

Extracción del Sintagma Nominal

Noun Phrase Extraction

Solana Zulema, Rodrigo Andrea, Méndez Bárbara

Grupo Infosur
Universidad Nacional de Rosario
Rosario, Argentina

Abstract

The purpose of this paper is to extract noun phrases (NPs) by focusing on the nucleus noun phrase as the starting point (NNP). In this first stage the most frequent structures of the noun phrase will be described but narrowing the subject to NPs subject, to the left of the verb and starting the sentence. For the automatic analysis, the SMORPH and MPS tools will be used. SMORPH enables the morphological analysis of chain of characters, allowing the categorial assignment corresponding to each word according to the features declared; and MPS has the SMORPH output as its input. Based on the established recomposition rules, the structures to analyze -in this case the NPs- will be extracted.

Keywords: Noun phrases (NPs), description, automatic analysis, extraction

Resumen

El objetivo que nos proponemos en este trabajo es la extracción de sintagmas nominales (SN's) partiendo del sintagma nominal núcleo (snn). Cabe aclarar que en esta primera etapa, se describen las estructuras más frecuentes del SN pero circunscribiendo el tema a los SN's sujetos que están a la izquierda del verbo y que inician la oración. Respecto del análisis automático, se recurre a las herramientas informáticas SMORPH y MPS. SMORPH permite analizar morfológicamente las cadenas de caracteres, dando como salida la asignación categorial y morfológica correspondiente a cada palabra, de acuerdo con los rasgos que se declaran; y MPS, tiene como input la salida de Smorph. A partir de reglas de recomposición establecidas, se extraen las estructuras a considerar, en este caso, los SN's.

Palabras claves: SN's, descripción, análisis automático, extracción

1. INTRODUCCIÓN

En el Grupo Infosur, en el camino abierto por el GRIL (Groupe de Recherche dans les Industries de la Langue) de la Universidad Blais Pascal de Clermont Ferrand, se han investigado los sintagmas núcleos, sintagmas que comienzan en el inicio de la construcción y finalizan en el núcleo. Dichos sintagmas integran una unidad mayor que los contiene. Así:

- 1- El sintagma nominal núcleo integra al Sintagma Nominal: [(*el lluvioso día*)₁ *de invierno*]₂ 1: sintagma núcleo 2: SN;
- 2- El sintagma adjetivo núcleo integra el Sintagma Adjetivo: [(*demasiado bueno*)₁ *para ser verdad*]₂ 1: sintagma núcleo 2: SAdj;
- 3- El sintagma adverbial núcleo integra el Sintagma Adverbial: [(*poco después*)₁ *de las diez*]₂ 1: sintagma núcleo 2: SAdv;
- 4- El sintagma verbal núcleo integra el Sintagma Verbal: [(*defendieron*)₁ *a sus hijos*]₂ 1: sintagma núcleo 2: SV

El objetivo que nos proponemos ahora es lograr la extracción de sintagmas nominales (SN's) para lo que se parte del sintagma nominal núcleo (snn) y se recurre a las herramientas informáticas SMORPH y MPS.

SMORPH desarrollado por Aït-Mokhtar [1] permite analizar morfológicamente las cadenas de caracteres, dando como salida la asignación categorial y morfológica correspondiente a cada palabra, de acuerdo con los rasgos que se declaran.

MPS, implantado por Faiza Abbaci [2], tiene como input la salida de Smorph. A partir de reglas de recomposición establecidas, se extraen las estructuras a considerar, en este caso, los SN's.

Se trabaja a partir de un corpus de 100.000 palabras de periódicos argentinos que forma parte del banco de datos del grupo Infosur y se estudian las estructuras más frecuentes del Sintagma Nominal. Cabe agregar que esta investigación retoma algunas cuestiones desarrolladas en el artículo acerca del snn de Solana y Rodrigo [3], y en la tesis de maestría de Andrea Rodrigo [4], fundamentalmente las reglas de reagrupamiento de los snn's.

2. LAS DIFICULTADES

Partimos de la siguiente observación:

- A continuación de cada snn puede haber otro sintagma núcleo que integra o no con él un sintagma nominal (SN)

Por consiguiente, se dan dos casos:

- a) Que el sintagma núcleo que sigue al snn integre con él un SN.
- b) Que el sintagma núcleo que sigue al snn se refiera al verbo o a otro elemento anterior de la oración.

2.1. Caso A: El sintagma núcleo siguiente pertenece al snn inmediatamente anterior y se refiere a su núcleo.

2.1.1 Sintagmas adjetivos núcleo (sadjn's) pospuestos al snn:

En los textos analizados, siempre que aparece un sadjn después de un snn (en posición de sujeto) integra con él un SN.

Ej: *(el hermoso día)*¹ *(soleado)*²

Como puede verse en el ejemplo, la totalidad forma un SN cuyo núcleo es '*día*'.

2.1.2 Sintagma Preposicional (SP's) después de un snn:

a) Un SP iniciado por '*de*', que sigue a un snn, puede integrar con él un SN.

Ej.: *(el hermoso día)*¹ *(de verano)*²

Como en el caso anterior, la totalidad forma un SN cuyo núcleo es '*día*'.

b) Puede haber más de dos SP's encadenados, integrantes del mismo SN.

Ej: *(las páginas)*¹ *(del libro)*² *(de historia)*³

En 1 está el núcleo al que se refiere 2 y en 2 está el núcleo al que se refiere 3.

c) Puede haber más de dos sintagmas núcleos NO encadenados, pero integrantes del mismo SN.

Ej.: *(festival)*¹ *(iberoamericano)*² *(de teatro)*³ *(de Bogotá)*⁴ ¹

Aquí 2 (sadjn) se refiere a 1, 3 también a 1 y 4 también a 1.

2.2. Caso B: El sintagma núcleo siguiente NO pertenece al snn inmediatamente anterior

a) Se trata de un SP

- *Se desplegó (un gran operativo)*¹ *(ante posibles choques)*² *(entre simpatizantes y opositores)*³

Aquí 2 no se refiere a 1, sino que depende del verbo y 3 se refiere a 2.

b) Se trata de un sadjn

- *Juan come (la carne)*¹ *(muy cocida)*²

En este caso, el sadjn 2 se refiere al verbo, porque es un predicativo objetivo (cf. Bosque y Demonte cap.38). Para trabajar este tema en una primera etapa, nos vamos a limitar a los SN's sujetos de la

¹ Diario Clarín, 13/04/04

oración, que están a la izquierda del verbo y que inician la oración, es decir, nos enfrentaremos al caso A. Vamos a identificar los SN's que estén al comienzo de oración, ya que nos valdremos previamente de la separación en oraciones, para esto nos apoyamos en la tesis doctoral de Celina Beltrán [5].

3. METODOLOGÍA

Para determinar el comienzo y el final de los SN's se parte de los snn's, es decir, los pasos son:

- 1) Pasar el texto a analizar por SMORPH.
- 2) Aplicar las reglas de formación de snn.
- 3) Aplicar las reglas de formación de SN.

Se formularán reglas para las siguientes estructuras:

Snom1:

- snn + sadjn

Ej.: *un clima tenso*

Snom2:

- snn + sp1 (iniciado por la preposición 'de' o como variante 'del')

Ej.: *un gran operativo de seguridad*

- snom1 + sp1

Ej.: *un operativo estricto de seguridad*

Snom3:

- snn + sp2 (iniciado por una preposición distinta de 'de')

Ej.: *la casa en la montaña*

- snom1 + sp2

Ej.: *la casa blanca en la montaña*

Snom4:

- snn + snn

Ej.: *mi hermano Pedro*

A continuación, se presenta la Tabla 1 que muestra los tipos de SN's trabajados. Las reglas de formación se exponen en los puntos 3.1. a 3.3.

Tabla 1: Tipos de SN's

Tipo de SN	Estructura	Ejemplos
snom1	snn + sadjn	<i>un clima tenso; los hermanos muy unidos</i>
snom2	snn + sp1 snom1 + sp1 snom1 + sp1	<i>un gran operativo de seguridad; un gran operativo estricto de seguridad; la arista más conflictiva de esta reforma presupuestaria;</i>
snom3	snn + sp2 snom1 + sp2	<i>la casa en la montaña la casa blanca en la montaña</i>
snom4	snn + snn	<i>mi hermano Pedro</i>

3.1. Regla de formación de snom1

Se llamó snom1 al que concatena un snn y un sadjn.

S1 [L1, 'EMS', 'snn'] S2 [L2, 'EMS', 'sadjn'] --> S1+S2[L1+L2, 'EMS', 'snom1'].²

A partir de:

'un clima'.

['un clima', 'EMS', 'snn'].

'tenso'.

['tenso', 'EMS', 'sadjn'].

² Según la sintaxis que requiere MPS: S1, S2, son variables sobre las ocurrencias.

L1, L2, son variables sobre los lemas.

EMS etiqueta morfosintáctica, allí aparece 'snn', sintagma nominal núcleo y 'sadjn', sintagma adjetivo núcleo, finalmente 'snom1', Sintagma Nominal 1.+ indica la operación de concatenación.

Se tendrá

'un clima tenso'.

['un clima tenso', 'EMS', 'snom1'].

A partir de

'los hermanos'.

['los hermanos', 'EMS', 'snn'].

'muy unidos'.

['muy unidos', 'EMS', 'sadjn'].

Se tendrá:

'los hermanos muy unidos'.

['los hermanos muy unidos', 'EMS', 'snom1'].

3.2. Reglas de formación de snom2 y snom3

Para realizar las reglas de snom2 y snom3, es preciso formular antes las reglas de formación de sp1 (sintagma preposicional iniciado por 'de' o 'del') y de sp2 (iniciados por otras preposiciones).

3.2.1. Reglas de formación de Sp

Tanto para los Sp1 como para los Sp2, se trabaja con dos estructuras:

- a) Sp formados con prep y snn.
- b) Sp formados con prep y snom1.

3.2.1.1. Regla de formación de Sp1

Se distinguirá entre los SP iniciados por 'de' y seguidos por snn; y los sp iniciados por 'de' y seguidos por snom1.

$S1 [L1, 'TPREP', 'prep1'] S2 [L2, 'EMS', 'snn'] \rightarrow S1+S2[L1+L2, 'EMS', 'sp1a']$.

de seguridad

$S1[L1, 'TPREP', 'prep1'] S2 [L2, 'EMS', 'snom1'] \rightarrow S1+S2[L1+L2, 'EMS', 'sp1b']$.

de un clima tenso

A partir de:

'de'.

['de', 'TPREP', 'prep1'].

'seguridad'.

['seguridad', 'EMS', 'snn'].

Se tendrá

'de seguridad'.

['de seguridad', 'EMS', 'sp1a'].

A partir de:

'del'.

['del', 'TPREP', 'prep1'].

'país'.

['país', 'EMS', 'snn'].

Se tendrá

'del país'.

['del país', 'EMS', 'sp1a'].

A partir de:

'de'.

['de', 'TPREP', 'prep1'].

'un clima tenso'.

['un clima tenso', 'EMS', 'snom1'].

Se tendrá:

'de un clima tenso'.

['de un clima tenso', 'EMS', 'sp1b'].

3.2.1.2. Reglas de formación de Sp2.

Como en el caso anterior, se distinguirá dos casos se sp2, el primero conformado por otras preposiciones distinta de 'de' y un snn; el segundo con otras preposiciones distinta de 'de' y un snom1.

S1 [L1, 'TPREP', 'prep2'] S2 [L2, 'EMS', 'snn'] --> S1+S2 [L1+L2, 'EMS', 'sp2a'].

S1 [L1, 'TPREP', 'prep2'] S2 [L2, 'EMS', 'snom1'] --> S1+S2 [L1+L2, 'EMS', 'sp2b'].

A partir de

'por'.

['por', 'TPREP', 'prep2'].

'su presunta responsabilidad'.

['su presunta responsabilidad', 'EMS', 'snn'].

Se tendrá

'por su presunta responsabilidad'.

['por su presunta responsabilidad', 'EMS', 'sp2a'].

A partir de:

'En'.

['en', 'TPREP', 'prep2'].

'un clima tenso'.

['un clima tenso', 'EMS', 'snom1'].

Se tendrá

'En un clima tenso'.

['en un clima tenso', 'EMS', 'sp2b'].

3.2.2. Reglas de formación de snom2

El Snom2 puede tener las siguientes estructuras:

snn + sp1 (introducido por 'de')

snom1 + sp1 (introducido por 'de')

S1 [L1, 'EMS', 'snn'] S2 [L2, 'EMS', 'sp1'] --> S1+S2[L1+L2, 'EMS', 'snom2'].

S1 [L1, 'EMS', 'snom1'] S2 [L2, 'EMS', 'sp1'] --> S1+S2[L1+L2, 'EMS', 'snom2'].

A partir de

'un gran operativo'.

['un gran operativo', 'EMS', 'snn'].

'de seguridad'.

['de seguridad', 'EMS', 'sp1'].

Se tendrá:

'un gran operativo de seguridad'.

['un gran operativo de seguridad', 'EMS', 'snom2'].

A partir de

'un operativo estricto'.

['un operativo estricto', 'EMS', 'snom1'].

'de seguridad'.

['de seguridad', 'EMS', 'sp1'].

Se tendrá

'un operativo estricto de seguridad'.

['un operativo estricto de seguridad', 'EMS', 'snom2'].

3.2.3. Reglas de formación de snom3

Aquí distinguimos:

snn + sp2 (introducido por una preposición distinta de 'de')

snom1+ sp2 (introducido por una preposición distinta de 'de')

Por tanto hay dos reglas de formación:

S1 [L1, 'EMS', 'snn'] S2 [L2, 'EMS', 'sp2'] --> S1+S2[L1+L2, 'EMS', 'snom3'].

S1 [L1, 'EMS', 'snom1'] S2 [L2, 'EMS', 'sp2'] --> S1+S2[L1+L2, 'EMS', 'snom3'].

A partir de:

'la casa'.

['la casa', 'EMS', 'snn'].

['en la montaña', 'EMS', 'sp2'].

Se tendrá

['la casa en la montaña'].

['la casa en la montaña', 'EMS', 'snom3'].

A partir de:

['la casa blanca', 'EMS', 'snom1'].

['en la montaña', 'EMS', 'sp2'].

Se tendrá

['la casa blanca en la montaña'].

['la casa blanca en la montaña', 'EMS', 'snom3'].

3.3. Reglas de formación de snom4:

Se llamó snom4 al sintagma nominal que concatena un snn y un snn.

S1 [L1, 'EMS', 'snn'] S2 [L2, 'EMS', 'snn'] --> S1+S2[L1+L2, 'EMS', 'snom4'].

A partir de:

'mi hermano'.

['mi hermano', 'EMS', 'snn'].

'Pedro'.

['Pedro', 'EMS', 'snn'].

Se tendrá

'mi hermano Pedro'.

['mi hermano Pedro', 'EMS', 'snom4'].

4. CONCLUSIONES Y PROYECCIONES

En este trabajo hemos logrado un primer paso en la extracción de sintagmas nominales partiendo del sintagma nominal núcleo. Por el momento nos hemos limitado a los SN's sujetos de la oración, que están a la izquierda del verbo y que inician la oración, esté el snn seguido por sintagma adjetivo, un sintagma preposicional o dos o más sintagmas preposicionales encadenados entre sí o dependientes del núcleo del snn.

Hemos recurrido a las herramientas informáticas SMORPH y MPS y la extracción se ha realizado mediante reglas de recomposición. También han sido clasificados los SN's según su estructura.

Referencias

- [1] Aït-Mokthar S. L'analyse présyntaxique en une seule etape. Tesis doctoral dirigida por GabrielG. Bès en el GRIL. Université Blaise-Pascal. Francia, 1998.
- [2] Abbaci F. Développement du Module Post-Smorph. Memoria del DEA de Linguistique et Informatique. Universidad Blaise-Pascal/GRIL, Clermont-Fd, 1999.
- [3] Solana, Zulema y Rodrigo, Andrea. El sintagma nominal núcleo. en Desarrollo, implementación y uso de modelos para el procesamiento automático de textos (ed. Víctor Castel) Facultad de Filosofía y Letras, UNCuyo 2005.
- [4] Rodrigo, A. Análisis automático de textos, el sintagma nominal núcleo. Tesis de Maestría, Escuela de Posgrado, Facultad de Humanidades y Artes, UNR, 2006.
- [5] Beltrán, C. Modelización Lingüística e Información Estadística en el Análisis Automático de Textos. Tesis Doctoral. Facultad de Humanidades y Artes, UNR, 2008.