Revista Signos
2009, 42(70)
153-169

Ambigüedad en nombres hispanos

Grettel Barceló
Eduardo Cendejas
Igor Bolshakov
Grigori Sidorov
Instituto Politécnico Nacional
México

Resumen: La constitución de los nombres hispanos presupone en muchos casos un grado de ambigüedad. La estructura de las secuencias denominativas en países hispanos conlleva a la presencia de cinco problemas fundamentales que obstaculizan su interpretación: (1) la doble deducción de sexo en nombres personales, por ejemplo 'Guadalupe'; (2) la asociación de nombres y/o apellidos en un solo elemento, como en 'Jorge Luis', cuyos constituyentes existen aisladamente; (3) la composición de los elementos mediante un conectivo; (4) la dualidad nombre/apellido; y (5) la omisión permitida de alguno de los elementos en la secuencia denominativa. Nuestro estudio está orientado a detectar y analizar la ambigüedad mencionada de manera automática. Se desarrolló una gramática formal que determina las interpretaciones válidas de las cadenas nominales, por medio de un etiquetado automático de todos los elementos que la componen. Además, se presentan gráficas que muestran la distribución de los nombres y apellidos, de lo cual, el hallazgo más importante es que la frecuencia de estos cumple con la ley de Zipf. Se utilizó como fuente de conocimiento un corpus con 745.084 registros personales, de los cuales se extrajeron 93.998 nombres y 13.779 apellidos únicos, entre simples, compuestos y asociados. Partiendo de estos, se detectaron 77.162 fuentes de ambigüedad en nombres y 2.739 en apellidos, lo que representa el 82% y el 20% respectivamente. Del total de los registros personales estudiados, 241.922 presentan al menos dos interpretaciones válidas en la denominación, lo cual corresponde al 33% de la muestra.

Palabras Clave: Ambigüedad, secuencia denominativa, gramática generativa, asociación, composición.

Recibido:
25-IV-2008
Aceptado:
20-XI-2008

Correspondencia: Grettel Barceló (gbarceloa07@sagitario.cic.ipn.mx). Centro de Investigación en Computación, Instituto Politécnico Nacional. Avenida Juan de Dios Bátiz s/n, Col. Nueva Industrial Vallejo, C. P. 07738, México D. F., México.

Ambiguity in Hispanic names

Abstract: The constitution of Hispanic names assumes a degree of ambiguity in many cases. The structure of the denominative sequences in Hispanic countries presents five fundamental problems that obstruct their interpretation: (1) the double sex deduction in personal names, as in *Guadalupe*; (2) the association of names and/or surnames in one name, as in *Jorge Luis*, whose components exist separately; (3) the composition of the elements by means of a connector; (4) the name/surname duality; and (5) the accepted omission of some of the elements of the denominative sequences. This study focuses on the automatic detection and analysis of these types of ambiguities (uncertainties). A formal grammar that determines valid interpretations of the nominal chains was developed by means of the automatic labeling of all the elements of which this grammar is composed. Furthermore, graphs of the distribution of the names and surnames are presented, the most important of which reveals that the frequency abides by Zipf's law. A corpus of 745,084 personal records was used as a data source. From these records, 93,998 type names, and 13,779 type surnames, including simple, compound, and associate ones, were taken. From these, 77,162 (82%) ambiguity sources in names and 2,739 (20%) ambiguity sources in surnames were detected. From all of the personal records analyzed, 241,992 (33%) present at least two valid interpretations in the denomination.

Key Words: Ambiguity, denominative sequence, generative grammar, association, composition.

INTRODUCCIÓN

Los nombres propios han ocupado un lugar importante en textos de varios géneros y en la constante automatización de procesos. En tiempos actuales, la mayoría de las empresas, instituciones y gobiernos poseen bases de datos con información personal de proveedores, alumnos, afiliados, trabajadores, socios, pacientes, etc. Esto sitúa a los nombres de las personas, como uno de los instrumentos de identificación esenciales, por lo cual es necesario minimizar los problemas de ambigüedad que de ellos se derive y evaluar la frecuencia en que estos se presentan.

Por otra parte, la búsqueda y recuperación de información compartida a través de Internet ha sido otro factor que proporciona los medios para el intercambio y análisis de textos en ámbitos científico-tecnológicos y culturales. La ambigüedad que podemos encontrar en nombres de autores dedicados a la investigación, la literatura y las artes, dificulta la adjudicación acertada de las obras concebidas. Este es un problema que se presenta frecuentemente a pesar del constante quehacer de los gobiernos por la uniformidad en la nominación, bajo la cual trata de regirse, por ejemplo, el actuar jurídico de los institutos nacionales de los derechos de autor.

En la actualidad, las computadoras ayudan a resolver muchos de los problemas relacionados con la ambigüedad en lenguaje natural y evaluar casos donde se presente, mediante técnicas algorítmicas y estadísticas de análisis del lenguaje (Gelbukh & Sidorov, 2006).

Por lo general, la estructura de los nombres hispanos supone un nombre simple, compuesto o

asociado, más dos apellidos, el primero que coincide con el primero de los que ostenta el padre, y el segundo que es el primero de los que ostenta la madre (aunque este orden puede ser modificado por previo acuerdo o por desconocimiento de alguno de los padres). Ejemplos de nombres hispanos son: María de los Ángeles Hernández García y Carlos Manuel González Fernández.

Sin embargo, en muchos países hispanohablantes se hace caso omiso de estas reglas básicas, lo que dificulta la correcta identificación de las personas mediante documentos oficiales. A esto se suma la validez de la omisión de alguno de los nombres u apellidos por parte del autor o persona en la situación denominativa actual para los documentos redactados en español.

Esta omisión permitida provoca ambigüedad ante la presencia de nombres y apellidos compuestos y/o asociados (unión de dos nombres simples), dado que en muchas ocasiones resulta difícil distinguir el inicio y término de estos elementos en una secuencia nominal.

Otro inconveniente que es común en estos países es la utilización de nombres que pueden ser considerados también como apellidos. Uno podría preguntarse, por ejemplo, si en 'Santiago García' se hace referencia al nombre y primer apellido o a los dos apellidos de la persona.

Aspectos como la asociación de más de dos nombres, la unión por medio de guiones y cuestiones sociales no se consideraron en este análisis. En el sentido social, son prácticas aún muy comunes la adopción por parte de la mujer del apellido paterno del hombre, tras el matrimonio; y el fenómeno de abolengo, en el cual se presenta la continuidad de un apellido a través de varias generaciones con ascendencia especialmente ilustre.

El propósito de este estudio es lograr determinar automáticamente las secuencias determinativas válidas, interpretarlas en registros de texto y evaluar el grado de ambigüedad presente en los nombres hispanos, por medio de la aplicación de una gramática generativa establecida.

1. Marco de referencia

1.1. Ambigüedad en nombres propios hispanos

El nombre de una persona consta de un nombre y de uno o varios apellidos, según las costumbres de cada idioma y país. El nombre lo dan los padres a los hijos cuando nacen o en el bautizo. En cambio, el apellido o nombre familiar, comúnmente el del padre o el del padre y el de la madre, pasa de una generación a otra.

La Figura 1 representa la formación generalizada de nombres hispanoamericanos, compuesta de tres elementos fundamentales:

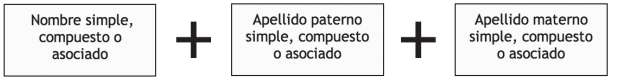


Figura 1. Estructura de la secuencia nominal hispana.

De ella, surgen cinco problemas esenciales que propician la ambigüedad (véase la discusión detallada de los mismos más adelante):

1. La doble interpretación de género en nombres.
2. La asociación de nombres y/o apellidos.
3. La composición de nombres y/o apellidos.
4. La dualidad nombre/apellido.
5. La omisión permitida de uno de los elementos.

A continuación se presentan las descripciones y algunos ejemplos de cada uno de los problemas tratados.

La **doble interpretación de género** se presenta en aquellos nombres propios que pueden ser entendidos como femeninos o masculinos, como Guadalupe o Hiram.

La **asociación** en los nombres se manifiesta con la unión de dos nombres simples, como en: Juan Carlos, Víctor Manuel, Sofía Miranda, etc. Un apellido asociado consiste en la concatenación de dos apellidos simples, como en: Alarcón Ruiz, Arteaga Jiménez, etc.

La **composición** en los nombres surge de la concatenación de nombres simples y la preposición 'de' a menudo articulada. Son ejemplos de nombres compuestos: María de los Ángeles, María del Carmen, etc. La composición de apellidos se presenta con la unión de dos apellidos simples mediante un conector (preposición 'de' o conjunción 'y'), como por ejemplo en Montes de Oca o en Montes y Gómez.

Existen muchos nombres y apellidos que presentan una **dualidad**, en cuanto a que pueden pertenecer a cualquiera de las categorías, como es el caso de: Santiago, Alfonso, etc.

Finalmente, la flexibilidad por **omisión** aumenta la ambigüedad apoyada en los casos anteriores. Por omisión se entienden aquellos casos en donde se prescinde de alguno de los tres elementos que conforman la secuencia nominal. Por ejemplo, en Francisco García se ha omitido el apellido materno.

Algunos de estos cinco problemas se presentan únicamente en las fronteras de la secuencia.

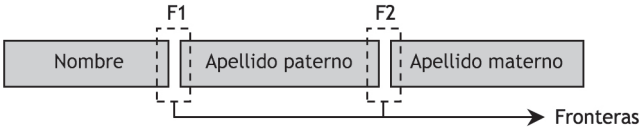


Figura 2. Fronteras en la secuencia nominal hispana.

En la frontera F1 se manifiesta la dualidad nombre/apellido, favorecida por la asociación y composición de nombres. Esto se presenta en secuencias nominales formadas por al menos dos elementos, donde el segundo de ellos pueda ser interpretado como nombre o apellido (dualidad). En este caso si no se permitieran nombres compuestos o asociados, dicho elemento siempre se entendería como primer apellido, evitando así los problemas de ambigüedad.

En la frontera F2 existe la omisión de uno de los elementos, asistida por la asociación y composición de apellidos. De esta forma se pueden presentar secuencias nominales con dos interpretaciones válidas del conjunto de apellidos: (1) apellido paterno + apellido materno o (2) apellido paterno compuesto o asociado.

1.2. Impacto del análisis de ambigüedad

La identidad que proporcionan los nombres propios es una necesidad del ser humano. El uso de los mismos suprime la colectividad y evidencia las implicaciones afectivas y sociales de las personas.

A pesar de todos los problemas de ambigüedad que presentan en la actualidad, las secuencias denominativas siguen siendo un instrumento determinante en la identificación de las personas. Las implicaciones que produce una interpretación errónea de estas secuencias, afectan no solo a las personas involucradas, sino también a la veracidad en información almacenada y toma de decisiones de instituciones gubernamentales y organizaciones no gubernamentales que manejan grandes volúmenes de registros personales.

Muchas investigaciones se han realizado hasta la fecha para la extracción de nombres propios en textos electrónicos (Galicia-Haro, Gelbukh & Bolshakov, 2004; Mani & MacMillan, 1996; Stevenson & Gaizauskas, 2000) y los resultados de esta tarea intermedia son empleados en aplicaciones que requieren comprensión del lenguaje, como: traducción automática (Chen, Huang, Ding & Tsai, 1998), recuperación de información (Paik, Liddy, Yu & McKenna, 1993a; Paik, Liddy, Yu & McKenna, 1993b; Thompson & Dozier, 1997), procesamiento de textos (Coates-Stephen, 1991; Huang & Waible, 2002), etc. Sin embargo, el énfasis en la interpretación de las secuencias y la

estructura de las mismas ha sido escaso, pues a pesar de que algunos trabajos consiguen una tipificación de los nombres mediante análisis de constitución (Chen & Lee, 1996), no persiguen el objetivo de ofrecer un sentido a cada uno de sus elementos y brindar todas sus posibles significaciones. Solo en Castro, Vera, Bolshakov y Sidorov (2004) se propone una gramática generativa para formar cadenas nominales completas para los nombres hispanos, sin embargo, no se estudian los problemas de la ambigüedad, ni se considera válida la asociación de apellidos.

En el presente trabajo se determinan las fuentes productoras de ambigüedad de los nombres, examinando la estructura de las secuencias permitidas oficialmente. Se detalla un analizador que detecta de manera automática estos problemas, cuyo uso presupone una disminución en el número de casos que humanamente deben ser examinados en documentos electrónicos. Por otra parte, se incluye un estudio de la distribución de las cadenas denominativas hispanas, del cual se deriva el descubrimiento del tipo de comportamiento que exhiben los nombres y apellidos según su presencia en el corpus empleado: sus frecuencias se relacionan con sus respectivos rangos de acuerdo con las predicciones de la ley de Zipf.

1.3. Ley de Zipf

Esta ley empírica, formulada usando estadística matemática, se refiere al hecho de que muchos tipos de datos estudiados tanto en textos, como en física y ciencias sociales pueden ser descritos por una distribución de Zipf (Gelbukh & Sidorov, 2001).

La ley de Zipf plantea que la frecuencia de cualquier palabra en un texto es inversamente proporcional a su rango; es decir, existe un pequeño número de palabras que son utilizadas con mucha frecuencia, mientras que hay un número muy grande de palabras que son poco empleadas. Esta afirmación, expresada matemáticamente tiene la forma: $P_n \sim 1 / x^{\alpha}$, donde P_n representa la frecuencia de una palabra ordenada $x^{\text{ésima}}$ y α es casi 1. Esto significa que el segundo elemento se repetirá aproximadamente con una frecuencia de 1/2 de la del primero, y el tercer elemento con una frecuencia de 1/3 y así sucesivamente.

2. Marco metodológico

Derivado de los problemas que las secuencias nominales hispanas pueden presentar, existe un alto grado de ambigüedad en las mismas. Partiendo de esta hipótesis, se ha implementado una gramática, que considera dichos problemas, en un analizador sintáctico automático.

Para dar sustento a esta hipótesis, se ha aplicado el analizador para la obtención de datos cuantitativos exploratorios sobre un corpus de estudio con más de 745.084 registros personales homogéneos, conseguido de una aproximación demográfica de un estado mexicano.

2.1. Gramática generativa

Siguiendo la estructura de la Figura 1, se planteó la gramática que a continuación se muestra, para determinar las posibles interpretaciones válidas de una secuencia nominal en un texto electrónico.

- Símbolos no terminales:
  - Símbolo inicial - Secuencia Denominativa (SD),
  - Secuencia de Nombres (SN),
  - Secuencia de Nombres Masculinos (SNM),
  - Secuencia de Nombres Femeninos (SNF),
  - Nombre Masculino (NM),
  - Nombre Femenino (NF),
  - Secuencia de Apellidos (SA),
  - Apellido Paterno (AP) y
  - Apellido Materno (AM).
- Símbolos terminales:

  Cadenas de letras del alfabeto hispánico (en nuestro caso, nombres y apellidos).

- Reglas de reescritura:

  Etapa 1

  $SD \rightarrow SN \mid SA \mid SN\ SA$

  $SN \rightarrow SNM \mid SNF$

  Etapa 2

  $SNM \rightarrow NM \mid NM\ NM \mid NM\ NF$

  $SNF \rightarrow NF \mid NF\ NF \mid NF\ NM$

  $SA \rightarrow AP \mid AP\ AM$

  Etapa 3

  $NM \rightarrow nm \in$ {Juan, Francisco, Pedro, José,…}

  $NF \rightarrow nf \in$ {Juana, María, Margarita, Rosa,…}

  $AP \rightarrow ap \in CA$

  $AM \rightarrow am \in CA$

  CA = {Hernández, Martínez, García, Pérez,…}

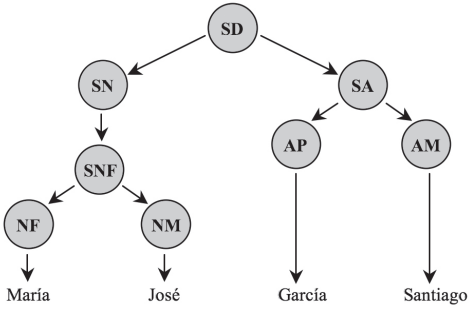
Los símbolos terminales de la gramática propuesta, fueron extraídos de una base de datos con información de nombres y apellidos admitidos en el sistema nominal hispano.

2.2. Analizador sintáctico

Se diseñó un analizador sintáctico automático especializado que utiliza la gramática propuesta para realizar un etiquetado de todos los elementos que forman parte de la secuencia denominativa en un texto; tomando en consideración durante el análisis, las cinco causas principales de ambigüedad que las cadenas nominales pueden presentar.

El siguiente es un ejemplo de una secuencia denominativa después de ser extraída de un documento. En este caso la cantidad de elementos que la componen resulta en una interpretación sin ambigüedad.

SD: María José García Santiago



Salida del analizador para una secuencia SIN ambigüedad:

María José García Santiago

$[\ [\ [\text{María}]_{NF}\ ,\ [\text{José}]_{NM}\ ]_{SNF}\ ,\ [\ [\text{García}]_{AP}\ ,\ [\text{Santiago}\ ]_{AM}\ ]_{SA}\ ]_{SD}$

Figura 3. Árbol sintáctico de una Secuencia Denominativa.

La primera etapa en el analizador sintáctico consiste en la delimitación de los elementos que conforman la Secuencia Denominativa (SD). En ella se pueden obtener hasta tres elementos (nombre, apellido paterno y apellido materno). Según la estructura propuesta en la Figura 1 y la composición de los apellidos si SD está formada por más de 6 palabras (nombre asociado/

compuesto + apellido paterno asociado/compuesto + apellido materno asociado/compuesto), se descarta.

El Algoritmo 1 muestra el proceso de formación de elementos, formando grupos para representar las asociaciones o composiciones. Los grupos obtenidos determinan la cantidad de elementos en la secuencia.

Algoritmo 1. Agrupamiento de los elementos que conforman la secuencia denominativa

```
Entrada: La secuencia denominativa, SD
Salida: La lista de los elementos agrupados, elementos
Delimita_elementos (SD)
tamaño = número de elementos en SD
for cada elemento e en SD do
  añade (e, elementos)   {añadir e a la lista elementos}
if tamaño > 6 then {Entrada inválida}
else if tamaño > 3 then
  grupo = tamaño – 3 {Cantidad de grupos para tener 4 elementos}
  {Se agrupan de derecha a izquierda viendo que existan en la BD}
  for cada elemento e en elementos y grupo do
    if esta_en_BD ((e – 1) + e) { e – 1 es elemento anterior} then
      e - 1 = e - 1 + e {Se concatenan}
      borrar(e) {Se borra e de elementos pues se concatenó a e – 1}
```

Con los grupos formados en el Algoritmo 1 se realizan todas las combinaciones válidas. Se utilizan dos vectores, uno SN y SA, que se van marcando según corresponda si el elemento de entrada está en la base de datos como nombre, como apellido o ambos (ambigüedad de dualidad nombre/apellido).

Algoritmo 2. Análisis de los elementos que conforman la secuencia denominativa

```
Entrada: La lista de los elementos agrupados, elementos
Salida: Interpretaciones válidas de la secuencia
Analiza_elementos (elementos)
cuentaIntersec = 0;
seDescarta = false;
for cada e en elements do
  SN(e) = añade (esNombre(e))
  SA(e) = añade (esApellido(e))
{Si hay más de 2 nombres (hacia la derecha) los siguientes se ponen en falso}
for cada e en elementos do
  if posición de e > 1 then
    SN(e) = 0
{Si hay más de 2 apellidos (hacia la izquierda) los siguientes se ponen en falso}
for cada e en elementos do
  if posición de e < tamaño - 2 then
    SA(e) = 0
{Si hay un elemento que no es nombre ni apellido se descarta la combinación}
for cada e en elementos do
  if NOT SN(e) AND NOT SA(e) then
    seDescarta = true
  else if SN(e) AND SA(e) then {Se cuentan las intersecciones}
    cuentaIntersec = cuentaIntersec + 1
```

La función *esNombre* además de verificar que el nombre exista, regresa el género extraído de la base de datos (*M* - masculino, *F* - femenino, *I* - indefinido). La función *esApellido* solo indica si está registrado o no en la lista de apellidos hispánicos con que se cuenta.

Finalmente con el número las secuencias *SN* y *SA* y el número de intersecciones *cuentaIntersec* es posible proponer todas las interpretaciones válidas para la *SD* de entrada.

Salida del analizador para secuencias CON ambigüedad:

- Juan Alfonso Rivera

En esta secuencia, Alfonso representa la fuente de ambigüedad por la dualidad que presenta al poder ser interpretado como nombre y apellido.

A continuación se muestran las dos posibles interpretaciones:

1. [[[Juan]$_{NM}$, [Alfonso]$_{NM}$]$_{SNM}$, [[Rivera]$_{AP}$, [∅]$_{AM}$]$_{SA}$]$_{SD}$
2. [[[Juan]$_{NM}$]$_{SNM}$, [[Alfonso]$_{AP}$, [Rivera]$_{AM}$]$_{SA}$]$_{SD}$

- Raúl Felipe Santiago Domínguez

En esta secuencia, no hay omitido ningún elemento, por tanto tiene una sola interpretación:

[[[Raúl]$_{NM}$, [Felipe]$_{NM}$]$_{SNM}$, [[Santiago]$_{AP}$, [Domínguez]$_{AM}$]$_{SA}$]$_{SD}$

Pero, ¿qué pasaría si se omite por ejemplo el apellido materno?

- Raúl Felipe Santiago

La secuencia denominativa tendría dos interpretaciones:

1. [[[Raúl]$_{NM}$, [Felipe]$_{NM}$]$_{SNM}$, [[Santiago]$_{AP}$, [∅]$_{AM}$]$_{SA}$]$_{SD}$
2. [[[Raúl] $_{NM}$] $_{SNM}$, [[Felipe] $_{AP}$, [Santiago] $_{AM}$] $_{SA}$] $_{SD}$

¿Y si ahora se descartara el primer nombre?

- Felipe Santiago

La secuencia denominativa tendría tres interpretaciones:

1. [[[Felipe]$_{NM}$, [Santiago]$_{NM}$]$_{SNM}$, [[∅]$_{AP}$, [∅]$_{AM}$]$_{SA}$]$_{SD}$
2. [[[Felipe]$_{NM}$]$_{SNM}$, [[Santiago]$_{AP}$, [∅]$_{AM}$]$_{SA}$]$_{SD}$
3. [[[∅]$_{NM}$]$_{SNM}$, [[Felipe]$_{AP}$, [Santiago]$_{AM}$]$_{SA}$]$_{SD}$

2.3. Corpus de análisis

Para nuestro análisis se consideró como fuente de información un corpus con 745.084 registros de personas residentes en el estado de Hidalgo, México (CEH). Esta cantidad representa casi la tercera parte (31,77%) del total de la población total de la región, según el último censo de población y vivienda realizado por el Instituto Nacional de Estadística, Geografía e Informática (INEGI) en el 2005.

Del CEH se extrajeron 93.998 nombres y 13.779 apellidos únicos, entre simples, compuestos y asociados.

3. Resultados

3.1. Distribución de nombres y apellidos hispanos

El conjunto de nombres y apellidos hispanos, como muchas otras recopilaciones amplias de palabras en los textos de algún lenguaje, cumple con la ley de Zipf.

Las distribuciones de los nombres y apellidos se representan en la Figura 3 y en el Gráfico 1. Cada uno de los nombres está representado por un punto, con su rango en el eje *x* y su frecuencia en *y*; ambos ejes son logarítmicos. El rango se refiere a la posición que ocupa dentro de los valores de frecuencia; de esta forma el nombre más frecuente tiene rango 1, como se observa en el Gráfico 1.

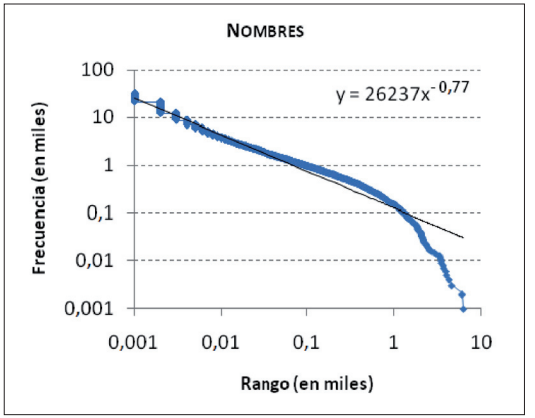


Gráfico 1. Representación de la distribución de nombres.

Los primeros diez puntos en el rango de los nombres equivalen a: Juan, Juana, María, Francisco, Margarita, Pedro, José, Alejandro, Antonio y Jesús.

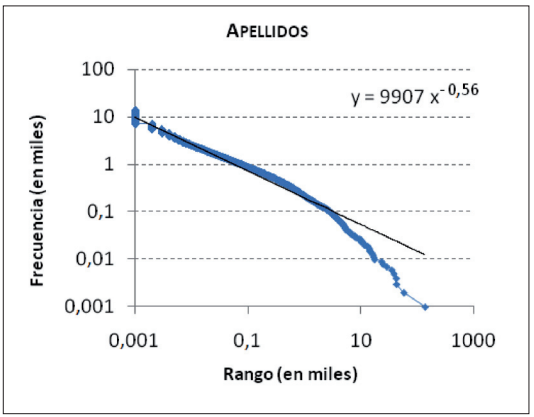


Gráfico 2. Representación de la distribución de apellidos.

Para el caso de los apellidos, los diez más frecuentes corresponden a: Hernández, Martínez, García, Pérez, Cruz, López, González, Ramírez, Sánchez y Mendoza.

Las gráficas de tendencia muestran que las distribuciones obtenidas son ejemplos de leyes de potencia discretas, familia a la cual pertenecen las distribuciones de Zipf. Nótese que en la ley de Zipf la frecuencia está expresada en términos matemáticos como $P_n \sim 1 / x^{\alpha}$, lo que es equivalente a $P_n \sim x^{-\alpha}$. Como la relación es de proporcionalidad puede incluirse una constante a, denominada constante de proporcionalidad en una relación en forma de ley potencial entre dos escalares, quedando formulada como: $P_n \sim ax^{-\alpha}$.

La expresión anterior puede interpretarse como una línea recta en un gráfico doble-logarítmico (como el empleado en las figuras), ya que se puede enunciar como: $\log(P_n) = -\alpha\log(x) + \log(a)$, la cual presenta la misma forma que la ecuación de una línea recta $y = -\alpha x + a$. El coeficiente a corresponde al punto de intersección con el eje y, y muestra la mayor frecuencia en los datos procesados.

Una ley de potencia está caracterizada por un exponente que se obtiene restando uno a la pendiente del ajuste de la gráfica de rango/frecuencia. En este caso de estudio se obtiene un coeficiente de -1,56 para los apellidos. En estudios similares realizados se produjeron los siguientes resultados: para USA, y Berlín -2, Japón -1,75 y Tailandia -1,9 (Reed & Hughes, 2003). Esto indica que el rango promedio del aumento en el tamaño de apellidos de familias excede el rango en el que los nuevos nombres se desarrollan a partir de los existentes.

3.2. Distribución de nombres y apellidos ambiguos

El Gráfico 3 muestra la distribución de los nombres y apellidos ambiguos, por concepto de dualidad entre ellos. Los ejes *x* e *y* son logarítmicos y la medida de ambigüedad ha sido determinada tomando en cuenta la cantidad de apariciones tanto en la lista de nombres, como en la de apellidos:

$$A_n = | (A_{nNombres} - A_{nApellidos}) | \times (A_{nNombres} + A_{nApellidos})$$

Por ejemplo, el nombre Santiago aparece en 858 registros como nombre y en 6.502 como apellido, por lo que su medida de ambigüedad se determina como:

$$A_n = | (858 - 6.502) | \times (858 + 6.502) = 41.539.840$$

De esta forma se obtiene un índice de ambigüedad cuyos valores miden la proporcionalidad en que un nombre puede ser también un apellido. Además, la estructura de fórmula implica que se le otorguen valores del índice más altos a aquellos elementos que posean una cantidad significativa de registros. Así, un elemento cuyos conteos de registros sean 5.000 y 2.000 para nombres y apellidos respectivamente, tendrá una medida de ambigüedad mayor que un elemento con presencia en 5 y 2 registros. Si se hubiera utilizado el cociente en la fórmula y no el producto de la diferencia y la suma, estos ejemplos hubiesen tenido los mismos valores de ambigüedad, siendo que el primero ocurre más comúnmente y es de mucho más interés en nuestro estudio.

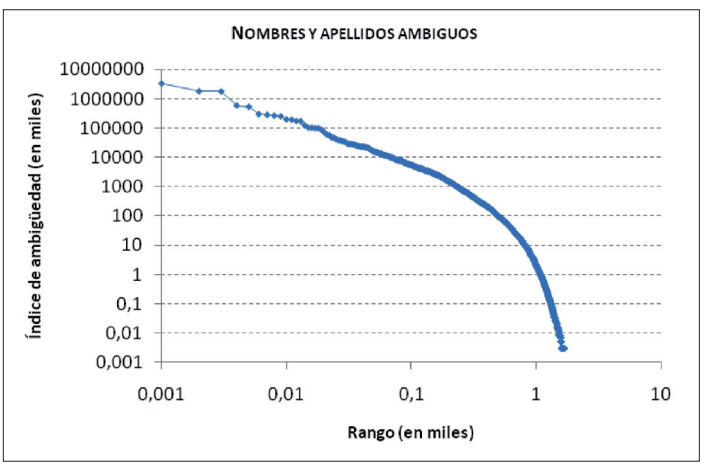


Gráfico 3. Representación de la distribución de los nombres y apellidos ambiguos.

Del gráfico puede observarse que el conjunto ambiguo obtenido aún cumple con la ley de Zipf. En este caso, Ángeles y Guadalupe ocupan las primeras posiciones en el rango.

3.3. Fuentes de ambigüedad en el corpus

Tras analizar todos los nombres y apellidos del CEH se encontraron los resultados que se muestran en la Tabla 1, que en su totalidad constituyen 79.901 fuentes productoras de ambigüedad en cadenas nominales.

En la segunda columna se consideran los elementos sin repetición de los totales (93.998 nombres y 13.779 apellidos); a pesar de que los mismos se encuentran en múltiples registros, información que se presenta en la tercera columna, con un total de 33% de registros ambiguos.

Tabla 1. Cantidad de fuentes de ambigüedad en CEH.

CAUSA DE AMBIGÜEDAD	CEH CANTIDAD / PORCENTAJE		REGISTROS CANTIDAD / PORCENTAJE	
Nombres asociados	75.581	80%	205.864	28%
Nombres compuestos	1.581	2%	14.845	2%
Apellidos asociados	363	3%	7.377	1%
Apellidos compuestos	315	2%	7.975	1%
Dualidad nombre/apellido	2.061	15%	5.861	1%

Reflexiones finales

Resulta evidente, tras el estudio realizado, la complejidad de las secuencias denominativas hispanas; aun cuando los resultados obtenidos no muestran la totalidad de las diversificaciones que pueden presentarse. Las libertades para la asignación de nombres en los países hispanos y la inexistencia de entidades reguladoras son, entre otras, causantes de los altos grados de ambigüedad.

Los resultados obtenidos arrojan que un 33% (241.922) de los registros personales analizados presentan al menos dos interpretaciones válidas, resultado de contener una o más de las causas de ambigüedad estudiadas. Se detectaron 77.162 (80%) fuentes de ambigüedad de los nombres analizados y 2.739 (20%) de los apellidos analizados.

La composición nativa de algunos nombres y/o apellidos y su doble significación provocan la mayoría de los problemas de interpretación, que en muchas ocasiones, en los países en que se originan, pueden ser resueltos con formatos oficiales en los que, por lo general, se incluye un campo para cada elemento del nombre. En este mismo sentido, las indeterminaciones que se presentan cuando las cadenas nominales son empleadas en otras lenguas, deben ser remediadas con la intervención humana por las diferencias que se presentan en sus estructuras.

Es posible considerar además, que los análisis se han realizado en un corpus que contiene registros de personas que viven en el estado de Hidalgo (México) únicamente, por tanto las proporciones en otros estados de México u otros países pueden presentar un comportamiento ligeramente diferente. Esto se debe a que los nombres y apellidos se distribuyen geográficamente, consecuencia de la dinámica ancestral y cultural de las regiones.

REFERENCIAS BIBLIOGRÁFICAS

Castro, N., Vera, J., Bolshakov, I. & Sidorov G. (2004). Formalización del sistema de nombres hispanos. En M. Arias & A. Gelbukh (Eds.), *Memorias del 5º Encuentro Internacional de Computación, Colima, México* (pp. 289-295). Washington, DC: IEEE Computer Society.

Chen, H., Huang, S., Ding, Y. & Tsai, S. (1998). Proper name translation in cross-language information retrieval. En W. Hoeppner (Ed.), *Proceedings of the 17th international conference on computational linguistics, Quebec, Canada* (pp. 232-236). Morristown, NJ: Association for Computational Linguistics.

Chen, H. & Lee, J. (1996). Identification and classification of proper nouns in Chinese texts. En T. Seely (Ed.), *Proceedings of 16th International Conference on Computational Linguistics, Copenhagen, Denmark* (pp. 222-229). Copenhagen: Ministry of Research Denmark.

Coates-Stephen, S. (1991). Automatic lexical acquisition using within-text descriptions of proper nouns. En E. Brennan & A. Renear (Eds.), *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research, Ontario, Canada* (pp. 154-169). Oxford: Oxford University Press.

Galicia-Haro, S., Gelbukh, A. & Bolshakov, I. (2004). Recognition of named entities in Spanish texts. En R. Monroy, G. Arroyo-Figueroa, L. Sucar & J. Sossa (Eds.), *Proceedings of International Conference MICAI, Lecture Notes in Artificial Intelligence, Mexico City, Mexico* (pp. 420-429). Berlin: Springer-Verlag.

Gelbukh, A. & Sidorov, G. (2001). Zipf and heaps laws coefficients depend on language. En A. Gelbukh (Ed.), *Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Mexico City, Mexico* (pp. 332-335). Berlin: Springer-Verlag.

Gelbukh, A. & Sidorov, G. (2006). *Procesamiento automático del español con enfoque en recursos léxicos grandes*. Ciudad de México: IPN.

Huang, F. & Waible, A. (2002). An adaptive approach to named entity extraction for meeting applications. En M. Marcus (Ed.), *Proceedings of the 2nd International Conference on Human Language Technology Research, California, USA* (pp. 165-170). San Francisco, CA: Morgan Kaufmann.

Mani, I. & MacMillan, R. (1996). Identifying unknown proper names in Newswire Text. En B. Boguraev & J. Pustejovsky (Eds.), *Corpus processing for lexical acquisition* (pp. 41-59). Cambridge, MA: MIT Press.

Paik, W., Liddy, E., Yu, E. & McKenna, M. (1993a). Interpretation of proper nouns for information retrieval. En M. Bates (Ed.), *Proceedings of the ARPA workshop on human language technology, New Jersey, USA* (pp. 1-5). San Francisco, CA: Morgan Kaufmann.

Paik, W., Liddy, E., Yu, E. & McKenna, M. (1993b). Categorizing and standardizing proper nouns for efficient information retrieval. En B. Boguraev & J. Pustejovsky (Eds.), *Proceedings of the workshop on acquisition of lexical knowledge from text, Ohio, USA* (pp.154-160). Cambridge, MA: MIT Press.

Stevenson, M. & Gaizauskas, R. (2000). Using corpus-driven name lists for name entity recognition. En S. Nirenburg, D. Appelt, F. Ciravegna & R. Dale (Eds.), *Proceedings of 6th Applied Natural Language Processing and 1st North American Chapter of the Association for Computational Linguistics, Washington, USA* (pp. 290-296). San Francisco, CA: Morgan Kaufmann.

Reed, W. & Hughes, B. (2003). On the distribution of family names. En H. Capel, K. Dawson, J. Indekeu, H. Stanley & C. Tsallis (Eds.), *Physica A, 319* (pp. 579-590). New York, NY: Elsevier.

Thompson, P. & Dozier, C. (1997). Name searching and information retrieval. En C. Cardie & R. Weischedel (Eds.), *Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing, Rhode Island, USA (*pp. 134-140). Somerset, NJ: Association for Computational Linguistics.