ONOMÁZEIN
Revista semestral de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE
FACULTAD DE LETRAS

La estructura de la enumeración. Análisis, descripción y propuesta de detección automática

Enumeration structure. Analysis, description and automatic detection proposal

Walter Koza
Pontificia Universidad Católica de Valparaíso
Chile

ONOMÁZEIN 35 (marzo de 2017): 173-194
DOI: 10.7764/onomazein.35.10

Walter Koza: Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso, Chile. Proyecto Fondecyt 11130469 | Correo electrónico: walter.koza@ucv.cl

Fecha de recepción: septiembre de 2015
Fecha de aceptación: diciembre de 2016

Resumen

En el presente artículo se indaga acerca de la estructura de la enumeración a partir de un enfoque gramatical y desde la perspectiva de la lingüística computacional. Para ello, se exponen algunas consideraciones teóricas del fenómeno sobre la base de la naturaleza de los elementos que componen la enumeración, la relación que esta entabla con la matriz que la contiene y el elemento sintáctico que posibilita su aparición en la cláusula. Posteriormente, y sobre la base de dicha descripción, se realiza una modelización que permite una implantación en máquina, a fin de desarrollar un método de detección automático. Para el trabajo computacional se recurrió al programa NooJ. Dicho método fue probado en un corpus compuesto por entradas de Wikipedia relacionadas con el dominio médico, logrando 88,40% de precisión, 90,19% de exhaustividad (recall) y 89,82% de Medida F. Finalmente, se plantean los aportes del presente estudio para los estudios gramaticales y la lingüística computacional, a la vez que se establecen nuevos lineamientos de trabajo.

Palabras clave: enumeración; enumerador; enumerando; enumeratema; análisis automático.

Abstract

In this paper, the enumeration structure is inquired from a grammatical aproach and a computational linguistics perspective. For this, some theoretical aspects based on the nature of the elements that compose the enumeration, the relation among the enumeration, the matrix that contains it and the syntactic element that allows the enumeration in the sentence are exposed. Afterwards, a modelization that allows a computational implementation are made. NooJ program was used for the computational work. The method has been checked in a corpus composed by Biomedical Wikipedia entries, and it got 88,40% of precision, 90,19% of recall and 89,82% of F-Measure. Finally, the contributions of this paper for grammatical studies and computational linguistics are presented, and new lines of work are proposed.

Keywords: enumeration; enumerator; enumerando; enumeratheme; automatic analysis.

1. Introducción

Dentro de las posibilidades de análisis que ofrece la lingüística computacional, una de sus áreas está abocada a la comprobación de hipótesis y descripciones lingüísticas, tarea que se lleva a cabo en dos momentos fundamentales. En primer lugar, se identifica un conjunto de rasgos propios de una estructura determinada y se describe exhaustivamente la manera en que estos se presentan. En segundo lugar, sobre la base de esa descripción, se realiza una implantación en máquina con el objeto de reconocer el fenómeno estudiado en textos de lenguaje natural, o bien, generar automáticamente dicha estructura. En este marco, se presenta un análisis de la enumeración dentro de la oración en español, a partir de la caracterización del fenómeno desde una perspectiva formal y de una implantación en máquina para su reconocimiento automático en textos de lenguaje natural. Esta investigación se realiza en el marco del proyecto Fondecyt 11130469, cuyo objetivo es la elaboración de un método de extracción automática de terminología médica, mediante el procesamiento de información lingüística.

La enumeración puede ser definida, de acuerdo con los planteos de Cortés (2008 y 2012), Ho-Dac, Péry-Woodley y Tanguy (2010), Fauconnier, Kamel y Rothenburger (2013), entre otros, como un conjunto de elementos que se presentan como una ilación o coordinación de ítems, que forman un todo y poseen una función sintáctica análoga. Este fenómeno ha sido tratado desde diversos enfoques, como la retórica (Marchese y Forradellas, 1986), el análisis del discurso (Bras, Prévot y Vergez-Couret, 2008; López Samaniego, 2006; Vergez-Couret, Bras, Prevot, Vieu y Attalah, 2011), la gramática textual (Porhiel, 2007), el análisis del discurso oral (Cortés, 2012) y desde la lingüística computacional (Luc, 2001; Facounnier, Kamel y Rothenburger, 2013). Para el caso del español, el trabajo más exhaustivo es el coordinado por Cortés (2008), *La serie enumerativa en el discurso oral en español*. En este volumen, se incluyen trabajos desarrollados por el grupo ILSE, de la Universidad de Almería, quienes enfocan el fenómeno de la enumeración en su relación con el texto oral y desde perspectivas retóricas, semánticas y pragmáticas. En el presente trabajo, se toman en cuenta algunos de los estudios contenidos en dicho volumen, tanto en la descripción de la estructura enumerativa como así también en los distintos casos de enumeración.

La implementación computacional de detección automática fue realizada en un corpus de textos escritos, en donde la separación de ítems estaba dada por la puntuación y conjunciones. En esta etapa se limita a la enumeración de elementos simples y quedan para futuras investigaciones las enumeraciones complejas, es decir, aquellas que enumeran elementos que poseen coma en su interior y son separados por punto y coma del primero al penúltimo, y coma y conjunción, el penúltimo y el último, como se observa en (1).

(1) [Los alumnos destacados son: Alsina, Pedro; Bertuccelli, María, y Leónidas, Matilde].

A tales efectos, se tuvieron en cuenta el estudio de Garat (2006), quien analiza las funciones de la coma desde una perspectiva computacional e incluye un breve apartado para la enu-

meración, y el de Facounnier, Kamel y Rothenburger (2013), quienes presentan un método de reconocimiento automático de enumeraciones, con el propósito de identificar relaciones término-ontológicas.

Aquí se propone un estudio de la enumeración en la oración desde una perspectiva formal, en la que se observan tres elementos: (i) la enumeración propiamente dicha, concebida aquí como un conjunto ordenado de "enumerandos", (ii) un elemento que permite la aparición de la enumeración, al que se denominará "enumerador" y que se relaciona sintácticamente con la enumeración, y (iii) un enumeratema, una partícula de la misma categoría gramatical de los enumerandos que actúa, en cierta medida, de manera análoga a un hiperónimo de cada enumerando, y que puede estar explícitamente o no. A modo de ejemplo:

(2) [Juan trabaja los días lunes, martes, miércoles y sábados].

En este caso, se pueden apreciar los siguientes elementos de:

a) Enumerandos: 'lunes', 'martes', 'miércoles' y 'sábados'.

b) Enumerador: 'trabaja'.

c) Enumeratema: 'los días'.

Partiendo de esta base, se propone un análisis que contemple la naturaleza de los enumerandos, la relación sintáctica entre estos y el enumerador. A partir de este, se realiza una modelización que permite la posterior implantación en máquina para la elaboración de un método de detección automática de estructuras enumerativas en textos de lenguaje natural. Este fue probado en un corpus compuesto por entradas de Wikipedia relacionadas con el dominio médico, el cual sumaba un total de 57.632 palabras y se lograron porcentajes adecuados de precisión (88,40%), exhaustividad (recall) (90,19%) y medida F (89,82%).

El artículo se organiza de la siguiente manera. En la sección 1, se presentan algunos antecedentes de los estudios de la enumeración. En 2, se analiza la relación entre la enumeración el enumeratema y el enumerador. En 3, se presenta la clasificación de las enumeraciones. En 4, se describirá la implantación en máquina realizada sobre la base de dicha clasificación y, por último, en 5, se discuten los resultados obtenidos y se presentan las conclusiones derivadas de la investigación.

2. Estado de la cuestión

La enumeración, como se mencionó más arriba, ha sido abordada desde diversos puntos de vista. Se destacan, particularmente, los llevados a cabo en el terreno de la retórica y, para el

caso de los estudios literarios, de la estilística (Dammame, 1989).

En cuanto a su definición, Damamme (1989), en su ya clásico estudio sobre la serie enumerativa, plantea esta construcción como:

> Toda expresión lingüística formada por un número mínimo de tres términos (palabras, sintagmas, unidades de enunciados) que pertenecen a categorías morfológicas o gramaticales idénticas o equivalentes, que ocupan una función idéntica en la sintaxis del enunciado y que conectadas lado a lado, se coordinan o conectan mediante un signo de puntuación (Dammame, 1989: 37; la traducción es mía).

Al respecto, quizá resultaría conveniente señalar que dicha "función sintáctica idéntica" que plantea la autora se suele dar en la mayoría de los casos, aunque, en ocasiones, los elementos cumplen una misma función a nivel general, pero difieren en lo específico. Tal sería el caso siguiente:

(3) [Comió en su casa, rápido y sin ganas].

Este tipo de enumeración es mencionado en la Nueva Gramática de la Lengua Española, de la RAE (2009), dentro de los casos de conjunción. Allí señala que, excepcionalmente, y por razones de énfasis, puede darse el caso de agrupaciones de elementos con funciones sintácticas distintas, como el caso de (3), que contiene dos complementos circunstanciales, de lugar ('en su casa') y de modo ('sin ganas'), y un predicativo ('rápido'). No obstante, más allá de esta especificidad, las tres pueden coexistir y enumerarse en la medida en que son adjuntos verbales.

Por otro lado, como fenómeno retórico, la enumeración se caracteriza por tratarse de una enunciación sucesiva de elementos de un conjunto que conforman un todo y ser el elemento fundamental de la serie enumerativa. Esta última es definida por Cortés (2008) de la siguiente manera:

> [...] un conjunto de elementos en relación, generalmente, de yuxtaposición, de adición o disyunción, con los que se pretende, mediante la formulación parafrástica de un fragmento discursivo anterior, elemento común al que vamos a denominar matriz, la progresión temática del discurso materializada en distintos remas que se van asignando a un mismo tema provisional (Cortés, 2008: 9).

Lo que se puede observar aquí es que la serie enumerativa incluye la enumeración y, a la vez, contiene un elemento rector denominado matriz. La matriz se puede identificar con lo que algunos autores (Fauconnier y otros, 2013; Ho-Dac, Péry-Woodley y Tanguy, 2010) han denominado 'enumeratema'. Al respecto, Berbel Rodríguez (2008) menciona que, en la enumeración, cada ítem se relaciona con los otros en la medida en que están en un idéntico nivel discursivo y establecen, al igual que ellos, una relación de dependencia con respecto a la matriz. A esto se lo denomina "principio de coenumerabilidad" y cada ítem, es decir, cada elemento coenu-

merado, presenta una relación de igualdad con respecto a un enumeratema, concibiéndose a este como una entidad de orden superior respecto de los ítems, a los que les asigna una etiqueta. Así, la interpretación de una enumeración consistiría en identificarla con un enumeratema; en caso de que este no esté explícito, se lo podría inferir a partir del contenido de los elementos de la lista (Ho-Dac, Péry-Woodley y Tanguy, 2010).

El concepto estructura enumerativa implica referirse a una categoría que puede situarse en el plano oracional o textual. Con respecto a este último, se han venido llevado a cabo estudios que, desde una perspectiva discursiva, analizan esta estructura a partir de los marcadores introductorios (Laippala, 2010) y el orden jerárquico de los elementos enumerados (Bras, Prévot y Vergez-Couret, 2008). En este trabajo, se focaliza en la relación sintáctica que se establece entre la enumeración y el resto de la oración. Para ello, se propone la categoría de enumerador para remitir al elemento que dispara los elementos de la enumeración y se relaciona sintácticamente con esta. Dicho enumerador puede coincidir o no con el enumeratema. Cuando se da esta situación, entre la enumeración y el enumerador hay una relación de aposición, en donde el enumerador es el núcleo y la enumeración el término apositivo. Esto será tratado en profundidad en la sección siguiente.

En cuanto a la clasificación de las estructuras enumerativas, se tuvieron en cuenta las propuestas de Garat (2006) y Fauconnier, Kamel y Rothenburger (2013). El primero analiza las funciones de la coma e incluye en ellas la separación de elementos de la enumeración. Las enumeraciones, de acuerdo con el autor, se dividen en simples y proposicionales. Las simples se corresponden con la enumeración de sintagmas y se clasifican en nominales, adjetivales y otras (por ejemplo, enumeraciones adverbiales). Las proposicionales, por su parte, incluyen oraciones yuxtapuestas y subordinadas. Por otro lado, Fauconnier, Kamel y Rothenburger (2013) proponen una tipología multidimensional de las estructuras enumerativas con el propósito de establecer y detectar automáticamente relaciones término-ontológicas. Dicha clasificación se basa en cuatro dimensiones: visual, retórica, intencional y semántica. El propósito es identificar los paradigmas establecidos en las estructuras enumerativas, a fin de explotarlos y construir recursos semánticos.

3. Relaciones entre la enumeración y los demás elementos oracionales. El enumeratema y el enumerador

En el análisis de las relaciones textuales entre la matriz y la enumeración, Camacho (2008) define la primera como un segmento discursivo que se amplía, "multiplicándose, diversificándose o expandiéndose en otros segmentos" (Camacho, 2008: 129). Esto implica que mientras la matriz es la expresión a expandir, la enumeración es la expansión propiamente dicha.

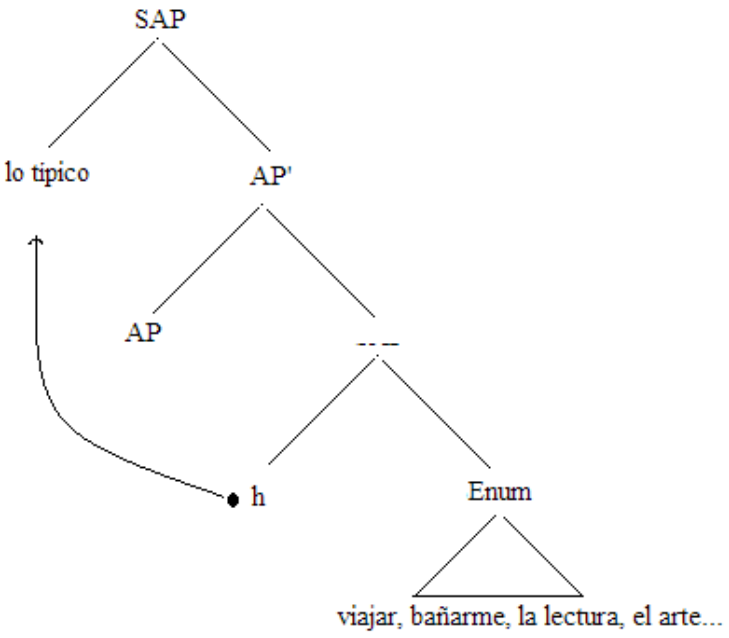
Desde esa perspectiva, la matriz constituiría la base desde la que se posibilita la enumeración. A modo de ejemplo, la autora presenta el siguiente fragmento:

(4) me gusta lo típico
viajar ehh
bañarme
la lectura
el arte
salir con amigos
y poco más (Camacho, 2008: 128).

En esta estructura, 'me gusta lo típico' es la matriz, mientras que 'viajar' es el primer elemento de la enumeración, 'bañarme', el segundo, y así sucesivamente. En este caso, dentro de la matriz se encontraría lo que se ha denominado enumeratema, en este caso, 'lo típico'.

No obstante, resulta conveniente señalar que la amplificación a la que alude la autora no afectaría a la matriz en su totalidad, sino a un elemento, que, como se verá más adelante, puede ser el enumeratema, pero no siempre es así. Si se entiende la amplificación como una relación no solo de tipo textual, sino también sintáctica, entonces, como se puede apreciar en el ejemplo de la propia Camacho (2008), la enumeración se relaciona con 'lo típico' de manera similar a una aposición, en donde 'lo típico' forma aposición con la enumeración en su totalidad. Gráficamente, y siguiendo las propuestas de Moro (1997) y Muñoz (2012), se podría representar de la siguiente manera:

FIGURA 1
Estructura del sintagma aposición con una enumeración como término apositivo



A partir de esto, se puede suponer que el enumeratema, cuando está explícito, se encontraría dentro de la matriz. Sin embargo, la autora también señala que, en ocasiones, la amplificación tiene lugar de manera adyacente e independientemente de la matriz. Esto se da cuando no se evidencia en esta última ningún elemento "significativo común" (Camacho, 2008: 129), es decir, que la matriz no contiene palabra, fragmento o posibilidad alguna de inferir sobre la información que despliega la enumeración, por ende, aquí el enumeratema debe ser inferido. Uno de los ejemplos que propone Camacho para esta particularidad es el siguiente:

(5) hay alcohólicos a los que les cuesta [ø] dejarlo
trabajo
esfuerzo
sangre
sudores y
dinero (Camacho, 2008: 130).

En lo que atañe a este ejemplo, según la autora, no se puede inferir el elemento que se expande, pero, y volviendo al plano sintáctico, lo que resulta evidente desde un primer momento es que la enumeración consiste en una ilación de sujetos del verbo 'cuesta'. En este caso, verbos como 'costar' pertenecen a la categoría de los denominados pseudo-impersonales y que, de acuerdo con Melis y Flores (2007), seleccionan sujetos inanimados, no agentivos, ubicados a la derecha del verbo, y suelen regir un objeto con el rasgo [+humano], que se codifica como objeto directo y se ubica en una posición de tópico, cobrando apariencia de sujeto.

Sin entrar en detalles de índole textual, como ser la inclusión de elementos catafóricos o anafóricos en la matriz, a los que alude Camacho (2008), puede observarse, entonces, la presencia de un elemento dentro de la oración que se haya ligado directamente por la enumeración. A dicho elemento se lo denominará enumerador y a los ítems enumerados, enumerandos:

(7) [Juan visitó parques, museos y restoranes].

Aquí, la función de enumerador la posee el verbo 'visitó y los sintagmas determinantes que actúan como complementos directos son los enumerandos; el enumeratema ('lugares') no está explícito. A la vez, se puede dar el caso de enumeradores que, a su vez, formen parte de una enumeración anterior:

(8) [Juan visitó parques, museos y restoranes argentinos, peruanos y mexicanos].

Aquí, el último elemento de la primera enumeración, 'restoranes', es el enumerador que inicia la segunda.

4. Clasificación de las enumeraciones

A continuación, se presenta una clasificación de las enumeraciones constituida sobre la base de tres aspectos:

a) Naturaleza de los enumerandos.
b) Grado de cierre de la enumeración.
c) Elementos separadores de enumerandos.

En a) se reconocen enumeraciones sintagmáticas, conformadas por sintagmas, y clausales, compuestas por cláusulas subordinadas. A la vez, pueden encontrarse también enumeraciones mixtas que combinen distintos tipos de sintagmas o sintagmas y subordinadas:

TABLA 1
Tipos de enumeración según naturaleza de los enumerandos

Naturaleza de los enumerandos		
Sintagmática	Nominales	Compré carne, pan y pescado.
	Adjetivales	María es buena, linda y simpática.
	Verbales	Quiero reír, cantar y bailar.
	Preposicionales	La casa es de mi madre, de mi padre y de mis hermanos.
Clausales	Subordinadas	Quiero que me entiendan, que no me juzguen y que me esperen.
Mixtas	Combinación de sintagmas	María es rubia, alta y de cabello largo.
	Combinación de sintagmas y cláusulas subordinadas	Quiero una casa linda, con buena vista y que tenga calefacción.

En el grado de cierre se distinguen enumeraciones abiertas y cerradas:

TABLA 2
Tipos de enumeración según grado de cierre

Grado de cierre		
Cerradas	Con conjunción	Platero es pequeño, peludo y suave.
Abiertas	Con etcétera	Platero es pequeño, peludo, suave, etcétera.
	Con puntos suspensivos	Platero es pequeño, peludo, suave…
	Con expresiones del tipo ‘entre otras cosas’, ‘entre otros’, etcétera	Platero es pequeño, peludo y suave, entre otras cosas.

Finalmente, en c), se determina el elemento que separa a los enumerandos, como ser coma, conjunciones (copulativas, disyuntivas, etcétera), contracción ‘ni’, etcétera.

TABLA 2
Tipos de enumeración según grado de cierre

Elementos ilativos		
Coma y conjunción		Compró cuadernos, lápices y gomas de borrar.
Coma	Asíndeton	Compró cuadernos, lápices, gomas de borrar.
Polisíndeton	Copulativa	Compró cuadernos y lápices y gomas de borrar.
	Copulativa negativa	No compró ni cuadernos ni lápices ni gomas de borrar.
	Disyuntiva	O compraba cuadernos o lápices o gomas de borrar.

La clasificación propuesta sirvió de base para una modelización y posterior implantación computacional. En la sección siguiente, se detallan las tareas llevadas a cabo para la detección automática.

5. Implantación computacional

Se pretendió desarrollar un procedimiento para el reconocimiento automático de las enumeraciones en textos de lenguaje natural. Esto se realizó con el objeto de corroborar la descripción de las enumeraciones realizada y, al mismo tiempo, obtener una herramienta que se pudiera aplicar a diversas tareas, como, por ejemplo, la extracción automática de candidatos a término (Koza, 2015). La modelización y la posterior implantación en máquina se focalizaron en enumeraciones de sintagmas separados por coma, dejando para futuras investigaciones las que incluyen enumerandos complejos, como ser oraciones subordinadas o aquellos separados por punto y coma.

Para esto, se recurrió al *software* libre NooJ[1], una herramienta de estados finitos desarrollada por Silberztein (2005), que cuenta con diversas utilidades para el tratamiento de lenguaje natural, según Bonino (2015), estas son:

- Gramáticas morfológicas y derivacionales (archivos .nof): modelos de flexión y derivación.
- Diccionarios (archivos .dic): listas de palabras con diversos tipos de información lingüística.
- Gramáticas productivas (archivos .nom): sistemas regulares o gráficos útiles para el tratamiento cadenas de caracteres con determinadas propiedades formales.

1 http://www.nooj4nlp.net/

- Gramáticas sintácticas (archivos .nog): sistemas regulares o gráficos útiles para el tratamiento de cadenas de caracteres formadas por dos o más unidades léxicas, generalmente, separadas por espacios en blanco.

Las gramáticas morfológicas permiten generar las variaciones de una palabra a partir de una sola entrada del diccionario. Por ejemplo, una palabra como 'médico' está listada de la siguiente manera:

```
médico,N+FLX=N4
```

Eso significa que 'médico' pertenece al grupo de nombres ('N') y tiene una flexión ('FLX') correspondiente con el modelo 4 de nombres ('N4'), especificado en la gramática morfológica:

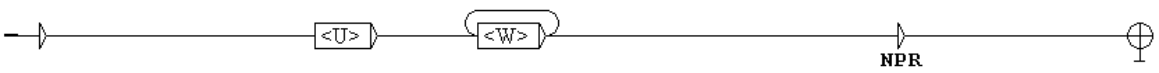
```
N4 = (<B>o/masc+sg | <B>a/fem+sg ) | <B>s/pl;
```

Este procedimiento resulta mucho más eficaz y económico, debido a que el mismo modelo permite flexionar numerosas palabras ('perro', 'abogado', 'sobrino', etcétera), en el caso de los verbos en español, que tienen una enorme riqueza flexiva; las ventajas de las gramáticas flexivas son indiscutibles (Bonino, 2015).

En esta ocasión, en el proyecto Fondecyt 11130469, se compiló en NooJ un diccionario conformado por lemas del *Diccionario Esencial de la Lengua Española* (RAE, 2006) más los nombres y adjetivos propios del dominio médico extraídos de los diccionarios *Mosby* (2005) y *Terminología médica* (Cárdenas, 2012). Se crearon modelos morfológicos para verbos, nombres y adjetivos.

Con las gramáticas productivas se pueden reconocer secuencias de caracteres específicas, mediante propiedades formales previamente declaradas. A modo de ejemplo, se puede etiquetar como nombre propio cualquier secuencia compuesta por una letra mayúscula (‹U›) seguida de un número indefinido de minúsculas (‹W›):

FIGURA 2
Gramática gráfica para el etiquetado de nombre propios



Finalmente, las gramáticas sintácticas pueden operar aisladamente o en interacción con gramáticas productivas y diccionarios que, a su vez, se integran con gramáticas flexivas, de-

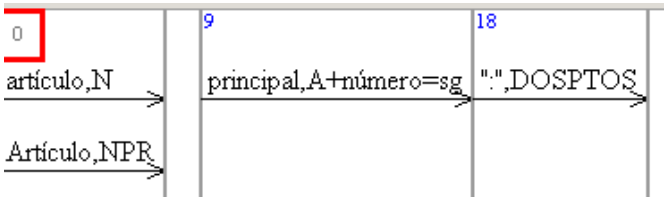
rivativas y productivas. En el primer caso, las entradas de la gramática serían palabras, por ejemplo, SD = el --> médico. Otra opción puede ser categorías declaradas en el diccionario, de este modo, SD = DET + N; esta gramática va a reconocer como sintagma determinante toda secuencia determinante y sustantivo siempre que estos hayan sido previamente reconocidos por un diccionario o una gramática productiva.

A continuación, se describe la implantación computacional realizada.

5.1. Análisis léxico y reconocimiento de los signos de puntuación

Si bien el diccionario compilado en NooJ es amplio, obviamente resulta imposible mantener diccionarios y bases terminológicas completamente actualizados de manera manual. A tales efectos, para mitigar el riesgo de palabras sin analizar, se establecieron gramáticas productivas. Una de ellas, como se mencionó más arriba, fue para etiquetar como nombre propio cualquier secuencia de letras iniciada con mayúscula, si bien, con este procedimiento, toda palabra que inicie la oración también se etiqueta como nombre propio, en caso de que tenga un etiquetado doble, el programa permite jerarquizar las gramáticas. Así por ejemplo, en el fragmento del corpus: "Artículo principal: (...)", en un etiquetado preliminar, NooJ arrojó el siguiente resultado:

FIGURA 3
Fragmento del *output* del análisis léxico de NooJ



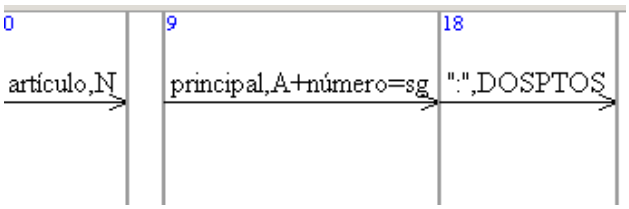
A tales efectos, se especifica en las preferencias del programa que el diccionario 'Diccionario Enumeración.nod' tenga una prioridad alta (H1), mientras que la gramática productiva de nombres propios 'NPR.nom' posea una prioridad baja (L1):

FIGURA 4
Jerarquías de los recursos de análisis léxico de NooJ

Priority	Resource
H1	Diccionario Enumeracion.nod
L1	NPR.nom

De este modo, el nuevo resultado de análisis fue el siguiente:

FIGURA 5
Nuevo *output* de NooJ



De manera similar se procedió con pistas morfológicas como, por ejemplo, toda palabra terminada en '-ción' es un nombre femenino singular, o toda palabra terminada en '-ó' es un verbo de tercera persona del pretérito perfecto simple del indicativo.

Luego del análisis léxico, se procedió a la conformación de sintagmas y de enumeraciones mediante la conformación de gramáticas sintácticas específicas.

5.2. Gramáticas sintácticas: reconocimiento de sintagmas y enumeraciones

Se establecieron reglas de reagrupamiento para el reconocimiento de sintagmas léxicos (nominales, adjetivales, verbales y preposicionales). A fin de disminuir el riesgo de ambigüedades el análisis se inició con la conformación de los sintagmas núcleos (Abney, 1994). Los sintagmas núcleos son sintagmas conformados por categorías morfosintácticas fijas, que posibilitan determinar dónde comienzan, dónde terminan, cómo están compuestos y cuál es el núcleo. A partir de determinadas propiedades de linealidad, se restringen las posibilidades combinatorias de sus elementos:

- Sintagma nominal núcleo (snn): 'los primeros reconocidos personajes'.
- Sintagma adjetival núcleo (sadjn): 'más importantes'.
- Sintagma verbal núcleo (svn): 'no los hacían'.

Abney (1994) justifica el análisis en sintagmas núcleos, tanto por razones prosódicas y psicolingüísticas como porque permiten un análisis sintáctico automático con menores dificultades. En el caso de los ejemplos, es posible determinar cuando 'los' se corresponde con un artículo neutro (snn) y cuándo, con un pronombre clítico (svn).

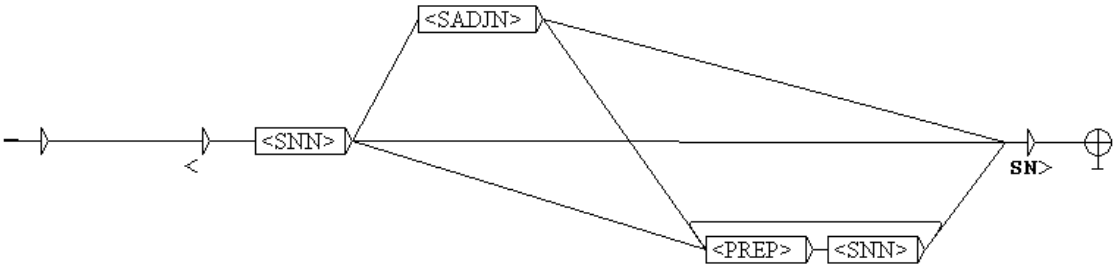
Una vez determinados los sintagmas núcleos, se procedió a la conformación de sintagmas completos. En esta ocasión no se tuvieron en cuenta las cláusulas subordinadas, como

así tampoco, en el caso de los sintagmas verbales (SV), el especificador SDET. A modo de ejemplo, se presenta la estructura del SN:

- SNN + SADJN + [(PREP + SNN)≥1] = SN
- SNN + SADJN = SN
- SNN = SN

Gráficamente:

FIGURA 6
Gramática para el reconocimiento del SN



Una vez establecidos los sintagmas, se procedió a la conformación de gramáticas para el reconocimiento de enumeraciones. De este modo, si se toma, por ejemplo, la cláusula (2) ('Juan trabaja los días lunes, martes, miércoles y sábados'), una modelización pertinente para la enumeración sería:

SN + SN + coma + SN + cop. + SN = ENUMSN

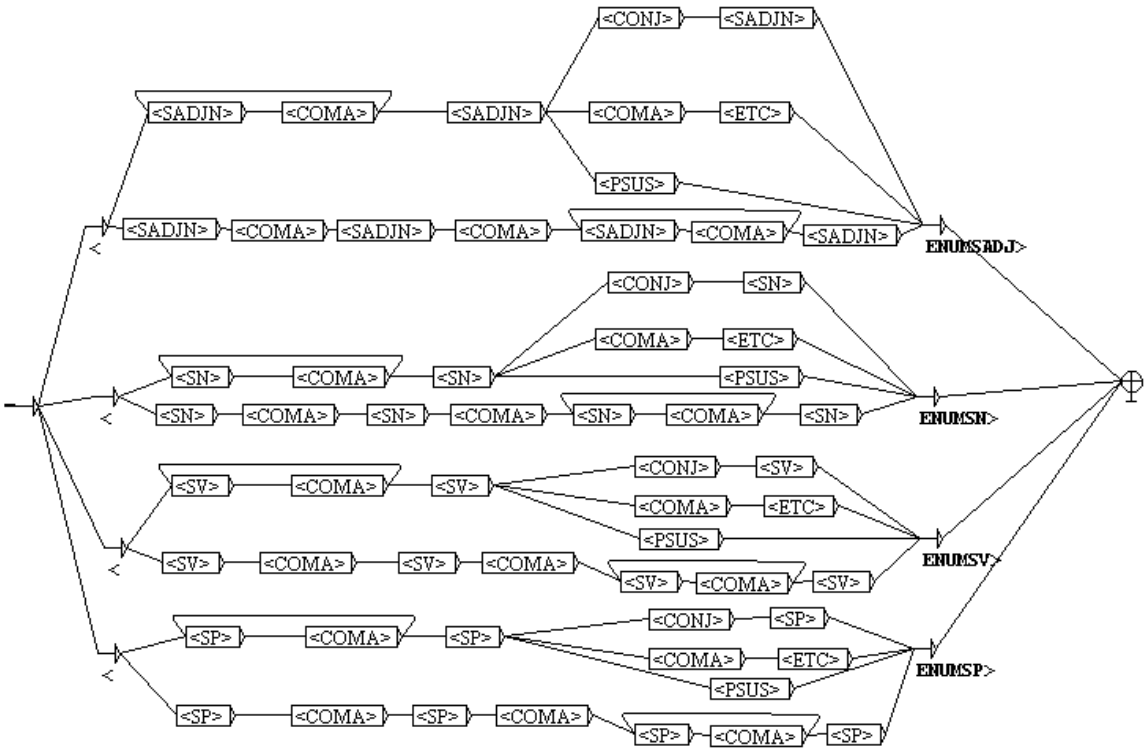
Esto quiere decir que, si en el texto se encuentra un SN seguido de coma, más otro SN seguido de una conjunción copulativa (cop), más otro SN, entonces hay una enumeración de sintagmas nominales.

De esta manera, se desarrolló un conjunto de reglas que modelan los casos posibles de enumeraciones sintagmáticas:

- (SX + coma) ≥ 1 + SX + conjunción + SX = ENUMSX
- (SX + coma) ≥ 3 + punto suspensivos = ENUMSX
- (SX + coma) ≥ 3 + 'etcétera' = ENUMSX
- (SX + coma) ≥ 3 = ENUMSX (para los casos de asíndeton)

Estas estructuras se implantaron computacionalmente mediante la siguiente gramática:

FIGURA 7
Gramática para el reconocimiento de enumeraciones



C:\Users\Water\Documents\NooJ\sp\Syntactic Analysis\enumeracion.nog
viernes, 11 de diciembre de 2015

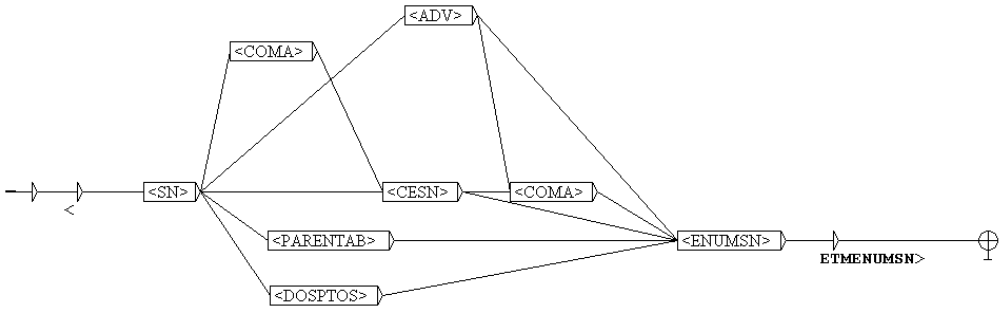
De este modo, se logró reconocer las enumeraciones sintagmáticas mencionadas. A modo de ejemplo, se presentan fragmentos del *output* generado por NooJ.

> (...) estas sustancias eran utilizadas en rituales mágicos por ‹ENUMSN›chamanes, sacerdotes, magos, brujos, animistas, espiritualistas o adivinos‹/ENUMSN› (...).
>
> (...) valor ‹ENUMSADJ›legal, educacional, informativo y científico‹/ENUMSADJ›, (...)
>
> (...) los trastornos del crecimiento ‹ENUMSP›de las células, de los tejidos y de los órganos‹/ENUMSP› (...)
>
> (...) suficientemente peligrosa como para ‹ENUMSV›retrasar, modificar o contraindicar la operación‹/ENUMSV› (...)

Finalmente, para el caso de las enumeraciones de SN, se elaboró una última gramática sintáctica para la detección de la secuencia 'Enumeratema – Enumerando – Enumeración'. Se establecieron como conectores de enumeración expresiones del tipo 'tales como', 'como por ejemplo', 'son', 'los que incluyen' y 'los cuales incluyen'. Para casos en los que no se contó con

un enumerador explícito, se incluyeron los signos de puntuación dos puntos y el paréntesis de apertura.

FIGURA 8
Gramática para la detección de la secuencia 'Enumeratema – conector – Enumeración'



De este modo, fue posible la detección de estructuras como las siguientes:

> (...) diferentes **culturas** *como* la medicina Ayurveda de la India, el antiguo Egipto, la antigua China y Grecia. Uno de los primeros recorridos (...).
>
> (...) las **especialidades quirúrgicas de la medicina**: la cirugía general, la urología, la cirugía plástica, la cirugía cardiovascular y la ortopedia entre otros. Pediatría, (...).
>
> (...) de salud, así como de las **ciencias básicas** (Física, Estadística, Historia de la Medicina, Psicología, Bioquímica, Genética...). El tercer año se dedica (...).

En la sección siguiente se presentan los resultados obtenidos.

6. Análisis de resultados

El método presentado se probó en un corpus compuesto por 64 entradas de Wikipedia relativas al dominio médico. En total, contenía 57632 palabras y, luego de una revisión manual, se reconocieron 357 enumeraciones, que se distribuían de la siguiente manera:

- Nominales: 319 (105 de ellas con enumeratema explícito)
- Adjetivales: 26
- Verbales: 7
- Preposicionales: 5

Una vez realizado el análisis automático, se extrajeron 370 expresiones etiquetadas como enumeración, presentando los siguientes resultados:

- Enumeraciones reconocidas correctamente: 322
- Expresiones etiquetadas erróneamente como enumeración: 35
- Enumeraciones no reconocidas: 42

Esto implicó 88,46% de precisión; 90,20% de exhaustividad, y 89,82% de medida F.

Específicamente, los resultados para cada tipo de enumeración se dividen de la siguiente manera:

TABLA 4
Resultados obtenidos

Enumeración	Totales en el corpus	Reconocidas	No Reconocidas	Etiquetados erróneos	Precisión	Exhaustividad	Medida F
Nominales	319	288	31	36	88,88%	90,28%	89,57%
Adjetivales	26	23	3	4	85,18%	88,46%	86,79%
Verbales	7	7	0	2	71,43%	100%	55%
Preposicionales	5	4	1	0	100%	83%	91%

Como se puede observar, el mayor número de enumeraciones se concentró en las nominales (86% del corpus). Esto se justifica en que en los SN se establece la mayoría de los conceptos de un área de conocimiento, en este caso, la medicina.

Ahora bien, los resultados obtenidos son lo suficientemente adecuados para validar la descripción propuesta. No obstante, se pudieron observar ciertas regularidades en los etiquetados erróneos. La naturaleza de estos se debió, en el plano léxico, principalmente, a ambigüedades:

> (...) formación sanitaria especializada para ‹ENUMSADJ›Médicos, Farmacéuticos y otros graduados‹/ENUMSADJ›/licenciados universitarios (...)

Aquí, la enumeración fue etiquetada como adjetiva, debido a que ‘médico’ y ‘farmacéutico’ están en el diccionario con las etiquetas nombre y adjetivo, en cambio, ‘graduado’ aparece como adjetivo y verbo (en calidad de participio).

Por otro lado, también se evidenciaron problemas con los nombres propios, por ejemplo:

> (...) personajes tales como Rudolf Virchow, Wilhelm Conrad Röntgen, Alexander Fleming, Karl Landsteiner, Otto Loewi, Joseph Lister, Francis Crick, Florence Nightingale, Maurice Wilkins,Howard Florey, Frank Macfarlane Burnet, William Williams Keen, William Coley, James D. Watson, Salvador Luria, Alexandre Yersin,Kitasato Shibasaburo, Jean-Martin Charcot, Luis Pasteur, Claude Bernard, Paul Broca, Nikolái Korotkov, William Osler y Harvey Cushing como los más importantes entre otros (...).

Como se mencionó anteriormente, se estableció una gramática productiva que reconociera como nombre propio toda secuencia iniciada por una letra mayúscula seguida de una o varias letras minúsculas. No obstante, se puede observar que el enumerando 'James D. Watson' contiene solo la inicial del segundo nombre, seguida de un punto, y 'Jean-Martin Charcot', que incluye un guion. Esto ocasionó que no fueran incluidos como SN por NooJ y que la enumeración fuera dividida en tres partes:

> personajes tales como ‹ENUMSN›Rudolf Virchow, Wilhelm Conrad Röntgen, Alexander Fleming, Karl Landsteiner, Otto Loewi, Joseph Lister, Francis Crick, Florence Nightingale, Maurice Wilkins,Howard Florey, Frank Macfarlane Burnet, William Williams Keen, William Coley, James‹/ENUMSN› D. Watson, Salvador ‹ENUMSN›Luria, Alexandre Yersin,Kitasato Shibasaburo, Jean‹/ENUMSN›-‹ENUMSN›Martin Charcot, Luis Pasteur, Claude Bernard, Paul Broca, Nikolái Korotkov, William Osler y Harvey Cushing‹/ENUMSN› como los más importantes entre otros.

Asimismo, también se puede observar un caso de ambigüedad en 'Salvador Luria', en donde 'Salvador' fue analizado como adjetivo y, por lo tanto, no incluido en la enumeración.

Por otro lado, en el plano sintáctico, se evidencian dos tipos de errores: aposiciones conformadas por un sintagma nominal seguido de otro compuesto (con dos núcleos coordinados) y casos en los que un adjunto verbal (generalmente de lugar o de tiempo) aparece dislocado a la izquierda, seguido de una coordinación de SN:

> (…) ‹ENUMSN›Dos investigadores, Roger Guillemin y Andrew Schally‹/ENUMSN› , observaron (…).
>
> (…) en ‹ENUMSN›un hospital de Edimburgo, el tocólogo James Simpson y su compañero Dunkan‹/ENUMSN› practicaron el primer parto sin dolor (…).
>
> (…) En ese ‹ENUMSN›mismo año, Esau y Schliephake‹/ENUMSN› iniciaron la radioterapia (…).

En el caso de las enumeraciones verbales, vale aclarar que hay que diferenciar la enumeración, que conforma un todo, de una secuencia de acciones del tipo 'restauramos, reparamos y volvemos a hacer esas partes'. Esto se justifica en que, desde una perspectiva sintáctica, una sucesión de acciones implicaría diferentes oraciones[2]. A tales efectos, solo se consideraron dentro de las enumeraciones verbales los verboides (participios, infinitivos y gerundios). Los dos etiquetados erróneos corresponden a una superposición de etiquetas:

> (…) no filiada mediante ‹ENUMSN›el ‹ENUMSV›legrado, cepillado y lavado‹/ENUMSV› bronquial‹/ENUMSN›, con aspiración citológica (…).
>
> (…) analizadores clínicos ‹ENUMSV›‹ENUMSADJ›automatizados, computarizados y especializados‹/ENUMSADJ› en diferentes campos analíticos ‹/ENUMSV› como hematología (…).

2 Lo que excede los propósitos del presente trabajo.

En relación con las enumeraciones preposicionales, solo se contabilizaron 5, de las cuales se lograron reconocer 4. La quinta no pudo ser analizada, dada la complejidad de uno de los SN que actuaba como complemento de la preposición:

> (…) distintas técnicas: de osteosíntesis, de traslado de tejidos mediante colgajos y trasplantes autólogos de partes del cuerpo sanas a las afectadas, etc. (…).

En lo que atañe a la secuencia 'Enumeratema -(Enumerador)- Enumerando', se lograron detectar 50 estructuras de este tipo y no hubo detecciones erróneas, lo que implica 100% de precisión y 52,50% de exhaustividad (68,65% de medida F). Si bien los resultados son, en cierta medida, positivos, este aspecto de la enumeración se pretende desarrollar en próximas etapas de trabajo.

Finalmente, en relación con las omisiones, estas se debieron, principalmente, a palabras que no pudieron ser reconocidas y ambigüedades. Por ejemplo, la enumeración: '(…) hábitos tóxicos: beber, fumar y drogas' no pudo ser reconocida, dado que 'drogas' no pudo ser desambiguada como nombre, debido a que estaba antecedida por dos verbos y no tenía complementos (determinantes, adjetivos, etcétera). No obstante, asimismo, se pudo constatar que, en varios casos, la detección de enumeraciones ha permitido la desambiguación de ciertas palabras. Tal es el caso de estructuras como "(…) dedicada al <ENUMSN>diagnóstico, cuidado preoperatorio, operación y manejo postoperatorio de los problemas</ENUMSN>(…)", en donde 'manejo' es analizado como nombre y no como verbo.

7. Conclusiones

Se presentó un estudio de la enumeración desde el punto de vista de la formalización y con el objetivo de establecer una modelización para que sea implantada en máquina, con el objeto de reconocer este tipo de construcciones en textos de lenguaje natural. Para ello, se analizaron algunas cuestiones referentes a la estructura de la propia enumeración, concebida como una sucesión de enumerandos, y a la denominada Estructura Enumerativa, conformada por una matriz, que suele contener un elemento que temáticamente permite la enumeración, como ser el enumeratema; un elemento sintáctico que la designa, el enumerador; una enumeración, y, opcionalmente, un elemento de cierre:

[matriz (enumeratema) enumerador] + [enumeración] + [(cierre)]

En relación con el enumeratema, se determinó que este, desde una perspectiva sintáctica, entabla con la enumeración una relación similar a la de dos sintagmas determinantes que conforman una aposición. Esto se entiende en la medida en que se considera la relación entre el enumeratema y los enumerandos similar a la de hiperónimo-hipónimos.

En segundo lugar, se propuso una clasificación de las enumeraciones de acuerdo con la naturaleza de sus componentes, el grado de cierre de sus elementos (enumeraciones abiertas o cerradas) y los elementos separadores. Dicha clasificación, además de dar cuenta del fenómeno de la enumeración, también resultó acorde con el trabajo de formalización con vista a una implantación en máquina, a fin de establecer un método de detección automática de enumeraciones en textos de lenguaje natural. Este se evaluó en un corpus compuesto por entradas de Wikipedia del área médica y se lograron adecuados porcentajes de precisión, exhaustividad y medida F. No obstante, se pudieron observar una serie de inconvenientes en la detección, por lo cual, se espera afinar algunas reglas, a fin de mejorar los resultados.

El presente trabajo pretende ser un aporte, tanto a los estudios gramaticales como a la lingüística informática, en la medida en que propone un análisis de la enumeración desde una nueva perspectiva como la gramática formal, mediante la propuesta de una clasificación y de la categoría de enumerador y una implementación computacional. Se espera que esta investigación propicie futuros estudios sobre el fenómeno. En lo que atañe al método computacional, se pretende que sea una herramienta de ayuda para las tareas de creación automática de ontologías, en la elaboración de gramáticas restrictivas, etcétera.

El trabajo a futuro se organiza en torno a los siguientes ejes: (i) continuar con la modelización de las enumeraciones, a fin de ampliar el método de detección automática e incluir enumeraciones de mayor complejidad; (ii) establecer un análisis de la enumeración dentro de los estudios de la interfaz sintaxis-discurso; (iii) ampliar el análisis a enumeraciones de elementos complejos, separadas por punto y coma, y (iv) establecer un paralelo con las investigaciones de Fauconnier, Kamel y Rothenburger (2013), mediante experimentos para el reconocimiento automático de enumeratema-enumeración en español.

8. Bibliografía citada

Abney, Steven, 1994: "Parsing by chunks" [http://www.vinartus.net/spa/90e.pdf, fecha de consulta: 12 de noviembre de 2015].

Berbel, José, 2008: "La serie enumerativa en los estudios de retórica y lingüística. Estado de la cuestión" en Luis Cortés (coord.): *La serie enumerativa en el discurso oral en español*, Madrid: Arco Libros, 35-74.

Bonino, Rodolfo, 2015: "Una propuesta para el tratamiento de los enclíticos en NooJ", *Infosur Revista* 7, 31-40.

Bras, Miriam, Laurent Prévot y Marianne Vergez-Couret, 2008: "Quelle(s) relation(s) de discours pour les structures énumératives?" en Jacques Durand, Benoît Habert y Bernard Laks (eds.) : *Ac-*

tes du Colloque Mondial de Linguistique Française CMLF'08, Paris, Institut de Linguistique Française, 1945-1964.

Camacho, María, 2008: "Relaciones textuales entre serie y matriz" en Luis Cortés (coord.): *La serie enumerativa en el discurso oral en español*, Madrid: Arco Libros, 127-155.

Cárdenas, Enrique, 2012, *Terminología Médica*, México D. F.: Mc Graw Hill.

Cortés, Luis, 2008: *La serie enumerativa en el discurso oral en español*, Madrid: Arco Libros.
Cortés, Luis, 2012: "La serie enumerativa en el cierre de los discursos", *Estudios filológicos* 49, 39-57.

Damamme, Béatrice, 1989: *La série enumerative*, Genève-Paris : Librairie Droz.

Diccionario Mosby (versión electrónica), 2005, Madrid: Harcourt.

Fauconnier, Jean, Kamel Mouna y Rothenburger Bernard, 2013: "Une typologie multi-dimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques" en *Conférence Internationale sur la Terminologie et l'Intelligence Artificielle (TIA 2013)*, Paris, Université Paris 13, 137-144.

Garat, Diego, 2006: *Análisis de superficie basado en puntuación*, Montevideo: PEDECIBA, Universidad de la República.

Ho-Dac, Lydia, Marie-Paule Péry Woodley y Ludovic Tanguy, 2010: "Anatomie des structures énumératives" [http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_26.pdf, fecha de consulta: 14 de octubre de 2015].

Koza, Walter, 2015: "Propuesta de extracción automática de candidatos a término del dominio medico procesando información lingüística. Descripción y evaluación de resultados", *Alfa. Revista de Lingüística* 59(1), 113-127.

Laippala, Veronika, 2010: "*0... Second... Finally...* Marking and unmarking of items in sequential text organization" en *MAD 2010* [http://w3.workshop-mad2010.univ-tlse2.fr/MAD_files/papers/Laippala.pdf, fecha de consulta: 15 de octubre de 2015]

López Samaniego, Ana, 2006: "Los ordenadores del discurso enumerativo en la sentencia judicial: ¿Estrategia u obstáculo?", *Revista de Llengua i Dret* 45, 61-87.

Luc, Christophe, 2001: "Une typologie des structures énumératives basée sur les structures rhétoriques et archictecturales du texte" en *Actes de TALN' 2001*, Tours, 263-272.

Marchese, Angelo y Joaquín Forradellas, 1986: *Diccionario de Retórica, Crítica y Terminología Literaria*, Barcelona: Ariel.

Melis, Chantal y Marcela Flores, 2007: "Los verbos pseudo-impersonales del español. Una caracterización sintáctica", *Verba* 34, 7-57.

Moro, Andrea, 1997: "Dynamic Antisymmetry: Movement as a symmetry breaking phenomenon", *Studia Lingüística* 51, 50-76.

Muñoz, Carlos, 2012: "Sobre la estructura sintagmática de la aposición explicativa", *Boletín de Filología* 47(2), 133-148.

Porhiel, Sylvie, 2007: "Les structures énumératives à deux temps", *Revue Romane* 42, 103-135.

Real Academia Española, 2006: *Diccionario esencial de la lengua española*, Bogotá: Espasa.
Real Academia Española, 2009: *Nueva gramática de la lengua española*, Buenos Aires: Asociación de Academias de la Lengua Española.

Silberztein, Max, 2005, NooJ: "A Linguistic Annotation System For Corpus Processing" [http://www.aclweb.org/anthology/H05-2, fecha de consulta: 20 de diciembre de 2015].

Vergez-Couret, Marianne, Myriam Bras, Laurent Prevot, Laure Vieu y Caroline Attalah, 2011: "Discourse contribution of Enumerative structures involving pour deux raisons" en *Proceedings of Constraints in Discourse 2011*, Agay-Roches Rouges [http://www.irit.fr/publis/LILAC/VCBPVA-pour2raisons-CID11.pdf, fecha de consulta: 19 de mayo de 2015].