Escritos, Revista del Centro de Ciencias del Lenguaje
Número 28, julio-diciembre de 2003, pp. 9-25

Análisis morfológico automático del español a través de generación

Alexander Gelbukh, Grigori Sidorov y Francisco Velásquez

La mayoría de los sistemas de análisis morfológico están basados en el modelo conocido como la morfología de dos niveles. Sin embargo, este modelo no es muy adecuado para lenguajes con alternaciones irregulares de raíz (por ejemplo, el español o el ruso). En este trabajo describimos un sistema computacional de análisis morfológico para el lenguaje español basado en otro modelo, cuya idea principal es el análisis a través de generación. El modelo consiste en un conjunto de reglas para obtener todas las raíces de una forma de palabra para cada lexema, su almacenamiento en el diccionario, la producción de todas las hipótesis posibles durante el análisis y su comprobación a través de la generación morfológica.

The majority of the systems of morphological analysis are based on the model known as two-level morphology. However, this model is not quite adequate for languages with irregular root alternations (for example, Spanish and Russian). In this work we describe a computational system of morphological analysis for the Spanish language based on another model whose principal idea is analysis through generation.The model consists of a set of rules for obtaining all of the roots of a form of the word for each lexeme, their storage in a dictionary, the production of all of the possible hypotheses during the analysis and its verification through morphological generation.

1. INTRODUCCIÓN

La morfología estudia la estructura de las palabras y su relación con las categorías gramáticas del lenguaje. El objetivo del análisis morfológico automático es llevar a cabo una clasificación morfológica de una forma de palabra. Por ejemplo, el análisis de la forma *gatos* resulta en:

gato+Noun+Masc+Pl,

que nos indica que se trata de un sustantivo plural con género masculino y que su forma normalizada (lema) es *gato*.

Los lenguajes, conforme a sus características morfológicas –básicamente, según la tendencia en la manera de la combinación de morfemas– se clasifican en *aglutinativos* y *flexivos*.

Se dice que un lenguaje es *aglutinativo* si:

· Cada morfema expresa un solo valor[1] de una categoría gramatical.

· No existen alternaciones de raíces o las alternaciones cumplen con las reglas morfonológicas que no dependen de la raíz específica, como, por ejemplo, armonía de vocales, etc.

· Los morfemas se concatenan sin alteraciones.

· La raíz existe como palabra sin concatenarse con morfemas adicionales algunos.

Algunos ejemplos de los lenguajes aglutinativos son las lenguas turcas (el turco, kazakh, kirguiz, etc.) o el húngaro.

Por otro lado, un lenguaje es *flexivo* si:

· Cada morfema puede expresar varios valores de las categorías gramaticales. Por ejemplo, el morfema *-mos* en español expresa cumulativamente los valores de las categorías *persona* (tercera) y *número* (plural).

· Alternaciones de raíces no son previsibles –sin saber las propiedades de la raíz específica no se puede decir qué tipo de alternación se presentará.

· Los morfemas pueden concatenarse con ciertos procesos morfonológicos no estándares en la juntura de morfemas.

· La raíz no existe como palabra sin morfemas adicionales. Por ejemplo, *escrib-* no existe como palabra sin *-ir*, *-iste*, *-ía*, etc.

Algunos ejemplos de los lenguajes flexivos son lenguas eslavas (ruso, checo, ucraniano, etc.) o románicas (latín, portugués, español, etc.).

Normalmente, esta clasificación de lenguajes refleja las tendencias, es decir, muy raras veces un lenguaje es absolutamente aglutinativo o flexivo. Por ejemplo, el finlandés es un lenguaje básicamente aglutinativo, aunque con algunos rasgos de un lenguaje flexivo –por ejemplo, varios valores de las categorías gramaticales pueden unirse en el mismo morfema.

---

1 Por ejemplo, *nominativo* es un valor de la categoría gramatical caso.

En este artículo sólo consideraremos los lenguajes flexivos, específicamente el español.

En teoría, ya que la morfología de cualquier lenguaje flexivo es finita, cualquier método de análisis basado en diccionario da resultados igualmente correctos. Sin embargo, no todos los métodos de análisis automático son igualmente convenientes en el uso y fáciles de implementar.

En este trabajo presentamos una implementación para el español de un modelo de análisis morfológico automático, basado en la metodología de análisis a través de generación. Esta metodología permitió realizar el desarrollo del sistema con un esfuerzo mínimo y aplicar el modelo gramatical del español que se presenta en las gramáticas tradicionales –el más simple e intuitivo.

En el resto del artículo se describen los modelos existentes del análisis morfológico automático y, especialmente, el modelo de análisis a través de generación, después se presenta el proceso de generación y análisis que se usó en el sistema desarrollado para el español, se explica el procedimiento de preparación de los datos, se presenta brevemente la implementación del sistema y finalmente se dan las conclusiones.

2. MODELOS DE ANÁLISIS MORFOLÓGICO AUTOMÁTICO

En la implementación de los analizadores automáticos morfológicos es importante distinguir:

· Modelo de análisis (el procedimiento de análisis).

· Modelo de gramática que se usa en el analizador (las clases gramaticales de palabras).

· Implementación computacional (el formalismo usado).

La razón de la diversidad de los modelos de análisis es que los lenguajes diferentes tienen la estructura morfológica diferente. Entonces, los métodos apropiados para lenguajes, digamos, con morfología pobre (como el inglés), o para los lenguajes aglutinativos, no son los mejores para los lenguajes flexivos, como el español o el ruso.

La complejidad del sistema morfológico de un lenguaje, para la tarea de análisis automático, no depende tanto del número de las

clases gramaticales ni de la homonimia de las flexiones, sino del número y tipo de las alternaciones en raíces, las cuales no se pueden saber sin consultar el diccionario, por ejemplo: *m**o**ver-m**ue**vo* vs *d**o**rmir-d**u**rmió vs c**o**rrer-c**o**rro*. En este caso, el tipo de alternación es una característica de la raíz, y el único modo de obtener esta información es consultar el diccionario, el cual contiene estos datos de antemano. Al contrario, para el caso de algún idioma aglutinativo, como el finlandés, existen alternaciones de raíces, pero normalmente son predecibles sin el uso del diccionario.

Un ejemplo de clasificación de los métodos de análisis morfológico es la propuesta por R. Hausser (1999a, 1999b), donde los métodos se basan en formas, morfemas y alomorfos. Para distinguir entre los sistemas basados en alomorfos y en morfemas también se usa el concepto de "procesamiento de raíces estático vs dinámico". De hecho, en el método de alomorfos se guardan todos los alomorfos de cada raíz en el diccionario, lo que es el procesamiento estático, mientras que en el método basado en morfemas es necesario generar los alomorfos de un morfema dinámicamente durante el procesamiento.

Consideremos esos métodos con más detalle. Como un extremo, se pueden almacenar todas las formas gramaticales en un diccionario, junto con su lema y toda la información gramatical asociada con la forma. Este método está basado en formas de palabras. Con esta aproximación, un sistema morfológico es sólo una gran base de datos con una estructura simple. Este método se puede aplicar para los lenguajes flexivos aunque no para los aglutinativos, donde se puede concatenar los morfemas casi infinitamente. Las computadoras modernas tienen la posibilidad de almacenar bases de datos con toda la información gramatical para grandes diccionarios de lenguajes flexivos –un aproximado de 20 a 50 megabytes para el español o el ruso. Para el español, un ejemplo de implementación del dicho método es el proyecto **MACO+** (Atseria *et al.*, 1998).

Sin embargo, tales modelos tienen sus desventajas. Por ejemplo, no permiten el procesamiento de palabras desconocidas. Otra desventaja es la dificultad de agregar las palabras nuevas al diccio-

nario –hay que agregar cada forma manualmente. Sin embargo, hacer este tipo de trabajo manualmente es muy costoso. Por ejemplo, los verbos españoles tienen, por lo menos, 60 formas diferentes (sin contar formas con enclíticos). Para evitar este tipo de trabajo manual se tienen que desarrollar los algoritmos de generación, lo que puede ser una parte significativa de desarrollo de los algoritmos de análisis, como se presenta en este artículo.

Otra consideración a favor del desarrollo de los algoritmos, en lugar de usar una base de datos de las formas gramaticales, es el punto de vista según el cual los algoritmos de análisis son un método de compresión del diccionario. Esos métodos permiten, por lo menos, 10 veces más la compresión. En nuestros experimentos hicimos la compresión de los diccionarios del ruso y del español en forma de una base de datos con una utilidad de compresión estándar (zip). El fichero del resultado para el ruso fue como 30 veces más grande que el diccionario de los sistemas de análisis. En el caso del español, la diferencia entre el tamaño de los ficheros fue como 10 veces a favor del diccionario para el algoritmo.

Una razón más para el uso de los algoritmos de análisis es que son necesarios para cualquier tarea que involucra el conocimiento morfológico. Por ejemplo, para dividir palabras con guiones. También, para la traducción automática o recuperación de datos, en ocasiones, es mejor tener la información de morfemas que forman la palabra y no únicamente a nivel de la palabra completa.

Otros tipos de sistemas se basan en almacenamiento de morfemas (de algún alomorfo que se considera el básico) que representan las raíces en el diccionario. Es decir, el diccionario tiene un solo alomorfo que representa cada morfema. Los demás alomorfos se construyen en el proceso de análisis. El modelo más conocido de este tipo es **PC-KIMMO**.

Muchos procesadores morfológicos están basados en el modelo de dos niveles, de Kimmo Koskenniemi (1983). Originalmente, el modelo fue desarrollado para el lenguaje finlandés, después se le hicieron algunas modificaciones para diferentes lenguajes (inglés, árabe, etc.). Poco después de la publicación de la disertación de Koskenniemi, donde se propuso el modelo, L. Karttunen y otras

personas desarrollaron una implementación en **LISP** del modelo de dos niveles y lo llamaron **PC-KIMMO**.

La idea básica del modelo **KIMMO** es establecer la correspondencia entre el nivel profundo, donde se encuentran solamente los morfemas, y el nivel superficial, donde hay alomorfos. Eso explica por qué, en este modelo, es indispensable construir los alomorfos dinámicamente.

Además, otra idea detrás del modelo **PC-KIMMO** es enfocarse en el formalismo –los autómatas finitos (transductores)– y de tal modo no pensar en implementación de los algoritmos (Beesley and Karttunen, 2003). Es decir, los transductores son una implementación del algoritmo. La complejidad del modelo gramatical no se toma en cuenta. No obstante, es necesario desarrollar las reglas para poder construir una raíz básica –el alomorfo que se encuentra en el diccionario– de cualquier otro alomorfo. Si en el idioma no existe una compleja estructura de alternaciones de raíces, entonces sí se puede usar el modelo **KIMMO**, como se muestra, por ejemplo, en (Karttunen, 2003). Para los idiomas donde hay muchas alternaciones de raíces no predecibles –como el ruso (Bider and Bolshakov, 1976)– es posible aplicar este modelo, pero el desarrollo de los algoritmos resulta mucho más difícil, aunque no imposible, porque todo el sistema es finito. Cabe mencionar que la complejidad de los algoritmos de este tipo, según algunas estimaciones, es **NP**-completa (Hausser, 1999b, 255).

Sin embargo, las ideas relacionadas con la implementación no deben considerarse como predominantes. Claro, si las demás condiciones son iguales, es preferible tener una implementación ya hecha (como un transductor), pero consideramos que la complejidad de los algoritmos es un costo demasiado elevado. Nuestra experiencia muestra que cualquier otro modo de implementación –interpretador de las tablas gramaticales o programación directa de las reglas– es igualmente efectivo tanto en el desarrollo como en el funcionamiento.

Hay otros modelos de análisis morfológico basados en morfemas. El español, Moreno y Goñi (1995) propone un modelo para el tratamiento completo de la flexión de verbos, sustantivos y adjetivos. Este modelo –**GRAMPAL**– está basado en la unificación de ca-

racterísticas y depende de un léxico de alomorfos tanto para las raíces como para las flexiones. Las formas de las palabras son construidas por la concatenación de alomorfos, por medio de características contextuales especiales. Se hace uso de las gramáticas de cláusulas definidas (DCG) modeladas en la mayoría de las implementaciones en Prolog. Sin embargo, según los autores, el modelo no es computacionalmente eficiente, es decir, el análisis es lento.

La desventaja común de los modelos basados en el procesamiento dinámico de las raíces es la necesidad de desarrollar los algoritmos de construcción de la raíz básica (la que se encuentra en el diccionario) y aplicarles muchas veces durante el procesamiento, lo que afecta la velocidad del sistema. Por ejemplo, en el español para cualquier *i* en la raíz de la palabra se tiene que intentar cambiarla por *e*, por la posible alternación de raíz tipo *pedir* vs *pido*. Esa regla puede aplicarse muchas veces durante el análisis para posibles variantes de la raíz, como *pido*, *pid-*, *pi-*, etc.

Los algoritmos de construcción de la raíz básica, a partir de las otras, no son muy intuitivos; por ejemplo, normalmente, no se encuentran en las gramáticas tradicionales de los idiomas correspondientes. Al contrario, en las clasificaciones de alternaciones, que están en las gramáticas, se usa la raíz básica y de ésta se construyen las demás. Asimismo, los algoritmos de construcción de raíz básica son bastante complejos; por ejemplo, para el ruso el número de las reglas se estima en alrededor de 1000 (Malkovsky, 1985).

Para evitar la construcción y aplicación de este tipo de algoritmos, se usa el enfoque estático (basado en alomorfos) del desarrollo de sistemas, siempre y cuando todos los alomorfos de raíz estén en el diccionario con la información del tipo de ésta. Este modelo de sistemas es más simple para el desarrollo (véase la sección 3).

Para construir el diccionario de tal sistema hay que aplicar el algoritmo de la generación de raíces a partir de la básica. Este procedimiento se hace una sola vez. Los algoritmos de la generación de raíces, a partir de la primera, son más claros intuitivamente, por lo regular se encuentran en las gramáticas de idiomas y se basan en el conocimiento exacto del tipo de raíz.

El único coste del almacenamiento de todos los alomorfos, con la información correspondiente en el diccionario es el aumento del tamaño del mismo, el cual no es muy significativo. En el peor caso –si todas las raíces tuvieran alternaciones– el aumento correspondiente sería al doble (o triple si todas las raíces tienen 3 alomorfos, etc.). Sin embargo, las raíces con alternaciones, usualmente son, como máximo, el 20 por ciento de todas las palabras y, además, la mayoría de las raíces sólo tiene 2 alomorfos, entonces el aumento del tamaño de diccionario es relativamente pequeño. Incluso, se puede aplicar algún método de compresión del diccionario reduciendo así los datos repetidos (Gel'bukh, 1992).

Otra consideración importante, para los modelos de análisis basados en diccionarios de morfemas o alomorfos, es el tipo de modelos gramaticales que se usan.

La solución directa es crear una clase gramatical para cada tipo de alternación de raíz posible, aunado al conjunto de flexiones que la caracterizan. De tal modo, cada tipo de palabra tiene su propia clase gramatical. Sin embargo, el problema es que tales clases, orientadas al análisis, no tienen correspondencia alguna en la intuición de los hablantes y, además, su número es muy elevado. Por ejemplo, para el ruso son alrededor de 1000 clases (Gel'bukh, 1992), y para el checo son alrededor de 1500 (Sedlacek y Smrz, 2001).

Otra posible solución es el uso de los modelos de las gramáticas tradicionales, las cuales son orientadas a la generación y clasifican las palabras según sus posibilidades de aceptar un conjunto determinado de flexiones (su paradigma). La clasificación, según las posibles alternaciones de las raíces, se da aparte porque son independientes. De tal modo, las clases corresponden muy bien a la intuición de los hablantes y su número es el mínimo posible. Así pues, en la clasificación según los paradigmas para el ruso, con su morfología bastante compleja, hay alrededor de 40 clases; para el español se aplica el modelo estándar de las tres clases para los verbos (con las finales *-ar*, *-er*, *-ir*) y una para sustantivos y adjetivos. Las diferencias en las flexiones dependen completamente de la forma fonética de la raíz, las peculiaridades adicionales como, por ejemplo, *pluralia tantum*, se dan aparte. Esos modelos son

intuitivamente claros y normalmente ya están disponibles –se encuentran en las gramáticas y diccionarios existentes.

Sin embargo, no es muy cómodo usar la clasificación orientada a generación para el análisis directo. Para eso proponemos usar una metodología conocida como "análisis a través de generación". Nuestra idea es tratar de sustituir el procedimiento de análisis con el procedimiento de generación. Es bien conocido que análisis es mucho más complejo que generación; pues podemos comparar los logros de generación de voz con los de reconocimiento de voz.

3. MODELO DE ANÁLISIS A TRAVÉS DE GENERACIÓN

Como hemos mencionado, un aspecto crucial en el desarrollo de un sistema de análisis morfológico automático es el tratamiento de las raíces alternas regulares (*deduc-ir – deduzc-o*). El procesamiento explícito de tales variantes en el algoritmo es posible pero requiere del desarrollo de muchos modelos y algoritmos adicionales, que no son intuitivamente claros ni fáciles de desarrollar. Para la solución de esta problemática, el sistema que se describe a continuación implementa el modelo desarrollado en (Gel'bukh, 1992, Sidorov, 1996] y generalizado en [Gelbukh y Sidorov, 2002]. Este modelo consiste en la preparación de las hipótesis durante el análisis y su verificación usando un conjunto de reglas de generación. Las ventajas del modelo de análisis, a través de la generación, son la simplicidad –no hay que desarrollar algoritmos de construcción de raíces, ni los modelos gramaticales especiales– y la facilidad de implementación. El sistema desarrollado para el español se llama **AGME** (Analizador y Generador de la Morfología del Español).

Como hemos mencionado, hay dos asuntos principales en el desarrollo del modelo para el análisis morfológico automático:

1) Cómo tratar los alomorfos –estática o dinámicamente.

2) Qué tipo de modelos gramaticales se usa.

La idea básica del modelo propuesto de análisis, por medio de generación, es guardar los alomorfos de cada morfema en el diccionario –procesamiento estático, que permite evitar el desarrollo de complejos algoritmos de transformaciones de raíces inevitable en el procesamiento dinámico– y usar los modelos de las gramáticas tradicionales intuitivamente claros.

3.1. PROCESO DE GENERACIÓN

El proceso de la generación se desarrolla de la siguiente manera: tiene como entrada los valores gramaticales de la forma deseada y la cadena que identifique la palabra (cualquiera de las posibles raíces o el lema).

· Se extrae la información necesaria del diccionario.

· Se escoge el número de la raíz necesaria según las plantillas (véase la sección 4.2).

· Se busca la raíz necesaria en el diccionario.

· Se elige la flexión correcta según el algoritmo desarrollado. El algoritmo es bastante simple y obvio: por ejemplo, para el verbo de la clase 1 en primera persona, plural, indicativo, presente, la flexión es *-amos*, etc.

· La flexión se concatena con la raíz.

3.2. PROCESO DE ANÁLISIS

El proceso general de análisis morfológico, usado en nuestra aplicación, es bastante simple: dependiendo de la forma de palabra de entrada se formula alguna(s) hipótesis de acuerdo con la información del diccionario y la flexión posible, después se generan las formas correspondientes para tal(es) hipótesis, si el resultado de generación es igual a la forma de entrada, entonces la hipótesis es correcta. Por ejemplo, para la flexión *-amos* y la información del diccionario para la raíz que corresponde al verbo de la clase 1, se genera la hipótesis de primera persona, plural, indicativo presente (entre otras), etc.

Las formas generadas, según las hipótesis, se comparan con la original, en caso de coincidencia las hipótesis son correctas.

Más detalladamente, dada una cadena de letras (forma de palabra), para su análisis se ejecutan los siguientes pasos:

1) Quitar una por una sus últimas letras, así formulando la hipótesis sobre el posible punto de división entre la raíz y la flexión (también siempre se verifica la hipótesis de la flexión vacía «» = ∅).

2) Verificar si existe la flexión elegida. Si no existe la flexión, regresar al paso 1.

3) Si existe la flexión, entonces hallar del diccionario la información sobre la raíz y llenar la estructura de datos correspondiente; si no existe la raíz, regresar al paso 1. En este momento no verificamos la compatibilidad de la raíz y la flexión –esto se hace en generación.

4) Formular la hipótesis.

5) Generar la forma gramatical correspondiente de acuerdo a la hipótesis y la información del diccionario.

6) Si el resultado obtenido coincide con la forma de entrada, entonces la hipótesis se acepta. Si no, el proceso se repite desde el paso 3 con otra raíz homónima (si la hay) o desde el paso 1 con otra hipótesis sobre la flexión.

Nótese que es importante la generación porque, de otro modo, algunas formas incorrectas serían aceptadas por el sistema, por ejemplo, **acuerdamos* (en lugar de *acordamos*). En este caso existe la flexión *-amos* y la raíz *acuerd-*, pero son incompatibles, lo que se verifica a través de la generación.

En caso del español es necesario procesar los enclíticos. Se ejecuta un paso adicional antes de empezar el proceso de análisis – los enclíticos se especifican en el programa como una lista (*-me*, *-se*, *-selo*, *-melo*, etc.). Siempre se verifica la hipótesis de haber un enclítico al final de la cadena.

4. MODELOS USADOS

En el español, los procesos flexivos ocurren principalmente en los nombres (sustantivos y adjetivos) y verbos. Las demás categorías gramaticales (adverbios, conjunciones, preposiciones, etc.), presentan poca o nula alteración flexiva. El tratamiento de estas últimas se realiza mediante la consulta directa al diccionario.

4.1. MORFOLOGÍA NOMINAL

La variedad de designaciones a que aluden los dos géneros y la arbitrariedad, en muchos casos, de la asignación del género (masculino o femenino) a los sustantivos impiden determinar con exactitud lo que significa realmente el género. Es preferible considerarlo como un rasgo que clasifica los sustantivos en dos categorías

diferentes, sin que los términos *masculino* o *femenino* prejuzguen ningún tipo de sentido concreto (Llorac, 2000).

No existen reglas estándar para la flexión de género en sustantivos. Por lo tanto, en nuestro programa, se almacenan todas las formas de sustantivos singulares en el diccionario, como *gato* y *gata*. Los adjetivos siempre tienen ambos géneros, entonces sólo una raíz se almacena en el diccionario: por ejemplo, *bonit-* tanto para *bonito* como *bonita*. Ahora bien, el tratamiento de la flexión del número puede ser modelado mediante un conjunto de reglas, así que es suficiente tener la única clase gramatical, porque las reglas dependen de la forma fonética de la raíz.

Por ejemplo, las formas nominales que terminan en una consonante que no sea /*s*/, agregan *-es* en su pluralización (por ejemplo, *árbol – árboles*). Por otra parte, los nombres que terminan en vocal *-á*, *-í*, *-ó*, *-ú* tienden a presentar un doble plural en *-s* y *-es* (*esquí*, *esquíes*; *tabú*, *tabúes*; la información de doble plural se da con una marca en el diccionario), aunque algunos de ellos sólo admiten *-s* (*mamás*, *papás*, *dominós*, etc.). La información sobre el plural no estándar se representa a través de las marcas en el diccionario para las raíces correspondientes.

4.2. MORFOLOGÍA VERBAL

Clasificamos a los verbos en regulares (no presentan variación de raíz, como *cantar*), semiirregulares (no más de cuatro alomorfos de raíces, como *buscar*) e irregulares (más de cuatro variantes de raíz, como *ser*, *estar*).

Afortunadamente, la mayoría de los verbos en español (85%) son regulares. Para éstos, usamos los tres modelos de conjugación tradicionales (representados por los verbos *cantar*, *correr* y *partir*).

Se usan doce modelos de conjugación verbal diferentes para los verbos semiirregulares según sus alternaciones de la raíces. Nótese que esta clasificación es independiente de los modelos de paradigmas. Cada modelo tiene su tipo de alternación y su plantilla de raíces. Por ejemplo, en el modelo A1 se encuentra el verbo *buscar* (entre otros). Tiene dos raíces posibles, en este caso *busc-*, *busqu-*; la segunda raíz se usa para el presente de subjuntivo, pri-

mera persona del singular del pretérito indefinido del indicativo y, en algunos casos, del imperativo; la primera raíz se usa en todos los demás modos y personas.

Se usa una plantilla (cadena de números) para cada modelo de conjugación semiirregular. Cada posición representa una conjugación posible; por ejemplo, la primera posición representa la primera persona del singular del presente del indicativo, las últimas posiciones hacen referencia a las formas no personales. Los números usados en la plantilla son de 0 a 4, donde 0 indica que no existe la forma correspondiente; 1 indica el uso de la raíz original; 2, 3 y 4 son las otras raíces posibles. Por ejemplo, para el modelo A1 se usa la siguiente plantilla:

1111111111112111111111111111112222221111111111111111111221

Esta estructura nos facilita el proceso de generación de las formas verbales. Nótese que son 61 posibles formas, ya que no tomamos en cuenta las formas verbales compuestas, como *haber buscado* porque cada parte se procesa por separado.

Al ser mínimo el número de los verbos completamente irregulares (como *ser, estar, haber*), su tratamiento consistió en almacenar todas sus formas posibles en el diccionario. El proceso de análisis para estas palabras consiste en generar la hipótesis de un verbo irregular con la flexión vacía; esta hipótesis se verifica a través de la generación, la cual consiste en buscar la palabra en el diccionario, obtener todas sus variantes y desplegar el campo de la información.

5. PREPARACIÓN DE LOS DATOS

La preparación preliminar de datos consiste de los siguientes pasos principales:

· Describir y clasificar todas las palabras del lenguaje (español) en las clases gramaticales y las marcas adicionales, como pluralia tantum. Esta información se toma completamente de los diccionarios existentes.

· Convertir la información léxica disponible en un diccionario de las raíces. Sólo la primera tiene que ser generada en este paso.

· Aplicar los algoritmos de generación de raíces para generar todas, copiándoles la información de la primera y asignando el número de la generada.

Se diseñó una estructura de almacenamiento de datos como se muestra en la tabla 1. Para los datos mostrados, el campo *Palabra* contiene el lema, el campo *Base* contiene la raíz, el campo *Info* contiene la clase gramatical, los campos *Marca1, Marca2* contienen las marcas gramaticales adicionales. Por ejemplo, el campo *Marca1* del registro 2 (*P*) indica que se trata de una *pluralia tantum* (es decir, al tratar de generar su singular, obtendríamos un error) y para los últimos dos registros indica el modelo de conjugación semiirregular al que pertenece el verbo. El campo *Marca2,* para los últimos dos registros, indica la raíz original (1) y la segunda raíz posible (2).

Tabla 1. Estructura del diccionario de raíces

Base	Palabra	Info	Marca1	Marca2
gato	gato	N		
gafa	gafas	N	P	
acert-	acertar	VI	M1	1
aciert-	acertar	VI	M1	2

La información para la preparación de este diccionario se tomó de los diccionarios existentes explicativos y bilingües.

6. IMPLEMENTACIÓN

La base de datos (el diccionario) es una tabla de Paradox donde se almacenan las raíces y otra información de las mismas, como se mostró en la sección 4.

El sistema se desarrolló en C++. Cuenta con una interfaz que permite escoger la forma gramatical para generar o introducir la palabra a analizar. También existe una versión del sistema que lee y procesa un fichero grande de texto.

6. CONCLUSIONES

Se presentó un sistema para el análisis morfológico, que implementa el modelo de comprobación de hipótesis a través de generación.

Las ventajas de este modelo de análisis, reflejadas en su implementación, son su simplicidad y claridad, lo que resultó en muy reducido tiempo: el desarrollo de los algoritmos principales sólo tomó algunos días.

El diccionario actual tiene un tamaño considerable: 40,000 lemas, incluyendo 23,400 sustantivos, 7,600 verbos y 9,000 adjetivos.

Es importante mencionar que el sistema AGME no sobregenera ni sobreanaliza, es decir, sólo se procesan las formas correctas.

Se está trabajando sobre la forma de tratar las palabras desconocidas (una aproximación inicial es la de formular una heurística del más parecido). Como trabajo futuro se sugiere considerar los procesos de derivación (*bella* $\Rightarrow$ *belleza*) y composición (*agua* + *fiesta* $\Rightarrow$ *aguafiestas*).

El sistema está disponible como un fichero EXE o DLL de Windows, sin costo alguno para el uso académico.

AGRADECIMIENTOS

Este trabajo fue realizado con el apoyo parcial del gobierno de México (CONACYT, SNI), IPN (CGEPI-IPN, COFAA, PIFI) y RITOS-2 del Subprograma VII de CYTED. El primer autor actualmente se encuentra realizando una estancia sabática en la Universidad Chung-Ang.

BIBLIOGRAFÍA

Atserias, J., J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Márquez, M.A. Martí, L. Padró, R. Placer, H. Rodriguez, M. Taulé, J.
1998 *Turmó Morphosyntactic analisis and parcing of unsrestricted Spanish texts*. Proc of LREC-98.

Beesley, K. B. and L. Karttunen
2003 *Finite state morphology*. Palo Alto, CA.: CSLI publications.

Bider, I. G. y I. A. Bolshakov
1976 *Formalization of the morphologic component of the Meaning - Text Model. 1. Basic concepts (in Russian with a separate translation to English)*. ENG. CYBER. R., núm. 6, p. 42-57.

Gel'bukh, A.F.
1992 *Effective implementation of morphology model for an*

inflectional natural language. *J. Automatic Documentation and Mathematical Linguistics*, Allerton Press, vol. 26, N 1, 1992, pp. 22–31.

Gelbukh, A. y G. Sidorov

2002 "*Morphological Analysis of Inflective Languages through Generation*", en revista *Procesamiento de lenguaje natural*. España, vol. 29, pp 105-112.

González, B. M. y C. Ll. Vigil

1999 *Los Verbos Españoles*. 3ª edición. España: Ediciones Colegio de España, 258 pp.

Hausser, Roland

1999a *Three Principled Methods of Automatic Word Form Recognition*. Venice, Italy: Proc. of **VEXTAL**: *Venecia per il Tratamento Automatico delle Lingue*, pp. 91-100.

1999b *Foundations of Computational linguistics*. Springer, 534 p.

Karttunen, L.

2003 *Computing with realizational morphology*, en: A.Gelbukh (ed.) Proc. of **CICL**ing-2003 (4th *International Conference on Intelligent Text Processing and Computational Linguistics*), 15–22 de febrero, ciudad de México. *Lecture Notes in Computer Science N* 2588, *Springer-Verlag*, pp. 203–214.

Koskenniemi, K.

1983 *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Tesis Doctoral. Universidad de Helsinki, 160 pp.

Llorac, E.

2000 *Gramática de la Lengua Española*. España: Ed. Espasa, 406 pp.

Malkovsky, M. G.

1985 *Dialogue with an artificial intelligence system* (in Russian). Moscow, Russia: Moscow State University, 213 pp.

Moreno, A. y J. Goñi

1995 ***GRAMPAL**: A Morphological Processor for Spanish*

Implemented in ***PROLOG*****,** en: Mar Sessa y María Alpuente, editores, *Proceedings of the Joint Conference on Declarative Programming* (**GULP-PRODE**'95), pp. 321-331, Marina di Vietri (Italia).

Santana, O., J. Pérez, *et al*.

1999 ***FLANOM****: Flexionador y Lematizador Automático de Formas Nominales*. Universidad de las Palmas de Gran Canaria. Lingüística Española Actual **XXI**, 2. Ed. Arco/ Libros, S.L. España.

Sedlacek R. y P. Smrz

2001 *A new Czech morphological analyzer* ***AJKA***. Proc. of ***TSD****-2001*. **LNCS** 2166, Springer, pp. 100–107.

Sidorov, G. O.

1996 *Lemmatization in automatized system for compilation of personal style dictionaries of literature writers (in Russian)*. In: *Word by Dostoyevsky* (in Russian), Moscow, Russia, Russian Academy of Sciences, 1996, pp. 266–300.

Sproat, R.

1992 *Morphology and computation*. Cambridge, **MA**: **MIT** Press, 313 pp.

PALABRAS CLAVE DEL ARTÍCULO Y DATOS DE LOS AUTORES

análisis morfológico automático - generación computacional - español
Alexander Gelbukh, Grigori Sidorov y Francisco Velásquez
Centro de Investigación en Computación
Instituto Politécnico Nacional
Av. Juan de Dios Bátiz, esq. con Miguel Othón de Mendizábal
Zacatenco, **CP** 07738
México, **DF**
gelbukh@cic.ipn.mx, sidorov@cic.ipn.mx, fcastillov@ipn.mx