IDENTIFICACIÓN DE SINTAGMAS NOMINALES NÚCLEO EN CORPUS DE APRENDIENTES BRASILEÑOS DE ESPAÑOL MEDIANTE EL SISTEMA NOOJ

Tramallino, Carolina- Arnal, Romina

(UNR- CONICET)

RESUMEN

En la gramática del español, la presencia o ausencia de determinantes interesa a nivel sintáctico pero también semántico, ya que reduce o amplía la referencia. Esto se hace aún más evidente en las construcciones de interlengua producidas por aprendientes de español como segunda lengua.

El objetivo del presente trabajo será realizar una descripción y lograr la automatización de los sintagmas nominales núcleo que se hallan en corpus de aprendientes brasileños de español, analizando tanto las construcciones coincidentes con la lengua estándar, así como también las que se desvían de ésta.

Para ello, en primer lugar se realizará una descripción lingüística de los snn en la lengua estándar, luego se describirán y analizarán las estructuras propias de interlengua y por último, se reconocerán automáticamente ambas construcciones.

Para el análisis automático utilizaremos la herramienta informática NooJ, que se encuentra disponible on line y ha sido creada para el análisis de lenguas naturales. Este sistema, desarrollado por Max Silberztein, permite analizar morfológica y sintácticamente las estructuras presentes en textos, dando como resultado la asignación categorial y morfológica correspondiente a cada palabra, y además agrupando las categorías a partir de la generación de gramáticas sintácticas de acuerdo con los rasgos que se declaren en sus archivos.

Palabras clave:
SINTAGMAS NOMINALES NÚCLEO- CORPUS ESPAÑOL – SISTEMA NOOJ

Introducción

El objetivo de este trabajo es poder reconocer sintagmas nominales núcleo no coincidentes con los del español estándar, hallados en corpus de estudiantes de español como segunda lengua. Particularmente, se estudian los casos de dos determinantes para un núcleo. En publicaciones anteriores (Tramallino, Arnal: 2019) se han estudiado los sintagmas nominales que presentan una asignación errónea de género y los sintagmas nominales con dos determinantes. (Tramallino, 2012).

Para lograr el objetivo propuesto se emplea el sistema NooJ que es un programa de libre acceso creado en 2002 y una herramienta para el tratamiento de lenguas naturales desarrollada por Max Silberztein que se encuentra disponible on line. [1]

El corpus está conformado por cincuenta producciones escritas, pertenecientes a estudiantes que tienen como lengua materna al portugués y se encuentran en un nivel inicial según el Marco Común de Referencia Europeo (MCRE). La investigación, enmarcada en el proyecto "La enseñanza de español como L2 a partir de herramientas informáticas" tiene como objetivo indagar sobre las problemáticas en el nivel sintáctico observadas en la interlengua de aprendientes brasileños de español y lograr mediante el uso del sistema NooJ que los estudiantes detecten y corrijan formas y estructuras desviadas en sus producciones escritas. Cabe aclarar que para una enseñanza de forma más rápida y eficaz, se hace necesario el uso de herramientas que faciliten el aprendizaje de los alumnos como: el empleo de recursos diversos y nuevas tecnologías, todo lo que puede llevar al éxito de la

[1] www.nooj-association.org

enseñanza y aprendizaje del alumno para que desarrolle las habilidades comunicativas en el nuevo idioma.

La hipótesis de trabajo es que pueden reconocerse sintagmas propios de la interlengua de aprendientes de español mediante la herramienta informática mencionada, a partir de la creación de gramáticas sintácticas productivas que analicen y etiqueten a dichas construcciones, distinguiéndolas de las coincidentes con el español estándar.

Para ello es necesario explicar en primer lugar a qué refiere el concepto de interlengua y desde qué enfoque se analiza el corpus. En segundo lugar, se explicará qué son los determinantes, qué valores poseen, en qué tipo de construcciones pueden aparecer y si permiten la presencia o no de otro determinante. Por último, se expondrán brevemente los alcances de la lingüística computacional y el funcionamiento del software con el que se identificarán automáticamente las estructuras mencionadas.

1. Concepto de interlengua

El término interlengua es utilizado por Selinker en 1972, con el cual propone un enfoque psicolingüístico y coloca a esta lengua en un lugar intermedio entre la lengua materna y la meta, es decir, la que se quiere aprender. Por lo tanto, posee elementos comunes con ambas pero es un sistema lingüístico independiente. Dicho de otro modo, cuenta con una gramática propia que se encuentra en continuo cambio; sus oraciones son idiosincrásicas, no erróneas y se constituye como un sistema peculiar de un estudiante individual o de un grupo de estudiantes que se halla en el mismo nivel de aprendizaje.

Con los estudios de la interlengua el investigador considera toda la producción del aprendiz, es decir, tanto las formas coincidentes como las no coincidentes con la lengua meta. La relevancia de los estudios desde esta perspectiva, coincidiendo con Alexopoulou (2010: 2)

"… surge de la necesidad de descubrir los principios generales que determinan el aprendizaje de la lengua extranjera, según los cuales el cerebro humano de los aprendientes procesa los datos del lenguaje a los cuales están expuestos."

Es importante aclarar que la interlengua contiene reglas propias y no debe tomarse como una mezcla entre los dos sistemas lingüísticos. La literatura especializada le otorga dos características que sin embargo parecen contradictorias: sistematicidad y variabilidad.

La primera refiere al conjunto de reglas coherentes con las que los aprendientes realizan hipótesis, sin embargo éstas son variables, se modifican a medida que el estudiante avanza en la adquisición y pasa a otra etapa. A partir del nuevo estadio se reestructuran esas reglas e hipótesis.

La sistematicidad está dada por la coherencia interna que presenta esa interlengua en un determinado momento de su desarrollo, se vislumbra en la aplicación de reglas lingüísticas que responden a estrategias y procesos que activan los aprendientes. En esta gramática singular, encontraremos oraciones o construcciones "desviadas" que sin embargo, serán adecuadas desde el punto de vista del estudiante.

Es importante destacar que ese sistema lingüístico tiene características y peculiaridades propias que coinciden con las halladas en el sistema de todo estudiante que está adquiriendo una segunda lengua, si se toma en cuenta un mismo nivel de aprendizaje.

2. Los determinantes en el español estándar

Los determinantes, también llamados cuantificadores existenciales, poseen características propias que han llevado a numerosos estudios. Bosque – Gutiérrez Rexach (2008) retoman las observaciones de Milsark (1974, 1977) sobre el tema, quien distingue entre determinantes fuertes (distributivos, demostrativos y posesivos) y determinantes débiles, y advierte que, mientras los primeros no pueden aparecer en las construcciones existenciales, los últimos sí pueden hacerlo:

*Hay ese estudiante en el jardín.
Hay muchos libros sobre la mesa.

Según Milsark los determinantes débiles (algún/ algunos, numerales cardinales, muchos, pocos, varios, etc.) no son cuantificadores, sino marcadores de cardinalidad. Por lo tanto, al

estar dichas oraciones cuantificadas existencialmente de forma inherente, no pueden aparecer cuantificadores pero sí marcadores de cardinalidad.

2.1. Los determinantes indefinidos

El determinante indefinido se utiliza para señalar que lo designado por el grupo nominal no es identificable por el oyente, aunque de esta significación se desprenden ciertas restricciones:

- El pronombre indefinido uno no puede ir precedido de artículos determinados, excepto: …el uno, el otro…".
- El indefinido no constituye por sí solo un grupo nominal: (*de un a otro lado)

2.2. El término otro

El término "otro" cuenta con propiedades que lo asemejan tanto a los adjetivos, como a los determinantes y a los cuantificadores. Acepta complementos partitivos y se acerca a los adjetivos por el hecho de que puede ir precedido de determinante, por ejemplo: *las otras cuestiones*, sin embargo no puede ir precedido del artículo indefinido, es decir que por ejemplo la combinación *un *otro* se rechaza.

Podemos atribuirle dos valores semánticos a *otro:* a) alteridad: Juan se mudó a otra casa y b) aditiva: Juan editó otro disco.

2.3. Los posesivos y demostrativos prenominales

Los posesivos se encuentran en distribución complementaria respecto de los otros determinantes en el español contemporáneo. Por lo tanto, puede decirse: el, este, mi, algún, ningún/ problema, pero no **el mi problema*, o * *algún su problema.*

3. El sintagma nominal núcleo

El sintagma núcleo es una secuencia de categorías que se encuentran en un orden, por lo tanto las combinaciones entre ellas son limitadas. Steven Abney (1991)

Este segmento se inicia con la primera categoría del sintagma y finaliza en su núcleo, por lo tanto, el sintagma nominal núcleo (SNN) es una construcción que va desde el comienzo hasta el núcleo del sintagma nominal. Solana y Rodrigo (2005).

En este trabajo se considerarán a los snn presentes en la interlengua deteniéndonos en la clase de determinantes que modifican al núcleo.

4. SNN propios de interlengua

En el corpus se presentan estructuras idiosincrásicas que no coinciden con los sintagmas nominales núcleo del español estándar. Se agrupan en tres casos:

I. Ausencia de artículos en construcciones encabezadas por "todos":
   a. "Recibiré **todas notas**…"
   b. "Me gusta mucho **todas días** en Rosario."
   c. "…son diferentes en **todos aspectos**"

II. Artículo seguido de posesivo más sustantivo común:
   a. "Los compañeros de **la mi nueva clase** son muy simpáticos…"
   b. "… **el nuestro** pueblo."
   c. "Yo soy mucho grato al pueblo de Rosario por **la su hospitalidad**."
   d. "**La mi profesora** se llama Carolina"
   e. "Yo soy grato al pueblo por **la su hospitalidad**"
   f. "…proponemos que **la nuestra empresa**…"
   g. "Para cumplir otra etapa **del mi doctorado**".

III. Artículo seguido de pronombre posesivo más sustantivo común:

a. “Es (…) mui linda tanto cuanto **la suya ciudad** Rosario.”

b. “…pero **la mía ciudad natal**…”

En esta oportunidad nos ocuparemos de generar gramáticas sintácticas para reconocer el segundo caso.

5. Lingüística computacional

En las últimas dos décadas se ha acuñado el término tecnologías del lenguaje para referirse a todas aquellas tareas en las que se aplica el conocimiento sobre la lengua para desarrollar sistemas informáticos capaces de reconocer, analizar, interpretar y generar lenguaje (Lavid, 2005). La lingüística computacional es un área interdisciplinaria que toma saberes de la Lingüística, la Informática y la Estadística y su tarea consiste en crear sistemas informáticos capaces de procesar el lenguaje humano y emular[2] la capacidad lingüística humana. De esta forma, las diversas ramas de la lingüística clásica tales como la psicolingüística, la neurolingüística, la dialectología, la sociolingüística y la lexicografía, entre otras, se proyectan renovadamente gracias a los instrumentos digitales que han venido en su complemento.(Parodi, 2004).

Con respecto a los programas necesarios para llevar a cabo investigaciones en lo que se denomina *lingüística de corpus* y *lingüística computacional* se han realizado para la lengua española en algunos centros académicos de Europa y, en otros casos, en países no hispanoparlantes (Parodi, 2004). En efecto, la mayor disponibilidad de ellos se halla en lenguas diferentes al español. En la actualidad, existen diversos traductores automáticos, diccionarios electrónicos monolingües y multilingües que responden a las tecnologías del habla, además de correctores automáticos y variados softwares educativos disponibles en Internet. Para el español se encuentran en la web flexionadores léxicos de sustantivos y

[2] Emular no significa comprender cómo funciona el cerebro humano si no intentar construir sistemas que comprendan y produzcan el lenguaje de manera similar a un humano.

adjetivos,[3] desambiguadores como el del Grupo de Estructuras de Datos y Lingüística computacional[4] y analizadores sintácticos como Freeling que es un analizador multilingüe[5].

5.1. Reconocimiento automático mediante el sistema NooJ

La utilización de este sistema permite la extracción de información en grandes corpus de textos y requiere de una formalización lingüística por parte de los usuarios.

Incluye herramientas para crear y mantener fuentes lexicales así como gramáticas sintácticas y morfológicas.

El software cuenta con dos tipos de recursos, los diccionarios y las gramáticas que pueden ser morfológicas o bien, sintácticas. En el módulo *Spanish* (Argentina) correspondiente al español[6], los diccionarios de las palabras variables como nombres, adjetivos y verbos, se encuentran asociados a gramáticas morfológicas en donde se declaran los modelos flexivos. (Tramallino, 2013)

En el archivo correspondiente a las propiedades definidas está declarada la siguiente información lingüística acerca de los cuantificadores:

CATEGORÍAS = V | N | ADJ | ADV | PREP | DET | CUANT | CONJ | CONTR | INTERJ | CL | PRON | ART | REL | INTEXC | LOCLAT;

DET_clase = artdet | artindet | dem | pos;

DET_género = masc | fem | neutro;

---

[3] Banco de datos SENSEM para el español Disponible en: https://www.cs.upc.edu/~nlp/papers/padro11.pdf

[4] Disponible en: http://www.gedlc.ulpgc.es/investigacion/desambigua/morfosintactico.htm

[5] Disponible en: http://nlp.lsi.upc.edu/freeling/index.php/node/1

[6] Spanish Module (Argentina) disponible en:

http://www.nooj-association.org/index.php?option=com_k2&view=item&id=6:spanish-module-argentina&Itemid=611

DET_número = sg | pl;

DET_persona = 1era | 2da | 3era;

#Género y número de los posesivos en cuanto al objeto poseído

DET_númerop = psg | ppl;

DET_génerop = pmasc | pfem;

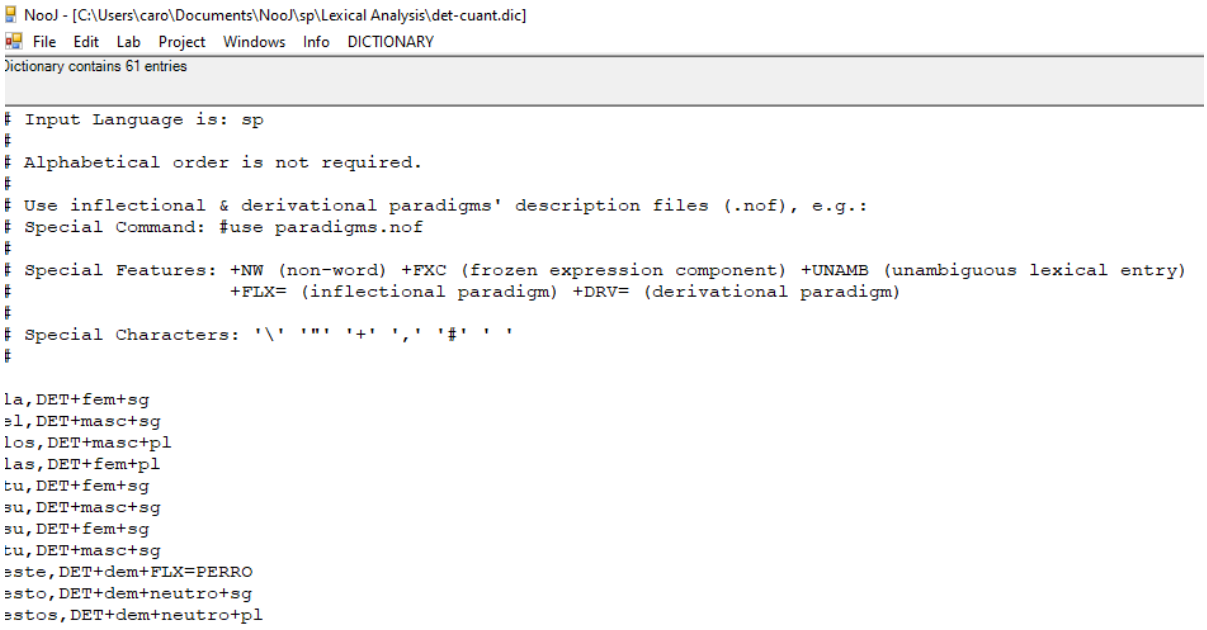
CUANT_clase = def | indef | num;

CUANT_tipo = apócope;

CUANT_género = masc | fem | neutro;

CUANT_número = sg | pl;

Respecto del diccionario determinantes-cuantificadores, se consiga la información categorial en mayúsculas con la etiqueta morfosintáctica correspondiente (DET para determinante) y los rasgos en minúscula (fem-femenino-; masc –masculino-), como se observa en la siguiente imagen:



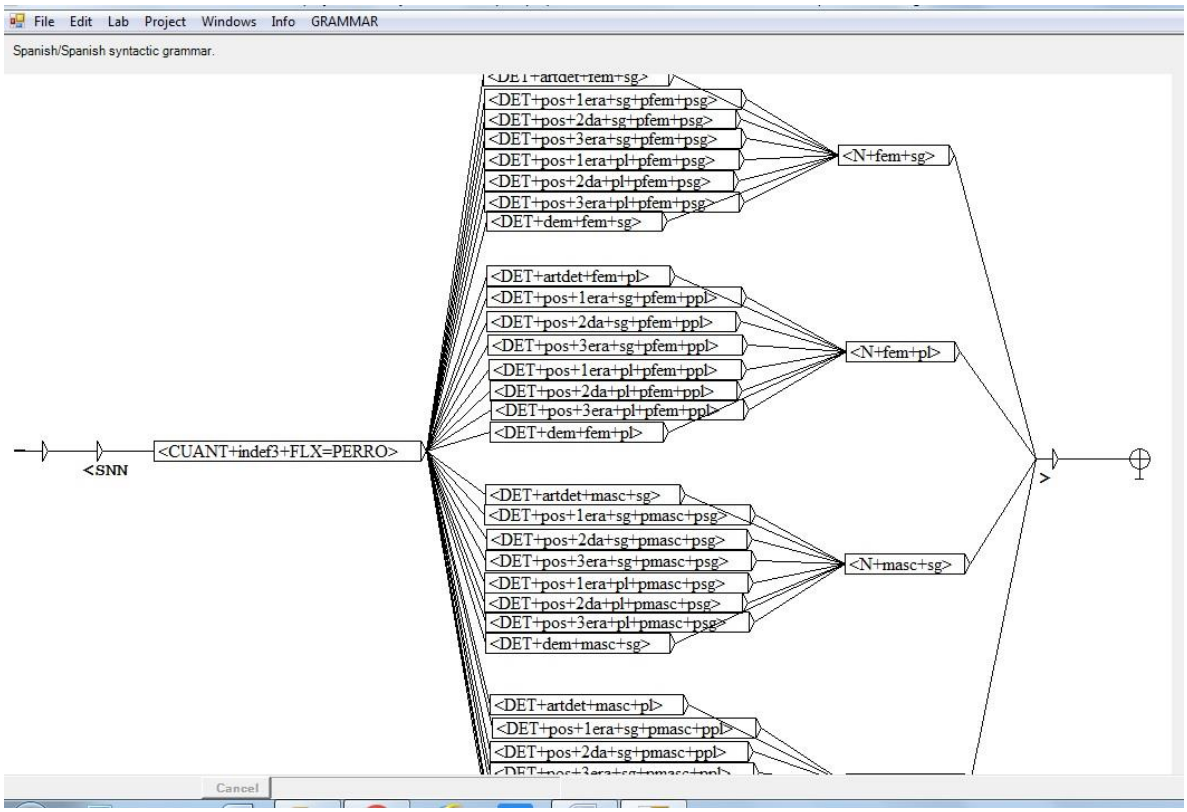
Captura de pantalla N° 1: diccionario determinantes-cuantificadores

En cuanto a las gramáticas, éstas se usan para representar fenómenos lingüísticos, desde niveles ortográficos y morfológicos a niveles sintagmáticos. Contiene tres tipos de gramáticas:

a. Gramáticas derivacionales (o archivos con terminación .nof) que se emplean para representar las propiedades flexivas, por ejemplo, conjugaciones o derivaciones como las nominalizaciones de las entradas léxicas. Estos modelos pueden declararse de forma gráfica o mediante reglas.
b. Gramáticas lexicales, ortográficas, morfológicas o terminológicas (archivos con terminación .nom). Éstas se usan para representar conjuntos de tipos de palabras y asociarlas con información léxica; por ejemplo: estandarizar la ortografía de palabras o de variantes, reconocer neologismos, asociar expresiones sinónimas.
c. Gramáticas semánticas o sintácticas (archivos con terminación .nog) que son utilizadas para reconocer o anotar expresiones en textos como, por ejemplo, etiquetar tanto a frases nominales como a ciertas expresiones sintácticas o idiomáticas. Se emplean para extraer o desambiguar palabras filtrando algunos ítems léxicos o anotaciones sintácticas en el texto.

Los paradigmas inflexión-derivación son formalizados como bibliotecas de gráficos estructurados o reglas basadas en textos. Se realizan mediante grafos.

En la captura de pantalla N°1 que se presenta a continuación, se observa una gramática del sintagma nominal núcleo propio del español estándar (SNN).



Captura de pantalla N° 2: "Gramática del SNN"

5.1.Reconocimiento automático de SNN de interlengua

Como se mencionó en el apartado anterior, el sistema Nooj permite a los usuarios crear, mediante grafos, *gramáticas productivas sintácticas*.

Productivas, porque se emplean categorías. Por ejemplo: DET para "determinante"; N para "nombre"; ADJ para "adjetivo".

Sintácticas, porque agrupa dos o más palabras y las etiqueta. Por ejemplo: SNN para "Sintagma Nominal Núcleo".

Precisamente, al ser los usuarios los que tienen la posibilidad de crear las etiquetas de las expresiones sintácticas, se pueden realizar adaptaciones para reconocer sintagmas propios de la interlengua.

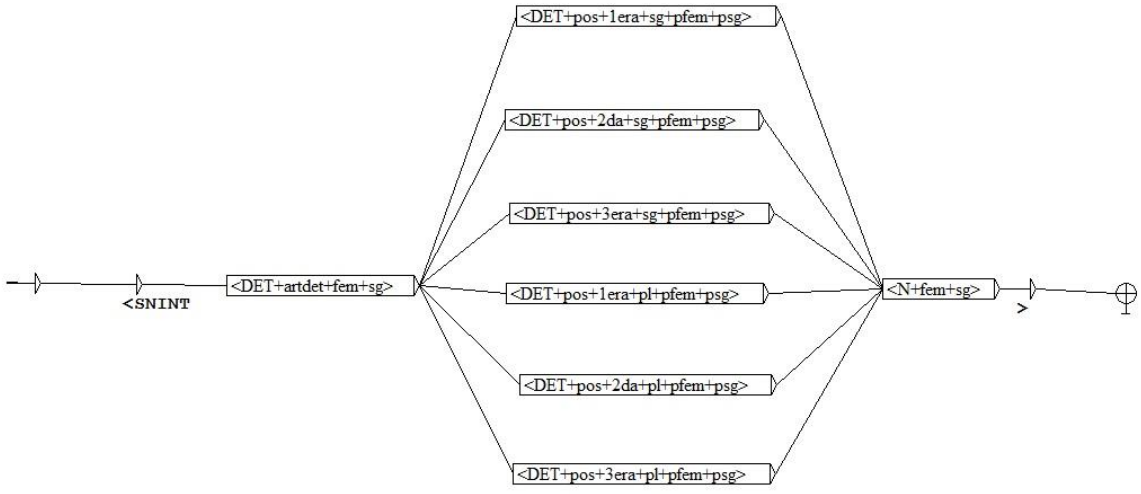
El programa proporciona el nodo inicial (▷) y el nodo final (⊕), por lo que son los usuarios los que deben crear los nodos intermedios y unirlos.

Entonces, para crear una gramática del sintagma nominal núcleo perteneciente al español estándar, se pueden crear nodos de determinantes y nombres con sus rasgos de género y número y unirlos estableciendo la concordancia entre ellos.

5.2. Gramáticas sintácticas propias de interlengua

Como se explicó en el apartado anterior, de la misma forma en que se generan gramáticas de grafos para el español, puede crearse una gramática con la etiqueta SNINT (Sintagma Nominal de Interlengua).

Para los ejemplos agrupados en el caso N°II se debe indicar que la suma de un DET (artículo) seguido de un DET (posesivo) y un N (nombre) dé SNINT, como se muestra en la captura de pantalla que sigue:

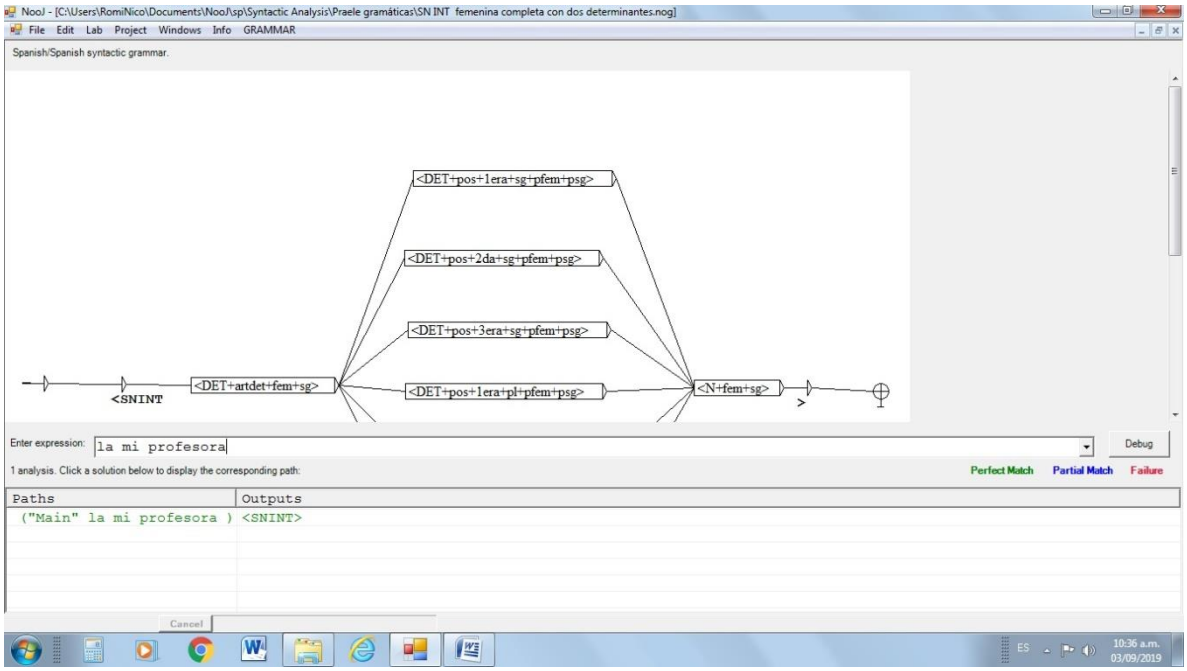


Captura de pantalla N°3: "Gramática del SN INT femenino con artículo y posesivo"

6. Resultados obtenidos

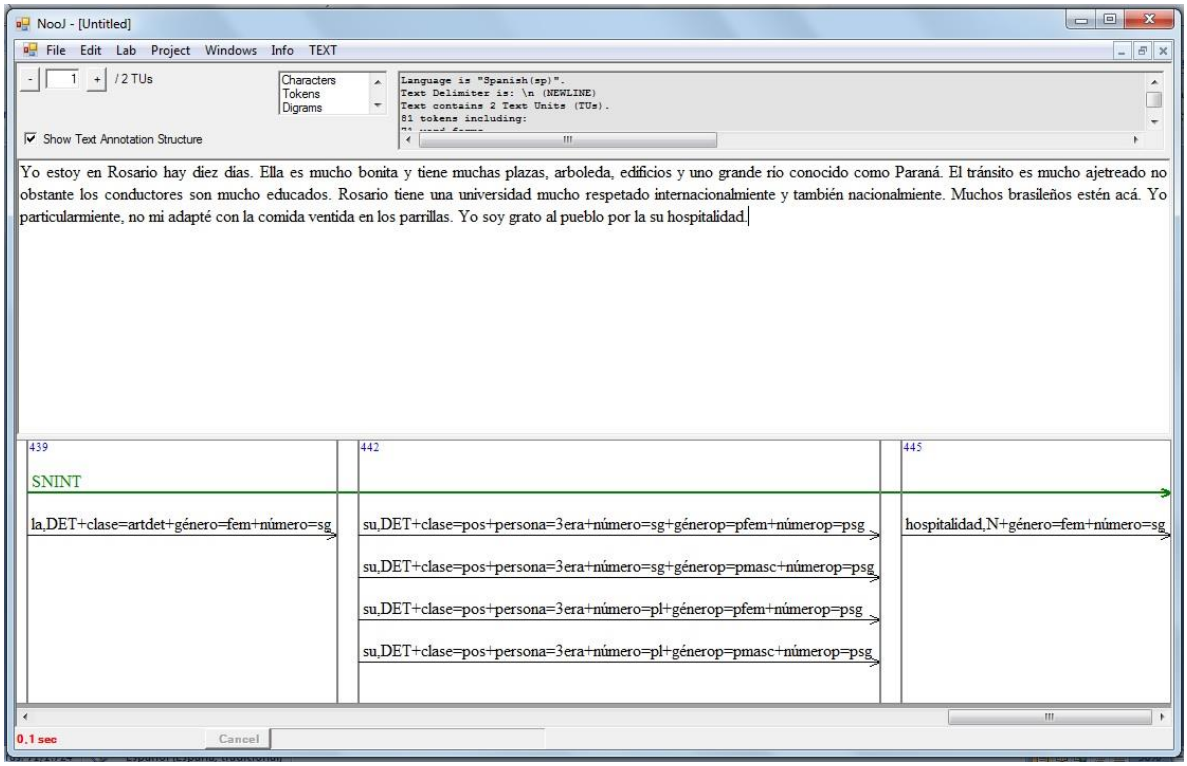
Nooj también permite verificar si la gramática está bien creada, con la acción "Show Debug", que se aprecia en la ventana inferior de la captura de pantalla anterior, donde se lee que el sintagma "la mi profesora" es reconocido como SN INT.

Esta gramática permite reconocer los ejemplos que se han clasificado como del tipo II, femeninos. Para identificar los ejemplos masculinos se emplea una gramática similar con el rasgo +masc en lugar de +fem.



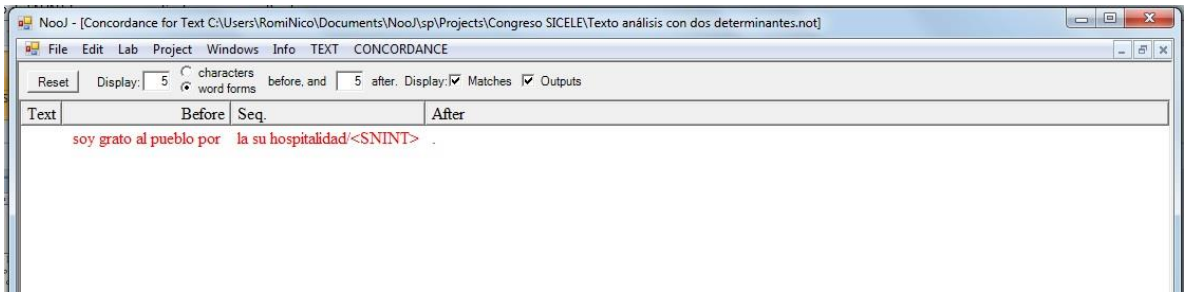
Captura de pantalla N°4: "Reconocimiento de SNINT femenino con artículo y posesivo"

Una vez creadas las gramáticas, se pueden reconocer o analizar estas estructuras en un texto. Para esto, se deben asociar estas gramáticas así como los diccionarios de categorías léxicas que se necesitan para que el programa reconozca las palabras que debe analizar (N, DET).



Captura de pantalla N° 5 “Análisis sintáctico de SN INT femenino con artículo y posesivo”

De esta manera, en grandes corpus, los investigadores tienen la posibilidad de extraer datos estadísticos con respecto a la cantidad de estructuras, como en este caso, o de términos o categorías.



Captura de pantalla N° 6 "Localización de SN INT femenino con artículo y posesivo"

Respecto a la medición de los resultados de acuerdo al índice de cobertura y precisión, la precisión se define como el cociente entre el número de ítems correctamente analizados y el número total de ítems analizados, mientras que la cobertura es el cociente entre el número total de ítems y el número de ítems correctamente analizados.

La cobertura alcanza un alto porcentaje, cercano al 90 por ciento, dado que la gramática confeccionada para sintagmas con determinante artículo más determinante posesivo contempla todas las posibilidades de combinación entre nombre y determinante. En cuanto a la precisión, ésta también arroja un porcentaje cercano al 90 por ciento.

7. Consideraciones finales

En esta investigación se analizó un corpus de producciones escritas de estudiantes brasileños de español como segunda lengua pertenecientes a un nivel inicial de aprendizaje. Se hallaron sintagmas nominales núcleos que presentaban divergencias respecto a los de la lengua estándar. Luego de establecer que estas anomalías residían en la elección y sobre todo, combinación de los determinantes, se especificó a estos como artículos, posesivos, demostrativos. Se agruparon las estructuras propias de interlengua en tres casos de los cuales se eligió para analizar automáticamente las que contenían artículo seguido de posesivo más sustantivo común.

Se trabajó con el software libre y de acceso gratuito NooJ

Con el objeto de distinguir e identificar a estas estructuras se crearon las gramáticas sintácticas mediantes grafos que indicaran que estos sintagmas correspondían a la interlengua y no al español estándar.

En síntesis, este artículo cumplió con el objetivo propuesto de identificar y distinguir las estructuras con núcleos nominales y determinantes a través del sistema NooJ tanto para los sintagmas nominales núcleo coincidentes con el español como para las estructuras idiosincrásicas encontradas en el corpus analizado. Asimismo, permitió contribuir al ámbito de los estudios de interlengua.

Las mediciones de los resultados arrojan un alto porcentaje respecto a la cobertura y precisión que validan la hipótesis en cuanto la posibilidad de detectar y etiquetar los sintagmas nominales núcleo pertenecientes a la interlengua de estudiantes brasileños de español como segunda lengua mediante el software de libre acceso NooJ.

8. Referencias

ALEXOPOLOU, A. (2010) "La función de la interlengua en el aprendizaje de lenguas extranjeras". Revista Nebrija de Lingüística aplicada 2010, 9.

BOSQUE, I. – GUTIÉRREZ REXACH, J. (2008) *Fundamentos de sintaxis formal*. Akal**,** Madrid.

LAVID, J.(2005). *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI.* Cátedra: Madrid.

PARODI, G. (2004). Textos de especialidad y comunidades discursivas técnico-profesionales: una aproximación basada en corpus computarizado**.** *Revista Estudios filológicos*,(39),7-36.

REAL ACADEMIA ESPAÑOLA (2010). *Nueva gramática de la lengua española*. Espasa, Buenos Aires.

SELINKER, L. (1972) "Interlanguage", en International Review of Applied Linguistics, num 10.3 pp.209-31.

STEVEN A. (1991) Parsing By Chunks. en: Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing.* Kluwer Academic Publishers, Dordrecht.

SOLANA, Z y RODRIGO, A. (2005). "El sintagma nominal núcleo", en *Desarrollo, implementación y uso de modelos para el procesamiento automático de textos* (ed. Victor Castel) Facultad de Filosofía y letras, UNCUYO.

TRAMALLINO, C. (2012) "Análisis automático de la interlengua de aprendientes de español: la presencia de determinantes indefinidos en el sintagma nominal núcleo" en Revista *INFOSUR* Nº 6, págs. 75-87. ISSN 1851-1996, diciembre 2012.

TRAMALLINO, C. (2013) Análisis morfológico con herramientas informáticas. Reconocimiento de nombres en textos de español con el sistema Nooj en *Revista Lingüística y Literatura*, (63), 33-48.

TRAMALLINO, C., ARNAL, R. (2019) "Reconocimiento automático de sintagmas nominales en producciones escritas de aprendientes brasileños de español" e-Universitas Jornal UNR. Vol. 2, Núm. 22 (11): Junio 2019. ISSN:1852-0707.