Propuesta recibida el 15 de octubre 2017 y aceptada para publicacin el 3 de diciembre 2017.

Detección automática de nombres eventivos no deverbales en castellano: un enfoque cuantitativo basado en corpus

Automatic detection of non-deverbal eventive nouns in Spanish: a quantitative, corpus-based approach

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Rebeca Soto
Pontificia Universidad Católica de Valparaíso
rebecasotoriveros@gmail.com

Karen Urrejola
Pontificia Universidad Católica de Chile
kuc@uc.cl

Resumen

Presentamos un estudio en el campo de la detección de nombres eventivos no deverbales, que son aquellos nombres que designan eventos pero que no han pasado por un proceso de derivación a partir de verbos, como *fiesta* o *cóctel*, y no presentan por ello las pistas morfológicas típicas de los nombres deverbales, como los afijos *-ción*, *-miento*, etc., por lo que son justamente los más difíciles de detectar.

En el presente artículo continuamos y extendemos el trabajo iniciado por Resnik (2010), quien ya ofrece pistas para la detección automática de este tipo de unidades léxicas. A las sugerencias de Resnik añadimos otras, entre ellas el análisis inductivo de corpus, analizando con qué tipo de palabras suele coocurrir el nombre eventivo, y utilizándolas como predictores de esta condición. Además, simplificamos considerablemente el algoritmo de detección y aplicamos los experimentos a un corpus de mayor tamaño, el EsTenTen (Kilgarriff & Renau, 2013), de más de 9 mil millones de palabras. Finalmente, presentamos los primeros resultados de nombres eventivos extraídos automáticamente, incluyendo numerosos no deverbales.

Palabras clave

análisis inductivo de corpus, lexicografía computacional, sustantivos eventivos no deverbales

Abstract

We present a study in the field of the automatic detection of non-deverbal eventive nouns, which are those nouns that designate events but have not experienced a process of derivation from verbs, such as *fiesta* ('party') or *cóctel* ('cocktail') and, for this reason, do not present the typical morphological features of deverbal nouns, such as *-ción*, *-miento*, and are therefore more difficult to detect.

In the present research we continue and extend the work initiated by Resnik (2010), who offers a number of cues for the detection of this type of lexical unit. We apply Resnik's ideas and we also add new ones, among them, the inductive analysis of the words that tend to co-occur with eventive nouns in corpora, in order to use them as predictors of this condition. Furthermore, we simplify the classification algorithm considerably, and we apply the experiments to a larger corpus, the EsTenTen (Kilgarriff & Renau, 2013), comprising more than 9 billion running words. Finally, we present the first results of the automatic extraction of eventive nouns from the corpus, among which we find plenty non-deverbal nouns.

Keywords

computacional lexicography, inductive corpus analysis, non-deverbal eventive nouns

1 Introducción

En el contexto de la clasificación de los tipos de nombres o sustantivos, en los últimos años ha resultado de interés la distinción entre los nombres que designan eventos (por ej. *ceremonia*) de aquellos que hacen referencia a entidades en lugar de eventos (ej. *silla*). Sin embargo, el examen de la bibliografía revela que el estudio de este fenómeno ha sido abordado principalmente desde una mirada teórica e introspectiva, y los criterios para la caracterización de los nombres eventivos que se han establecido en la investigación actual se han basado predominantemente en una lógica deductiva (Graña López, 1993; Bosque, 1999; De Miguel, 2006; Real Academia Española, 2010; Fábregas, 2010). En otras palabras, los investigadores, a partir de su conocimiento de la lengua,

DOI: 10.21814/lm.9.2.253

dan cuenta de las particularidades del comportamiento sintáctico-semántico de este tipo de sustantivos y establecen criterios para su identificación. Son todavía pocos los estudios que adoptan una mirada empírica o que intentan contrastar con el corpus las propiedades establecidas deductivamente, como es el caso de Resnik (2010).

A partir de esta constatación, el presente trabajo busca ampliar el estudio pionero de Resnik para aportar a la caracterización de nombres eventivos por medio del análisis de su comportamiento en un corpus, con el fin de establecer criterios que permitan su identificación de forma objetiva y sistemática. Creemos que avanzar desde la teoría y los métodos puramente introspectivos hacia un análisis empírico es un paso fundamental, ya que el corpus representa el uso efectivo que los hablantes hacen de la lengua.

Concretamente, aplicamos los criterios de detección aportados por Resnik e incluimos otros que obtuvimos de manera inductiva; es decir, planteamos un método mixto, combinando pistas inductivas con aquellas encontradas en la bibliografía. Además, hemos conseguido simplificar considerablemente el algoritmo de clasificación, desde un método computacionalmente intensivo como el aprendizaje automático a uno basado en simples cálculos de coocurrencia. Esta simplificación metodológica permite la aplicación a un mayor volumen de datos de manera más rápida, lo que nos permite aplicar el método al corpus EsTenTen, que supera los 9 mil millones de palabras, y obtener grandes cantidades de nombres eventivos con precisión suficiente para ser de utilidad práctica en el campo del análisis lexicográfico.

El artículo se estructura de la siguiente manera: en la sección 2 revisamos la caracterización de los nombres eventivos en español que se ha realizado en los últimos años. En la sección 3 presentamos nuestra metodología de trabajo, que consiste en un algoritmo de clasificación de nombres eventivos a partir del estudio de sus contextos de coocurrencia. En la sección 4, en tanto, presentamos los resultados de la aplicación de este método primero con un listado de 100 nombres eventivos (no deverbales) y 100 nombres no eventivos compilados previamente por Resnik, para luego ofrecer también los resultados de la aplicación del método al corpus EsTenTen. El resultado es evaluado de manera manual examinando una muestra de 400 candidatos.

Este artículo es acompañado además de un sitio web[1] en el que se ofrecen los resultados del análisis y todo el código fuente del proyecto. Pensamos que este algoritmo de clasificación y su implementación pueden ser reaprovechados para realizar otro tipo de clasificaciones en el campo de la lexicología y lexicografía computacional.

[1] http://www.tecling.com/neven

2 Marco teórico

Las clasificaciones principales aportadas por la gramática

La gramática tradicional ha clasificado las palabras en diferentes clases sintácticas: artículo, sustantivo, pronombre, verbo, adverbio, preposiciones y conjunciones (Real Academia Española, 2010). Dentro de los denominados sustantivos o nombres, se han distinguido diferentes tipos a partir de criterios diversos: contables e incontables, abstractos y concretos, comunes y propios, individuales o colectivos, etc. La clase de los sustantivos es heterogénea en cuanto al comportamiento sintáctico-semántico de las unidades léxicas que la conforman y, de este modo, constituye un área de interés para los investigadores caracterizarlas desde diversos enfoques.

Esta investigación en particular se centra en la diferencia de los nombres eventivos respecto de los no eventivos. Los primeros corresponden a "un tipo de sustantivos individuales (por tanto, contables) que no designan objetos físicos, sino acontecimientos o sucesos" (Bosque, 1999, p. 55), por ejemplo: *fiesta*. Los no eventivos, al contrario, designan entidades, contables y no contables, que no se corresponden con sucesos o acontecimientos, por ejemplo: *gato*.

Dentro de la categoría de los nombres eventivos es posible diferenciar, por un lado, entre aquellos que derivan de verbos, proceso que puede ser acusado por la presencia de un morfema nominal (ej. *inaugura-ción*) o puede no presentar dicha marca (ej. *desfile*), y, por otro lado, los que no provienen de un verbo (ej. *boda*) (Fábregas, 2010). A su vez, dentro de la clase de los sustantivos deverbales, algunos pueden denotar un acontecimiento (eventivos) o bien el resultado del proceso implicado en el mismo (resultativos) (Grimshaw, 1990; Pustejovsky, 1995; Picallo, 1999; Alexiadou, 2001; Alonso Ramos, 2004). Grimshaw (1990) señala que las nominalizaciones eventivas no son contables pero las resultativas sí, como se verá en detalle más adelante (Cuadro 1).

La investigación en este ámbito ha establecido diferencias en el comportamiento sintáctico de los nombres eventivos con el fin de caracterizar esta clase de sustantivos, que aunque no derive de un verbo, de todos modos expresa un evento (Fábregas, 2010). Esto los distingue de los nombres pu-

ramente designativos, ya que los eventivos tienen una capacidad predicativa (De Miguel, 2006) o estructura argumental (Grimshaw, 1990). Al respecto, De Miguel (2006) señala que los nombres eventivos suelen aparecer con verbos de soporte o de escaso contenido léxico como en *dar una cena* o *hacer una fiesta*, aunque no de forma exclusiva, ya que también son comunes construcciones como *dar un golpe* o *hacer un pastel*.

Bosque (1999) destaca que los nombres eventivos pueden ser sujeto de predicados como *tener lugar* y también complemento directo de verbos como *presenciar*. Por otro lado, y dado que poseen límites temporales, también se acompañan de verbos como *empezar* y *concluir* y aparecen como complemento preposicional de *durante*: *durante la clase/el eclipse/la ocupación alemana*; *antes* y *después* (o *tras*): *después de la cena, antes de la conferencia*.

En este contexto, sin embargo, pueden también producir una lectura eventiva nombres que en principio no son eventivos (como *después del último autobús* o *antes del cigarrillo*). De hecho, la lectura eventiva de nombres no eventivos no es infrecuente. El nombre *libro* también puede resultar ambiguo en enunciados que pueden admitir una interpretación eventiva (como en *empecé el libro esta tarde*) y una interpretación objetual (como *el libro está sobre la mesa*). Esta ambigüedad entre una lectura eventiva y una objetual en los sustantivos eventivos es sistemática y observable en muchos otros casos, tales como *cena* o *concierto*, y entraría en el ámbito de lo que Apresjan (1974) y Pustejovsky (1995) denominan polisemia sistemática o regular. En este caso particular, podríamos hablar de coerción de tipos, o bien de tipos complejos (dotted types) en la terminología de Pustejovsky (1995). En castellano, el estudio de estos casos ha sido desarrollado por Adelstein et al. (2012), quienes contrastan los usos locativos y eventivos que puede mostrar un mismo sustantivo.

El sustantivo eventivo no deverbal como una clase autónoma

Un precedente importante en el estudio de los nombres eventivos no deverbales en lengua castellana es el estudio de Resnik (2010). En este se presenta un panorama completo de la investigación en el ámbito hasta esa fecha y, además, ofrece una propuesta de caracterización que, si bien se basa en el método introspectivo, es contrastada con un análisis de corpus.

La autora propone que los nombres eventivos no deverbales serían una clase autónoma, con un comportamiento sintáctico distinto al de las otras categorías, como los eventivos deverbales, de proceso y de resultado, y, por supuesto, de los sustantivos no eventivos. Así, si bien los eventivos no deverbales se suelen clasificar dentro de las nominalizaciones resultativas (Grimshaw, 1990), la autora propone, a partir de distintas pruebas, que los no deverbales serían una clase diferente.

Resnik parte de las propuestas de Grimshaw (1990), quien distingue entre nominalizaciones eventivas y nominalizaciones resultativas en función de la presencia o ausencia de una estructura eventiva compleja, y Picallo (1999), que lleva la distinción de Grimshaw al plano sintáctico y afirma que las nominalizaciones eventivas se realizan con una construcción pasiva, mientras que las nominalizaciones resultativas lo hacen con una construcción activa. Picallo sostiene que las nominalizaciones eventivas/pasivas se distinguen de las resultativas/activas por la presencia de elementos como la expresión del agente, que aparece en un sintagma preposicional introducido con *(por parte) de* en las nominalizaciones eventivas pero con un nombre genitivo con *de* en las resultativas. Las nominalizaciones eventivas aparecerían así en función de sujeto de predicados como *tener lugar, durar u ocurrir*; las resultativas, en cambio, serán sujeto de predicados tales como *ser inconsistente, ser considerado incorrecto*, *ser publicado*, etc. En respuesta a esta idea, Resnik sostiene que "la interpretación de la diferencia entre lectura eventiva y lectura resultativa de las nominalizaciones en términos de construcción pasiva y construcción activa se vuelve extraña al incorporar el caso de la nominalización creada a partir de una base inacusativa: es cierto que se trata de una construcción sin argumento externo, con un tema como sujeto, pero está claro que no es una construcción pasiva (no se puede incluir un agente como adjunto) y en ese sentido sería más adecuado, en todo caso, hablar de las nominalizaciones eventivas en general como construcciones ergativas" (Resnik, 2010, p. 75).

Para Alexiadou (2001), en tanto, se distingue entre nominalizaciones y nombres no deverbales en función de la estructura morfológica. Los últimos, a su vez, se diferencian de los nombres resultativos en que no se interpretan como eventivos. Así, las nominalizaciones eventivas tendrían parte de la estructura funcional de los verbos, a diferencia de las nominalizaciones resultativas, que carecen de estas proyecciones verbales. Resnik (2010) adapta esta propuesta para los nombres eventivos simples que, si bien carecen de morfología verbal, sí tienen propiedades aspectuales.

Teniendo en cuenta las clases aspectuales de Vendler (1967) (estados, actividades, logros, realizaciones), Resnik también distingue clases aspectuales de nombres no deverbales en función de los modificadores que admiten. El caso del nombre *clase*, por ejemplo, en tanto que actividad, corresponde a un evento durativo, por tanto atélico, mientras que otros, como *accidente*, ya son un evento puntual y, por tanto, no admiten el modificador durativo. Esto, en definitiva, lleva a pensar que la categoría de aspecto léxico no es una propiedad intrínseca de las raíces verbales, sino una categoría funcional que puede aparecer tanto con núcleos verbales como nominales.

Así, a partir de la clasificación de los nombres eventivos en español en nominalizaciones eventivas, nominalizaciones resultativas y nombres eventivos no deverbales, Resnik se centra en el análisis de las propiedades de los eventivos no deverbales y demuestra que estos no son equiparables a las nominalizaciones resultativas, ya que tienen una estructura funcional específica que incluye propiedades aspectuales.

El Cuadro 1, adaptado de Bel et al. (2010), resume las propiedades léxico semánticas de los diferentes sustantivos descritos en este apartado. En este cuadro se muestran, por un lado, las clases distinguidas por Grimshaw (1990): los sustantivos no eventivos, los sustantivos eventivos de proceso y los sustantivos resultativos; y por otro, los sustantivos eventivos no deverbales como una clase autónoma, tal como propone Resnik (2010). De esta manera, es posible observar cómo los no deverbales presentan propiedades específicas en relación con los demás.

La detección automática de sustantivos eventivos

La dificultad principal de la detección automática de sustantivos eventivos es, como se ha explicado, que el nombre eventivo no deverbal no ofrece las marcas morfológicas que son propias de los eventivos deverbales, como *accidente* o *guerra*, y por tanto no pueden ser detectados con afijos como *-ción*, *-miento*, etc[2].

Resnik (2010) no recurre a pistas morfológicas porque trabaja específicamente con nombres eventivos no deverbales, y es, en cambio, el comportamiento sintáctico-semántico de estos nombres lo que permite su detección automática. Entre los rasgos predictores para esta clasificación, la autora destaca los siguientes:

- Los nombres son complementos de *durante, hasta el final de, desde el principio de*, entre otras
- Son argumento de verbos tales como *ocurrir, producir, celebrar* y verbos aspectuales como *empezar* o *durar*
- Admiten cuantificadores aspectuales como *dos días/semanas de* o *una etapa/un período de*, entre otros
- Predicados aspectuales como *ocurrir, producir, desatar, desencadenar, celebrar*
- Cláusulas sustantivas: proceso/hecho/actividad/evento de + construcción nominal
- Referencia anafórica con *esto*
- Selección de ser/estar: el verbo ser selecciona un evento como sujeto cuando presenta complementos locativos o temporales
- Paráfrasis con *hecho, actividad, evento*
- Argumento del verbo *presenciar*
- Complemento de preposiciones aspectuales (*en medio de, durante*)
- Modificación con adjetivos aspectuales
- Modificación con cláusula temporal al + infinitivo

Utilizando pistas de este tipo, Resnik (2010) intenta la clasificación de los nombres en las listas que aparecen en el Cuadro 2.

El corpus utilizado en su experimento está conformado por textos de prensa de El País y La Vanguardia con un total de 21 millones de palabras, parte del corpus técnico del IULA o CT-IULA, (Cabré et al., 2006). Utilizando este corpus y el software Weka (Hall et al., 2009), entrenó un clasificador del tipo árbol de decisión (Decision Tree) C4.5 (Quinlan, 1993). La clasificación se llevó a cabo por medio del "10-fold cross-validation", que implica repetir el experimento diez veces utilizando cada vez un 90 % del set como entrenamiento y un 10 % como test, garantizando que cada elemento haya estado en ambos conjuntos, es decir, en el entrenamiento y en el test.

Con este método la autora reporta una tasa de éxito importante para ser un estudio pionero en el área: una precisión de 0.84 y 0.82 clasificando los nombres eventivos y no eventivos, y una

[2] Esto no quiere decir que la detección de nombres eventivos mediante pistas morfológicas sea una tarea sencilla, como señalan Balvet et al. (2011): no todo nombre con *-ción* será eventivo: no lo es *población*, aunque sí *poblamiento*.

	Sustantivo Eventivo no-deverbal	**Sustantivos de Proceso**	**Sustantivos Resultativos**	**Sustantivos no-eventivos**
Ejemplo	*guerra*	*construcción = evento*	*construcción = objeto resultativo*	*mapa*
Argumento interno obligatorio	No	Sí	No	No
Realización del argumento externo	Genitivo frase determinante	Frase preposicional 'por'	genitivo frase determinante	genitivo frase determinante
Sujeto de verbo aspectual (comenzar, terminar)	Sí	Sí	No	No
Cuantificador aspectual (un periodo de)	Sí	Sí	No	No
Complemento de durante...	Sí	Sí	No	No
Contable/ /no-contable (determinantes, formas plurales)	contable/ /no-contable	no-contable	contable	contable/ /no-contable

Cuadro 1: Clasificación de sustantivos adaptado de Bel et al. (2010, p. 47)

cobertura de 0.82 y 0.84, respectivamente. La limitación está en el aspecto empírico, ya que se opera con un listado de 100 nombres eventivos y 100 no eventivos, y en un corpus de tamaño limitado.

En un estudio posterior, Bel et al. (2010) presentan el análisis de los sustantivos eventivos no deverbales y un experimento de detección automática para el inglés y el español. Allí reproducen los experimentos y se habla de un "accuracy" de 80 % para el castellano. Se presume que con ese término se refieren en realidad a la precisión, definida como (1) junto con la cobertura (2) (Baeza-Yates & Ribeiro-Neto, 1999), y no a la "accuracy" definida en (3), donde *tp* o *true positive* sería el sustantivo correctamente detectado como eventivo, el *fp* o *false positive* el sustantivo no eventivo incorrectamente seleccionado como eventivo, el *fn* o *false negative* el sustantivo eventivo no detectado y *tn* o *true negative*, el sustantivo no eventivo correctamente descartado.

$$precision = \frac{tp}{fp + tp} \tag{1}$$

$$recall = \frac{tp}{fn + tp} \tag{2}$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

En este segundo estudio, Bel et al. (2010) intentan elevar la precisión clasificando solo los elementos más seguros, y consiguen llegar a un 95 %, aunque entonces la cobertura disminuye a 43 %. Más allá de la tasa de éxito en la detección, estos dos trabajos son de importancia porque se propone una caracterización que, aunque basada en el método introspectivo, es contrastada con un análisis de corpus. Las limitaciones, sin embargo, son el tamaño de la muestra de sustantivos (Cuadro 2) y que el corpus con el que trabajan es de un tamaño muy reducido.

El trabajo que presentamos ahora comparte el objetivo de Resnik (2010) y Bel et al. (2010) en cuanto el relevamiento de características contextuales que permitan identificar de manera automática en un corpus sustantivos eventivos no deverbales y diferenciarlos de sustantivos no eventivos. Sin embargo, la metodología seguida es ahora un enfoque mixto con elementos que provienen del análisis de corpus basado en la observación del comportamiento de los sustantivos eventivos no deverbales en un corpus. En esta investigación sistematizamos y simplificamos el esquema de elementos predictores y utilizamos un

Categoría	Sustantivos
Eventivos	*fiesta, feria, festival, boda, funeral, velorio, velatorio, ceremonia, evento, picnic, cóctel, té, banquete, festín, ágape, tertulia, campaña, cónclave, cumbre, asamblea, sesión, misa, vacaciones, receso, excursión, trayecto, travesía, clase, conferencia, curso, taller, workshop, congreso, simposio, jornadas, tumulto, coloquio, entrevista, audiencia, concierto, ópera, serenata, espectáculo, show, programa, película, ciclo, discurso, sermón, torneo, campeonato, carrera, rally, tormenta, tempestad, temporal, borrasca, terremoto, sismo, huracán, maremoto, sequía, catástrofe, cataclismo, desastre, tragedia, holocausto, drama, incendio, accidente, impacto, siniestro, caos, crisis, guerra, batalla, conflicto, paz, silencio, ruido, escándalo, lío, follón, problema, motín, huelga, incidente, boicot, pánico, miedo, pasión, furor, rabia, siesta, frío, calor, hambre, pereza, dolor, fiebre, gripe*
No eventivos	*mapa, antología, característica, droga, plasma, teléfono, montaña, tubo, estética, cliente, escena, colectividad, canal, arquitectura, cara, levedad, estadio, batuta, súbdito, ciudad, madera, cifra, habitación, fotocopia, vivienda, gas, literatura, especie, paisaje, diferencia, carretera, seguridad, red, contraseña, rodilla, virus, cantidad, provincia, detalle, público, garganta, maqueta, dato, volcán, cárcel, familia, dinero, estereotipo, tarifa, compañía, justicia, humo, balneario, paquete, prensa, vehículo, dueño, prejuicio, banda, consorcio, economía, figura, mar, pancarta, grupo, arma, informe, diario, trama, zona, misterio, facultad, cadáver, nivel, pista, columna, combustible, estructura, ruta, alimento, herramienta, factura, miembro, forma, tema, fuente, temperatura, euro, ilusión, punto, batería, silueta, unidad, organismo, norma, vía, planta, autobús, perspectiva, antena*

Cuadro 2: Listados de sustantivos eventivos y no eventivos compilados por Resnik (2010)

algoritmo más sencillo que el árbol de decisión, lo que nos permite trabajar con un corpus de tamaño mucho mayor.

3 Materiales y Métodos

Para conseguir el objetivo de establecer criterios que permitan identificar de forma objetiva y sistemática los nombres eventivos, planteamos el problema como una tarea de clasificación y aplicamos para ello un método basado en corpus. En nuestro caso, este corpus es el EsTenTen (Kilgarriff & Renau, 2013), un corpus de gran tamaño (9 mil millones de palabras), constituido por páginas web de distintos países de habla castellana. Entendemos que este corpus cumple con la definición de Sinclair (1991): una colección de textos que manifiestan ocurrencias de lenguaje natural y que han sido escogidos para caracterizar un estado o variedad de lenguaje.

La idea principal

El examen de las concordancias de un sustantivo eventivo no deverbal como *fiesta* en el corpus EsTenTen revela rápidamente una serie de pistas sistemáticas ofrecidas por el contexto inmediato, infraoracional. El Cuadro 3 ofrece algunos ejemplos extraídos de este corpus.

Ya en el examen visual de estas concordancias, antes de aplicar ningún criterio de contabilización de frecuencia de aparición de las palabras del contexto (la metodología básica para este tipo de problema), podemos encontrar con facilidad una serie de elementos que son consistentes con la condición de eventivo del sustantivo analizado.

Como primera aproximación, lo que salta a la vista es que el sustantivo tiene más de un significado, ya que se utiliza para designar un tipo particular de automóvil, el *Ford Fiesta*. Difícilmente se puede hablar de un caso de polisemia en este caso, ya que no se puede confundir el sustantivo con la condición de nombre propio que tiene cuando se usa para designar el coche. Las concordancias con este uso aparecen, en el Cuadro 3, agrupadas en las líneas 1 a 8. En el resto, sin embargo, advertimos la interpretación eventiva y encontramos elementos como días de la semana, especificadores como *durante*, sustantivos utilizados para las medidas de tiempo: *hora, día, semana, año* y los adverbios utilizados para el orden secuencial: *antes* y *después*.

Estos elementos específicos aparecen con una alta frecuencia en los contextos de los nombres eventivos, lo que sugiere que puede ser

1	ejemplo , asique no se si Ford permitirá que el	Fiesta	supere al Focus en ese aspecto , o mejorarà el
2	ese aspecto , o mejorará el Focus cuando venga el	Fiesta	Un autazo A ver si entendi bien por
3	y pensar que van a hacerle un restyling al pedorrisimo	Fiesta	que venden aca no El fiesta actual es un
4	restyling al pedorrisimo Fiesta que venden aca no El	fiesta	actual es un desastre mi novia tenia uno y los
5	Hace meses que estoy esperando este	Fiesta	porque me estaba por comprar un Agile y despues de
6	jamas compraría un 207 c y si compraría un	Fiesta	KD si ahora estuviera al precio de lanzamiento ,
7	seras ignorante , no es para gente con familia el	fiesta	KD es para gente q le gustan los AUTOS ,
8	cosa , la clase de consumidor que apunta a un	Fiesta	KD dudo mucho que tenga en cuenta un Fluence
9	En esta ciudad , **se lleva a cabo la**	fiesta	en honor a su patrona , con actos litúrgicos ,
10	folclórico Herencia Gaucha brilló **el domingo** en la	Fiesta	del Gaucho Carlos Andina , el profesor que los
11	comienza en el año 1997 **El domingo** en la	Fiesta	del Gaucho vimos en su plenitud al ballet folclórico
12	lo que saben hacer Pero no es la única	fiesta	, **este año** fueron a competir a Berazategui y a
13	un sábado muy especial en la XXXIX edición de la	Fiesta	Nacional del Gaucho , con la incorporación en la
14	más destacados de la **jornada** con que **se inicia la**	fiesta	El clima acompaño la primera jornada de la Fiesta
15	que propone la fiesta**dieron inicio a la** XXXIX	Fiesta	Nacional del Gaucho Con el usual espectáculo
16	registran distintos **momentos de la**	fiesta	con las nuevas tecnologías , haciendo un uso diferencial
17	encender la vela correspondiente **durante la**	fiesta	La torta Existen en plaza muchos diseños Los
18	bar mitzvá , con cintas de acuerdo al color de	fiesta	Se puede contratar con el servicio de catering y
19	en exposición **durante las horas que dure la**	fiesta	El candelabro y encendido de velas Se contratará el
20	de anticipación para que esté disponible **el día de la**	fiesta	Se lo puede pedir decorado con un arreglo de
21	una carpeta forrada en raso (del tono de la	fiesta	) que puede estar bordado con hilos plateados
22	se definen una o **dos semanas antes de la**	fiesta	en una reunión con el DJ y el grupo familiar
23	, y no al revés **Después que pasó la**	fiesta	Todo salió perfecto , como lo habían soñado El
24	camino de los Picos de Europa donde **tiene lugar la**	Fiesta	, muy cerca del lago Enol , siendo la única

Cuadro 3: Ejemplos de concordancias del sustantivos *fiesta* en el corpus. Las líneas 1 a 8 representan usos del nombre propio que refiere al automóvil *Ford Fiesta*. En negrita los elementos que consideramos predictores de la condición de nombre eventivo en el caso de *fiesta*.

conveniente, en el sentido de seguir el principio de parsimonia, atender primero a este grupo de predictores y no otros que, si bien pueden ser también confiables, van a ocurrir con menor frecuencia. Procedemos de esta manera también porque si es posible obtener el resultado esperado despejando algunas variables del problema, ello puede ser también beneficioso desde el punto de vista computacional, porque se traduce directamente en mayor capacidad para procesar más texto en menor tiempo.

La intuición general es entonces que los sustantivos eventivos tenderán a coaparecer más frecuentemente con elementos predictores eventivos en comparación con los no eventivos. En este punto es importante recalcar que los elementos predictores no son verdaderos diagnósticos de eventividad. Esa es precisamente la diferencia entre el pensamiento cuantitativo y el simbólico o basado en reglas: la presencia de un predictor en un determinado contexto no indica que esa sea una ocurrencia concreta de un nombre eventivo. La que cuestión es que si analizamos una gran cantidad de contextos de aparición del sustantivo podemos determinar si está asociado o no con esos predictores. Estos elementos predictores se pueden compilar en un listado que luego se coteja con el vocabulario de los contextos de ocurrencia del sustantivo analizado, lo que lo hace un procedimiento bien sencillo.

Clasificación de pistas contextuales

Mediante el análisis de concordancias de nombres eventivos y no eventivos, procedimos a analizar y clasificar los distintos elementos léxicos del contexto por frecuencia decreciente, reteniendo aquellos que consideramos predictores confiables de la categoría de nombre eventivo. Esto significa que utilizamos solo rasgos positivos, es decir, solo rasgos que son indicativos de la clase de eventivos y no tenemos en cuenta características que serían propias únicamente de los no eventivos. Disponemos de cuatro categorías de rasgos:

1. **Días de la semana y meses** (*lunes, martes, miércoles, jueves, viernes, sábado, domingo, enero, febrero, marzo, abril, mayo, junio, julio, agosto, septiembre, octubre, noviembre, diciembre*)
2. **Medidas temporales** (*semana, día, mes, año, hora, minuto*)
3. **Verbos aspectuales** (*ocurrir, comenzar, iniciar, efectuar, celebrar, (hubo, hubieron, habrán*))
4. **Otros ítems léxicos aspectuales** (*durante, antes, después, duración, constante, menudo, frecuente, rápido, lento*)

Como se puede apreciar, llevamos a cabo una considerable simplificación de los rasgos clasificadores con respecto al trabajo de Resnik (2010),

coincidiendo también con una simplificación importante del algoritmo de clasificación, al estar basado únicamente en frecuencias de coocurrencia y no en aprendizaje automático.

El algoritmo de clasificación

Implementamos un algoritmo que acepta como entrada un conjunto de sustantivos, que denotaremos como $X = \{X_1, ...X_n\}$ y el resultado es un conjunto tal que $(\forall x \in X)\ E(x)$, donde $E(x)$ resulta en un valor o ponderación que servirá para ordenar el conjunto X en un listado, en función de la probabilidad del candidato x de ser un nombre eventivo.

El algoritmo está basado en el análisis de los contextos de aparición de los sustantivos analizados en un corpus de gran tamaño. Por tanto, por cada sustantivo analizado $x \in X$, extrae una muestra aleatoria de 5.000 contextos de aparición de x, de extensión oracional, y recorre estos contextos para encontrar ocurrencias a derecha o izquierda de los elementos predictores descritos en el apartado 3.2.

Podemos representar cada contexto de aparición de x a su vez como un conjunto $C_j = \{t_1, ...t_{|C_j|}\}$, es decir ignorando el orden de aparición. Cada unidad léxica $t \in C_j$ es una palabra apareciendo a derecha o izquierda del elemento analizado x.

Una vez clasificados los predictores o pistas contextuales, que definimos aquí como un conjunto L, el algoritmo de clasificación puede detectar aquellos sustantivos que muestren una alta proporción de estos elementos predictores en sus contextos, y acusarlos como candidatos a nombres eventivos.

Definimos una función que asigna un valor $E(x)$ en (4), que medirá la cantidad de apariciones de cualquier elemento predictor de la condición de nombre eventivo en los contextos de x.

$$E(x) = \sum_{j=1}^{|C|} \begin{cases} 1 & \exists t \in C_j \wedge t \in L \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

La medida $E(x)$ pondera entonces al candidato x para indicar la frecuencia relativa de dichos elementos predictores en sus contextos. En función de esta estimación, podemos exponer al candidato a un ordenamiento más cercano a las primeras posiciones, o bien directamente tomar una decisión binaria de tipo eventivo/no eventivo. Esto último también puede ser llevado a cabo por medio de la aplicación de una ordenación de todos los candidatos y la aplicación posterior de un umbral de corte arbitrario, pero también se puede aplicar una regla de eliminación como $K(x)$, como se indica en la ecuación (5) y eliminar, así, a los nombres simples. Esto disminuye drásticamente el tamaño de los listados resultantes, lo cual facilita su posterior examen y procesamiento computacional. Reservamos un valor "I" para los casos indefinidos, tal que el elemento no se puede clasificar debido a que no hay suficientes contextos de aparición, ignorándolo como un elemento no analizable si se encuentra por debajo un umbral de frecuencia arbitrario u.

$$K(x) = \begin{cases} \text{I} & |C| \leq u \\ \text{E} & E(x) > p \\ \text{N} & \text{otherwise} \end{cases} \tag{5}$$

Añadimos además un criterio penalizador consistente en determinar si en alguno de los parámetros 1 a 4, descritos en la sección 3.2 se encuentra que en total esas unidades aparecen en menos del 2 % de la muestra. Si este es el caso, entonces el sustantivo se considera no eventivo.

4 Resultados

Para nuestros experimentos, procedimos en primer lugar a reproducir la clasificación del conjunto de datos elaborado por Resnik (2010), presentado en el Cuadro 2.

Los resultados obtenidos en la prueba con la muestra de 200 sustantivos eventivos y no eventivos tomada de Resnik, con los parámetros definidos en la sección anterior, revelan una precisión de 95 %, cobertura de 63 % y F1 de 75, por tanto un aumento significativo de la cobertura –casi 20 puntos porcentuales– con respecto a Bel et al. (2010).

Luego, y con el objeto de superar una de las limitaciones del trabajo de Resnik, que era el trabajar con una muestra pequeña y no aleatoria, intentamos reproducir el experimento con una muestra de sustantivos sensiblemente más grande. Como ya hemos indicado, aprovechamos por un lado la simplificación de nuestra metodología, que no requiere la utilización de software de aprendizaje automático, y por otro lado el avance en materia de hardware de los últimos siete años, más la disponibilidad actual de un corpus de gran tamaño como el EsTenTen.

Como listado de sustantivos a analizar, tomamos 65.000 sustantivos de la taxonomía *open source* ofrecida por Nazar & Renau (2016). A partir de ese listado, obtuvimos un reordenamiento de los sustantivos de ese listado en función de la ponderación que recibieron con nuestra medida $E(x)$ como probablemente eventivos.

Candidato	Evaluación
duración	0
pascuilla	1
bisemana	1
día	1
semanada	1
mesta	1
triduo	1
anaplastia	1
chaquetía	1
ramadán	1
nisán	0
prejornada	1
interescuadra	0
madrigada	0
novenario	1
conmemoración	1
crismal	1
vendimiario	1
...	...

Cuadro 4: Algunos ejemplos de los sustantivos obtenidos y su evaluación

A partir de estos nuevos resultados, examinamos manualmente una muestra y encontramos que, como era de esperar, no todos son eventivos y de los eventivos no todos son no deverbales. La precisión es variable en función de la ponderación que recibieron, pero incluso cuando esta es alta, la tasa de error en la clasificación es mayor que la obtenida en el primer experimento con la lista de 200 unidades tomada de Resnik (2010).

El Cuadro 4 muestra algunos ejemplos de sustantivos en la muestra analizada. En el examen de esta muestra nos limitamos a evaluar la condición de eventivo del sustantivo, suspendiendo por el momento la distinción con el sustantivo deverbal. Esto es porque lo que nos interesa evaluar de momento es el hecho de que el sustantivo ha sido acusado como eventivo por los elementos de su contexto, y no porque hayamos tenido en cuenta aspectos morfológicos. La morfología del castellano ofrece la posibilidad de detectar (y, si cabe, eliminar) los sustantivos deverbales, ya que los morfemas que indican tal condición pertenecen a una categoría cerrada.

Examinando los resultados encontramos casos de nombres eventivos deverbales, como *conmemoración*, casos indiscutibles de nombres eventivos como *ramadán, crimal* o *pascuilla* (cabe destacar la alta contribución de sustantivos eventivos de los diferentes ritos religiosos) y otros que, a pesar de su morfología, como *duración*, no pueden ser considerados sustantivos eventivos. En el caso de este error, la explicación es sencilla: el sustantivo *duración* también aparece acompañado, en gran medida, de aquellos elementos que consideramos predictores en el apartado 3.2. Otros casos son más discutibles. En el caso de *interesescuadra*, su presencia allí se debe a su uso como *torneo interesescuadra*, que es terminología perteneciente al campo del deporte, y que en ese sentido, sí puede ser leído como eventivo. Pero creemos que no podemos considerarlo, en forma aislada, como un sustantivo eventivo, al menos desde un criterio lexicográfico.

Para tener una medida más precisa de la calidad general de los resultados obtenidos en el segundo experimento, la Figura 1 muestra la precisión acumulada. Allí, el eje vertical presenta el porcentaje de precisión y en el horizontal están los candidatos ordenados según la ponderación dada por el algoritmo. Tal como cabía esperar, la precisión disminuye a medida que se consideran más candidatos. La pendiente de ese descenso no es excesivamente acusada, pero permite prever una tasa de disminución del desempeño bastante significativa.

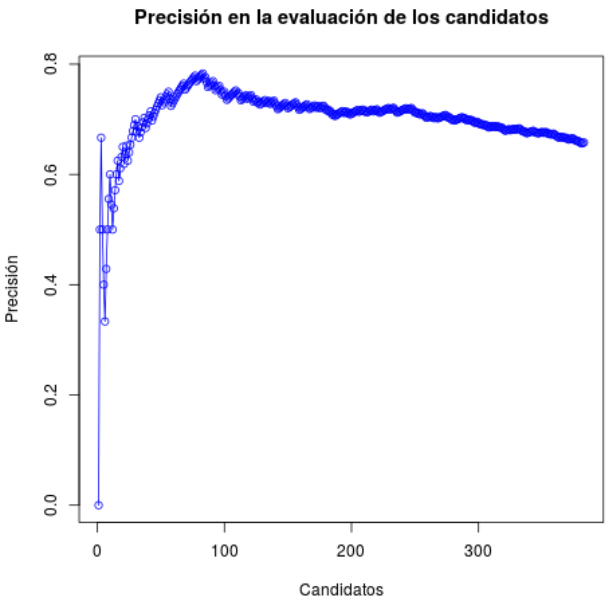

Figura 1: Precisión acumulada en una muestra de 400 candidatos seleccionados como nombres eventivos

Lo primero que salta a la vista en el examen del segundo experimento, y que apunta a explicar la diferencia en desempeño con el anterior, es que los sustantivos son en su mayoría extremadamente raros. Esto no resulta sorprendente ya que es lo que dicta la ley zipfeana de la distribución de frecuencias del vocabulario. La mayor parte del vocabulario estará constituido por

hapax legomena y *dislegomena*, y en muchos casos los autores nos encontramos examinando sustantivos que no conocíamos pero que, en general, se encuentran documentados en al menos un diccionario de la lengua. En el conjunto del Cuadro 2, en cambio, la frecuencia era una variable controlada. Pero justamente por esta diferencia entre el primer y el segundo resultado, creemos que es mejor estimación la del segundo experimento, por resultar más realista.

Otro aspecto relevante a tener en cuenta es que, al examinar la muestra, no solamente nos encontramos con palabras que no conocíamos: también encontramos casos en que no estábamos de acuerdo en si el sustantivo podía o no ofrecer una lectura eventiva. Con el objeto de cuantificar este desacuerdo, tomamos una submuestra de 100 resultados que fueron evaluados por los tres autores. En total, cada uno revisó 200 unidades, pero la mitad de estas unidades eran las mismas en los tres casos, por tanto se revisaron 400 unidades léxicas en total, y utilizamos la intersección de 100 unidades para el cálculo del acuerdo entre anotadores. En el 82 % de los casos existe acuerdo entre los tres anotadores, lo que resulta en un Kappa de 0.526, que se puede considerar un "acuerdo moderado" según Artstein & Poesio (2008). No es infrecuente que exista desacuerdo en materia de clasificaciones lingüísticas, pero al menos confirmamos que tenemos en común una intuición que nos permite reconocer un nombre eventivo en la mayoría de los casos.

5 Conclusiones

En este trabajo hemos propuesto un análisis cuantitativo de corpus –de tipo inductivo y deductivo– para la identificación de las palabras que suelen coocurrir con nombres eventivos no deverbales, con el fin de caracterizar e identificar de forma automática esta clase de sustantivos estudiada por Resnik (2010). Un argumento a favor del enfoque que proponemos en este trabajo es que la metodología es considerablemente más simple y menos costosa desde el punto de vista computacional en comparación con un enfoque basado en aprendizaje automático.

A partir del trabajo realizado, se advierte que la revisión en un corpus de los contextos de aparición de los nombres eventivos ofrece información empírica sobre las relaciones sintácticas que son frecuentes en esta clase de palabras y que, por ende, pueden constituir una fuente confiable para su caracterización e incluso de validación y/o confrontación de las propuestas que se han realizado sobre la base de la introspección.

Reconocemos, sin embargo, las limitaciones de nuestra investigación y encontramos que aún queda mucho trabajo por realizar. Como primera medida, es necesario revisar el algoritmo de clasificación identificando las causas de error en los resultados. Posiblemente no todos los sustantivos puedan ser tratados de la misma manera, y la variable frecuencia es de seguro un factor que debe ser controlado. Otras vías de acción están en la exploración de nuevos elementos predictores y ahondar en mayor complejidad, de ser necesario, en el tratamiento de la información contextual. Otra posibilidad sería incluir también rasgos predictores negativos, es decir rasgos que predicen la condición de no eventivo o nombre simple.

En cualquier caso, creemos que la presente investigación ofrece un precedente más en una línea de trabajo que merece seguir abierta, y de la que existen pocos referentes además de los citados, no ya en castellano sino también en el resto de las lenguas. Esperamos, además, que el presente trabajo resulte un aporte desde el punto de vista metodológico, ya que las herramientas que hemos desarrollado, que son abiertas y de muy sencilla aplicación y uso, pueden alentar a otros investigadores a probar con otros elementos predictores y así mejorar, posiblemente, las tasas de éxito que hemos conseguido hasta ahora. La simplicidad de la propuesta, a su vez, es suficiente motivación como para intentar reproducir los experimentos en otras lenguas, al menos las lenguas europeas, en las que cabría esperar resultados similares.

Por el momento, la utilidad práctica inmediata que tiene para nosotros este trabajo es el de poder enriquecer, por medio de este método, a una taxonomía de sustantivos en castellano que alberga en sí la categoría de "eventos", como es el caso de la ya mencionada taxonomía *open source* de Nazar & Renau (2016).

Agradecimientos

Este proyecto ha sido posible gracias a la financiación de Fondecyt Iniciación (Ref. 11140686), adjudicado por la agencia Conicyt del Gobierno de Chile al primer autor como Investigador Principal. Además, ha recibido también financiación por parte de Conicyt en forma de las becas CONICYT-PCHA/Doctorado Nacional/2016-21160915, concedida a la segunda autora, y CONICYT-PCHA/Doctorado Nacional/2016-21161057, concedida a la tercera autora.

Referencias

Adelstein, Andreína, Marina Berri & Victoria Boschiroli. 2012. Polisemia regular y representación lexicográfica: los nombres locativos en español. *Terminàlia* 5. 33–41.

Alexiadou, Artemis. 2001. *The functional structure in nominals: Nominalization and ergativity*. John Benjamins.

Alonso Ramos, Margarita. 2004. *Las construcciones con verbos de apoyo*. Visor Libros.

Apresjan, Juri. 1974. *Lexical semantics. user's guide to contemporary russian vocabulary*. Karoma Publishers.

Artstein, Ron & Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596.

Baeza-Yates, Ricardo & Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. Addison-Wesley.

Balvet, Antonio, Lucie Barque, Marie-Helene Condette, Pailne Haas, Richard Huyghe, Rafael Marín & Aurélie Merlo. 2011. Nomage: an electronic lexicon of French deverbal nouns based on a semantically annotated corpus. En *International Workshop on Lexical Resources (WoLeR'2011)*, 8–15.

Bel, Núria, Maria Coll & Gabriela Resnik. 2010. Automatic detection of non-deverbal event nouns for quick lexicon production. En *23rd International Conference on Computational Linguistics*, 23–27.

Bosque, Ignacio. 1999. El nombre común. En *Gramática descriptiva de la lengua española*, 3–75. Espasa.

Cabré, M. Teresa, Carme Bach & Jorge Vivaldi. 2006. 10 anys del corpus de l'IULA. barcelona. Informe técnico. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

De Miguel, Elena. 2006. Tensión y equilibrio semántico entre nombres y verbos: El reparto de la tarea de predicar. En *XXXV Simposio Internacional de la Sociedad Española de Lingüística*, 1289–1313.

Fábregas, Antonio. 2010. Los nombres de evento: clasificación y propiedades en español. *Pragmalingüística* 18. 54–73.

Graña López, Benilde. 1993. La prominencia del argumento externo: el diagnóstico de los nombres eventivos. *Revista española de lingüística aplicada* 9. 85–96.

Grimshaw, Jane. 1990. *Argument structure*. The MIT Press.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11(1). 10–18.

Kilgarriff, Adam & Irene Renau. 2013. esTenTen, a vast web corpus of Peninsular and American Spanish. En *V International Conference on Corpus Linguistics (CILC2013)*, 12–19.

Nazar, Rogelio & Irene Renau. 2016. A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. En *10th International Conference on Language Resources and Evaluation (LREC'2016)*, .

Picallo, M. Carme. 1999. La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales. En *Gramática descriptiva de la lengua española*, 363–393. Espasa.

Pustejovsky, James. 1995. *The generative lexicon*. MIT Press.

Quinlan, Ross. 1993. *C45: Programs for machine learning*. Morgan Kaufmann.

Real Academia Española. 2010. *Diccionario de la lengua española*. Espasa-Calpe 22ª ed.

Resnik, Gabriela. 2010. *Los nombres eventivos no deverbales en español*: Universidad Pompeu Fabra. Tesis Doctoral.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Vendler, Zeno. 1967. *Linguistics in philosophy*. Cornell University Press.