Una propuesta para la extracción automática del sintagma adverbial

A proposal for the automatic extraction of the adverbial syntagm

Andrea Rodrigo
Grupo INFOSUR, U.N.R.
Rosario, Argentina
andreafrodrigo@yahoo.com.ar

Rodolfo Bonino
Grupo INFOSUR, U.N.R.
Rosario, Argentina
rodolfobonino@yahoo.com.ar

Abstract

The aim of this work is to postulate a linguistic formalism for the adverbial syntagm, understood it as a unit constituted by a head adverbial syntagm (SADVN) and a complement- to be implemented into Nooj [2]. In order to do so, the differences between the head adverbial syntagm (SADVN) and the adverbial syntagm (SADV) are stated and some SADVN which frequently occur in the formation of SADV are taken into consideration by observing authentic texts. Thus, we can find sequences such as ***más abajo*** *de la mesa* ***(further under*** *the table)*, ***más lejos*** *que lo esperable* ***(farther away*** *than expected),* etc., where the SADVN takes an SPN (head prepositional syntagm) as a complement, a gerund construction or a comparative complement which, in turn, also contain an SNN or another SPN as complements.

Previous works are taken as a point of departure, in which the description of the SADVN was formalized according to the properties of the 5P Paradigm (Bès, G.G. 1999) and implanted into other software tools. NooJ is a computational device developed by **Max Silberztein (2002)**, used for the formalization of linguistic phenomena and text analysis of natural languages. Given that it only makes concatenating operations, Nooj does not allow to implant 5P's requirement and exclusion properties but is highly efficient to operate with linearity properties, thus making it feasible to corroborate the linguistic hypothesis on the SADV previously stated with other computational formalism.

Key words: computational linguistics, Spanish, adverbs, head syntagm, adverbial syntagm, Nooj application into Spanish

Resumen

En este trabajo, se trata de postular un formalismo lingüístico que permita que el sintagma adverbial (SADV), entendido como una unidad constituida por SADVN (sintagma adverbial núcleo) y un complemento, pueda ser implantado en NooJ [2]. Para ello, se diferencian las nociones de sintagma adverbial núcleo (SADVN) y sintagma adverbial (SADV) y se consideran algunas de las

combinaciones de SADVN que intervienen más frecuentemente en la conformación de los SADV, según la observación de textos reales. Así se pueden encontrar secuencias como: ***más abajo*** *de la mesa,* ***más lejos*** *que lo esperable,* donde se ve que el SADVN se complementa con un SPN (sintagma preposicional núcleo) o un complemento comparativo que, a su vez, contienen un SNN u otro SPN como complementos.

Se toman como punto de partida trabajos anteriores, donde la descripción del SADVN se formalizó según las propiedades del Paradigma 5P (Bès, G.G. 1999) [4] y se implantó en otras herramientas informáticas. NooJ es un dispositivo computacional desarrollado por Max Silberztein (2002), que se utiliza para la formalización de fenómenos lingüísticos y el análisis de textos en lenguas naturales. Dado que solo efectúa la operación de concatenación, no permite implantar las propiedades de exigencia y de exclusión de 5P; pero tiene gran eficacia para operar con propiedades de linealidad, por ello hace factible corroborar con otro formalismo computacional las hipótesis lingüísticas planteadas previamente en torno al SADV.

Palabras claves: linguistica computacional, español, adverbios, sintagma núcleo, sintagma adverbial, aplicación Nooj en español

1. INTRODUCCIÓN

Teniendo en cuenta la noción de sintagma núcleo[1], sintagmas que comienzan en el inicio de la construcción y finalizan en el núcleo; en este trabajo, se trata de postular un formalismo lingüístico que permita que el sintagma adverbial (SADV), entendido como una unidad constituida por SADVN (sintagma adverbial núcleo) y un complemento, pueda ser implantado en NooJ. Para ello, se diferencian las nociones de sintagma adverbial núcleo (SADVN) y sintagma adverbial (SADV) y se consideran algunas de las combinaciones de SADVN que intervienen más frecuentemente en la conformación de los SADV, según la observación de textos reales. Se pretende abordar las siguientes combinaciones[2]:

- **SADVA** [SADVN + SPN (SNN)]: Ej: *más abajo de la mesa, arriba de la cama, bien enfrente de tu casa*, *poco después de las diez*[3]*, más allá del gobierno*.[4]
- **SADVB** [En una comparativa: SADVN + complemento comparativo (CONJ+ SNN) (CONJ+SNN)]: Ej: *más lejos que lo esperable, más aquí que allá*

Se toman como punto de partida trabajos anteriores, donde la descripción del SADVN se formalizó según las propiedades del Paradigma 5P (Bès, G.G. 1999) [4] y se implantó en otras herramientas informáticas. Aquí utilizamos NooJ, que es un dispositivo computacional desarrollado por Max Silberztein (2002) para la formalización de fenómenos lingüísticos y el análisis de textos en lenguas naturales. Dado que solo efectúa la operación de concatenación, no permite implantar las propiedades de exigencia y de exclusión de 5P; pero tiene gran eficacia para operar con propiedades de linealidad, por ello hace factible corroborar con otro formalismo computacional las hipótesis lingüísticas planteadas previamente en torno al SADV.

[1] Según la investigación que lleva a cabo el Grupo Infosur, con los lineamientos trazados por el GRIL (Groupe de Recherche dans les Industries de la Langue) de la Universidad Blaise Pascal de Clermont Ferrand.

[2] Se da cuenta de las estructuras más frecuentes, según el banco de datos del Grupo Infosur que dispone de un corpus de 100.000 palabras de periódicos argentinos.

[3] Ejemplo citado en Rev. Infosur N° 5, Extracción del Sintagma Nominal, (Solana, Rodrigo, Méndez), p. 18

[4] Pag. 12, 4/01/04

Nuestro objetivo es crear diccionarios y gramáticas que, aplicados a secuencias del lenguaje natural, posibiliten la extracción de sintagmas adverbiales (SADV), a partir de reglas establecidas. Para lograrlo se crean diccionarios donde las entradas léxicas se asocian con rasgos que permiten calcular su comportamiento sintáctico. El rasgo categorial (adverbio, nombre, determinante, preposición) en muchos casos debe ser complementado con otros que identifican propiedades específicas de subconjuntos de elementos integrantes de la categoría; de modo que en los diccionarios que elaboramos las palabras no se definen por su contenido semántico sino por los conjuntos (de uno o más elementos) de rasgos que se asocian entre sí y caracterizan la proyección sintáctica de la entrada léxica.

2. DESCRIPCIÓN DEL SADV

El SADV (sintagma adverbial) está formado por los siguientes sintagmas núcleos:

a. El SADVN (sintagma adverbial núcleo)

Está conformado por adverbios, lo que son clasificados por sus combinaciones, según Rodrigo (2011) [1].

ADV

ADV1 Ej: *alrededor, actualmente*

ADV 2

ADV2a

ADV2am[5] Ej: *aproximadamente,*

ADV2a0 Ej: *abajo, lejos*

ADV2b

ADV2bm Ej: *absolutamente,*

ADV2b0 Ej: *casi, menos, muy*

ADV2c

ADV2cm Ej: *admirablemente, inmediatamente,*

ADV2c0 Ej: *más*

Los adverbios pueden aparecer en grupos de dos o bien, solos. Entre las combinaciones más frecuentes de adverbios se observa:

[ADV2c0 + ADV2a0] Ej: *más abajo*, *bien enfrente, más allá*

[ADV2b0 + ADV2c0] Ej: *muy bien*

[cuant + ADV2a0] Ej: *poco después*

A su vez, en algunos SADV, el SADVN está seguido por un complemento comparativo.

[5] Se llaman 2am y 2a0 en lugar de 2a1 y 2a2 porque no era posible "repetir el rasgo 1 y 2" según la sintaxis que requiere NooJ. Lo mismo en 2b y 2c.

b. El SPN (sintagma preposicional núcleo)

Está conformado por una preposición núcleo que se ubica al principio del sintagma[6], seguido de un SNN (sintagma nominal núcleo) en su interior.

(2) [*de* (*las diez*) SNN] SPN

El SNN está constituido por[7]:

determinante + nombre. Ej:

(3) *las* ***diez***

En la categoría determinante, se incluye: artículo, posesivo, según Solana/Rodrigo (2005) [6] y Rodrigo (2006) [3].

También puede observarse que el SPN incluye al SADVN en su interior,

(4) [*en* (***adelante***) SADVN] SPN

Es factible que el SPN se integre en una unidad mayor, el SP (sintagma preposicional).

3. LA HERRAMIENTA INFORMÁTICA

NooJ [2] es una herramienta informática para el tratamiento de las lenguas naturales desarrollada por Max Silberztein a partir del año 2002; analiza textos digitalizados mediante la aplicación de diccionarios y gramáticas creadas previamente; es de libre acceso y, actualmente, es utilizado por investigadores de varias universidades del mundo para la modelización de diversas lenguas. Sus usuarios intercambian conocimientos a través de un foro de Internet y realizan congresos anuales. El autor colabora activamente con los proyectos que utilizan el programa, asesorando a los investigadores y efectuando las modificaciones necesarias para la resolución de problemas específicos de cada investigación.

El procedimiento de implantación requiere la creación de varios archivos:

a) **Definición de propiedades** (.def): se declaran los rasgos que se utilizarán para etiquetar las entradas de los diccionarios. Estos rasgos pueden ser utilizados por separado o en forma conjunta en las gramáticas sintácticas. Por ejemplo: si un sintagma requiere la presencia de cualquier adverbio solo se utilizará el rasgo categorial ADV, que incluye a todos los adverbios (ADV1, ADV2am, ADV2a0, ADV2bm, ADV2b0, ADV2cm y ADV2C0), en cambio si requiere la presencia de cualquier adverbio en *–mente* al rasgo categorial se adicionará el rasgo +m, que incluye a los ADV2am, ADV2bm y ADV2bm.

b) **Gramáticas morfológicas** (.nof): se utilizan para obtener automáticamente las variaciones morfológicas flexivas o derivacionales de cada entrada léxica, de modo que en los diccionarios NooJ se declara el lema y el modelo flexivo o derivacional que le corresponde y el sistema genera automáticamente todas las variaciones que se indican en las gramáticas morfológicas que el diccionario utiliza. El tratamiento de la morfología es muy eficiente, porque no solo efectúa las

[6] El SPN es el único sintagma núcleo que comienza con el núcleo, a diferencia de los demás, que terminan en el núcleo.
[7] Nos referimos a los snn´s más frecuentes según la observación de textos reales.

operaciones de sustracción y de concatenación al final de una cadena, sino también las de sustracción, cambio y duplicación en lugares que pueden ser determinados por el usuario (por ejemplo, al final de palabra, al principio de palabra, dos caracteres a la izquierda, tres caracteres a la derecha, etc.). En trabajos previos, [10] se propone una morfología flexional del verbo en español.

c) **Diccionarios** (.dic): en los diccionarios se declaran las gramáticas morfológicas que se van a utilizar y las entradas léxicas con la etiqueta categorial, el modelo de flexión o derivación que se le aplica y los demás rasgos sintácticos y semánticos que la caracterizan. Los diccionarios se compilan y el sistema genera automáticamente un nuevo diccionario .nod, con todas las variaciones morfológicas de cada entrada, que es el que utiliza el analizador.

d) **Gramáticas sintácticas** (.nog): se declaran las reglas sintácticas que se aplican en la formación de sintagmas. Tanto las gramáticas morfológicas como las sintácticas se pueden declarar en forma de reglas (rule editor) como en forma de gráficos (graphical editor) y son recursivas en tanto permiten utilizar en la definición el elemento definido. Por ejemplo: el sintagma preposicional núcleo se puede definir como: **SPN = P + :SN** y el sintagma nominal que aparece en el SPN como **SN = :SNN + :SPN;** como se ve, en la definición de SN (sintagma nominal) se utiliza la definición de SPN, que lo contiene; esto permite analizar secuencias como *en la casa de su hermano* [$_{SPN}$en [$_{SNN}$ la casa [$_{SPN}$de [$_{SNN}$su hermano]]]].

e) **Textos**: (.not) se cargan o se importan los textos que se pretende analizar.

Los pasos para efectuar el análisis son los siguientes:

a) Se abre el texto que se quiere analizar: *File → Open → Text* .

b) Se compilan los diccionarios que se van a utilizar en *Lab → Dictionary*. Este proceso se debe aplicar incluso en los diccionarios de categorías invariables porque, como se explicó más arriba, el analizador utiliza el diccionario .nod generado automáticamente por la compliación del archivo.dic

c) En Info → Preferences se visualizan tres pestañas (*General, Lexical Analysis y Syntactic Analysis*). En *General* se selecciona el idioma; en *Lexical Analysis* los diccionarios y en *Syntaxis Analysis* las gramáticas sintácticas. Las gramáticas morfológicas se aplican indirectamente a través de los diccionarios.

d) En *TEXT* se selecciona *Linguistic Analysis*, que aplica todos los diccionarios y gramáticas sintácticas seleccionadas. El análisis completo se puede ver seleccionando *Show Text Annotation Structure*; si se pretende encontrar una secuencia específica de palabras o etiquetas de palabras, o el resultado de una sola gramática; se debe seleccionar *TEXT → Locate* que da dos alternativas: *a NooJ regular expression*, que permite buscar palabras o secuencias de palabras (*abajo, más abajo de la mesa*) o etiquetas categoriales (<ADV>, <SADV>) y *a NooJ grammar*, que selecciona las secuencias lingüísticas que genera una gramática determinada. En ambos casos, los resultados de *Locate* se obtienen haciendo clic sobre las etiquetas coloreadas que se encuentran en la parte inferior derecha de la ventana. Los diferentes colores se utilizan únicamente para seleccionar el color de la fuente de salida.

4. IMPLANTACIÓN DE NOOJ PARA LA EXTRACCIÓN DEL SINTAGMA ADVERBIAL

Primero es preciso que Nooj reconozca cada uno de los sintagmas núcleos y las categorías que los integran. Por tanto en el archivo correspondiente a diccionario se declaran las etiquetas según se ve a continuación:

a. Diccionario de adverbios, (siempre se adosa un fragmento de cada diccionario)

```
tampoco,ADV+1
todavía,ADV+1
abiertamente,ADV+2+a+m
abnegadamente,ADV+2+a+m
abruptamente,ADV+2+a+m
```

b. Diccionario de cuantificadores

```
poco,CUANT
bastante,CUANT
un poco,CUANT
demasiado,CUANT
nada,CUANT
medio,CUANT
algo,CUANT
```

c. Diccionario de preposiciones

```
a,PREP
ante,PREP
bajo,PREP
con,PREP
contra,PREP
de,PREP
desde,PREP
en,PREP
entre,PREP
hasta,PREP
hacia,PREP
para,PREP
por,PREP
según,PREP
sin,PREP
sobre,PREP
tras,PREP
al,CONTR
del,CONTR
```

d. Diccionario de nombres

```
abrigo,N+FLX=ABRIGO
gobierno,N+FLX=ABRIGO
mesa,N+FLX=MESA
casa,N+FLX=MESA
cena,N+FLX=MESA
cama,N+FLX=MESA
altura,N+FLX=MESA
```

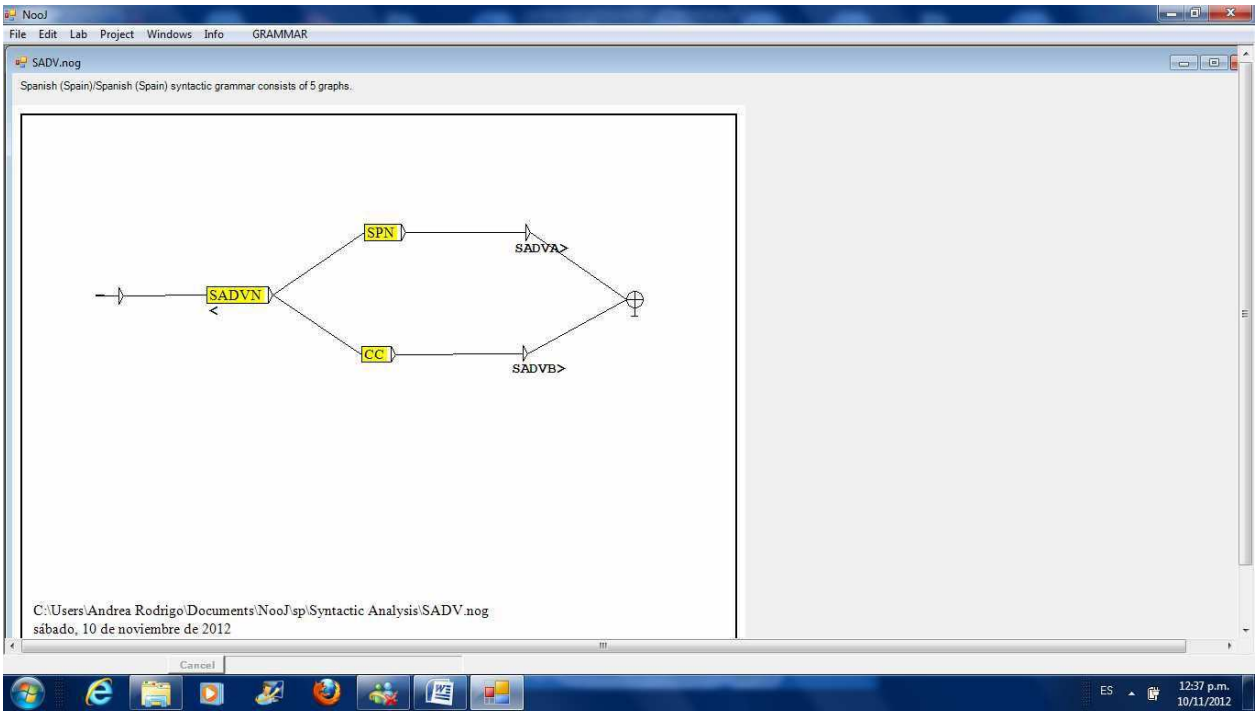
esquina,N+FLX=MESA
accionista,N+FLX=ACCIONISTA
niño,N+FLX=NIÑO
hermano,N+FLX=NIÑO
acumulador,N+FLX=ACUMULADOR
edad,N+FLX=EDAD
vez,N+FLX=VEZ

e. Diccionario de conjunciones

que,CONJ+subord
si,CONJ+subord
porque,CONJ+subord
conque,CONJ+subord
y,CONJ+coord
ni,CONJ+coord
pero,CONJ+coord
sino,CONJ+coord
e,CONJ+coord
u,CONJ+coord
o,CONJ+coord.

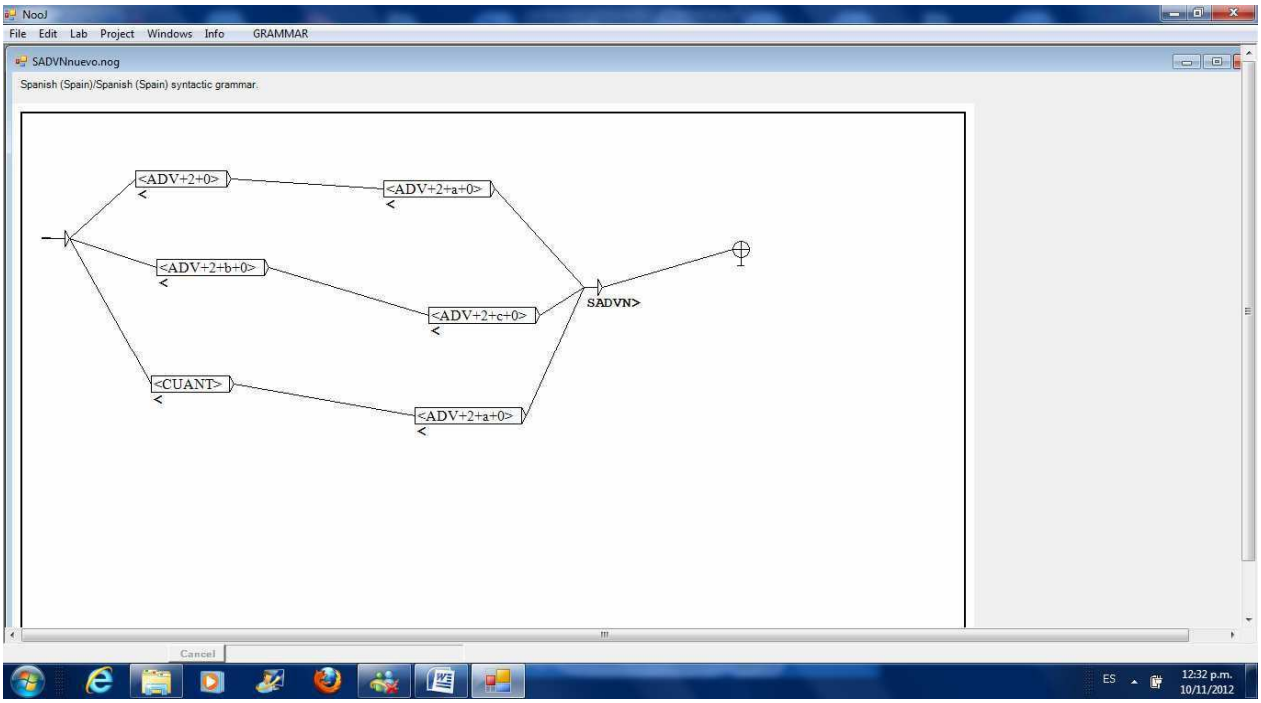
A continuación,

f. la gramática correspondiente al SADV como unidad mayor:

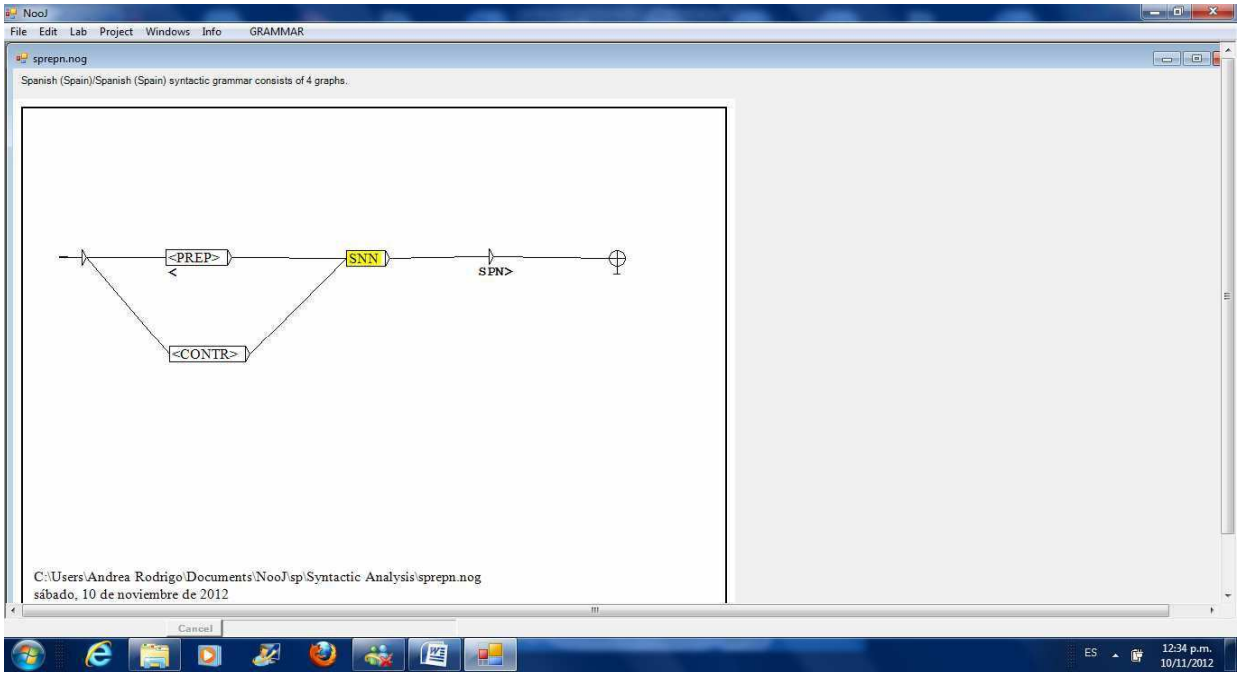


Y finalmente, las gramáticas correspondientes a cada uno de los sintagmas núcleos que lo integran:

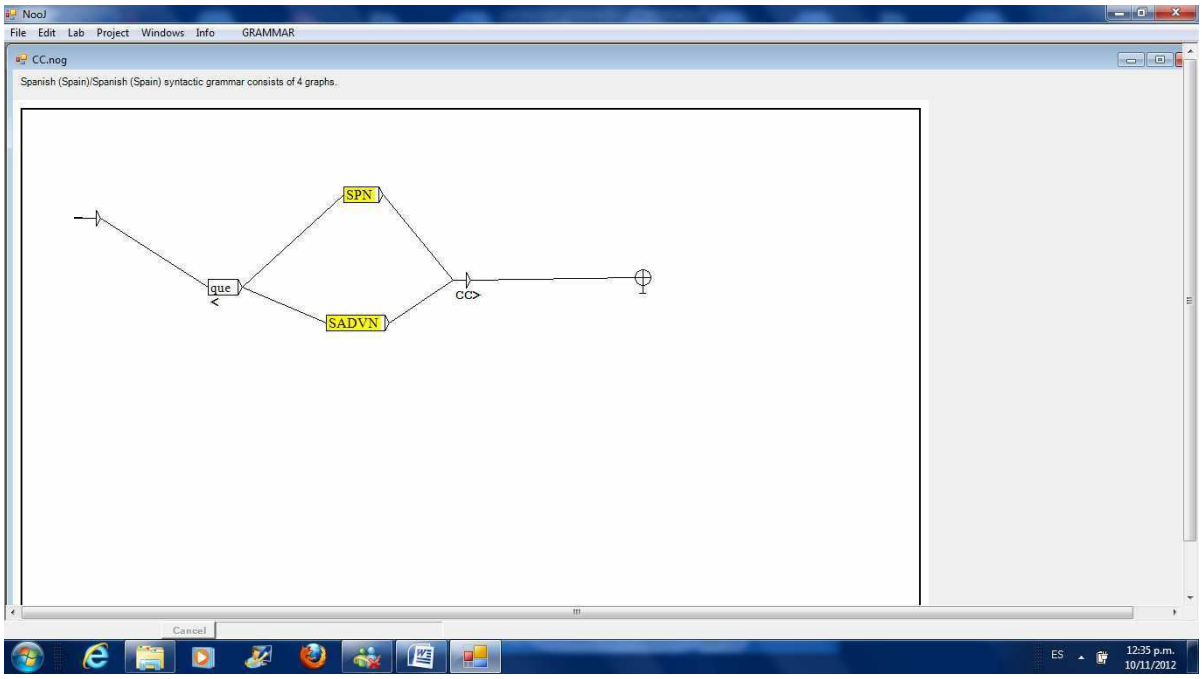
g. El SADVN



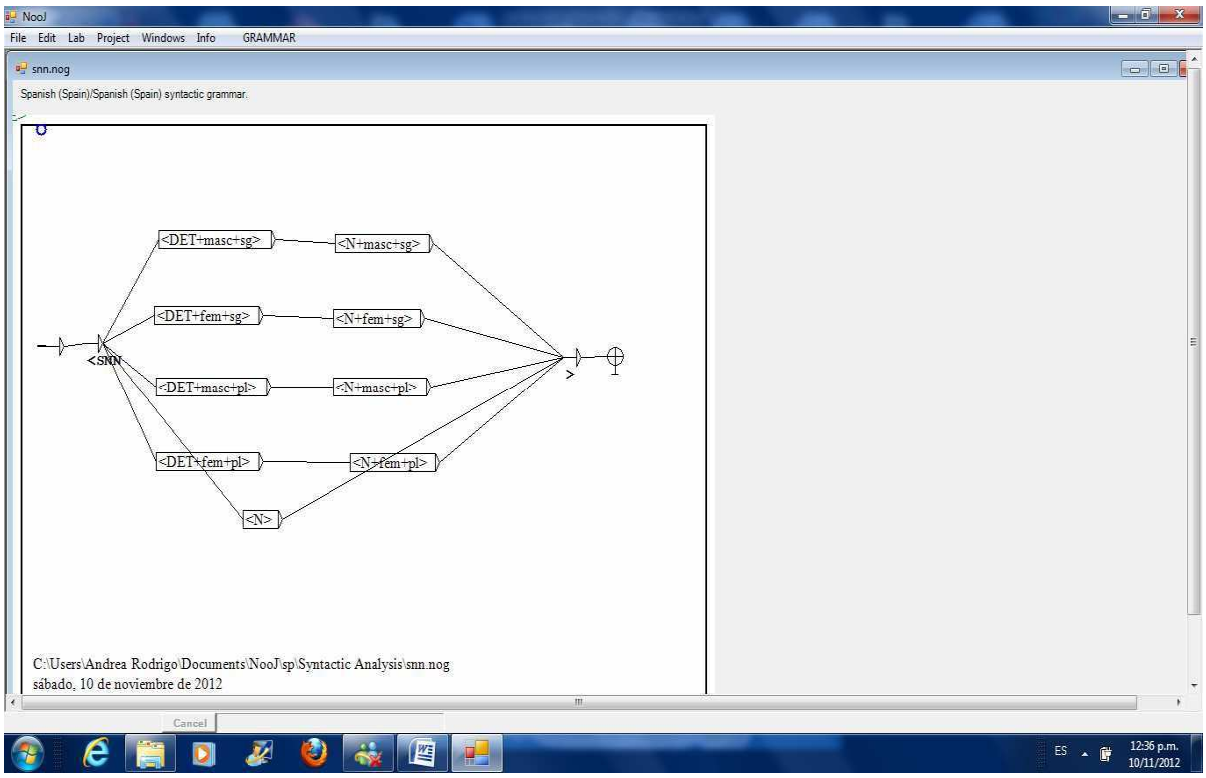
h. El SPN



6.8.La CC (construcción comparativa)



i. El SNN que integra tanto al SPN como a la CC



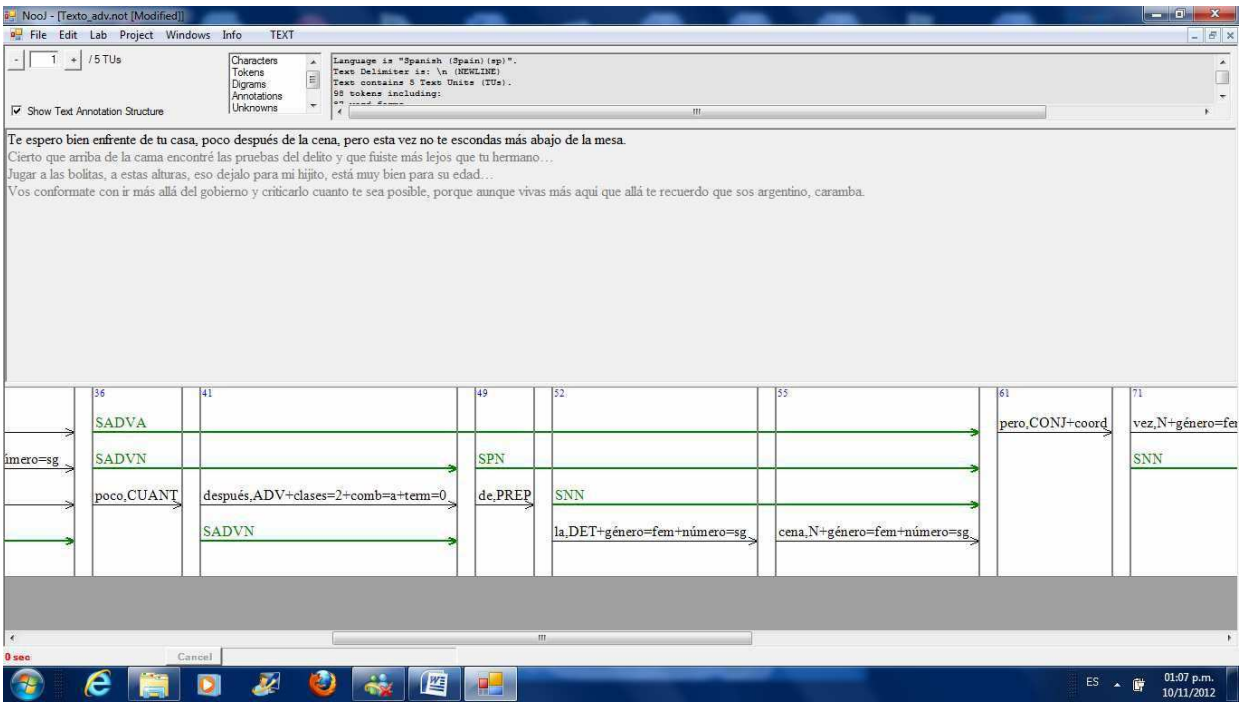
5. APLICACIÓN DE NOOJ EN EL ANÁLISIS DE UN TEXTO

5.1. El texto

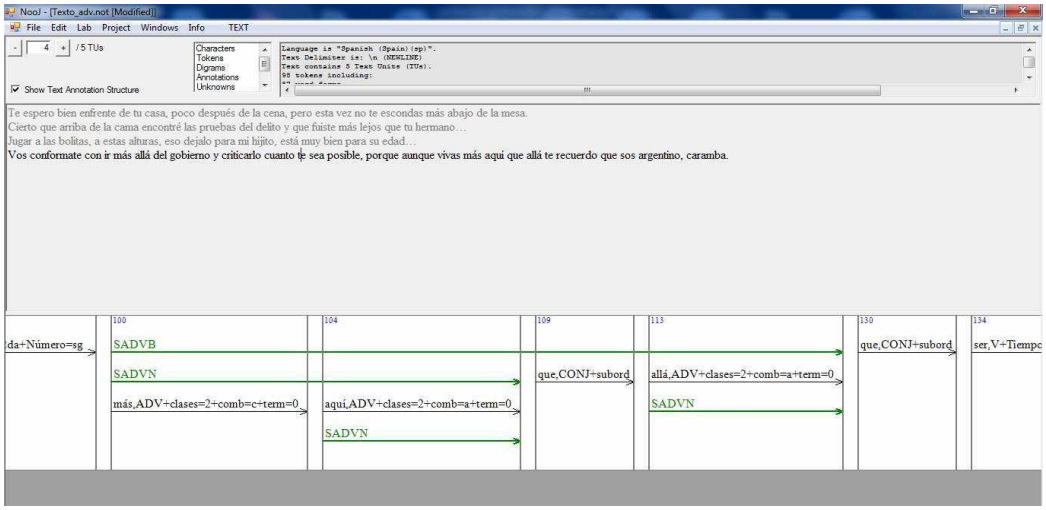
Te espero bien enfrente de tu casa, poco después de la cena, pero esta vez no te escondas más abajo de la mesa.
Cierto que arriba de la cama encontré las pruebas del delito y que fuiste más lejos que tu hermano…
Jugar a las bolitas, a estas alturas, eso dejalo para mi hijito, está muy bien para su edad… Vos conformate con ir más allá del gobierno y criticarlo cuanto te sea posible, porque aunque vivas más aquí que allá te recuerdo que sos argentino, caramba.

5.2 .El resultado que arroja NooJ

Se observa cómo extrae el SADVA (*poco después de la cena*):



Se observa cómo extrae el SADVB (*más aquí que allá*)[8]:

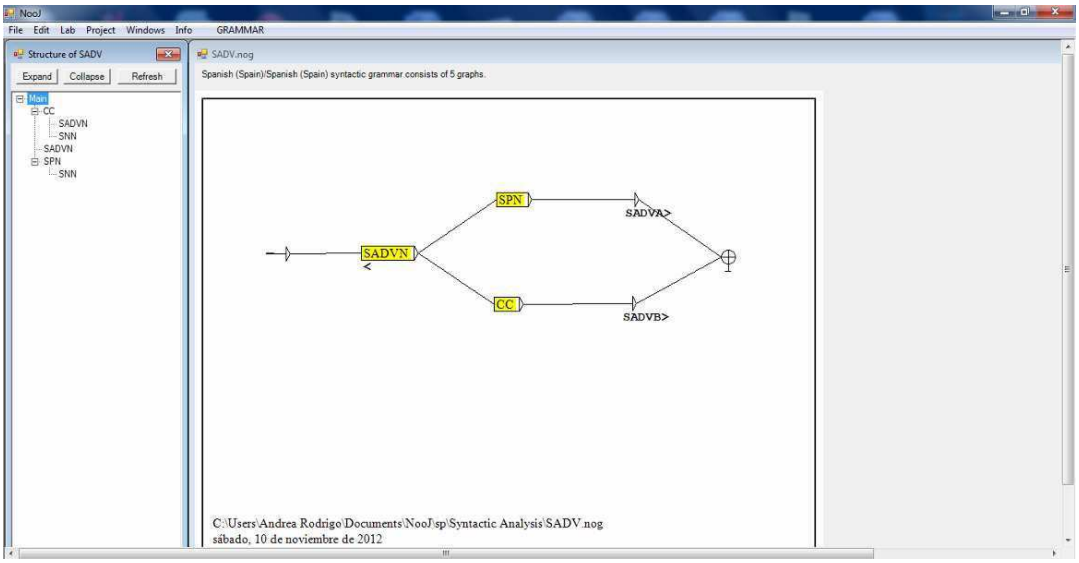


6. CONCLUSIONES

Según lo observado en 5.2., la herramienta logra extraer el SADV, aplicando la gramática, es decir, puede verse cómo el SADV aparece conformado por los sintagmas núcleos, según las estructuras planteadas en la Introducción:

- **SADVA** [SADVN + SPN (SNN)]: Ej: *más abajo de la mesa, arriba de la cama, bien enfrente de tu casa*, *poco después de las diez*[9]*, más allá del gobierno*.[10]
- **SADVB** [En una comparativa: SADVN + complemento comparativo (CONJ+ SNN) (CONJ+SNN)]: Ej: *más lejos que lo esperable, más aquí que allá*

Esto permitió corroborar las hipótesis lingüísticas planteadas en torno al SADV, a la vez que se puede visualizar cómo una estructura entra dentro de la otra:



[8] Se observa la extracción de uno SADV de cada clase, para no hacer tan extensa la exposición
[9] Ejemplo citado en Rev. Infosur N° 5, Extracción del Sintagma Nominal, (Solana, Rodrigo, Méndez), p. 18
[10] Pag. 12, 4/01/04

Se entiende así que en los SADV estudiados el SADVN puede ser seguido por un SPN, para los llamados clase SADVA o por una CC, para los llamados SADVB, según se ve en el rectángulo de la izquierda encabezado por **Structure of SADV**.

Referencias

[1] Rodrigo, A. Tratamiento automático de textos, el sintagma adverbial núcleo. Tesis doctoral, Escuela de Posgrado, Facultad de Humanidades y Artes, UNR, Ediciones Juglaría, 2011.

[2] Silberztein Max, 2003-. NooJ Manual. Available for download at: www.nooj4nlp.net

[3] Rodrigo, A. Análisis automático de textos, el sintagma nominal núcleo. Tesis de Maestría, Escuela de Posgrado, Facultad de Humanidades y Artes, UNR, 2006.

[4] Bès G.G. La phrase verbale noyau en français. En Recherches sur le français parlé, Volume 15, 1999.

[5] Hagège C. Analyse syntaxique automatique du portugais. Tesis de Doctorado GRIL, Univ.Blaise Pascal, 2000.

[6] Solana Z., Rodrigo A. El sintagma nominal núcleo. En Desarrollo, implementación y utilización de modelos para el procesamiento automático de textos. Compilado por Víctor Castel, Trabajos de las Segundas Jornadas de Lingüística Informática: Modelización e Ingeniería, Facultad de Filosofía y Letras, Universidad Nacional de Cuyo, 2005.

[7] Abney, S. Parsing by Chunks. En Berwick et al. Principle-Base Parsing. Kluwer Academic Publishers. Dordrecht, 1991.

[8] Aït-Mokhtar, S. L'analyse presyntaxique en une seule étape. Tesis de Doctorado, GRIL, Univ. Blaise Pascal Clermont-Ferrand, 1998.

[9] Bès G.G., Solana Z. Análisis morfológico y gramáticas locales: introducción y una aplicación concreta. En Jornadas Argentinas de Lingüística Informática Modelización e Ingeniería, JALIMI, Rosario, 2004.

[10] Bonino, R. "Una propuesta para la implantación de la morfología verbal del español en NooJ" en *Revista Infosur*. Nro. 5, [en línea] <http://www.infosurrevista.com.ar/>