

Descripción formal de secuencias nominales hispanas

Noé Alejandro Castro Sánchez, José Angel Vera Félix, Igor A. Bolshakov

Centro de Investigación en Computación,
Instituto Politécnico Nacional
México, D. F.
(ncaastro, javera)@sagitario.cic.ipn.mx
igor@cic.ipn.mx

Resumen. A las secuencias de nombres y apellidos que forman nombres oficiales para habitantes de los países hispanohablantes, las denominamos Secuencias Nominales (SN) Hispanas. Se propone la descripción de las SN en forma libre y con una gramática generativa. Esta gramática permite desarrollar un algoritmo y un programa de análisis de las SN en textos electrónicos. Se demuestra que el programa distingue la estructura de las SN de manera comparable a como lo realizaría un ser humano.

1 Introducción

A las secuencias de nombres y apellidos que forman nombres oficiales para habitantes de los países hispanohablantes, las denominamos Secuencias Nominales (SN) Hispanas.

Es bien conocido que las SN Hispanas son unas de las más complejas entre los lenguajes europeos. Se llegan a encontrar SN con tres nombres seguidas por los apellidos paterno y materno. Es decir, es posible encontrar SN con cinco o más palabras separadas; e.g., *Juan Luis Díaz de León Santiago*.

La estructura de las SN es clara para cada hispanohablante, por lo que puede ser estrictamente formalizada. Las aplicaciones de la formalización incluyen al menos:

- Reconocer los nombres humanos en textos electrónicos arbitrarios, que es un problema bien conocido para varios lenguajes en lingüística computacional [1-4], [6-10];
- Extraer de los títulos de trabajos científicos los apellidos de autores hispanos, solucionando el problema de atribuir trabajos únicamente a autores hispanos específicos, para después recuperarlos correctamente de bases de datos e Internet, e.g., identificar artículos de un determinado autor hispano, aún sin importar las posibles variantes con la que puede aparecer su nombre, i.e., Gregorio Pastor Aldama, ó Pastor Aldama Gregorio, ó G. Pastor A., etc.; ventaja que, por ejemplo, un motor de búsqueda no proporciona.
- Automatizar el control de los nombres de ciudadanos realizado por los organismos oficiales para identificar personas de manera única y no ambigua.

El objetivo de esta investigación es proponer una gramática generativa que estudie todos y cada uno de los elementos que participan en la estructuración de nombres hispanos más comunes y que establezca las reglas necesarias y suficientes que guíen la construcción de SN. De igual forma se considera el desarrollo de un algoritmo capaz de detectar, en textos electrónicos, SN, descomponiendo su estructura en sus elementos constituyentes.

2 Descripción libre de las secuencias nominales

Para analizar y comprender la estructura de una SN hispana, es necesario identificar todo elemento que la constituye.

Consideramos como una **Secuencia Nominal Completa** aquella que se encuentra constituida por una **Secuencia de Nombres** y una **Secuencia de Apellidos**.

2.1 Secuencia de nombres

La Secuencia de Nombres consiste de uno o varios elementos. Tradicionalmente solía constituirse por una larga serie de nombres. Actualmente, por razones prácticas, se limita su número y es más usual encontrar casos de nombres únicos, e.g., *Juan, Ana, Pedro, María*, etc. Dado que es muy raro encontrar casos con más de tres elementos, para efectos de nuestra gramática consideraremos como incorrecta aquella secuencia que rebase este número.

Denominamos **Nombres Solos** a los nombres antes mencionados. Los agrupamos como **Nombres Solos Masculinos** (*Noé, Ángel, Raúl*, etc) y **Nombres Solos Femeninos** (*Josefina, Carlota, Alina*, etc.).

Estructuras más complejas se forman al relacionar dos Nombres Solos a través de una preposición articulada, formación que denominamos **Nombre Compuesto**. Estrictamente hablando, esta estructura la concebimos como la unión de un Nombre Solo y un **Complemento Deado**, el cual está conformado por la preposición (*de*) y un artículo (no obligatorio), concatenados al segundo Nombre Solo. Ejemplos de Complementos Deado son *de la Luz, del Refugio, del Carmen, de Jesús*, etc. Los Nombres Compuestos son del tipo: *María del Refugio, José de Jesús, Antonio de la Caridad*, etc., teniendo la flexibilidad de combinar cualquier Nombre solo, independientemente de su sexo, con cualquier Complemento Deado.

Finalmente, identificamos estructuras que representan concatenaciones de dos Nombres Solos de sexo opuesto. Este tipo de agrupaciones presenta restricciones formativas motivadas seguramente por una estética de tipo fonético, es decir, no es compatible cualquier combinación arbitraria; de tal forma, encontramos un número de combinaciones muy limitado: *José María, María José, José Trinidad*, etc., pero no son aceptadas tuplas de tipo *Antonio Laura, Alejandro Fabiola, Gabriela Humberto*, etc. Este tipo de relaciones las denominamos **Nombres Combinados**, y en términos estrictos constituyen dos Nombres Solos.

2.2 Secuencia de apellidos

Un apellido hispano se conforma por dos secciones de elementos, conocidos como Apellido Paterno y Apellido Materno respectivamente. Esta diferencia presentada obedece únicamente a un orden preestablecido para señalar qué apellido de una persona corresponde al padre y cuál a la madre.

Aunque en un principio el sistema puede parecer confuso para alguien no habituado a él, ofrece muchas ventajas. Esta disposición de los apellidos ha dado origen a la isonomía o isonimia, que no es más que el estudio de la frecuencia y distribución de apellidos en poblaciones humanas, mediante el cual pueden establecerse relaciones de parentesco y origen. De esta forma, los métodos isonímicos que analizan la composición de apellidos en una población, proporcionan una válida información de la estructura genética de la misma, y constituyen un instrumento habitual en estudios de dinámica evolutiva de poblaciones humanas.

Las secuencias de apellidos más sencillas y usuales son las compuestas por un solo elemento, e.g., *Martínez, Castro, Rodríguez, etc.* A estas secuencias las denominamos **Apellidos Solos**.

Un conjunto que podemos señalar como extensión del anterior, es el que contempla Apellidos Solos anteceditos por la preposición *De*, acompañada o no por algún artículo. A diferencia de formaciones recíprocas vistas anteriormente, los **Apellidos Deado**, no son considerados como complementos, i.e., no requieren concatenarse a otro elemento para determinarse como apellidos. Como ejemplos podemos mencionar: *De León, Del Valle, De la Barrera, De los Cobos, etc.*

Una práctica dada por uso social y en determinados ambientes, permite a una mujer casada utilizar la fórmula "De + <apellido del esposo>"; por ejemplo., teniendo las secuencias *Alicia Juárez Moreno* y *Rigoberto Almodóvar Hernández*, la esposa, en actos sociales, puede utilizar la formación *Alicia Juárez De Almodóvar*; y en caso de enviudar, puede aparecer en una nota social de prensa como *Alicia Juárez Viuda de Almodóvar*. Sin embargo, estas fórmulas no se usan con frecuencia, y sobre todo, no pueden aparecer nunca en un documento oficial o legal, censo, acta de Registro Civil, etc.

Una formación más es la que alberga palabras unidas entre sí por la preposición *de*. Una de estas palabras puede existir como una de las variantes de Apellidos ya mencionados, pero no ambas. A esta formación la denominamos **Apellidos Deado Inseparables**. Como ejemplo podemos mencionar *Montes de Oca*, donde encontramos *Montes* como Apellido, pero no *de Oca*.

Formaciones caprichosas encontramos en apellidos muy poco conocidos, y de los cuales no mucho podemos mencionar. Los elementos peculiares de este conjunto pueden ser *La Rubia, La Chica, La Moneda, etc.*, Catalogamos este tipo de estructuras como **Apellidos Articulados**.

En algunos de nuestros conjuntos existe la intersección de elementos. Por ejemplo, es común utilizar, como Apellidos Solos, Nombres Solos (*Jorge, Santiago, Santos, Félix, Camilo, Miguel, etc.*), y Complementos Deado (*De Miguel, De la Cruz, De los Santos, etc.*). No nos parezca raro en algún momento tratar de precisar, de la secuencia *Jorge Miguel Félix*, qué elementos conforman el nombre, cuáles el apellido y en qué orden se encuentran escritos (¿cómo debemos nombrar a esta persona?

¿como *Jorge* o como *Félix*?). En nuestro modelo, a este conjunto de casos, lo denominamos **Apellidos existentes como Nombres**.

Las Secuencias de Apellidos pueden ser alteradas ajustándose a cánones regionales y/o temporales, o por simple arbitrariedad. De esta manera distinguimos Apellidos cuya estructura escapa de los modelos antes mencionados: *Ponce de León*, *González-Hermosillo*, *Molina y Vedia*, *Barrera-y-Fuentes*, *De Zárate y Urbina*, etc. Pero si descomponemos estas estructuras en sus elementos atómicos, notamos que a éstos los hemos ya descrito anteriormente. Para conservar la formación completa del apellido introducimos una nueva agrupación, que denominamos **Apellidos Compuestos**. Estos se forman combinando dos segmentos, constituidos cada uno por un elemento de los conjuntos ya mencionados en esta sección, unidos por simple concatenación, por alguna conjunción, guiones o ambos.

Los casos en los cuales nuestra descripción no funciona son bastante raros, y haremos mención de ellos en la conclusión.

3 Gramática generativa

Antes de iniciar con la descripción de nuestra gramática, es necesario anticipar algunas consideraciones.

1. Estos símbolos son terminales:
 - a) Los nombres propios (palabras escritas en letra cursiva).
 - b) El símbolo de puntuación – (guión).
 - c) El conjuntivo *y*.
 - d) Las preposiciones articuladas *De*, *De la*, *De las*, *Del* y *De los*.
2. Las denominaciones definidas más adelante, en el apartado 3.1.
3. A menos que se diga otra cosa, el lado izquierdo de la primera producción es el símbolo inicial.

3.1 Denominaciones utilizadas

Es necesario incorporar términos más compactos, a fin de tener referencias rápidas y prepararnos en la sintaxis de nuestra gramática. En la tabla 1 introducimos las expresiones mencionadas. En las columnas impares indicamos las denominaciones a los conceptos tratados y en las columnas pares hacemos mención de su significado correspondiente:

Tabla 1. Expresiones usadas en la gramática

SecNml	Secuencia Nominal	ApPat	Apellido Paterno
		ApMat	Apellido Materno
SecNom	Secuencia de Nombres	ApSolo	Apellido Solo
SecAp	Secuencia de Apellidos	ApDeado	Apellido Deado

		ApDeadoInsprble	Apellido Deado Inseparable
NomSolo	Nombre Solo	ApArtdo	Apellido Articulado
Ncmsto	Nombre Compuesto	ApExsteNom	Apellido existente como Nombre
NCmb	Nombre Combinado	ApCmsto	Apellido Compuesto
CplmtoDeado	Complemento Deado	Sgmto(j)	Segmento j-ésimo

3.2 Reglas estructurales

En la tabla 2 mostramos las reglas de la gramática que permite definir si una secuencia de palabras corresponde a una SN. Las reglas constan en su mayoría de símbolos no terminales.

Tabla 2. Reglas de la gramática

1	SecNml	→ SecNom(Sex) SecAp SecNom(Sex) SecAp
2	SecNom(Sex)	→ NomSolo(1, Sex) NomSolo(1, Sex) NomSolo(2, Sex) NomSolo(1, Sex) NomSolo(2, Sex) NomSolo(3, Sex) NCmsto(1, Sex) NomSolo(1, Sex) NCmsto(2, Sex) NomSolo(1, Sex) NCmsto(2, Sex ⁻¹) NCmb(Sex) NomSolo(1, Sex) NCmb(Sex)
3	NCmsto(j,Sex)	→ NomSolo(j, Sex) CplmtoDeado
4	SecAp	→ Ap(pat) Ap(pat) Ap(mat)
5	Ap(Prgtor)	→ ApSolo ApDeado ApDeadoInsprble ApArtdo ApExsteNom ApCmsto
6	ApCmsto	→ Sgmto(1) Sgmto(2) Sgmto(1) y Sgmto(2) Sgmto(1)-Sgmto(2) Sgmto(1)-y-Sgmto(2)
7	Sgmto(j)	→ ApSolo ApDeado ApExsteNom De ApExsteNom De la ExsteNom Del ExsteNom De los ExsteNom

Nota 1. j es variable con valores 1 o 2.

Nota2. Los elementos NomSolo(1, Sex), NomSolo(2, Sex) y NomSolo(3, Sex) concatenados entre sí son diferentes, es decir, NomSolo(1, Sex) ≠ NomSolo(2, Sex), NomSolo(2, Sex) ≠ NomSolo(3, Sex) y NomSolo(1, Sex) ≠ NomSolo(3, Sex)

Nota 3. Sex es variable de Sexo, con valores *mas* (masculino) o *fem* (femenino). En cada regla el valor de Sex, que puede encontrarse varias veces, es el mismo. Sex⁻¹ es complemento (valor opuesto) de Sex.

Nota 4. Prgtor es variable de Progenitor, con valores *pat* (paterno) o *mat* (materno).

Nota 5 En la regla 2 $\text{NomSolo}(1, \text{Sex}) \text{NCmsto}(2, \text{Sex}^{-1})$, el primer elemento del Nombre Compuesto (vea regla 3) guarda relación con $\text{NomSolo}(1, \text{Sex})$. Al estar éstos concatenados, siendo de sexo opuesto, se exige que formen un Nombre Combinado válido (vea ejemplo en sección 3.4).

3.3 Instancias a los símbolos no terminales

Definimos como instancia a un símbolo no terminal a aquellos símbolos básicos con los que se forman las SN, conocidos también como terminales –o bien, según los lenguajes de programación, componentes léxicos.

Tabla 3. Símbolos terminales y posibles valores

$\text{NomSolo}(i, \text{mas})$	→ Hugo Gregorio Rodolfo Humberto ...
$\text{NomSolo}(i, \text{fem})$	→ Rigoberta Dolores Antonieta Andrea ...
CplmtoDeado	→ de la Luz del Carmen del Refugio de Jesús ...
$\text{NCmb}(\text{mas})$	→ José María José Trinidad ...
$\text{NCmb}(\text{fem})$	→ María José María Jesús ...
ApSolo	→ Sánchez Miranda Vélez Landa ...
ApDeado	→ De la Barca De León De la O De los Cobos ...
ApDeadoInsprble	→ Montes de Oca Cabeza de Vaca ...
ApArdo	→ La Rubia La Chica La Moneda ...
ApExsteNom	→ Jorge Santiago Félix Camilo Alonso ...

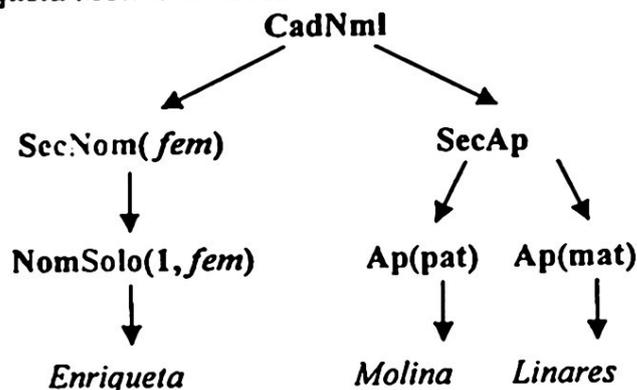
La variable i de la fórmula $\text{NomSolo}(i, \text{Sex})$ tiene valores 1, 2 ó 3.

3.4 Árboles de análisis sintáctico

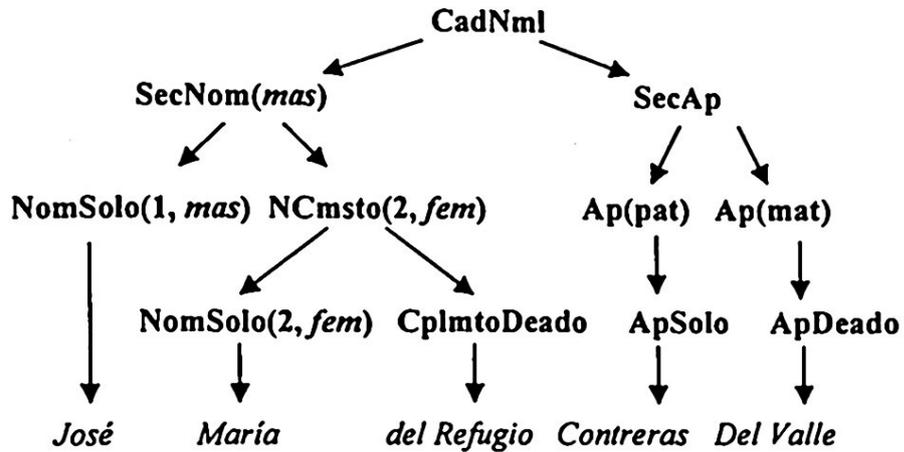
Un árbol de análisis sintáctico se puede considerar como una representación gráfica de una secuencia de sustituciones que no muestra la elección relativa al orden de sustitución.

Mostramos algunos árboles de análisis sintáctico, derivados de las siguientes SN.

- Se deriva de *Enriqueta Molina Linares*:



- El árbol a derivar de *José María del Refugio Contreras Del Valle* es:

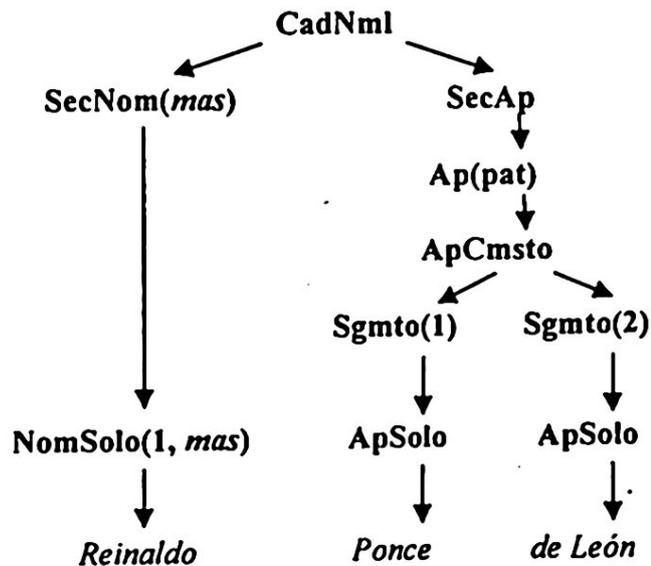


Dada una Secuencia de Nombres formada por un Nombre Solo de sexo *masculino* y un Nombre Compuesto de sexo *femenino*, los Nombres Solos (ver último nivel) deben considerarse como un Nombre Combinado, de lo contrario nuestro modelo aceptaría Secuencias de Nombres extrañas (erróneas) del tipo *Roberto Fabiola*, *Gerardo Patricia*, etc.

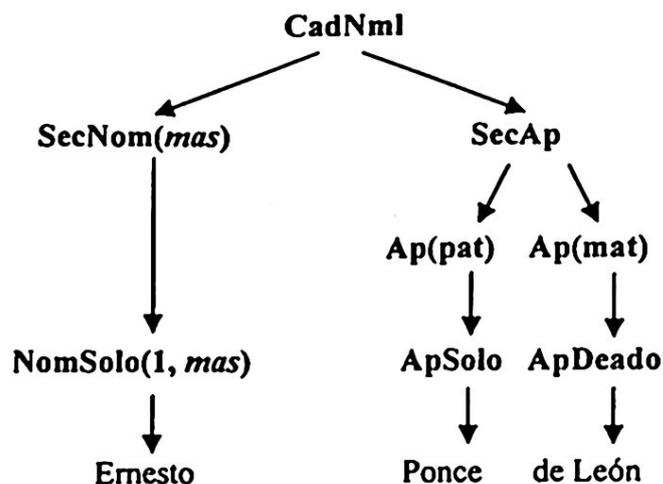
Como integrantes de los Lenguajes Naturales, los Sistemas de Nombres no pueden prescindir de ambigüedades. Específicamente, éstas recaen en la determinación de la estructura de una SN, que puede considerarse en más de una forma distinta. Mostremos la siguiente SN y los dos árboles sintácticos que pueden construirse a partir de ella.

Sea la secuencia *Reinaldo Ponce de León* y los siguientes dos casos:

Caso 1: Es posible determinar la Secuencia de Apellidos como un único Apellido.



Caso 2: Es posible determinar la Secuencia de Apellidos como dos Apellidos.



4 Analizador gramatical

Al construir la gramática que guiará el análisis, se debe construir también el analizador gramatical que determinará si se están cumpliendo o no las reglas que hemos elaborado.

El tipo de analizador que diseñamos es de tipo Ascendente, dado que tiene como objetivo construir el axioma de la gramática a partir de los elementos de léxico que recibe; es decir, es a partir de los elementos más simples que espera conformar el nivel más alto en la gramática: el axioma.

4.1 Denominaciones utilizadas en el programa

Es necesario determinar y precisar algunas expresiones que se utilizan como formato del programa en el resultado del análisis. Éstas se han tratado de ajustar lo más posible a la gramática que hemos formalizado. En la tabla 4 listamos las denominaciones utilizadas por el programa y las definidas en nuestra gramática.

Tabla 4. Correspondencia entre denominaciones de programa y gramática

Gramática	Programa
SecNom(<i>mas</i>)	(SecNom,M)
SecNom(<i>fem</i>)	(SecNom,F)
NomSolo(<i>i, mas</i>)	(nomi,M)
NomSolo(<i>i, fem</i>)	(nomi,F)
Ap(Prgtor)	(ap,Prgtor)

La forma en que el programa estructura una SN, es por medio de etiquetas que coloca al final de cada elemento encontrado. Las etiquetas quedan definidas en la tabla 5:

Tabla 5. Etiquetas de estructuración de SN

Etiqueta	Significado
(SecTot, x, y)	Etiqueta colocada al final de una SN.
(SecNom, x)	Variante 1. Etiqueta colocada al final de la Secuencia de Nombres
(SecNom, x, z)	Variante 2. Etiqueta colocada al final de la Secuencia de Nombres

El valor y significado de cada variable usada en las etiquetas se dan en la tabla 6:

Tabla 6. Variables para las etiquetas de estructuración de SN

Variable	Significado
SecTot	Secuencia Total.
SecNom	Secuencia de Nombres.
x	Sexo de la SecNom. Toma el valor M o F.
y, z	Consideraciones finales que se hacen al final de cada etiqueta, proporcionando información más específica del análisis.

La variable y, presente sólo en la etiqueta colocada al final de una SN, puede llegar a tomar los valores especificados en la tabla 7:

Tabla 7. Valores válidos para la variable y

Valor	Significado
Cmpl	[Secuencia] Completa. Valor que indica que la SN analizada cuenta con todos los componentes necesarios.
InCmpl	[Secuencia] Incompleta. En la SN, hace falta ya sea la Secuencia de Nombres o la Secuencia de Apellidos.
inddo	Indeterminado. En la secuencia analizada, los elementos presentes sí se reconocen como nombres pero no forman una estructura coherente.

En la tabla 8 se especifican los valores válidos para z:

Tabla 8. Valores válidos para la variable z

Valor	Significado
inddo	Indeterminado. Señala que en la secuencia analizada, los elementos presentes sí se reconocen como nombres pero no forman una estructura coherente.
sexoCntrio	Sexo Contrario. Indica que se ha encontrado un par de Nombres de sexo opuesto pero éstos no se encuentran definidos dentro del grupo de Nombres Combinados.
rep	Repetidos. Valor que indica que en la Secuencia de Nombres se han detectado nombres repetidos.
Larg	Larga. La Secuencia de Nombres es muy extensa (más de tres elementos).

4.2 Descripción del programa

En nuestro trabajo, para la identificación de las SN, se emplearon diccionarios (bases de datos) de nombres y de apellidos. El sistema es totalmente dependiente de estas bases, aún cuando se utiliza una pequeña heurística para determinar Nombres Combinados, Apellidos Compuestos, etc. A grandes rasgos, el proceso seguido en el análisis es el siguiente:

En primera instancia, el sistema realiza un parseo por el texto electrónico, determinando si cada palabra encontrada pertenece o no a la base de datos; cuando se ha terminado de extraer una determinada secuencia, se aplica a ésta un segundo proceso de análisis más específico (proceso de etiquetación), que consiste en asignar a cada elemento su denominación correspondiente a la gramática elaborada, según el lugar que ocupe dentro de la SN extraída.

Listamos a continuación una serie de ejemplos tomados del programa para mostrar los resultados arrojados:

Secuencia texto: *Angélica María de la Luz Hernández De Alba*

Resultado del programa:

[[[[*Angélica*(nom1,F)][*María de la Luz*(nom2,F)]](SecNom,F)]
[*Hernández*(ap,pat)][*De Alba*(ap,mat)]](SecTot,F,Cmpl)

Secuencia en texto: *Angélica María de la Lourdes Hernández De Alba*

Resultado del programa:

[[[[*Angélica*(nom1,F)][*María de la Lourdes*(nom2,F,inddo)]](SecNom,F)]
[*Hernández*(ap,pat)][*De Alba*(ap,mat)]](SecTot,F,Cmpl)

Secuencia en texto: *María Angélica Lorena Patricia De Alba*

Resultado del programa:

[[[[*María*(nom1,F)] [*Angélica*(nom2,F)] [*Lorena*(nom3,F)] [*Patricia*(nom4,F)]]]
(SecNom,F,larg)][*De Alba*(ap,pat)]] (SecTot,F,Cmpl)

Secuencia en texto: *María Angélica*

Resultado del programa:

[[[[*María*(nom1,F)][*Angélica*(nom2,F)]](SecNom,F)]](SecTot,F,Incpl)

Secuencia en texto: *De Alba*

Resultado del programa:

[[*De Alba*(ap,pat)]](SecTot,inddo,Incpl)

5 Conclusiones

Son muchos los factores que dificultan el desarrollo de un modelo que sea capaz de identificar y reconocer Secuencias Nominales Hispánicas dentro de textos electrónicos.

Hemos visto como la ambigüedad inherente a sistemas relacionados con el Lenguaje Natural dificultan el proceso de modelación. No todo lo presentado en este capítulo es concluyente, nuestro ámbito de análisis ha tenido que verse restringido por

las variaciones tan extensas que existen en la formación de SN Hispanas. Un ejemplo claro de esto se constata en Secuencias de Apellidos. Una práctica habitual hace tiempo era la concatenación de más de dos apellidos para constituir uno solo, e.g., *Sánchez de Anda y Martínez Salgado*, *Marti y Zayas-Bazán*, *Bancés y Fernández-Criado*, etc. Obviamente, el tratar de abarcar todos estos casos (y más) resultaría en un entre marañado de reglas, en una complejidad realmente enorme.

Notamos que es imposible aplicar análisis morfológico [5], porque los apellidos no se encuentran normalmente en las bases de datos de los sistemas de este tipo (por lo tanto usamos las mayúsculas para detectar presencia de los apellidos cuando uno de éstos no se haya en nuestra base de datos).

Estos son sólo algunas características que pueden dificultar alcanzar el objetivo que se planteó en este trabajo. La más clara conclusión de esto es que un sistema de clasificación de nombres no puede ser automatizado en su totalidad, siempre será necesaria la intervención del usuario.

Agradecimientos

Este trabajo fue realizado con el apoyo parcial de CONACyT y PIFI. Deseamos agradecer, asimismo, al Dr. Grigori Sidorov por sus valiosos comentarios y contribuciones.

Referencias

1. Chen, H. H., S. J. Huang, Y. W. Ding, S. C. Tsai. (1998). *Proper Name Translation in Cross-Language Information Retrieval*. Proceedings of the 17th International Conference on Computational Linguistics. pp 232 – 236. <http://acl.ldc.upenn.edu/P/P98/P98-1036.pdf>
2. Chinchor, N. (1997). *MUC-7 Named Entity Task Definition*. Proceedings of the 7th Message Understanding Conference. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html
3. Galicia-Haro, S. N., A. Gelbukh, I. A. Bolshakov. *Recognition of Named Entities in Spanish Texts*. Proceedings of International Conference MICAI'2004, Mexico City, Mexico. Lecture Notes in Artificial Intelligence, ISSN 0302-9843, Springer, 2004 (accepted).
4. Galicia-Haro, S. N., A. Gelbukh, I. A. Bolshakov. *Web-based Sources for an Annotated Corpus Building and Composite Proper Name Identification*. Proceedings of International Conference AWIC04, Cancun, Mexico. Lecture Notes in Computer Science, ISSN 0302-9843, Springer, 2004 (accepted).
5. Alexander Gelbukh and Grigori Sidorov. *Approach to construction of automatic morphological analysis systems for inflective languages with little effort*. In: Computational Linguistics and Intelligent Text Processing. Proc. CICLing-2003, 4th International Conference on Intelligent Text Processing and Computational Linguistics, February 15–22, 2003, Mexico City. LNCS, N 2588, Springer-Verlag, pp. 215–220.
6. Huang, F., A. Waibel. *An Adaptive Approach to Named Entity Extraction for Meeting Applications*. http://www.is.cs.cmu.edu/papers/speech/HLT2002/hlt2002_fei.pdf
7. Mani, I., R. McMillian, S. Luperfoy, E. Lusher, S. Laskowski. (1993). *Identifying Unknown Proper Names in Newswire Text*. Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text. pp 44 – 54. <http://acl.ldc.upenn.edu/W/W93/W93-0105.pdf>

8. Petasis, G., S. Petridis, G. Paliouras, V. Karkaletsis, S. J. Perantonis, C. D. Spyropoulos. (2000). *Symbolic and Neural Learning for Named-Entity Recognition. Proceedings of the Symposium on Computational Intelligence and Learning*. pp. 58-66.
<http://iit.demokritos.gr/~paliourg/papers/COIL2000.pdf>
9. Thompson, P., C. C. Dozier. (1997). *Name Searching and Information Retrieval. Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing*. Providence, Rhode Island. pp. 134-140.
<http://acl.eldoc.ub.rug.nl/mirror/W/W97/W97-0315.pdf>
10. Wacholder, N., Y. Ravin, M. Choi. (1997). *Disambiguation of Proper Names in Text. Proceedings of the 5th Conference on Applied Natural Language Processing*. pp. 202 – 208
<http://www.research.ibm.com/talent/documents/anlp97.pdf>