latín como intermediario, y porque su entrada es reciente.

Otro factor que influye en la adaptación de una palabra extranjera es la vía por la que llegue al español. Si la palabra está restringida al ámbito culto y académico es más probable que conserve las estructuras anómalas que si la palabra pasa al habla popular. Un ejemplo de este fenómeno son los dobletes castellanos (una palabra culta y otra popular) que derivan de un mismo término latino: por ejemplo, la palabra latina *episcopatus* nos ha dado *obispado* por la vía patrimonial y *episcopado* (mucho más parecida a la forma origina) por la vía culta.

3. Creación de AZRAEL

Basándonos en el estudio teórico que acabamos de exponer, hemos creado AZRAEL, un detector automático de textos en español, y los dos programas auxiliares de los que se sirve: el Silabeador y ESDRA (Etimólogo Selectivo Del Reconocedor Automático). El Silabeador es un programa de silabeo de palabras, construido sobre el estudio teórico de la estructura silábica del español y que puede usarse de forma independiente a AZRAEL. ESDRA se encarga de buscar posibles orígenes etimológicos para las palabras que el Silabeador no ha reconocido.

AZRAEL, a su vez, es un detector tanto de textos como de palabras sueltas que decide si algo es español o no a partir de la información que le suministra el Silabeador. Una de sus ventajas es que no decide de forma global si un texto es entero español o no, sino que analiza palabra a palabra, y devuelve en una lista aparte lo que no tiene estructura de español, es decir, lo que de ningún modo podría ser español.

AZRAEL en compañía de ESDRA va un paso más allá, y, además de devolver las palabras que no siguen la estructura silábica del español, afinan más la respuesta asignando posibles orígenes etimológicos a aquellas palabras que sean extranjerismos o cultismos. Los tres programas han sido probados con diversos textos y son eficaces.

9. BIBLIOGRAFÍA CONSULTADA MÁS IMPORTANTE

ALCARAZ VARÓ, ENRIQUE; MARTÍNEZ LINARES, ANTONIA (2004): *Diccionario de Lingüística moderna*, Barcelona, Ariel.

ÁLVAREZ, ALFREDO I. (2005), *Hablar en español: la cortesia verbal. La pronunciacion del español estándar. Las formas de expresion oral,* Oviedo, Ediciones Nobel.

BENIERS, ELISABETH (ed.) (2000): *Lecturas de morfología*. México. Instituto de Investigaciones Filológicas.

CANO, RAFAEL (coord.) (2004), *Historia de la lengua española*, Barcelona, Ariel.

CHOMSKY, NOAM, MILLER, GEORGE A. (1976): *El análisis formal de los lenguajes naturales*, Madrid, Alberto Corazón, D.L.

Gerdemann, Dale, "Evaluation of Language Identification Methods", disponible en:
http://www.sfs.nphil.uni-tuebingen.de/iscl/Theses/kranig.pdf

GREFENSTETTE, GREGORY (1995), "Comparing two languages identification schemes", en *JADT 3rd Interantional conference on Statistical Analysis of Textual Data*, Roma. Disponible en:
http://www.xrce.xerox.com/Publications/Attachments/1995-012/Gref---Comparing-two-language-identification-schemes.pdf

HÁLA, BOHUSLAV (1973): *La sílaba. Su naturaleza, su origen y sus transformaciones*, Madrid, Consejo Superior de Investigaciones Científicas.

HARRIS, JAMES W. (1983): *Syllable Structure and Stress in Spanish: A Nonlinear Analysis*, Boston, The Massachusetts Institute of Technology.

HUTCHINS, JOHN W. (2000), *Early years in machine translation. Memoirs and biographies of pioneers*, Philadelphia, John Benjamins Publishing Company.

LÓPEZ GARCÍA, ÁNGEL (2000), *Cómo surgió el español*, Madrid, Gredos.

MENÉNDEZ PIDAL, RAMÓN (2005), Historia de la Lengua española, Madrid, Fundación Menéndez Pidal.

Pena, Jesús (2008) "El cambio morfológico en el interior de las series de derivación" en *Revista de Investigación Lingüística*, Vol 11, No 1, Universidad de Murcia. Disponible en:
http://revistas.um.es/ril/article/view/53771/51791

REAL ACADEMIA ESPAÑOLA (1973), Esbozo de una nueva gramática de la lengua española, Madrid, Espasa.

VÄÄNÄNEN, VEIKKO (1988), *Introducción al latín vulgar*, Madrid, Gredos.

VARELA ORTEGA, SOLEDAD (2005), *Morfología léxica: la formación de palabras*, Madrid, Gredos.