LA REPRESENTACIÓN DE NOMBRES PROPIOS MULTIPALABRA EN CORPUS ANALIZADOS SINTÁCTICAMENTE[1]

ORSOLYA VINCZE

Universidad de Szeged – Universidad de La Coruña

The representation of multiword proper names in syntactically annotated corpora

The present paper constitutes a first approach to the representation of multi-lexemic proper names in syntactically annotated corpora. We provide a brief review of existing proposals on the description of the structure of multi-lexemic proper names in linguistics, followed by a discussion of how these expressions are treated in the field of Natural Language Processing, more specifically in annotated corpora. Finally, adopting the formalisms presented by the Meaning⇔Text Theory, we propose a syntactic representation for multi-lexemic proper names, concentrating on the semantic class of personal names in Galician language.

1. Introducción

La tarea de creación de recursos lingüísticos para aplicaciones informáticas como la extracción y la recuperación de información, el resumen automático, o los sistemas de pregunta-respuesta, requiere una descripción precisa y formalizada de las lenguas, basada en el uso real. Este tipo de investigación a menudo pone en relieve problemas específicos, varios de los cuales apenas han sido objeto de un estudio sistemático dentro de la lingüística. Así es el caso de los nombres propios, que tradicionalmente han ocupado una posición marginal en la lingüística – con excepción de la filosofía del lenguaje. Por esa razón, se han descuidado, al menos hasta recientemente, cuestiones como la descripción de su comportamiento morfológico y sintáctico.

Gran parte de los nombres propios están constituidos por más de una forma léxica. El tratamiento de estas expresiones multilexémicas por aplicaciones informáticas resulta problemático, dado que presentan cierto grado de idiosincrasia. Podemos observar que las expresiones como *Fundación Pedro Barrié de la Maza* u *Organización de las Naciones Unidas* son (parcialmente) no composicionales semánticamente y presentan idiosincrasias en cuanto a su estructura y material léxico:

[1] Este trabajo ha sido realizado en el marco de un proyecto de investigación financiado por el Ministerio de Ciencia e Innovación y los fondos FEDER (FFI2008-06479-C02-01). Me gustaría también agradecer a Margarita Alonso Ramos por sus correcciones y comentarios sobre las versiones anteriores de este texto.

(1) a. Trabaja en una organización/asociación de periodistas en la calle Panaderas.
b. Trabaja en la Organización/*Asociación de las Naciones Unidas.
c. Aquellos países siempre fueron naciones (muy) unidas.
d. Aquellos países son miembros de las Naciones (*muy) Unidas.

Al mismo tiempo, en muchos casos, la estructura sintagmática de los nombres propios multilexémicos se asemeja a la de sintagmas regulares, y, además, presenta cierta variación en el uso:

(2) a. Este mes hay una exposición interesante en la Fundación Pedro Barrié/Fundación Barrié/Barrié.
b. La Organización de Naciones Unidas/Naciones Unidas ha declarado el 2011 como Año Internacional de los Bosques.
c. Como reportero, Arturo Pérez-Reverte Gutiérrez/Arturo Pérez-Reverte/Pérez-Reverte/Arturo ha cubierto, entre otros conflictos, la guerra de Chipre.

En adelante nos referiremos a este tipo de expresiones como *nombres propios multipalabra.*

El objetivo de este trabajo es proporcionar un primer acercamiento a la representación de la estructura sintáctica de los nombres propios multipalabra en un *treebank*, es decir un corpus lingüístico anotado sintácticamente. En lo que sigue, empezamos por repasar algunas propuestas existentes para la descripción de la estructura sintáctica de los nombres propios multipalabra dentro de la lingüística, seguido de una descripción de la problemática del tratamiento que se hace de la misma dentro de la lingüística computacional, y más específicamente, en corpus lingüísticos anotados. Como veremos, una de las cuestiones pendientes es la de la identificación de las diferentes formas de un mismo nombre. Finalmente, tratando de abordar este problema, proponemos una posible representación de la estructura de los nombres propios multipalabra, centrándonos en la clase semántica de los nombres de persona en gallego. Para ello partimos del modelo lingüístico y los formalismos propuestos dentro de la Teoría Sentido⇔Texto[2], dado que estos nos permiten reflejar las características idiosincrásicas de las unidades multipalabra, y permiten un tratamiento de la variación formal que estas expresiones presentan en el uso real.

[2] Igor MEL'ČUK, *Dependency syntax: Theory and practice,* Albany, State University of New York Press, 1988. Igor MEL'ČUK, "Dependency in natural language", in: Igor MEL'ČUK y Alain POLGUÈRE (eds.), *Dependency in linguistic description,* Amsterdam/Philadelphia, John Bengamins, 2009, 1-110.

2. La descripción de la estructura del nombre propio

Los estudios que tienen como objetivo la descripción del nombre propio desde un punto de vista gramatical raras veces abordan la estructura interna de estas expresiones[3]. Así es el caso de la gramática española, donde este aspecto se ha ignorado completamente, mientras que sí se han tratado cuestiones como el género, el número, el uso del artículo y modificadores[4]. A continuación enumeramos brevemente las escasas consideraciones que hemos encontrado en la literatura sobre la estructura sintagmática del nombre propio.

En la gramática descriptiva inglesa, dentro de la clase de nombre propio, se hace una distinción entre una subclase llamada *proper noun* 'nombre propio', unidad monolexémica (p. ej. *John, Cambridge*) y *name* 'nombre', unidad bien mono- o polilexémica, con una forma normalmente invariable, cuya estructura interna puede ser gramaticalmente analizable (p. ej. *United States of America, University of Cambridge*)[5]. Se observa que las expresiones pertenecientes a esta última categoría, en ocasiones, se corresponden con la estructura regular de un sintagma nominal común; sin embargo constituyen formas lexicalizadas invariables que no siguen el comportamiento regular de los sintagmas nominales[6].

Allerton[7] observa que los nombres propios multilexémicos se oponen a la estructura de los sintagmas nominales regulares en que sus elementos léxicos no son regularmente contrastivos (p. ej. *United/*federated States*). El autor distingue cuatro tipos de nombres propios en inglés según su material léxico[8]: 1) los nombres propios *puros* contienen una o más formas léxicas especializadas en el papel de nombre propio (p. ej. *Sócrates, Pedro, Salamanca*); 2) los nombres propios *mixtos* se componen de nombres comunes y de nombres propios puros (p. ej. *Instituto Cervantes, Universidad de A Coruña*); 3) los nombres propios *de base común* están compuestos por los mismos elementos que un sintagma nominal común (p. ej. *Real Academia Española, Costa Brava*); y, por último, 4) los nombres propios *codificados* son siglas o abreviaturas utilizadas como nombres propios (p. ej. *AVE, RAE*).

---

[3] D. J. ALLERTON, "The linguistic and sociolinguistic status of proper names", in: *Journal of Pragmatics*, XI, 1987, 61-69.

[4] Algunos estudios que tratan las características gramaticales de los nombres propios son: María Jesús FERNÁNDEZ LEBORANS, "El nombre propio" in: Ignacio BOSQUE y Violeta DEMONTE (eds.), *Gramática descriptiva de la lengua española*, Madird, Espasa-Calpe, 1999, 77-128., Elena BAJO PÉREZ, *El nombre propio en español*, Madrid, Arco/Libros, 2002. y REAL ACADEMIA ESPAÑOLA, *Nueva gramática de la lengua española*, Madird, Espasa, 2009.

[5] Randolph QUIRK, Sidney GREENBAUM, Geoffrey LEECH y Jan SVARTVIK, *A contemporary gramar of the English language*, London/New York, 1985, 288.

[6] Ibidem, 294., Cf. Randolph QUIRK, Sidney GREENBAUM, Geoffrey LEECH y Jan SVARTVIK, *A gramar of contemporary English*, London/New York, 1972, 163-164. y Rodney HUDDLESTON, *An introduction to the gramar of English*, Cambridge, Cambridge University Press, 1984, 229-230.

[7] ALLERTON, op. cit., 64.

[8] Ibidem, 67-69.

Finalmente, Anderson[9] describe la estructura de los nombres de persona como yuxtaposición de sus elementos, argumentando que, dependiendo del contexto discursivo, tanto el nombre de pila como el apellido pueden funcionar como el núcleo del sintagma. El autor también observa que las subclases semánticas de nombres propios tienden a diferenciarse por su estructura sintagmática característica, como el uso del artículo y la presencia y el orden de elementos descriptivos[10].

3. La problemática de los nombres propios multipalabra en la lingüística computacional

Dentro del enfoque de la lingüística computacional, los nombres propios multilexémicos se mencionan entre las llamadas *unidades multipalabra* (*Multiword Entities*). Estas se conciben como expresiones idiosincrásicas que sobrepasan los límites de una palabra limitada por espacios en la escritura[11].

En un resumen sobre los problemas relacionados con el tratamiento de unidades multipalabra, los autores[12] observan que tanto el tratamiento de los nombres propios multipalabra como una única unidad léxica, como su representación como un conjunto de varias unidades léxicas, resultan problemáticos por diferentes razones. La primera estrategia no permite, por ejemplo, representar sistemáticamente el comportamiento sintáctico idiosincrásico que caracteriza una subcategoría semántica particular. Un ejemplo es el caso de nombres de equipos deportivos en donde el elemento descriptivo, es decir, el nombre del lugar u organización de procedencia puede ser elíptico: *(Los Angeles) Lakers*. El segundo tipo de representación, sin embargo, resulta en la admisión de elementos léxicos que no pueden combinarse libremente. Los nombres de equipos deportivos, por ejemplo, a menudo contienen elementos que no se utilizan en otro contexto: existen equipos llamados *San Francisco 49ers* y *Oakland Raiders* pero la combinación de los elementos es restringida, **Oakland 49ers*[13].

Dado que las tareas del reconocimiento de los nombres propios son capaces de funcionar con considerable éxito utilizando heurísticas simples[14], este enfoque del procesamiento de lengua natural (PLN), en general, no se ocupa de la representación y la descripción de las propiedades lingüísticas de estas expresiones. Esto no es verdad, sin embargo, en el caso de tareas más complejas que requieren la identificación de

---

[9] John ANDERSON, "On the structure of names", in: *Folia lingüística: Acta Societatis Linguisticae Europaeae*, XXVII/3-4, 1003, 347-398.

[10] Ibidem, 358-359.

[11] Ivan A. SAG, Timothy BALDWIN, Francis BOND, Ann COPESTAKE y Dan FLICKINGER, "Multiword Expressions: A pain in the neck for NLP" in: Alexander GELBUKH (ed.), *Proceedings of CICLING-2002, 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, Berlin/Heidelberg, Springer-Verlag, 2002, 1-15.

[12] Idem.

[13] Ibidem, 5-6.

[14] Cvetana KRSTEV, Duško VITAS, Denis MAUREL, Mickaël TRAN, "Multilingual Onthology of Proper Names", in: *Proceedings of the 2nd Language & Technology Conference*, 2005, 116-119.

aliases, las diversas formas en las que el nombre de una entidad puede aparecer en un texto[15], como pueden ser la traducción asistida y el alineamiento multilingüe[16].

Uno de los fenómenos que pueden complicar la tarea de la traducción es la existencia de varias formas, desde abreviaturas hasta las formas completas y sus formas truncadas. Así, por ejemplo, cualquiera de las formas inglesas *United Nations Organisation, United Nations* y *UNO* puede servir como equivalente de las expresiones *Organización de las Naciones Unidas, Naciones Unidas* y *ONU*. Al mismo tiempo, existen formas que sin contexto suficiente pueden resultar ambiguas. Por ejemplo, tanto el nombre de la capital gallega *Santiago de Compostela* como el de la capital chilena *Santiago de Chile* se simplifican normalmente en *Santiago*[17]. Suponen otra dificultad los casos en que un nombre propio se traduce por una expresión que contiene un adjetivo derivado y viceversa, véase 3. Existen además desajustes entre las diferentes lenguas en cuanto a la preferencia de uso y la existencia de adjetivos derivados de nombres propios[18].

(3) la nueva edición de la obra de Miguel de Cervantes/M. de Cervantes/Cervantes
la nueva edición de la obra cervantina

Un recurso concebido expresamente con el fin de proponer una solución para este tipo de dificultades es Prolexbase[19], una ontología multilingüe de nombres propios. La base de datos ofrece información sobre las formas variantes listándolas explícitamente. Así, por ejemplo se indican las formas *United States of America, US, USA* y *American* como posibles realizaciones de la forma más común *United States*. Este último se trata como la forma base, es decir el lema, y se vincula con un ID, una representación universal de la entidad referente, que a su vez está vinculada con los lemas correspondientes en otras lenguas, por ejemplo *États-Unis*, véase la Figura 1.

---

[15] David NADEAU y Satoshi SEKINE, "A survey of named entity classification", in: *Lingvisticae Investigationes*, Vol. 30, 1, 2007, 3-26.

[16] David MAUREL, "Prolexbase: A multilingual relational lexical database of proper names", in: Nicolas CALZOARI et al. (eds.), *Proceedings of LREC-2008*, París, ELRA, 2008, 334-338.

[17] Odile PITON, Thierry GRASS y Denis MAUREL, "Linguistic resource for NLP: Ask for «Die Drei Musketiere» and meet «Les Trois Mousquetaires»", in: Antje DÜSTERHÖFT y Bernhard THALHEIM (eds.), *Natural language processing and information systems, 8th International Conference on Applications of Natural Language to Information Systems*, 2003, 200-213.

[18] Duško VITAS, Cvetana KRSTEV y Denis MAUREL, "A note on the semantic and morphological properties of proper names in the Prolex Project, in: *Lingvisticae Investigationes, 30(1)*, 2007, 115-133.

[19] KRSTEV, VITAS MAUREL y TRAN, op. cit., MAUREL op. cit., Mickaël TRAN y Denis MAUREL, "Un dictionnaire relationnel multilingüe de noms propres", in: *Traitement Automatique des Langues*, XLVII(3), 2006, 115-133., etc.

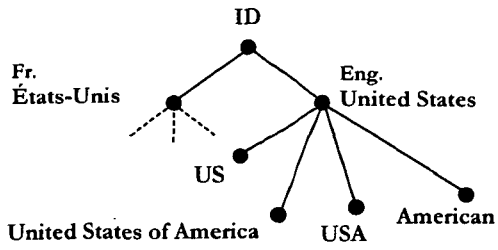

Figura 1: Representación de las diferentes formas de nombres propios en Prolexbase[20]

4. La representación de la estructura de los nombres propios en corpus anotados

La anotación, es decir, la adición de información lingüística a corpus electrónicos, es un recurso que facilita considerablemente la investigación lingüística. Se considera que los *treebanks*, corpus anotados sintácticamente, además de ser una fuente de información valiosa para el lingüista, constituyen un recurso especialmente importante para el desarrollo de software de PLN, en particular los analizadores sintácticos (*parsers*) para aplicaciones como la extracción y la recuperación de la información, el resumen automático y la traducción automática[21].

A propósito del presente estudio hemos examinado la representación de los nombres propios multipalabra en cinco *treebanks* y un corpus anotado morfológicamente. El *Penn Treebank*[22] y *AnCora*[23] que ambos utilizan un análisis de constituyentes, aunque hoy en día cada uno dispone de una versión con sintaxis de dependencias, el *Prague Dependency Treebank (PDT)*[24] que utiliza el formalismo de la

[20] Adapato de MAUREL, op. cit., 338.

[21] Geoffrey LEECH, "Introducing corpus annotation", in: Roger GARSIDE, Geoffrey LEECH, Anthony McENERY (eds.), *Corpus annotation: Linguistic information from computer text corpora,* New York, Longman, 1997, 1-18.

[22] Michell P. MARCUS, Beatrice SANTORINI y Mary Ann MARCINKIEWICZ, "Building a large annotated corpus of English: the Penn Treebank", in: *Computational Linguistics,* XIX/2, 1993, 313-330.

[23] Maria Antònia MARTÍ, Mariona TAULÉ, Manu BERTRAN y Lluís MÁRQUEZ, "AnCora: Multilingual and multilevel annotated corpora", 2007, accesible en: http://clic.ub.edu/ancora /ancora-corpus.pdf, fecha de consulta: 22 de agosto de 2011.

[24] Eva HAJIČOVÁ, Zdeněk KIRSCHNER y Petr SGALL: "A Manual for Analytical Layer Annotation of the Prague Dependency Treebank", 1999, accesible en: http://ufal.mff.cuni.cz /pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf, fecha de consulta: 22 de agosto de 2011., Alena BÖHMOVÁ, Silvie CINKOVÁ y Eva HAJIČOVÁ: "A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank" accesible en: http://ufal.mff.cuni.cz/pdt2.0 /doc/manuals/en/t-layer/pdf/t-man-en.pdf, 2005, fecha de consulta: 22 de agosto de 2011.

sintaxis de dependencias, el *TIGER*[25] y la *Floresta Sintá(c)tica*[26] que ambos utilizan una representación híbrida de constituyentes y dependencias, y, finalmente, el CORGA etiquetado[27], un corpus público de la lengua gallega analizado morfológicamente.

En resumen, podemos decir que en estos corpus se observan las dos estrategias de representación de unidades multipalabra mencionadas por Sag et al.[28] (véase arriba). Es decir, en algunos corpus, *AnCora, Floresta Sintá(c)tica* y *CORGA*, los nombres propios multilexémicos se analizan como una única unidad léxica compleja (véase en la Figura 2); y al mismo tiempo, otros corpus, *Penn Treebank, PDT* y *TIGER*, dividen los nombres propios multipalabra en sus unidades léxicas componentes, representando su estructura interior. Es interesante que en el caso de los últimos dos se utiliza una anotación especial para indicar la integridad semántica de la expresión analizada (véase en la Figura 3).

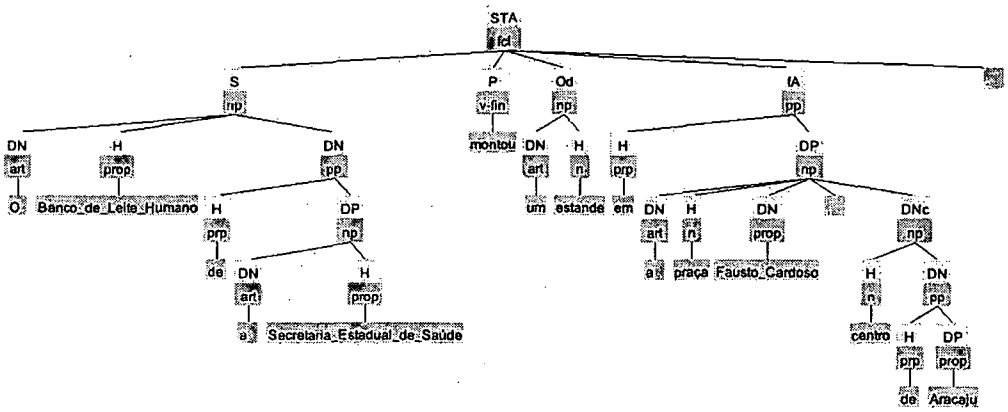

Figura 2: Los nombres propios multilexémicos *Banco de Leite Humano, Secretaria Estadual de Saúde* y *Fausto Cardoso* se representan como un único nodo sintáctico en la *Floresta Sintá(c)tica.*

[25] Sabine BRANTS y Silvia HANSEN, "Developments in the TIGER annotation scheme and their realization in the corpus", in: Mark MAYBURY (ed.), *Proceedings of LREC 2002*, Paris, ELRA, 2002, 1643-1649.

[26] Susana AFONSO, Eckhard BICK, Renato HABER y Diana SANTOS: "Floresta sintá(c)tica: a treebank for Portuguese", in: González Rodríguez, Manuel y Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002,* París, ELRA, 2002, 1698-1703.

[27] Eva DOMÍNGUEZ NOYA, Francisco Mario BARCALA RODRÍGUEZ, Miguel Ángel MOLINERO "Avaliación dun etiquetador automático estatístico para o galego actual: Xiada", in: *Cadernos de lengua* 30-31, 2009, 151-193.

[28] SAG, BALDWIN, BOND, COPESTAKE y FLICKINGER, op. cit.

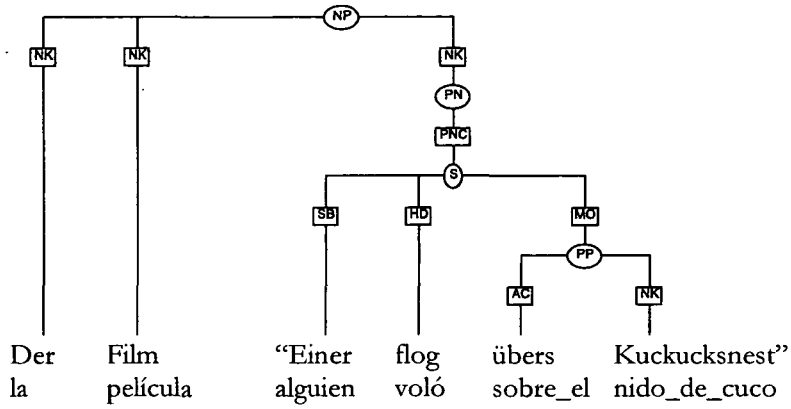

Figura 3: El nombre propio multilexémico *Einer flog übers Kuckucksnest* se representa como un subárbol sintáctico en el corpus *TIGER*[29].

5. Propuesta para la representación de la estructura de nombres propios

Tras ver las dificultades que supone el tratamiento de nombres propios multipalabra, podemos concluir que un factor importante a tener en cuenta a la hora de proponer un tipo de representación sintáctica es que esta debe permitir la identificación de las variantes formales de un nombre propio determinado. Pensamos que una representación adecuada puede servir para tratar estas variantes de una manera más económica, proporcionando una descripción de las regularidades del comportamiento de la estructura interna de las expresiones estudiadas, en vez de listarlas explícitamente, como hemos visto en el caso de Prolexbase.

En nuestra propuesta hemos optado por el tipo de representación sintáctica utilizada en el modelo lingüístico de la Teoría Sentido⇔Texto (TST)[30], dado que consideramos que su formalismo se adecua a nuestros propósitos. Dentro de este marco teórico se ha dedicado considerable atención al estudio y la descripción de las unidades fraseológicas o frasemas, definidas como expresiones compuestas de más de un elemento léxico cuyas propiedades semánticas, fonéticas y combinatorias no se derivan de las propiedades de los elementos constituyentes según las reglas de la lengua en cuestión. En la categoría de frasemas se hace la distinción entre colocaciones y locuciones[31]. Ahora nos centraremos en la representación sintáctica de estas últimas.

El modelo de la TST hace uso de dos niveles de representación sintáctica: la sintaxis profunda, más cercana a la semántica y la sintaxis superficial, orientada a la linealización y la representación morfológica. En ambos casos, se utiliza el formato de sintaxis de

[29] Ejemplo adaptado de BRANTS y HANSEN, op. cit., 34.
[30] MEL'ČUK, op. cit., 1988, MEL'ČUK, op. cit., 2009.
[31] Véase por ejemplo Igor MEL'ČUK, *Cours de morphologie générale.* Vol. 3., Montreal/París: Les Presses de l'Université de Montréal, 1997.

dependencias, donde cada relación de dependencia se representa mediante una flecha que apunta al dependiente sintáctico, y en cada caso se especifica el tipo de relación sintáctica.

La distinción entre los dos niveles sintácticos permite representar las locuciones de manera que se refleje que estas, desde un punto de vista semántico, constituyen una unidad léxica, y deben ser tratadas de igual manera que las unidades léxicas constituidas por una sola palabra. Desde un punto de vista sintáctico, sin embargo, se muestra que los frasemas, a la hora de la linearización, prosodización y morfologización se comportan como un sintagma libre[32]; véase la Figura 4. Se prevé además un conjunto de reglas que median entre los dos niveles. Este es el enfoque que, en nuestra opinión, se debe aplicar en el análisis y representación de los nombres propios multipalabra[33].

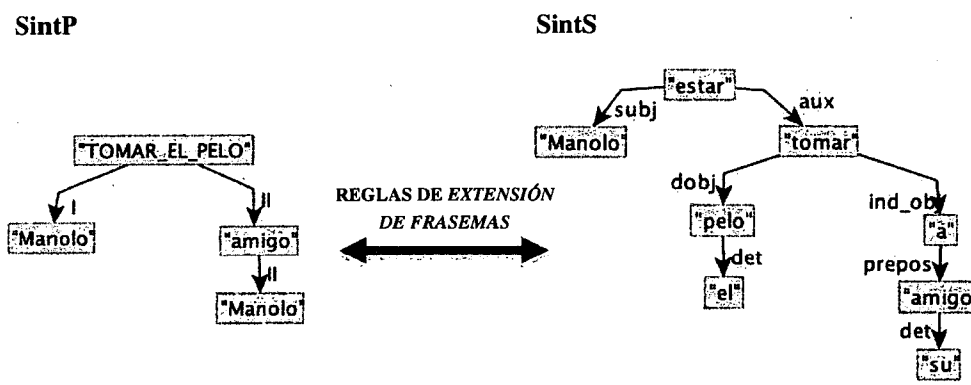

Figura 4: Representación aproximada de la estructura de sintaxis profunda (SintP) y de sintaxis de superficie (SintS).

6. La forma de los nombres de persona multipalabra

Siguiendo la terminología de Allerton[34], los nombres de persona multipalabra generalmente son nombres propios puros, ya que contienen exclusivamente elementos que a su vez son nombres propios: nombres de pila, hipocorísticos y apellidos, véase (4)[35].

(4) nombre + apellido: Manuel Rivas, Eduardo Noriega
nombre + nombre + apellido: *Celso Emilio Ferreiro*
nombre + nombre + apellido + apellido: *Xosé Luís Méndez Ferrín*
nombre + apellido + apellido: Emilio Pérez Touriño, Jaime Quesada Blanco

---

[32] Igor MEL'ČUK, "Parties du discours et locutions", in: *Bulletin de la Société de linguistique de Paris*, CI/1, 2006, 29-65.

[33] MEL'ČUK, op. cit., 1988, 27-28.

[34] ALLERTON, op. cit.

[35] Todos los ejemplos presentados en este apartado proceden de un corpus de textos periodísticos en gallego.

apellido + apellido: [Manuel] *Curros Enríquez*, [Xesús] *Alonso Montero*
hipocorístico + apellido: Manolo Escobar, Xabi Alonso, Chus Pato

En otros casos encontramos algún componente que en sí no es un nombre propio. El nombre del rey *Hassan II* está constituido por un nombre y un numeral; *José Luis G. C.* está compuesto por dos formas que representan nombres de pila y la abreviatura correspondiente a los apellidos, véase (5).

(5) nombre + numeral: *Hassan II*
nombre + nombre + abreviatura: José Luís G. C., José Manuel B. V.

Encontramos además apellidos que, por su parte, están formados por una unidad multipalabra constituida por una preposición o la combinación de una preposición con un artículo seguida de un nombre propio: *Rosalía de Castro*, *Ana Martínez de Aguila*, *Milagros del Corral*[36]. En cuanto a los nombres de pila, un ejemplo parecido es *María del Carmen*. En este caso, al contrario del anterior, se trata de una combinación que carece de autonomía, es decir no puede aparecer como elemento independiente, véase (6).

(6) a. *Milagros del Corral/ Del Corral* foi Subdirectora Xeral de Bibliotecas [...]
b. [...] esta mañá pasaron a disposición xudicial José Luís G.C. de 39 anos e *María del Carmen/ *del Carmen/ Carmen* F.R. de 39 anos, [...]

Hay que mencionar que resulta problemático decidir si estas formas complejas, tanto en castellano como en gallego, deben ser consideradas como un conjunto que forma un nombre, por tanto una unidad léxica, o si deben ser tratadas como dos nombres encadenados. En el caso concreto de *María del Carmen* o *María do Carme* en gallego, podemos observar la ya citada falta de autonomía de **do Carme* frente a las formas *María* y *Carmen*, y, además la existencia de un único hipocorístico para la combinación: *Maricarme/Mari Carme*. Otras combinaciones de nombres de pila a las que les corresponde un único hipocorístico son: *María Xosé=Chus, Xosé María=Chema, María Teresa=Maite*. Además, según la *Nueva gramática de la lengua española*[37], se observa una variación en el uso del plural de las formas establecidas de nombres compuestos, con cierta preferencia a pluralizar únicamente el segundo componente (p. ej. *los Juan Antonios, los José Manueles, los Pedro Pablos, los Francisco Javieres*)[38]. El fenómeno se interpreta como una consecuencia de la diferencia en la percepción de los hablantes sobre la segmentación de estas palabras.

[36] Somos conscientes de que en este caso se trata de formas que claramente pertenecen al español, sin embargo, debido a que existen las formas correspondientes en gallego (*do Corral, María do Carme*) que no aparecen en nuestro corpus, preferimos mencionarlos aquí.
[37] RAE op. cit., 160-161.
[38] En el caso del gallego, no encontramos información con respeto a estos casos.

7. Representación TST de la estructura sintáctica de los nombres de persona

En cuanto al análisis sintáctico de los nombres de persona multilexémicos, hemos encontrado tres propuestas específicas. En primer lugar, el PDT[39] propone un análisis en el cual los demás elementos son dependientes del último elemento del nombre, generalmente el apellido y se representan como modificadores adnominales; véase la Figura 6[40]. En segundo lugar, como hemos mencionado anteriormente, Anderson[41] propone representar la relación entre el nombre de pila y el apellido como yuxtaposición. En tercer lugar, desde el punto de vista de la TST, Bolshakov[42] sugiere representar los nombres de persona españoles como cadenas de dependencias, siguiendo el orden lineal de los elementos, y propone una relación sintáctica de superficie específica, llamada *apositiva de denominación* (*nomination-appositive*) para describir la dependencia entre los componentes; véase la Figura 7.

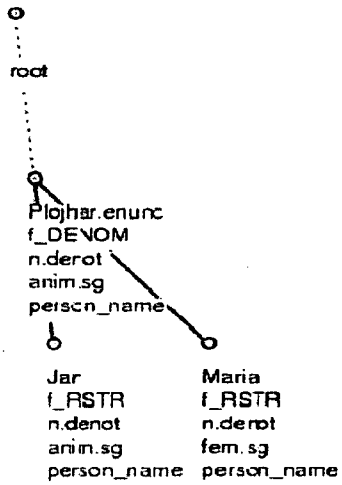

Figura 6: Representación del nombre *Jan Maria Plojhar* en el PDT

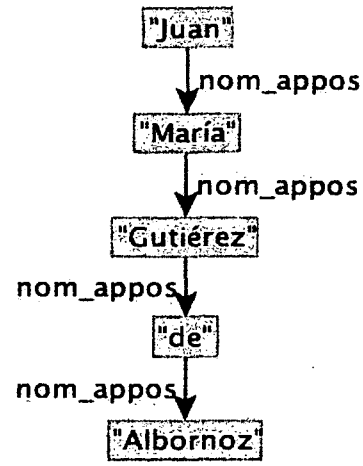

Figura 7: Representación del nombre *Juan María Gutiérez de Albornoz*

Veremos que cada una de estas tres propuestas pone de relieve algún aspecto problemático en cuanto a la representación sintáctica de los nombres de persona, una tarea que, según nuestra opinión, también supone un problema para la TST. Como hemos dicho, siguiendo las pautas que el marco teórico utiliza para la representación de

[39] Böhmová et al., op. cit., 836.
[40] Ejemplo adaptado de Ibidem, 837.
[41] ANDERSON, op. cit., 374.
[42] Igor BOLSHAKOV, "Surface syntactic relations in Spanish", in: Alexander GELBUKH (ed.), *Proceedings of CICLing-2002, 3rd International Conference on Intelligent Text Processing and Computational Linguistics,* Berlin/Heidelberg, Springer-Verlag, 2002, 210-219.

las locuciones, consideramos que los nombres propios multipalabra deben ser representados como un único nodo en la sintaxis profunda. En lo que sigue, estudiaremos la posibilidad de la aplicación de los criterios formales establecidos por la TST en la representación de la sintaxis superficial al caso de los nombres de persona gallegos. Los criterios que hemos tenido en cuenta se dividen en dos grupos: los que determinan la orientación de la relación de dependencia y los que conciernen el tipo de la relación sintáctica entre dos elementos.

7.1 La orientación de la relación sintáctica en la representación de nombres de persona

La TST dispone de tres criterios formales para establecer cuál de dos elementos es el regente y el dependiente[43]. El primero de ellos define el regente sintáctico como el elemento que determina en un mayor grado la valencia pasiva del sintagma, es decir, las funciones sintácticas que este puede tener dentro de la oración. Consideramos que este criterio no es aplicable en el caso de los nombres de persona, debido a que, los componentes, el nombre de pila y el apellido, y la forma compleja de un nombre de persona se caracterizan por la misma valencia pasiva. Representan una excepción algunas formas ya mencionadas que contienen un componente que no pertenece a la categoría de nombre propio. En estos casos, la orientación de la dependencia se puede determinar de igual manera que en otros sintagmas parecidos: *Hassan→II.*

Los apellidos que tienen la forma preposición+nombre representan un caso problemático para este análisis. Como hemos visto, estas formas pueden aparecer de manera autónoma, sin que la preposición esté regida por ningún elemento, en otras palabras, tienen la valencia pasiva de un nombre propio en vez de un subárbol regido por una preposición. Según Mel'čuk[44], una expresión multilexémica solo puede ser representada mediante un árbol sintáctico de superficie si este puede ser realizado siguiendo de manera usual las reglas de la sintaxis de superficie y la morfología de la lengua en cuestión. En el caso contrario, la expresión debe ser concebida como un *compuesto fraseologizado*, y representada por un único nodo: *de_Albornoz*[45].

El segundo criterio establece que el regente es el punto de contacto morfológico del sintagma, el elemento que impone o recibe inflexión morfológica en relación con los elementos exteriores a él. Su aplicación al caso de los nombres de persona también resulta problemático, debido a la escasa flexión morfológica en gallego, en comparación con otras lenguas; pues los nombres en gallego no presentan variación morfológica según el caso. El criterio sería aplicable, por ejemplo, en el caso del húngaro, dado que

[43] MEL'ČUK, op. cit., 2009, 27-34.

[44] MEL'ČUK, op. cit., 2006, 40.

[45] TRAN y MAUREL (op. cit., 120) mencionan un caso del francés en donde el apellido puede aparecer sin la preposición. Es posible derivar del nombre del escritor francés *François-René de Chateaubriand* las dos variantes *de Chateaubriand* y *Chateaubriand*, al contrario, en el caso del nombre *Charles de Gaulle*, solo es disponible la forma *de Gaulle* frente a **Gaulle*. En cuanto al caso de los apellidos en gallego y en español, no hemos encontrado información sobre la posible elipsis de la preposición.

en esta lengua siempre es el último componente del nombre el que recibe la flexión morfológica, independientemente de si se trata de un apellido o un nombre de pila, véase (7). Basándonos en estos ejemplos, podemos decir que el criterio de tomar el punto de contacto morfológico del sintagma como nodo dominante nos permite seleccionar el último elemento de los nombres de persona como regente en el húngaro, al igual que en el análisis del PDT (véase arriba).

(7) a. Tegnap láttam Szabó Máriát / Máriát / Szabót.
Ayer vi Szabó María$_{acusativo}$ / María$_{acusativo}$ / Szabó$_{acusativo}$.
Ayer vi a Szabó María/María/Szabó.

b. Julia Sáncheznek / Juliának / Sáncheznek ítélték oda az ösztöndíjat.
Julia Sánchez$_{dativo}$ / Julia$_{dativo}$ / Sánchez$_{dativo}$ concedieron la beca.
Le concedieron a beca a Julia Sánchez[46].

En gallego y castellano la única flexión que puede recibir un nombre propio es la del plural. No hemos encontrado información con respeto a la formación del plural de los nombres de persona en gallego, por tanto aquí nos remitimos a su comportamiento en castellano. Según la *Nueva gramática de la lengua española*[47], no existen normas establecidas para la formación del plural de los nombres propios de persona, de hecho, en el uso se observa una considerable variación. Se presentan ejemplos donde el nombre de pila aparece en la forma del plural, acompañado de la forma del singular del apellido, véase (8). A partir de estos casos podríamos concluir que únicamente el nombre de pila recibe la flexión. Debemos observar, sin embargo, que gran parte de los apellidos es invariable en cuanto a la marca del plural (p. ej. *los Fernández, los Sanz, los Valdés*), y que el resto de los apellidos oscila considerablemente entre el uso de la forma del plural y una forma invariable (p. ej. *los Ochoa~los Ochoas*). Al mismo tiempo, como hemos visto, también se observa alternancia en los nombres de pila compuestos, en el caso de los cuales a menudo se pluraliza únicamente el segundo componente. En conclusión, estos ejemplos no nos proporcionan un argumento firme para establecer el regente sintáctico en? base del criterio morfológico.

(8) "[...] nunca más volverá a haber en Nicaragua Adolfos Díaz, Emilianos Chamorro, José Marías Moncada, Anastasios Somoza en el poder [...]"[48]

El tercer criterio definitorio, según el cual el regente sintáctico es el elemento que aporta en mayor medida la carga semántica del sintagma, nos remite a la reflexión de Anderson[49].

[46] Nótese que en húngaro en los nombres de persona húngaros el nombre de pila precede al apellido, mientras en los nombres de persona extranjeros se mantiene el orden original. Sin embargo, independientemente de si se trata de un nombre húngaro o uno extranjero, es siempre el último elemento que recibe la flexión.

[47] RAE, op. cit., 160-164.

[48] Ejemplo presentado en RAE (op. cit.) procede de Sergio Ramírez: *El alba de oro. La historia viva de Nicaragua.*

Como vimos, este autor mantiene que tanto el nombre de pila como el apellido pueden funcionar como modificadores del otro elemento, y que este comportamiento semántico se determina en base del contexto discursivo en cada caso concreto y no supone estructuras sintácticas diferentes. Por ejemplo, en el caso del nombre *Eduardo Noriega*, podemos imaginar un contexto determinado en donde el nombre de pila *Eduardo* identifique a la persona en cuestión entre los miembros de la familia *Noriega*, o, al contrario, el apellido puede servir para identificarla entre las personas que llevan el nombre *Eduardo*. En consecuencia, según el autor, los nombres de persona complejos no tienen un núcleo sintáctico real y sus componentes se combinan simplemente por yuxtaposición.

Finalmente, observemos un criterio que, según Mel'čuk[50], no es definitorio a la hora de determinar la orientación de una relación sintáctica, el criterio de omisibilidad. Según el autor, en una relación de dependencia, el dependiente típicamente se puede omitir sin afectar la corrección sintáctica, mientras que el regente típicamente no es omisible. En el caso de los nombres de persona, tanto el nombre de pila como los apellidos son omisibles. Además, la omisibilidad de los componentes se condiciona por la convención social, el contexto discursivo, etc. Así, por ejemplo, en el caso de los nombres gallegos y españoles, comúnmente se utiliza el primer nombre de pila junto al primer apellido, aunque hay ciertas combinaciones de nombres de pila que muestran una tendencia del uso del segundo nombre (p. ej. *María Dolores, José Luis*, etc.). De igual modo, en general, se prefiere el uso del primer apellido, pero encontramos ejemplos contrarios: *José Luis Rodríguez Zapatero→Zapatero*. En un contexto familiar o informal tendemos a utilizar el nombre de pila, mientras que en un contexto formal como los textos de un periódico, como hemos podido observar en los ejemplos del corpus, utilizamos una combinación de nombres de pila y apellidos, y en algunos casos solo el apellido, pero raras veces el nombre de pila.

En conclusión, podemos decir que las características particulares de los nombres de persona complejos no se prestan a que establezcamos claramente relaciones de dependencia sintáctica utilizando los criterios de la TST. Por razones prácticas, optaremos por un análisis semejante al de Bolshakov[51], en el cual los componentes del nombre forman una cadena de dependencias respetando su orden lineal, dado que consideramos que esta representación ofrece una ventaja a la hora de la morfologización y linealización de la expresión. Nótese que en el análisis del PDT (véase arriba) la linealización puede ser problemática, dado que el árbol aparentemente no conserva ninguna información sobre el orden de los constituyentes, salvo el último elemento. En nuestra opinión, este aspecto es muy importante y debemos tenerlo en cuenta en el caso de los nombres de persona, ya que la determinación del orden de los elementos no depende (enteramente) de la gramática[52].

---

[49] ANDERSON, op. cit., 374.

[50] MEL'ČUK, op. cit., 2009, 42.

[51] BOLSHAKOV, op. cit.

[52] Una de las características importantes de la sintaxis de dependencias, en comparación con la sintaxis de constituyentes, es la falta de representar explícitamente el orden de palabras en el árbol sintáctico. En el marco de la TST, se considera que el orden lineal se determina en la morfología, por tanto, es en la representación morfológica donde se debe reflejar, mientras que en la sintaxis superficial queda implícita, pero obligatoriamente representado por las relaciones sintácticas (Cf.

7.2 Los tipos de relación sintáctica en la representación de nombres de persona

En la TST se definen tres criterios para determinar el número y el tipo de relaciones sintácticas de superficie[53]. Según el primero, no puede existir el mismo tipo de relación entre dos formas de palabras de dos sintagmas diferentes si en ambos casos se trata de la combinación de los mismos lexemas, pero los dos sintagmas difieren en su contenido semántico, o muestran diferencias formales relevantes en la sintaxis, como el orden de palabras o gramemas sintácticos, etc. (p. ej. *the visible*←*stars* = 'las estrellas generalmente visibles' vs. *the* stars→visible = 'las estrellas visibles en el momento'). Este criterio apunta a que entre dos componentes determinados de un nombre siempre debe existir el mismo tipo de relación sintáctica. Dado que hemos decidido establecer relaciones de dependencia entre los componentes consecutivos, teóricamente, la combinación de dos formas consecutivas (p. ej. *José→María*; *José→Pérez*) pueden hacer referencia a la misma persona en cualquier contexto, por tanto, pueden ser siempre vinculadas por la misma relación. Al contrario, en el caso de las combinaciones *José→María* y *María→José*, la orientación de la relación de dependencia es diferente.

El segundo criterio establece que cada relación sintáctica debe tener una categoría sintáctica como dependiente prototípico que pueda sustituir al dependiente actual en cualquier configuración con la misma relación. En consecuencia, para representar un nombre completo correctamente formado necesitamos más de una relación sintáctica: tenemos que distinguir entre, al menos dos tipos de relaciones[54]: una de ellas tiene como dependiente prototípico un nombre de pila o un apellido, mientras que el dependiente prototípico de la otra es un apellido. Un nombre de pila puede ir seguido tanto de otro nombre como por un apellido, en ambos casos resultando en formas correctas: *María→Camiño Noia Campos*, *María→Noia Campos*, *María→Campos*. Sin embargo, la sustitución del dependiente por un nombre de pila tras un apellido resulta en una estructura incorrecta[55]: *Noia→Campos*, **Noia→María*. Pensamos que es conveniente la introducción de un tercer tipo de relación para indicar la dependencia entre el nombre de pila (o un hipocorístico) y el apellido. La relación propuesta, en el caso del gallego (o el español), tendrá como dependiente prototípico un apellido.

Por último, según el tercer criterio, una relación sintáctica debe ser no repetible (p. ej. relaciones actanciales) o infinitamente repetible (p. ej. modificadores) con el mismo

---

MEL'ČUK op. cit., 1988, 15-16; Lidija IODANSKAJA e Igor MEL'ČUK, "Establishing an inventory of surface-syntactic relations: Valency-controlled surface-syntactic dependents of the verb in French", in: Igor MEL'ČUK y Alain POLGUÈRE (eds.), *Dependency in linguistic description*, Amsterdam/Philadelphia, John Benjamins, 2009, 151-234.

[53] MEL'ČUK, op. cit., 2009, 34-39.

[54] Cf. BOLSHAKOV, op. cit.

[55] Este análisis no abarca los casos cuando el nombre de pila y el apellido aparecen en un orden invertido (p. ej. *Noia Campos, María*). Consideramos que aquí la relación entre los apellidos y el nombre de pila es la de aposición, ya que se trata de una modificación semántica de los primeros elementos por el último.

regente sintáctico. Con respeto a este criterio, podemos observar que en la cadena de dependencias propuesta para el análisis de nombres de persona, por definición, cada relación es irrepetible con el mismo regente.

En cuanto a la denominación de los tres tipos de relaciones propuestas, adoptaremos el nombre *apositiva de denominación* de Bolshakov para la relación entre un nombre de pila (o hipocorístico) y el primer apellido. Sin embargo, preferimos denominar *junctive de nombre* y *junctive de apellido* la relación entre dos nombres de pila y la relación entre los dos apellidos, respectivamente. Nótese que la relación *junctive* se utiliza para designar la relación sintáctica existente entre los elementos de formas compuestas como numerales o los verbos preposicionales del inglés (p. ej. *ciento*←num-junctive-*treinta*, *give*-verb-junctive→*up*)[56].

Para concluir nuestra argumentación, hemos utilizado el modelo de la TST para presentar un ejemplo con respecto a las diferentes estructuras posibles en la sintaxis de superficie del nombre de persona *María Teresa Táboas Veleiro*; véase la Figura 8.

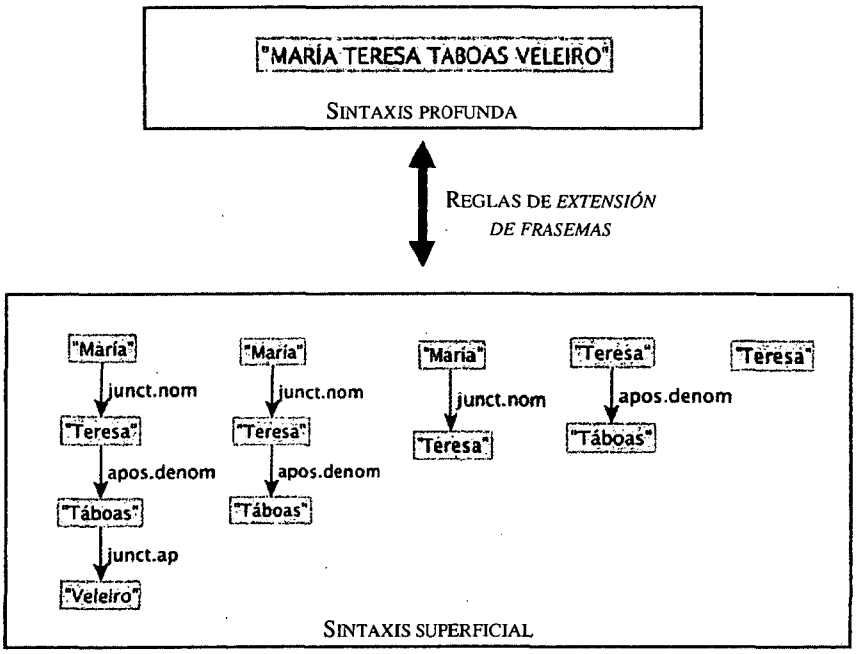

Figura 8: Representación de las posibles estructuras de un nombre de persona en la sintaxis de superficie.

[56] MEL'ČUK, op. cit., 2009, 52.

8. Conclusión

El objetivo del presente trabajo ha sido proporcionar un primer acercamiento a la representación de la estructura sintáctica de los nombres propios multipalabra. Hemos examinado algunas consideraciones sobre la descripción lingüística de este tipo de expresiones y su tratamiento en varios *treebank*s. Hemos concluido que las propuestas de corpus se reducen principalmente a dos tipos de análisis. Algunos de ellos conciben los nombres propios multipalabra como una única unidad léxica, y, en consecuencia, los representan como un único nodo sintáctico, mientras que otros corpus sintácticos los representan de manera semejante o igual a los sintagmas composicionales.

Desde el punto de vista de diversas tareas de PLN, se presenta la necesidad del tratamiento de las variantes formales de los nombres propios. Hemos propuesto que el formalismo de análisis sintáctico de la TST resulta adecuado para dar cuenta de esta variación, debido a que hace uso de dos niveles de representación sintáctica. Siguiendo el modelo de representación que se aplica en el caso de las locuciones, en el nivel de sintaxis profunda los nombres propios se pueden representar como un único nodo sintáctico, conforme a sus características semánticas, mientras que, en el nivel de sintaxis de superficie, se les puede asignar un subárbol de dependencias.

No obstante, como hemos podido observar en el caso de los nombres de persona multipalabra, las características especiales de estas expresiones plantean varios problemas en cuanto a la aplicación de los criterios formales de análisis sintáctico establecidos dentro del marco teórico de la TST. Con todo, la escasez de descripción lingüística detallada, frente a la gran utilidad que esta puede tener en el ámbito computacional, nos indica que la tarea de la descripción y representación de la estructura tanto de los nombres de persona como de los nombres propios multipalabra en general es un reto fundamental para la futura investigación.