http://web2.rosario-conicet.gov.ar/ojs/index.php/revistairice - ISSN: 0327-392X

Reconocimiento de sintagmas nominales construidos con indefinidos a través del sistema NooJ en corpus de español como segunda lengua

Carolina Tramallino*
Romina Paola Arnal**

Resumen

El propósito de este trabajo, que se enmarca dentro de un proyecto de investigación acerca de la enseñanza del español a través de herramientas informáticas, es lograr el reconocimiento automático de sintagmas nominales construidos con indefinidos, que se hallan en un corpus de producciones escritas de estudiantes brasileños de español como segunda lengua. Cabe aclarar que, la muestra se divide en dos grupos de textos de acuerdo al nivel de conocimiento de la lengua de los sujetos participantes. Para ello, se empleará el sistema NooJ, con el que se reconocerá tanto a las construcciones coincidentes con la lengua meta como también a las que se desvían de esta, distinguiéndolas mediante una etiqueta. La metodología de trabajo consistirá en realizar una descripción lingüística de los SN en español y proponer luego una clasificación para los indefinidos que servirá a los fines de poder emplear el recurso del software respecto de la generación de "gramáticas sintácticas productivas" que permitan detectar las estructuras idiosincrásicas presentes en el corpus. Los resultados obtenidos del análisis de los datos podrán contribuir a los estudios dentro de la perspectiva de la hipótesis de interlengua, demostrando que ciertas construcciones son comunes a las producciones de estudiantes que comparten un mismo nivel de instrucción, pero que estas pueden disminuir en cantidad e incluso desaparecer a medida que se avanza en el proceso de adquisición de una segunda lengua, debido al carácter transitorio del mencionado sistema lingüístico.

Palabras clave: sintagmas nominales, indefinidos, corpus español, sistema NooJ

* Doctora en Humanidades y Artes con mención en Lingüística, Universidad Nacional de Rosario (UNR). e-mail: carotramallino@hotmail.com
** Profesora en Letras, Universidad Nacional de Rosario (UNR). e-mail: arnalromina@gmail.com

Recognition of noun phrases formed by indefinite articles in Spanish corpus as a second language with the NooJ system

Abstract

The goal of this work, which is part of a research project on the teaching of Spanish through computer tools, is to show how the automatic recognition of noun phrases formed by indefinite articles, which are found in corpus of texts written by students of Spanish as a second language, can be carried out. It should be noted that the sample is divided into two groups of texts according to the level of knowledge of the language of the participating subjects. For this purpose, the NooJ system, not only will be used to recognize the constructions that coincide with the target language, but also those that deviate from it, distinguishing them by means of a label. The work methodology will consist of a linguistic description of the noun phrases in Spanish and then propose a classification for the indefinite ones that will permit to use the resource of the software in respect of generating "productive syntactic grammars" that allow to detect the idiosyncratic structures present in the corpus. The results obtained from the analysis of the data may contribute to the studies within the perspective of the interlanguage hypothesis, demonstrating that certain constructions are common to the productions of students who receive the same level of instruction, but these may decrease in quantity and even disappear as soon as the students make progress in the process of acquiring a second language, due to the transitory character of the mentioned linguistic system.

Keywords: noun phrases, indefinite articles, Spanish corpus, NooJ system

Introducción

El objetivo de este trabajo es poder reconocer mediante un software, que fue creado para el análisis de lenguas naturales, las construcciones nominales construidas con indefinidos que se hallan en textos escritos pertenecientes a estudiantes brasileños de español como segunda lengua (L2). El interés radica en que dichas estructuras presentan divergencias respecto de las de la lengua meta, que es el idioma que se intenta aprender.

La presente propuesta se enmarca en un proyecto de investigación que refiere a la enseñanza de español como L2 a partir de herramientas informáticas. Por lo tanto, se ubica dentro del campo de la lingüística computacional y de la adquisición de segundas lenguas. El objeto de implementar un software de acceso libre y gratuito, como es el sistema NooJ, es que los aprendices[1] puedan analizar sus propios escritos para detectar y posteriormente reformular aquellas estructuras que exhiben diferencias respecto del español. Para ello, será necesario explicitar a qué teoría adscribimos dentro del campo de estudio de adquisición y cuál es el campo de trabajo de la lingüística computacional como área multidisciplinar, para luego explicitar cuestiones referidas al corpus reunido y a los sujetos participantes.

La metodología de trabajo consistirá en realizar, primero, una descripción del corpus que está dividido en dos muestras según el nivel de aprendizaje de los estudiantes y, luego, de las construcciones nominales formadas a partir de un núcleo sustantivo, a las que se denominan sintagmas nominales (SN). A partir de esto, se propondrá una clasificación para los indefinidos a los fines de realizar la implantación en máquina de esta categoría en la herramienta informática que se emplea. A continuación, se hará referencia al funcionamiento del sistema Nooj, al inventario de sus archivos y a la generación de “gramáticas sintácticas productivas” mediante grafos, las cuales posibilitarán la detección y el etiquetado de las

[1] Se deja claro que en este trabajo se emplearán los términos estudiante, alumno, aprendiente y aprendiz como expresiones de igual significado.

estructuras idiosincrásicas presentes en el corpus, que se designarán como SNINT (sintagmas nominales de interlengua).

Finalmente, se exhibirán los resultados del reconocimiento automático respecto de los SN y los SNINT presentes en las dos muestras que conforman el corpus. El objetivo será establecer qué tipo de estructuras aparecen en ambos grupos y si estas se hallan en un mismo porcentaje o, por el contrario, presentan diferencias significativas.

Marco Teórico

En esta sección se realizará un recorrido por las diferentes teorías hasta llegar a la corriente de interlengua y se mencionarán algunas investigaciones actuales sobre el uso de determinantes en corpus de español como segunda lengua. Luego se explicará qué alcance tienen los estudios referidos al área de la lingüística computacional.

Teorías de Adquisición de Segundas Lenguas

En lo que refiere a la adquisición de segundas lenguas, en la década del 40 del siglo pasado, en el ámbito de la lingüística aplicada, comienzan las investigaciones que tienen como objetivo el estudio de la lengua del aprendiente. Se intenta predecir, mediante una gramática contrastiva, las dificultades con las que se encontrará el estudiante para poder evitar el error. Este campo de investigación tuvo su auge en las décadas de 1950 y 1960 con la hipótesis del análisis contrastivo, que se encargó de describir y comparar las lenguas involucradas en el proceso (la lengua nativa del aprendiente y la lengua meta) para hallar los puntos de contacto y las zonas divergentes entre uno y otro sistema en cada nivel del análisis lingüístico (ya sea en cuanto a fonología, morfología o sintaxis) con el objetivo de detectar y predecir las equivocaciones que cometerían los estudiantes.

Sin embargo, al advertir que la fuente de la causa de los errores no sólo respondía a la interferencia con la lengua materna (en-

tendida esta como el efecto de la lengua nativa sobre la lengua meta) surge un nuevo paradigma: el del análisis de errores. Corder (citado en Liceras, 1992), primer exponente de esta corriente, les atribuye un valor positivo. Sostiene que actúan como un instrumento mediante el cual los aprendientes pueden deducir nuevas reglas del sistema lingüístico que están adquiriendo.

El aporte de esta línea investigativa reside en despojar de negatividad al error y tomarlo como un indicador del proceso de aprendizaje llevado a cabo por el estudiante. Respecto de las clasificaciones de errores, se crean varias que centran su atención en las desviaciones respecto de la norma de la lengua meta. Para la enseñanza del español como segunda lengua se presenta, por ejemplo, entre otras investigaciones, la de Sonsoles Fernández (1997), quien analiza con corpus de producciones de aprendices que poseen cuatro lenguas de origen diferentes. Trabaja en todos los niveles de la lengua: fonológico, morfosintáctico, sintáctico, léxico y discursivo en tres niveles de competencia, con el objeto de extraer un corpus de errores universales o característicos de cada lengua materna.

Finalmente, surge la corriente de interlengua, la cual, coincidiendo con Alexopolou (2010), significa un giro metodológico porque se propone el estudio y análisis tanto de las formas idiosincrásicas como de las que coinciden con el sistema lingüístico que se está adquiriendo. Es decir, que esta línea de investigación toma en consideración la producción total de los estudiantes y demuestra que tanto unas como otras son centrales en el proceso de aprendizaje (Alexopolou, 2010). Cabe aclarar que el término es empleado por primera vez por Selinker (1972), quien, desde un enfoque psicolingüístico, coloca a este sistema en un lugar intermedio entre la lengua materna y la meta, teniendo elementos comunes con ambas. Así, sostiene que los estudiantes transitan por etapas o niveles de competencia que se van modificando a medida que van adquiriendo vocabulario y nuevas estructuras del idioma que están aprendiendo. Lo interesante de esta perspectiva

radica en el siguiente supuesto: aunque esta lengua idiosincrásica difiera en cada alumno en particular, presenta zonas de intersección en aprendientes con un nivel académico similar e idéntica lengua nativa. Corder (citado en Liceras, 1992) había denominado a este sistema como un *dialecto idiosincrásico* o *transitorio*, al que le atribuía como particularidad el poseer una gramática propia que se encuentra en continuo cambio con oraciones indiosincrásicas, no erróneas. Por lo cual, afirma que es un sistema peculiar de un estudiante individual o de un grupo de estudiantes que posee igual formación académica. Nemser (1971) se refería al mismo concepto pero con la denominación de *sistema aproximativo*, al que también le otorga una gramática y vocabulario particulares.

Algunos de los trabajos que han representado un aporte en lo que respecta a este ámbito son los de Liceras (1992, 2009), Rigamonti (2006) y Alba Quiñones (2009); este último retoma todas las taxonomías de errores existentes realizadas desde distintos enfoques, entre otros.

La presente investigación adscribe a la teoría de interlengua y, por lo tanto, contempla en su conjunto tanto a los SN coincidentes con el español como a los idiosincrásicos (SNINT). Propone la localización y contabilización de estos a través de la herramienta informática Nooj, en dos corpus diferentes: uno perteneciente a estudiantes brasileños de nivel inicial y otro, a estudiantes brasileños de nivel intermedio con el objeto de comparar qué sucede en cuanto a las estructuras sintácticas peculiares y si estas se efectúan con la misma frecuencia en ambas muestras.

Respecto de trabajos relacionados a la adquisición de determinantes en español, cabe mencionar el estudio de Valenzuela y Toledo (2019), quienes analizan el uso y frecuencia de artículos mediante test a estudiantes no nativos de español de niveles diferentes, para concluir que el artículo definido es el que muestra un alto índice de error, como estructura con mayor uso y mayor elección errónea, seguido del artículo neutro. Los resultados exhiben una mayor cantidad de sobre uso de los artículos en lugar de la omisión de estos y un mejor empleo cuanto mayor es el nivel de enseñanza. También observan un retroceso en relación con los

usos correctos de la lengua meta que puede explicarse por una tendencia a hipergeneralizar reglas o al uso de estructuras antes omitidas por falta de confianza en el uso (Torijano Pérez, 2008).

Por otra parte, Torijano Pérez (2008) trabaja la frecuencia y recurrencia de indefinidos en corpus de español de estudiantes luso hablantes. Específicamente, se ocupa de determinar: la omisión de elementos necesarios, la adición de partes innecesarias y la elección errónea y problemas afines (Torijano Pérez, 2008). Determina que se hallan más errores de adición que de omisión del artículo definido mientras que el indefinido no conlleva dificultades en su uso. Respecto de los posesivos, el mayor escollo se encuentra en la utilización de formas plenas antepuestas en lugar de las formas apocopadas, como por ejemplo: "suyas vacaciones, mía querida amiga" (Torijano Pérez, 2008, p. 253). Como conclusión, afirma que la corta distancia interlingüística, es decir entre la lengua nativa y la lengua meta, presentaría una mayor resistencia a la eliminación de las desviaciones, especialmente las referidas a los determinantes.

Lingüística Computacional

En los últimos años 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. En las últimas tres décadas surge la Lingüística Computacional como ciencia del lenguaje que contribuye al conocimiento de los procesos cognitivos de comprensión y producción del lenguaje. Esta 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. Siguiendo a

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

Parodi (2004), 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.

La ventaja de utilizar modelos computacionales del lenguaje es que permite la comprobación de teorías lingüísticas, ya que los procesos pueden inspeccionarse y experimentarse a través de la construcción de programas y bases de conocimiento (Lavid, 2005). Esta disciplina se caracteriza por ser teórico-aplicada; como explica Yllescas (2011, p. 341), no solo se trata de desarrollar una teoría propia que no necesariamente debe coincidir "con los presupuestos de la Lingüística teórica, sino que (…) ha de verse cristalizada en un producto informático". Así pues, siguiendo a Halvorsen (citado en Tordera Yllescas, 2011), se pueden establecer las siguientes ramas dentro de la teoría computacional: (a) Tratamiento del habla (síntesis de voz y reconocimiento del habla); (b) Análisis, Generación e Interpretación del lenguaje natural; (c) Traducción automática. En este caso, nos situamos en la segunda, ya que empleamos un programa diseñado para el tratamiento de veinte lenguas naturales diferentes.

Es importante mencionar que se trabajará a partir de una adaptación de la información lingüística declarada en los archivos que componen el módulo español. De esta forma, el sistema NooJ permitirá reconocer automáticamente estructuras sintácticas bien formadas y otras propias del sistema de interlengua de estudiantes que tienen como lengua materna al portugués. El objetivo general que se espera cumplir a futuro es que los aprendientes puedan analizar sus propios textos con el software y descubrir cuáles son las estructuras que se diferencian de la lengua que están aprendiendo y, por lo tanto, puedan corregirlas.

En cuanto a la detección automática de errores en corpus de español L2 podemos mencionar el estudio de Ferreira, Elejalde, y Vine (2014), quienes presentan un análisis de Errores Asistido por computador a partir de Corpus de Aprendientes de Lenguas en Formato Electrónico, compuesto de resúmenes realizados por

estudiantes de nivel intermedio. Para ello, diseñan un sistema de etiquetas de anotación de errores y los analizan en el corpus en diversos niveles de categorización a partir del programa Nvivo 10, que es un software de análisis de datos cualitativos. Los resultados a los que arriban señalan que la mayoría de errores observados son los gramaticales; que corresponden a las siguientes categorías: con mayor frecuencia los referidos a las preposiciones con un 39 %, a los verbos con un 23 % y a los artículos con un 17%.

Metodología

En esta sección se explicará cómo se reunieron las muestras que conforman el corpus; a continuación, se describirán las construcciones nominales halladas en los textos que lo integran y se las dividirá en seis casos diferentes. Por último, se hará referencia brevemente a cuestiones teóricas no solo en lo que atañe a los SN del español, sino también en lo que respecta a la categoría gramatical indefinidos para poder establecer una clasificación sobre este tipo de determinantes. Esta acción permitirá crear las gramáticas específicas, de acuerdo a la posibilidad de combinación entre los distintos tipos de indefinidos que permitan reconocer automáticamente a los SNINT.

Descripción del Corpus

El análisis propuesto será realizado sobre dos muestras que suman un total de 3953 palabras. La primera, a la que denominaremos corpus A, está compuesta por 22 textos escritos y 1654 palabras de aprendientes brasileños de español que se hallan entre un nivel A1 y A2 de acuerdo con el *Marco Común Europeo de Referencia* (Consejo de Europa, 2002). Los sujetos son adultos, poseen una edad promedio de 40 años y se encuentran al inicio de carreras de Posgrado en la Universidad Nacional de Rosario. Cada texto responde a una de dos diferentes consignas propuestas por el docente, estas son:

- Escriba un texto comparando su ciudad natal con la ciudad de Rosario
- Responder las siguientes preguntas mirando un retrato del pintor Molina Campos: ¿Qué hay? ¿Cómo es? ¿Cómo están?

El segundo corpus, al que llamaremos B, está integrado por 12 textos que suman 2299 palabras. Los sujetos escogidos para esta muestra son jóvenes brasileños cuya edad promedio es de 20 años y se encuentran entre el nivel B1 y B2. Las producciones escritas responden a una situación de examen de proficiencia que realizan para obtener la certificación de ese nivel. Son estudiantes que están comenzando diferentes carreras de grado en facultades pertenecientes a la Universidad Nacional de Rosario.

En el corpus, aún sin discriminar el nivel de aprendizaje de los sujetos, hallamos diferentes estructuras que hacen referencia a un núcleo sustantivo. Las clasificamos en seis casos diferentes:

1. Empleo de artículo neutro lo en lugar del artículo masculino:

   a. lo desayuno
   b. lo problema
   c. lo cuidado
   d. lo equipaje

2. Asignación incorrecta de género:

   a. la título
   b. la paisaje
   c. el dirección
   d. el salud público

3. 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”

4. 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. la nuestra empresa
   f. del mi doctorado

5. Artículo seguido de pronombre posesivo más sustantivo común:

a. la suya ciudad Rosario
b. la mía ciudad natal

6. SN coincidentes con la lengua española:

a. todas las regiones
b. mis vecinos
c. los baños eléctricos
d. el artículo publicado

En esta oportunidad, nos ocuparemos de generar las gramáticas necesarias para rastrear las estructuras identificadas con los números 3, 4, 5 y 6. En cuanto a los dos primeros casos ya han sido trabajados anteriormente (Tramallino & Arnal, 2019).

La Presencia de Indefinidos en los Sintagmas Nominales

Definiremos el sintagma nominal (SN) como la estructura que corresponde a la unión de un determinante y un sustantivo que será el núcleo (Núcleo, N). No entraremos en discusión acerca de la extensión de los sintagmas, tema largamente trabajado desde distintas disciplinas, ya que sólo nos interesa reconocer las estructuras presentes en el corpus que reúnan dos determinantes para un solo núcleo; constituyente en el que acabará nuestro análisis. Es decir, no consideraremos los sintagmas nominales llamados extensos o largos (Quiroz Herrera, 2005); únicamente nos proponemos reconocer las construcciones encabezadas con determinantes que finalizan en el núcleo sustantivo.

Aclarada esta cuestión, nos detendremos en la clase de determinantes que modifican al núcleo sustantivo para explicar las relaciones de combinación y exclusión entre las subclases. Para definir la clase gramatical correspondiente a los determinantes nos basamos en el estudio realizado por Tomás Jiménez Juliá (2007), quien expone que son palabras gramaticalizadas que forman un paradigma encabezado por el artículo y seguido por una serie de unidades que provienen del inventario de adjetivos demostrativos

y posesivos latinos, así como de ciertos indefinidos de creación más reciente:

> El término ”determinante” designa una unidad gramatical, no un valor semántico, lo que implica una distinción entre “valor determinativo”, que puede expresarse a partir de varios recursos (adjetivos, estructuras relativas o preposicionales, plurales o singulares genéricos), y “determinante”, que es una clase definida (…) que tiene como causa fundamental de su existencia la expresión de la determinación. (p. 3)

Siguiendo al autor, al integrarse los antiguos determinativos definidos del latín, que son los demostrativos y posesivos, con las unidades indefinidas de nueva creación en el mismo paradigma general pasan a ser unidades sintagmáticamente equivalentes y, por lo tanto, “mutuamente excluyentes” (Jiménez Juliá, 2007, p. 3).

Agregaremos que los determinantes refieren al núcleo del SN. Dentro de estos se pueden distinguir, por un lado, a los artículos que permiten delimitar la denotación del grupo nominal e informan de su referencia y, por otro, a los posesivos y demostrativos. Además, se agrupa como tales a los numerales cardinales y a los indefinidos, para los cuales, a continuación, se propone una clasificación de acuerdo a las posibilidades de combinación que presentan.

Clasificación propuesta para indefinidos

Los determinantes indefinidos, de acuerdo a la definición de la Real Academia Española (RAE, 2010), se utilizan para señalar que lo designado por el grupo nominal no es identificable por el oyente, por ejemplo: “Un niño se asomó por la puerta”. Si la construcción se encuentra en singular recibe interpretación genérica: “Un buen maestro siempre explica muchas veces”.

No obstante, diversos autores (Alcina Claudet, 1999; Leonetti, 1999) complejizan la alternancia de los artículos definido e indefinido en función de la (in)especificidad semántico-pragmática del referente, discusión teórica en la que no se incursionará, ya que la finalidad de la investigación no es analizar las causas que llevaron a hacer determinado uso de indefinidos a los aprendices de espa-

ñol sino poder detectar automáticamente dichas construcciones. Por ese motivo, tampoco se mencionan las distintas clasificaciones que se han realizado desde la lingüística tradicional.

En consecuencia, se propone una clasificación propia para los indefinidos que consiste en dividirlos en cuatro tipos, cuyos dos primeros grupos se subdividen. Esta categorización responde a un criterio metodológico que se realiza a los fines de poder establecer un diccionario específico de determinantes del español en NooJ. Este diccionario, a su vez, permitirá confeccionar las "gramáticas sintácticas productivas" (de las que nos ocuparemos más adelante), capaces de reconocer automáticamente las mencionadas construcciones propias de interlengua, y poder distinguirlas de las pertenecientes al español. En el cuadro que sigue se expone la taxonomía sugerida:

Cuadro N° 1. Clasificación de indefinidos

INDEFINIDOS 1		
Excluyen artículos, demostrativos y posesivos. Se combinan sólo con los indefinidos del tipo 2, precediéndolos.		
INDEFINIDOS 1 a	INDEFINIDOS 1 b	INDEFINIDOS 1 c
Preceden a los indefinidos 2b	Preceden a indefinidos 2a y 2b	Excluyen a indefinidos 2b y 3, preceden sólo a indefinidos 2a.
ningún / ninguna algún / alguna	algunas / algunos	un / una unos / unas

INDEFINIDOS 2	
Admiten a artículos, demostrativos y posesivos.	
INDEFINIDOS 2 a	INDEFINIDOS 2 b
Excluyen a indefinidos 1a	Se combinan con indef1a e indef1b. Preceden a indef2a pero no pueden preceder a los indefinidos 1a y 1b sino que van pospuestos.
poco / poca pocos / pocas	otro / otra otros / otras

INDEFINIDOS 3	INDEFINIDOS 4
Preceden a artículos, posesivos y demostrativos. Excluyen a indefinidos 1 y 2	No pueden combinarse con artículos, posesivos y demostrativos. Excluyen a todos los indefinidos.
Todo / toda Todos / todas	mucho / mucha muchos / muchas varios / varias demasiado / demasiada demasiados / demasiadas bastante / bastantes ciertos / ciertas sendos / sendas.

A partir de esta categorización, las estructuras posibles de los SN, según las distintas combinaciones, son las siguientes:

1. [Indef 1a + Indef 2b +N] (“una red de la que ninguna otra lengua del mundo dispone”)[3]
2. [Indef 1b + Indef 2a +N] (“con excepción de algunas otras poesías”)
3. [Indef 1b + Indef 2b +N] (“algunas pocas décadas más tarde”)[4] d) [Indef 1c + Indef 2a +N] (“Solo en unos pocos casos”)
4. [DET Art/Dem/Pos + Indef 2 (a / b) +N] (“los otros datos de interés”)[5]
5. Indef 3 + DET Art/ Pos / Dem + N] (“todas las mujeres”)

Herramienta informática: sistema NooJ

El sistema NooJ es un programa de libre acceso creado en 2002 por Max Silberztein, diseñado para el tratamiento de más de veinte lenguas naturales; no solo realiza el análisis morfológico, sintáctico y semántico de textos sino que también permite la extracción de información en grandes corpus. Cabe aclarar que requiere de una formalización lingüística por parte de los usuarios y que previamente hubo una labor de equipo que permitió llevar a cabo la implantación en máquina correspondiente al módulo español de Argentina (Tramallino, 2013).

Tanto para que el sistema reconozca, por ejemplo, los nombres en un texto, como para que analice morfológicamente enunciados, es necesario ingresar tanto los ítems léxicos como los modelos de flexión y categorías con las que deberá operar el software. Estos se consignan en tres archivos de diferente extensión: los modelos se ubican en el archivo *Gramática* (“Grammar”), que posee extensión

---

[3] Los ejemplos fueron extraídos del corpus CREA disponible on line.

[4] Ejemplo extraído Pereda, F. (2007). *Las imágenes de la discordia: política y poética de la imagen sagrada de la España del 400*. Madrid: Marcial Pons. (p. 142.)

[5] Con respecto a los elementos que se ubican después de estas construcciones, de interés, este sintagma será reconocido por las gramáticas sintácticas creadas para los Sintagmas Preposicionales del español. Sería factible generar una gramática amplia que considerara a los SP que funcionan como complementos pero no lo creemos productivo a los fines de los alcances de este trabajo.

.nog; los lemas se ingresan en el archivo *Diccionario* (“Dictionary”), con extensión .dic; por último, las categorías gramaticales con sus rasgos, por ejemplo, Nombre: género/ número, en el archivo *Propiedades* (“Properties ´ definition”), que tiene extensión .def.

Los diccionarios asocian palabras o expresiones a determinada categoría. Por ejemplo, los determinantes están relacionados a paradigmas de inflexiones, constituidos por los modelos flexivos correspondientes a género y número.

El recuadro que sigue contiene un fragmento del diccionario nombrado `determinantes-cuantificadores´

Cuadro N°2. Diccionario determinantes- cuantificadores

la,DET+fem+sg
los,DET+masc+pl
las,DET+fem+pl
poco,CUANT+indef
mucho,CUANT+indef
demasiado,CUANT+indef

En cuanto a las gramáticas, estas se usan para representar fenómenos lingüísticos, desde niveles ortográficos y morfológicos a niveles sintagmáticos. Dentro de los tipos de gramáticas, se emplearán las sintácticas.

Generación de Gramáticas Sintácticas Productivas

Este software incluye, además, herramientas para crear y mantener fuentes lexicales, así como gramáticas sintácticas y morfológicas. Para poder crear las gramáticas sintácticas capaces de reconocer las estructuras de interlengua, primero se debe consig-

nar la clasificación de indefinidos propuesta en el archivo correspondiente a las propiedades, de la siguiente manera:

CUANT_clase = def | indef | indef1a | indef1b | indef1c | indef2a | indef2b | indef3 | indef4 | num;

A continuación, se confeccionan tres *gramáticas sintácticas productivas* con el objeto de reconocer automáticamente a los sintagmas nominales no coincidentes con el español, agrupados en los casos N° 3, 4 y 5. Siguiendo a Silberztein (2003), son *Productivas*, porque se emplean categorías, por ejemplo: DET para "determinante"; N para "nombre"; ADJ para "adjetivo" y se denominan *Sintácticas* porque agrupan dos o más palabras y luego las etiquetan según la indicación del usuario. Por ejemplo: SN para "Sintagma Nominal".

El programa proporciona el nodo inicial que se llenará con la etiqueta que se le quiera dar y el nodo final que cerrará la estructura. Por lo tanto, para crear una gramática del sintagma nominal perteneciente al español, se pueden crear nodos de determinantes y nombres con sus rasgos de género y número y unirlos estableciendo la concordancia entre ellos.

Para detectar a los sintagmas nominales de interlengua (SNINT), cuya estructura está formada por [Indef 3 + N], (ej: "todas notas"), se creó la gramática 1 como se observa en la siguiente imagen:

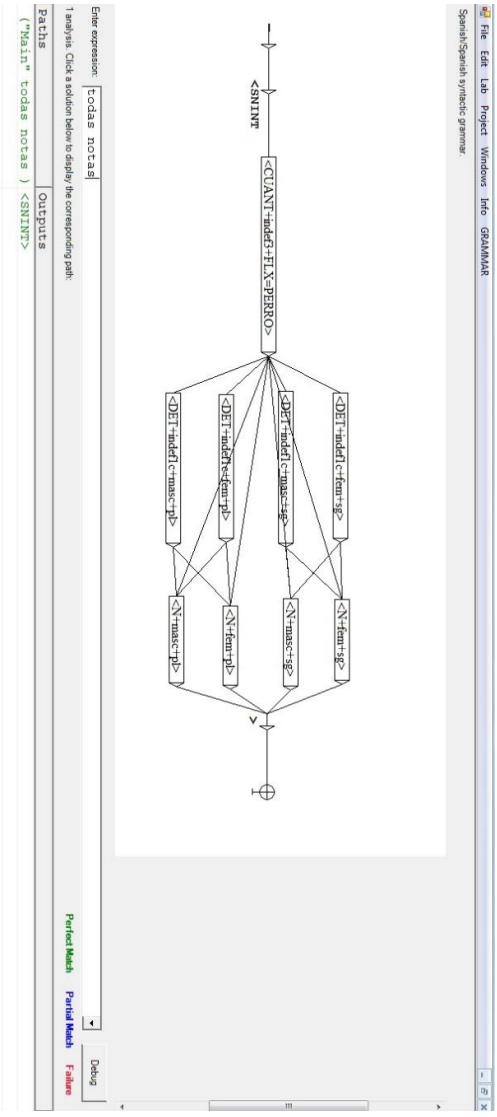


Figura 1. Gramática 1 [Captura de pantalla]

Para identificar los SNINT, cuya estructura está formada por [DET art + DET posesivo + N] (ej: “la mi nueva clase”), se realizó la gramática 2:

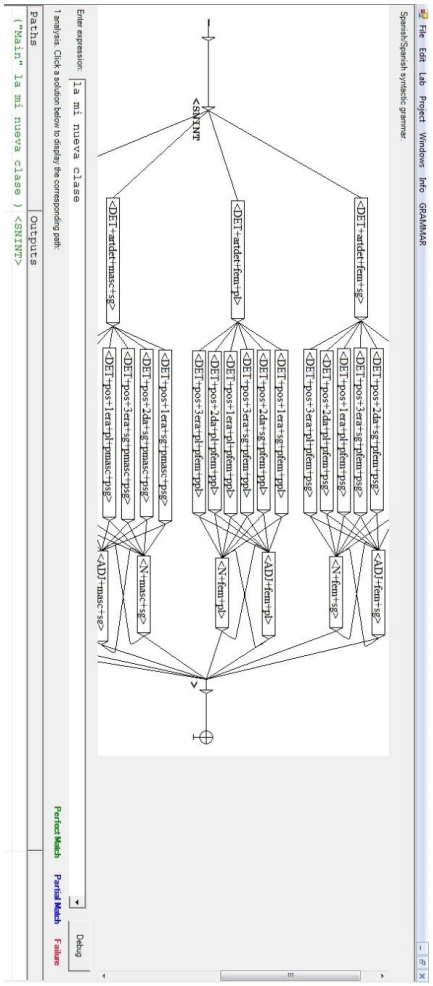


Figura 2. Gramática 2 [Captura de pantalla]

Por último, para identificar a las construcciones formadas por [DET art + PRON posesivo + N] (ej.: “la mía ciudad natal”), se confeccionó la gramática 3 como se exhibe en la imagen que sigue:

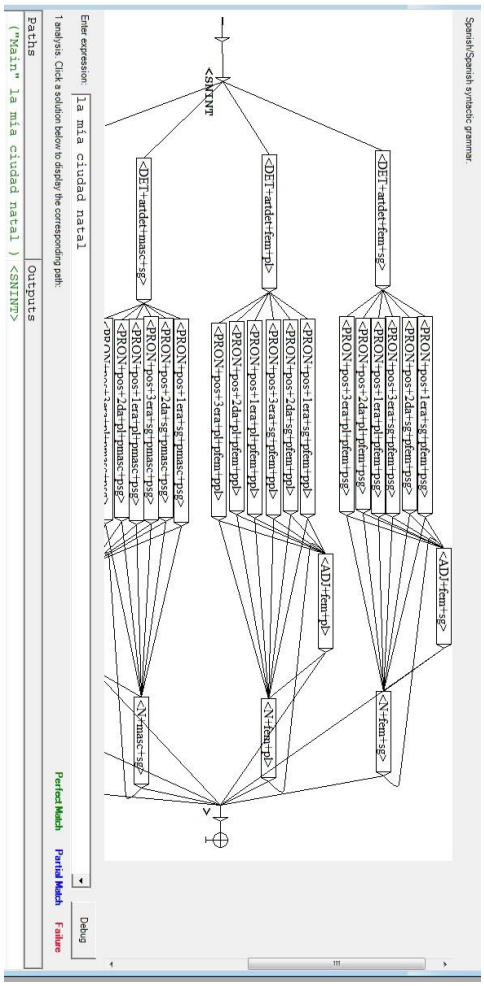


Figura 3. Gramática 3 [Captura de pantalla]

Es importante mencionar que, para localizar y contabilizar los SN de español, hubo que generar las gramáticas que siguen:

- Gramática del SN para sintagmas nominales encabezados con artículos, posesivos o demostrativos (SN):
  DET (ARTÍCULO / POSESIVO / DEMOSTRATIVO) + NOMBRE (ejemplo: “los hombres”)
  DET (ARTÍCULO / POSESIVO / DEMOSTRATIVO) + NOMBRE + ADJETIVO (ejemplo: “la ciudad linda”)
  DET (ARTÍCULO / POSESIVO / DEMOSTRATIVO) +ADJETIVO + NOMBRE (ejemplo: “el mismo edificio”)
- Gramática del SN con el cuantificador “todo/a/s”, que permite identificar los sintagmas nominales conformados por la estructura:
  INDEF 3 + DET (ARTÍCULO / POSESIVO / DEMOSTRATIVO) + NOMBRE (ejemplo: “todo el año”)
- Gramática del SN con pronombre posesivo, que identifica los sintagmas nominales cuya estructura puede ser:
  DET (ARTÍCULO) + NOMBRE + PRONOMBRE POSESIVO (ejemplo: “La ciudad mía”)
  DET (ARTÍCULO) + ADJETIVO + NOMBRE + PRONOMBRE POSESIVO (ejemplo: “La enorme ciudad mía”)

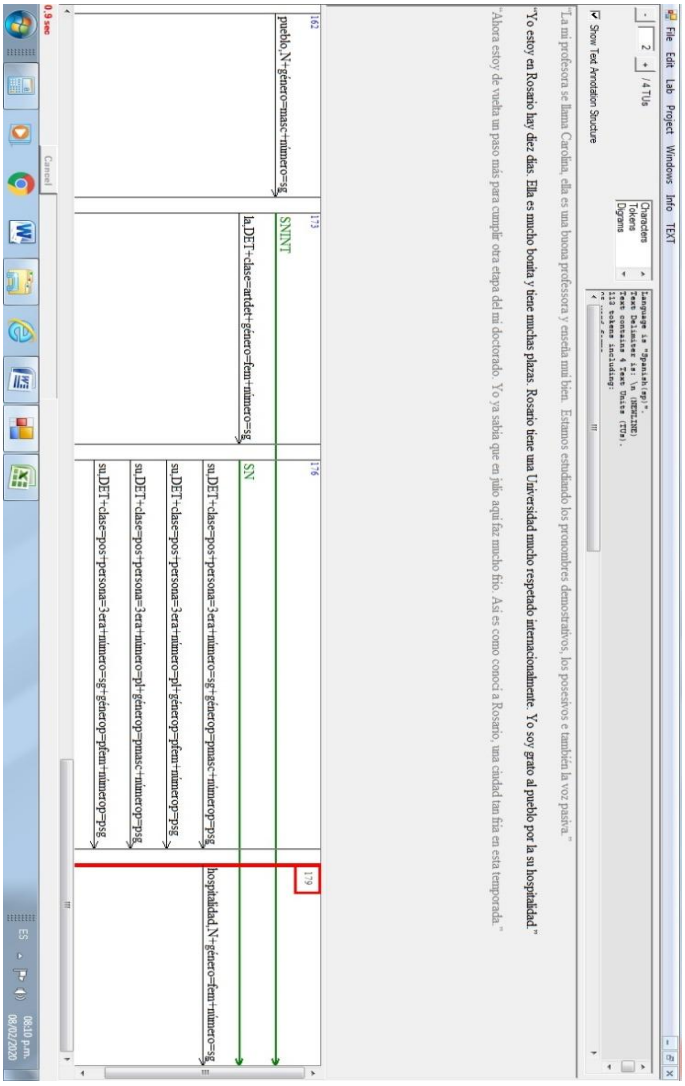
Reconocimiento Automático

A modo de ejemplo, se presenta el análisis realizado a partir de una producción perteneciente al corpus A en la captura de pantalla que se ubica a continuación:

Cuadro N° 3. producción textual del corpus A

"*La mi profesora* se llama Carolina, ella es *una buona professora* y enseña mui bien. Estamos estudiando *los pronombres demostrativos,* los posesivos e *también la voz pasiva."* "Yo estoy en Rosario hay *diez días*. Ella es mucho bonita y tiene *muchas plazas.* Rosario tiene una Universidad mucho respetado internacionalmente. Yo soy grato al pueblo por *la su hospitalidad*." "Ahora estoy de vuelta un paso más para cumplir *otra etapa del mi doctorado*. Yo ya sabía que en julio aquí faz mucho frío. Así es como conocí a Rosario, *una ciudad tan fría* en *esta temporada*."

Figura 4. Reconocimiento automático de SN y SNINT [Captura de pantalla]



Se observa el análisis automático morfológico realizado para cada palabra: se consigna en mayúscula la categoría gramatical y en minúscula, los rasgos flexionales. Además, marcado a través de flechas de color verde, se exhibe el análisis sintagmático, que diferencia los dos tipos de sintagmas a través de las etiquetas asignadas. Cabe aclarar que, previamente deben seleccionarse tanto los diccionarios como las gramáticas sintácticas con las que el programa analizará el texto.

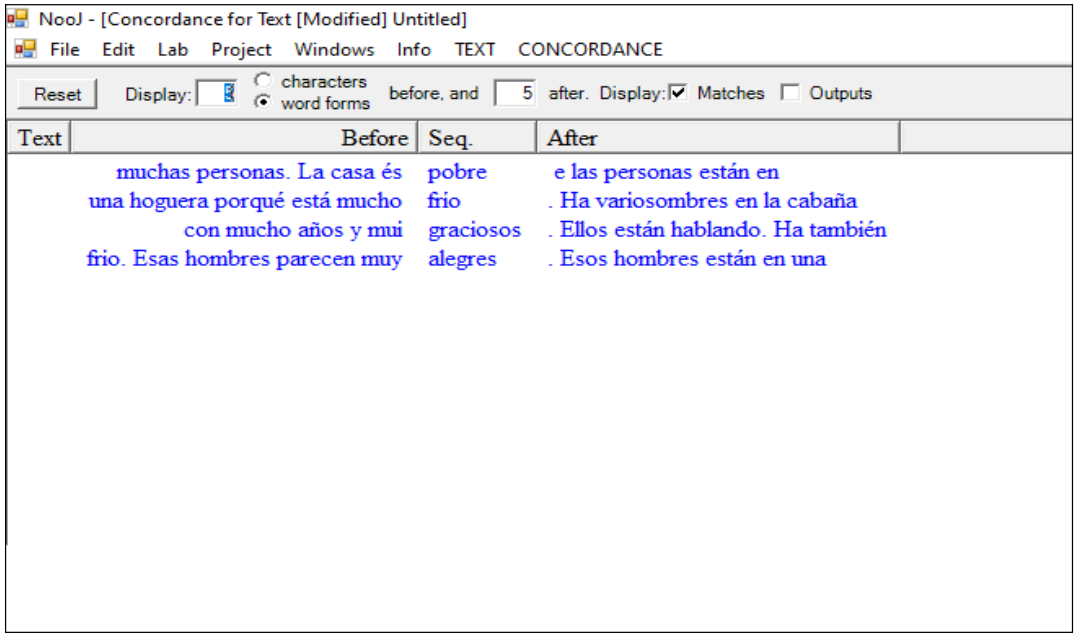
Localización y Contabilización de SN y SNINT

Este programa informático permite, también, localizar expresiones, terminaciones o estructuras presentes en un texto o en grandes corpus de textos, mostrando el fragmento anterior y el posterior. Además, puede contabilizarlas y arrojar la cantidad de dicha búsqueda en 100 ocurrencias o en todas, mediante la pestaña Locate. No solo eso, también ofrece la posibilidad de buscar todas las variantes morfológicas de un lema; por ejemplo, si se quiere localizar las formas flexionadas de un verbo se debe ingresar el verbo en infinitivo.

Asimismo, permite hallar palabras a partir de su categoría gramatical. Por ejemplo, para detectar los adjetivos, se escribirá ADJ y el programa rastreará en el texto todos los adjetivos que se encuentren declarados en el diccionario correspondiente.

La captura de pantalla que sigue muestra la localización de adjetivos en una producción del corpus A:

Figura 5. Localización de categorías gramaticales [Captura de pantalla]



Ahora bien, para contabilizar la cantidad de SN presentes en ambas muestras y la cantidad de SNINT, se deben seleccionar, a través de la opción *'a Nooj grammar´* del menú, las gramáticas confeccionadas para ese fin, de a una por vez. En las siguientes capturas de pantalla se exhiben algunos de los resultados obtenidos:

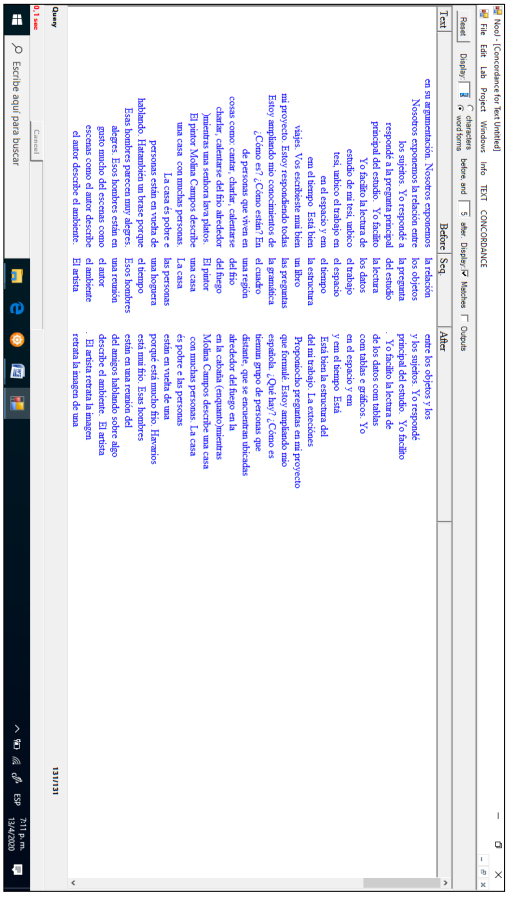


Figura 6. Reconocimiento de SN en el corpus A [Captura de pantalla]

Figura 7. Reconocimiento automático de SN en el corpus B [Captura de pantalla]

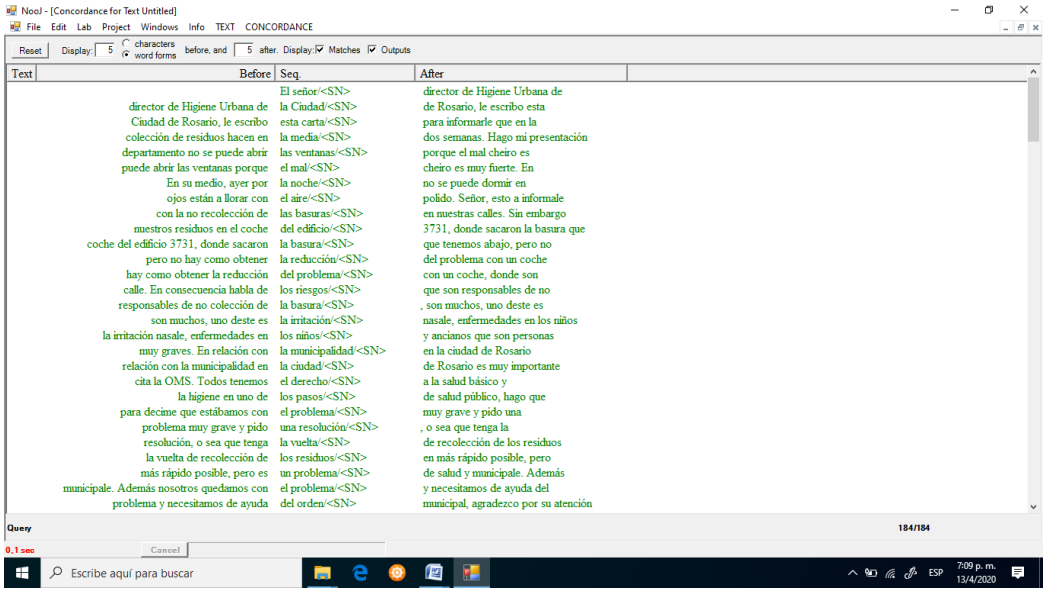
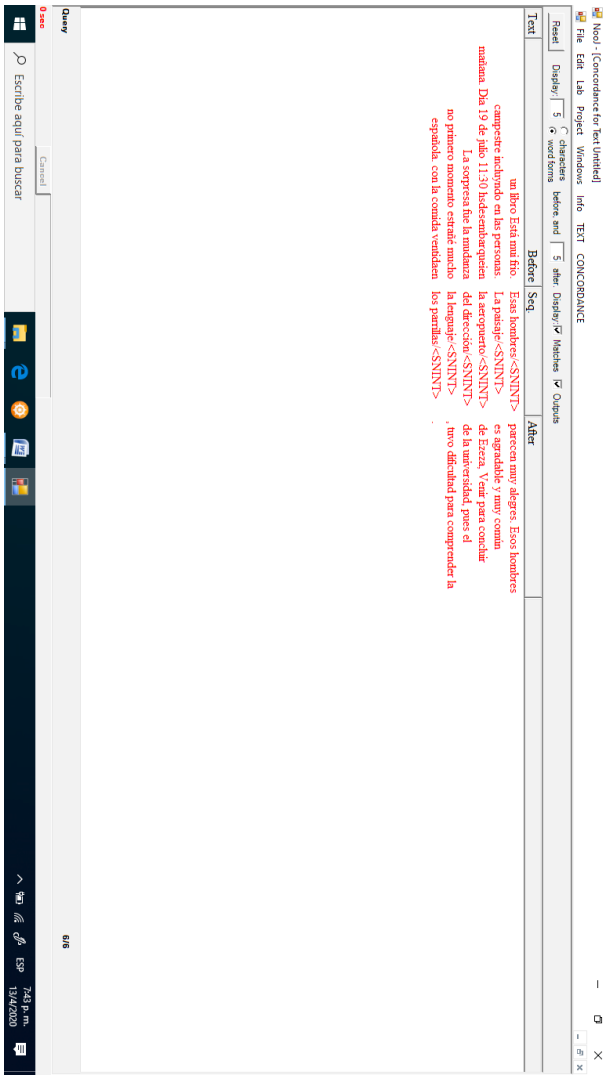


Figura 8. Reconocimiento automático de SNINT (asignación incorrecta de género) en corpus A [Captura de pantalla]



En la parte superior de la pantalla se selecciona la cantidad de palabras o caracteres del contexto anterior o posterior que se pretende visualizar (en este caso se colocó el número cinco). En la parte inferior derecha de la pantalla se muestran las coincidencias de la búsqueda.

Resultados

Del análisis de ambos corpus, resultó que en los textos de la muestra A, que pertenecían a aprendices de nivel A y sumaban un total de 1654 palabras, el programa contabilizó 131 SN; en cuanto a las producciones de la muestra B, correspondiente a estudiantes de nivel B, de un total de 2299 palabras, detectó 197 SN de español. Cabe aclarar que, el programa no reconoció sintagmas que incluyeran sustantivos con errores ortográficos, como por ejemplo: *una question, mi tesi, la única muyer, las clases*, ya que estas palabras no están declaradas en los diccionarios; este factor incide también en el porcentaje de SN coincidentes con la lengua española.

Con respecto a los SNINT encabezados por `todos´, compuestos por dos determinantes y un núcleo sustantivo y por un determinante seguido de pronombre, en el corpus A se detectó un porcentaje del 12 % mientras que en el corpus B no se halló ninguno. En cuanto a los sintagmas en donde se emplea el artículo neutro lo en lugar de artículo masculino singular, se halló una cantidad similar de casos en cada una de las muestras, presentándose casi en la misma proporción: 6% para el corpus A frente a 4 % en el corpus B.

Respecto a la asignación incorrecta de género de los nombres, en el corpus A representó un 11,5 %, a diferencia del corpus B en que sólo apareció en un 2,5 % del total de los sintagmas.

A continuación, se disponen todos los resultados en un cuadro con los porcentajes respecto del total de SN y de SNINT y, a la vez, se discrimina dentro de esta categoría a qué tipo de estructura corresponden:

Cuadro 4. Porcentajes de SN y SNINT

MUESTRAS	*Lo* en lugar de artículo masc. (caso 1)	Asignación incorrecta GÉNERO (caso 2)	Casos 3, 4 y 5	Total SNINT	SN	Total PALABRAS
Corpus A	6 %	11,5 %	12%	29,5 %	70,5 %	1654
Corpus B	4, %	2,5%	0%	6,5 %	93,5 %	2299

Consideraciones Finales y Discusiones

En este estudio se hallaron sintagmas nominales en las muestras A y B de producciones escritas que conforman el corpus de trabajo. Estas construcciones presentaban divergencias respecto a las construcciones propias de la lengua meta. Luego de establecer que dichas anomalías residían en la elección y sobre todo combinación de los determinantes, se especificó a estos como artículos, posesivos, demostrativos e indefinidos y se propuso una clasificación para los últimos, de acuerdo a las posibles combinaciones entre ellos y con otros determinantes. A continuación, se declaró dicha distinción en los diccionarios del Sistema NooJ vinculados a las gramáticas flexivas según el modelo derivacional, para los rasgos de género y número.

Se agrupó a las estructuras halladas en los textos de estudiantes, cuya lengua materna es el portugués, en seis casos: 1) empleo de artículo neutro *lo* en lugar de artículo masculino, 2) asignación incorrecta de género, 3) ausencia de artículos en construcciones encabezadas por `todos´ 4) artículo seguido de posesivo más sustantivo común, 5) artículo seguido de pronombre posesivo más sustantivo común; 6) SN coincidentes con la lengua española.

Para lograr reconocer automáticamente a los últimos cuatro tipos de construcciones se confeccionaron gramáticas sintácticas mediante grafos que pudieran identificarlos. Para los casos N° 3, 4 y 5 mediante la etiqueta SNINT y para los N° 6, con la etiqueta SN.

A continuación, se analizaron las dos muestras del corpus y se localizaron todas las estructuras mencionadas. Como resultados, se obtuvo que para los SN de español, se detectó un 70,5 % en el corpus de nivel inicial, frente a un 93,5% en el corpus de nivel intermedio, estableciendo un 29,5 % de SNINT en la primera muestra contra el 6,5% de SNINT en la segunda. Además, el grupo A presentó un 12 % de construcciones correspondientes a los casos 4, 5 y 6, que fueron reconocidos mediante las gramáticas sintácticas y etiquetados como SNINT, mientras que en el corpus B no se halló ninguna estructura de este tipo.

Con respecto a las asignaciones equivocadas de género del sustantivo, que determina una elección inadecuada del artículo (por ejemplo: *la paisaje*), en el corpus de nivel inicial representa el 11,5% de los sintagmas mientras que en el intermedio, sólo el 2,5% de ellos, es decir, que estas construcciones disminuyen notablemente de un nivel a otro. Por lo tanto, podemos arriesgar que un mayor aprendizaje léxico de la lengua meta influiría de forma positiva para hacer desaparecer dicha dificultad que se manifiesta en el corpus de nivel inicial, dada por el desconocimiento del género que posee el sustantivo.

La mayor diferencia, por lo tanto, está dada en los tipos de estructuras identificadas como SNINT, que reúne a los casos en donde hay ausencia de artículo en construcciones encabezadas por `todos´ (ej: *todos aspectos*), artículo seguido de posesivo más sustantivo común (ej: *la mi profesora*) y artículo seguido de pronombre posesivo más sustantivo común (ej: *la mía ciudad*), que no se hallan en la muestra de nivel intermedio.

Por el contrario, percibimos que la estructura que se genera casi en la misma proporción en los textos de aprendientes de nivel A1, A2 y B1, B2; es la utilización del artículo neutro *lo* en lugar del artículo masculino *el* para los sintagmas que poseen un sustantivo

de género masculino en singular.[6] A modo de discusión, esperamos poder ampliar el corpus de trabajo y analizar producciones de estudiantes brasileños que se encuentren en niveles avanzados de aprendizaje, para poder establecer si las construcciones con asignación incorrecta de género o con dicha sustitución del artículo masculino singular, propias de los primeros estadios en la adquisición del español como segunda lengua, aparecen y en qué medida o si, por el contrario, se han superado.

En síntesis, en este estudio se mostró cómo reconocer automáticamente sintagmas nominales propios de la interlengua de aprendices que tienen el portugués como lengua materna. De esta forma, pretendemos que signifique una pequeña contribución al campo de la enseñanza del español como L2, al proponer el empleo del software como recurso didáctico.

Referencias

Alba Quiñones, V. de. (2009). El análisis de errores en el campo del español como lengua extranjera. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas, 3*(5), 1-16. https://doi.org/10.26378/rnlael35103

Alcina Claudet, A. (1999). *Las expresiones referenciales. Estudio semántico del sintagma nominal* (Tesis doctoral). Universidad de Valencia, Valencia.

Alexopolou, A. (2010). La función de la interlengua en el aprendizaje de lenguas extranjeras. *Revista Nebrija De Lingüística Aplicada a La Enseñanza De Lenguas, 5*(9), 86-101. Recuperado de https://revistas.nebrija.com/revista-linguistica/article/view/157

Fernández, S. (1997). *Interlengua y análisis de errores en el aprendizaje de español como lengua extranjera*. Madrid: Edelsa.

Ferreira, A. , Elejalde, J., & Vine, A. (2014). Análisis de Errores Asistido por Computador basado en un Corpus de Aprendientes de Español como Lengua Extranjera. *Revista signos, 47*(86), 385-411. https://dx.doi.org/10.4067/S0718-09342014000300003

[6] Es importante señalar que la palabra *lo* coincide con dos categorías del español: con el clítico acusativo masculino singular y con el artículo neutro. En el sistema NooJ se encuentra declarado con esas dos entradas. Para poder desambiguar los resultados que arroja deben crearse gramáticas sintácticas.

Jiménez Juliá, T. (2007). *Aspectos gramaticales de la frase nominal en español*. Santiago de Compostela: Universidade de Santiago de Compostela.

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

Leonetti, M. (1999). *Los determinantes*. Madrid: Arco/Libros.

Liceras, J. (Ed.). (1992). *La adquisición de las lenguas extranjeras. Hacia un modelo de análisis de la interlengua*. Madrid: Visor.

Liceras, J. (2009). La interlengua del español en el siglo XXI. *Revista Nebrija de Lingüística Aplicada, 3*(5), 36-49. https://doi.org/10.26378/rnlael35107

Consejo de Europa (2002). *Marco Común de Referencia Europeo (MCRE)*. Madrid: Instituto Cervantes. Recuperado de http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf

Nemser, W. (1971). Approximative Systems of Foreign Language Learners. En J. C. Richards (Ed.), *Error Analysis* (pp.55-63). Londres: Longman.

Parodi, G. (2004). Textos de especialidad y comunidades discursivas técnico-profesionales: una aproximación basada en corpus computarizado. *Revista Estudios filológicos, 39*, 7-36. Recuperado de http://revistas.uach.cl/html/efilolo/n39/body/art01.html

Quiroz Herrera, G. (2005). *Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma*. Recuperado de https://repositori.upf.edu/bitstream/handle/10230/1310/05mon010.pdf?sequence=1&isAllowed=y

RAE (2010). *Nueva gramática de la lengua española*. Buenos Aires: ESPASA.

Rigamonti, D. (2006). *Problemas de lingüística de la adquisición y enseñanza del E/ELE a italófonos*. Recuperado de https://www.ledonline.it/ledonline/rigamonti/Rigamontiproblemas.pdf

Selinker, L. (1972). Interlanguage. *IRAL, 3,* 209-231.

Silberztein, M. (2003). *NooJ Manual*. Recuperado de www.nooj4nlp.net

Tordera Yllescas, J. (2011). Puentes entre la lingüística computacional y la Psicolingüística. *Revista de Lingüística y Lenguas Aplicadas, 6*, 341-352. https://doi.org/10.4995/rlyla.2011.914

Torijano Pérez, J. A. (2008). El estudio de los determinantes en aprendices lusohablantes de español, DICENDA. *Cuadernos de Filología Hispánica, 26*, 235-257.

Tramallino, C. (2013). Análisis morfológico con herramientas informáticas. Reconocimiento de nombres en textos de español con el sistema Nooj. *Revista Lingüística y Literatura, 64*, 33-48. Recuperado de http://aprendeenlinea.udea.edu.co/revistas/index.php/lyl

Tramallino,C., & Arnal, R. (2019). Reconocimiento automático de sintagmas nominales en producciones escritas de aprendientes brasileños de español. *e-universitas, 22*, 1-9. Recuperado de http://www.e-universitas.edu.ar/index.php/journal/article/view/188/11-2-87

Valenzuela, P., & Toledo, G. (2019). Uso y adquisición de artículos en español como segunda lengua. *Logos (La Serena), 29*(2), 268-285. https://dx.doi.org/10.15443/rl2922