Frecuencia y auto-clasificación de Verbos y Perífrasis Españolas: Un estudio exploratorio

Alejandro Carlos Renato
Instituto Genius de Tecnología
Manaus, Brasil
arenato@genius.org.br

Abstract

The frequency of words and structures can be important for both the study of human language processing as the computational linguistics. In the former field, several recent models of language comprehension (Roland et al, 2007) [1] have showed the important role played by the distributional frequencies in determining the accessibility associated with a particular lexical item or sentence structure. On another hand, in the linguistics engineering, the knowledge derived from frequency of words and structures permits establishing a necessary order to process a large corpus, from the most frequent words to lesser, from the most common structure to the infrequent ones, focusing in the characteristics of domain. Verb and phrasal verbs statistics were made in order to do an exploratory study, corpus driven, that illustrates the implications and limitations of using corpus data. Finally, a clustering of words using an automatic method was accomplished.

Keywords: Corpus; Verb Phrase; Frequency; Natural language processing, Clustering.

Resumen

La frecuencia de palabras y estructuras puede ser de importancia tanto para el estudio del procesamiento del lenguaje humano como para la lingüística computacional. En el campo de la psicolingüística, varios estudios (Roland, 2007) han demostrado la importancia del rol ejercido por la distribución de la frecuencia de ítems lexicales y estructuras oracionales en relación a su accesibilidad. Por otro lado, en la ingeniería lingüística, el conocimiento derivado de la frecuencia de las palabras y estructuras permite establecer un orden necesario a la hora de procesar corpora de grandes dimensiones, desde los eventos más frecuentes, hasta los menores, de esta manera focalizando las características del dominio. Con tal fin, se realizó un análisis estadístico de las formas verbales, tanto palabras como perífrasis. El estudio exploratorio basado en corpus, ilustra las implicaciones así como las limitaciones del trabajo empírico. Por último, se realizó un agrupamiento automático de los verbos mediante la técnica de clustering.

Palabras claves: Corpus, Frases verbales, Frecuencia, procesamiento del lenguaje natural, Agrupamiento, Clasificación automática de palabras.

1. INTRODUCCION

La subcategorización de verbos en términos de la estructura predicado-argumentos o roles-temáticos ha sido un tema de investigación y debate en las últimas décadas. Algunos proyectos de grandes dimensiones como WordNet [2], PropositonsBank [3], FrameNet [4], tanto en inglés como sus versiones en español [5] han procurado obtener, desde distintos enfoques, un cuadro completo de la subcategorización verbal. Esto ha sido de importancia también para la ingeniería lingüística, ya que la interpretación semántica para sistemas de comprensión "Undestanding Systems", requiere de la información brindada por la estructura argumental del verbo.

El conocimiento empírico acerca de la aparición de verbos y perífrasis verbales en un corpus puede convertirse en una fuente importante de información a la hora de construir un banco de estructuras gramaticales o una herramienta informática. En vez de poseer la información detallada de todas las posibles subcategorizaciones de los verbos estudiados, -los cuales debido al intenso trabajo que demanda la subcategorización no llegan a unos pocos centenares- puede obtenerse información acerca de las formas verbales más usadas y las subcategorizaciones más frecuentes. Un trabajo seminal en este sentido puede encontrarse en Roland y otros (2007) [1].

En la literatura puede encontrarse el dilema entre entre cantidad de formas verbales versus la complejidad y el detalle de la información.

El presente trabajo se propone como estudio exploratorio, es decir la construcción de un mapa, sobre el cual se toman las decisiones teóricas. A los fines de la lingüística informática, o más precisamente de la ingeniería del lenguaje, conocer cuáles son las formas verbales más usuales y sus respectivas subcategorizaciones en el dominio de estudio, puede tener mayor utilidad que poseer un conjunto paradigmático de subcategorizaciones verbales, debido a los esfuerzos en el mejor de los casos son suficientes para conocer las características del dominio de aplicación.

La idea subyacente es que el conocimiento de las formas más frecuentes puede ser de gran utilidad para derivar aquellas formas menos conocidas e infrecuentes.

El estudio se realizó sobre todos los artículos del diario Clarín [6] del año 2001. El corpus fue etiquetado en forma automática con un tagger perteneciente al conjunto de herramientas OpenNLP [6] que utiliza la técnica de MaxEnt "Máxima Entropía", el cual fue entrenado varias veces para alcanzar un grado aceptable de adaptación al corpus y de desempeño. Sobre el corpus normalizado y etiquetado, mediante la utilización de una herramienta básica de búsqueda de formas regulares "grep", se construyó un conjunto de patrones de búsqueda, que incluyen tanto palabras como estructuras creadas a partir de la información proveniente de las categorías de las palabras. Esto permitió etiquetar las distintas formas de perífrasis verbales y ordenar su búsqueda. Las formas encontradas permitieron la extracción de los verbos y la elaboración estadística.

Por último, se realizó una experiencia de clasificación automática de palabras, con atención a los verbos, a partir de la técnica de "Clustering" agrupamiento, al uso en sistemas de reconocimiento automático de habla, con el conjunto de herramientas automáticas disponibles HTK [7].

2. METODOLOGÍA

La metodología utilizada en la construcción del corpus como en la estadística de los verbos intenta ser simple por dos motivos. El primero, para que pueda ser reproducida con facilidad por colegas con similares intereses de investigación, pero que muchas veces encuentran obstáculos a la hora de obtener y manejar las herramientas adecuadas. Segundo, para documentar un procedimiento sencillo aún en etapas inciales, para luego poder automatizar teniendo en cuenta las dificultades halladas en cada paso.

2.1. Procedimiento

El corpus fue normalizado, -dividido en oraciones y en tokens, reemplazados los valores números- con el auxilio del conjunto de herramientas OpenNLP [6]. El mismo conjnto de herramientas fue utilizado para el etiquetado en clases de palabras. El programa posee facilidades para : 1) Segmentar el texto en oraciones - sentence splitter-, 2) Segmentar cada oración en tokens, - tokenization-, y 3) Etiquetar las palabras con clases lexicales - tagging -. Detrás de tales procedimientos subyace el modelo Maxent, de máxima entropía. El funcionamiento de las herramientas en español muestran buen desempeño pero siempre existen diferencias con respecto al corpus aplicado. La mayoría de las herramientas estadísticas tienen buen desempeño en domínios similares a aquellos donde fueron entrenadas. Por eso es necesario luego realizar ajustes, como corregir y reentrenar el programa para su adaptación a las características del corpus.

Se eligieron mil oraciones del corpus en forma aleatoria para utilizar como grupo testigo del desempeño del etiquetador. Cuando el etiquetador alcanzó el 95 % de acierto en la clasificación de palabras, se consideró aceptable para la tarea, ya que

Una forma de corregir es construir listas con ocurrencia de eventos en forma decreciente. Por ejemplo, co-ocurrencia de nombres propios [NOMBRE]/NP [NOMBRE]/NP, los cuales presentan un alto grado de error. Es posible que el tagger etiquete la primera, segunda o tercera ocurrencia de un nombre propio como otra clase de palabra. Una lista de palabras en forma decreciente es fácil de visualizar, permite ver la cantidad de veces que un evento ocurre y corregirlo de una sola vez para todo el corpus. La corrección se realiza entonces desde los errores más frecuentes a los de menor frecuencia. El ciclo de corrección, entrenamiento y evaluación permite ir ajustando el corpus de manera que los errores encontrados tengan cada vez menos incidencia en los resultados.

La herramienta 'grep' presente en todos los sistemas unix, como por ejemplo Linux, permite encontrar las ocurrencias de tokens como palabras, mediante la utilización de expresiones regulares. Otros trabajos, por ejemplo Roland et al (2007) [1], utilizan una versión mejorada para búsquedas en el formato del Tree Bank, pero combinando expresiones regulares en forma adecuada, es posible tener similares funcionalidades con la herramienta convencional. Por ejemplo, realizar una búsqueda regular de una perífrasis verbal, en un corpus etiquetado con categorías morfo-sintácticas, puede realizarse de la siguiente forma:

```
grep -o 'contin[a-z]\{1,\}/V[A-Z]* siendo/VG [a-z]\{2,\}/VP' clarin_2001_tagged_utf8.txt
```

donde -o, significa que sólo devuelve la ocurrencia de esa expresión y la expresión entre comillas simples es el patrón de búsqueda. Dicho patrón, por ejemplo, devuelve perífrasis como “continúa siendo evaluado”.

Se inicia la búsqueda desde los verbos de mayor longitud en número de tokens, hasta los de menor, que puede variar de 5 tokens a 1 token. Los verbos encontrados primero se van marcando con etiquetas diferentes y guiones bajos de manera que no aparezcan en la búsqueda siguiente. Por ejemplo, la frase anterior será etiquetada como ' continúa_siendo_evaluado/FPV'. Una forma sencilla de realizar sustituciones es por medio del programa 'sed' presente en toda versión Linux. Así puede empezarse por buscar la lista de perífrasis verbales, las frases verbales modales, las frases verbales pasivas, los verbos con auxiliares haber, etc, hasta llegar a las formas simples. Las listas resultantes son revisadas una y otra vez, porque se da el caso que muchas veces se incluyen elementos no buscados.

Una vez confeccionadas las listas, se realizan las sustituciones correspondientes.

En algunos casos se buscó adicionar preposiciones, sustantivos y otras palabras para obtener ocurrencias de verbos del tipo “tener conciencia de”, “tener lugar en” , “tener permiso”, “tener como”, “tener asiento en”, “tener seguridad de”, etc.

2.2. Corpus utilizado

Todos los artículos encontrados en la edición electrónica del diario Clarín [6] del año 2001 fueron utilizados para la prueba.

El corpus posee una extensión aproximada –dado que las frases verbales están marcadas como una sola palabra- de 12.051.359 palabras, entre las cuales 1559798 fueron etiquetadas como verbos o perífrasis verbales. El vocabulario utilizado asciende a 155.695 de palabras únicas, sin distinción de mayúsculas o minúsculas.

Debido a que el etiquetador es sensible a mayúsculas y minúsculas, y al bajo desempeño mostrado en los títulos escritos en letras mayúsculas -que en el texto original carecen de tilde-, fueron suprimidos todos los titulares del análisis. Por motivos similiares, también fueron suprimidas informaciones como fechas, copyright, etc, que se repiten en todas las páginas, tablas de posiciones deportivas, descripción de partidas de ajedrez, leyendas de imágenes, etc. Las partes suprimidas no forman parte de las estimaciones.

3. Frecuencia de las formas verbales y perífrasis

Las formas verbales consideradas se superponen. Una frase verbal como “puede haber sido considerado”, es una frase clasificada como modal por ser precedida por el verbo poder, y al mismo

tiempo una frase pasiva y poseedora de un verbo compuesto de la forma haber más participio. La estadística entonces surge a partir de la forma más envolvente a la izquierda. De ahí, que una frase como la anterior será considerada dentro de las frases verbales modales con poder, y no como una frase pasiva.

Sobre cada búsqueda realizada se construyó una lista con la frecuencia de cada elemento dentro de su tipo, y fue corregida manualmente desde los eventos más frecuentes a los menos infrecuentes, para cerciorarse de no incluir errores en elementos muy frecuentes que pueden distorsionar los resultados generales.

Tabla 1: Distribución de las formas verbales. * Dentro de las perífrasis aspectuales también existen otro tipo de frases.

Verbos				Cantidad	Porcentaje
	Sin perífrasis				
		No conjugados	Infinitivo	232907	14.93
			Participio	21661	1.38
			Gerundio	24640	1.57
		Conjugados	Simples	1059633	67.93
			Compuestos	31790	2.03
	Perífrasis				
		Pasiva	Haber+sido	2442	0.15
			Ser+participio	37025	2.37
		Aspectuales y otras*		75795	4.85
		Modales		37256	2.38
TOTAL				1559798	100

Dentro de las frases modales se han tomado las siguientes formas: deber, poder, haber que, y tener que. Las frases aspectuales contienen formas como “estar seguro de” que son consideradas como otras. Existe un conjunto numeroso de frases que al no estar presentes en las gramáticas consultadas no fueron tenidas en cuenta, y que según la búsqueda realizada, de combinaciones de formas conjugadas con formas no conjugadas podrían ascender a 40.000, que incrementaría en más del doble a las perífrasis individualizadas. Sin embargo, muchas de ellas posiblemente no serían consideradas como tales por las perspectivas gramaticales actuales. Uno de los criterios más utilizados es que la frase verbal debe tener unidad de acción. Por ejemplo, una frase como “pensaron colaborar” es considerada como formada por dos acciones por algunas gramáticas.

Las diferentes perífrasis encontradas alcanzan la suma de 109. En la Tabla 2 se muestra la frecuencia de aparición de las distintas frases verbales encontradas en el corpus.

Tabla 2. Frecuencia de aparición de perífrasis verbales.

N.	CANTIDAD	%	FRASE VERBAL	N.	CANTIDAD	%	FRASE VERBAL
1	12120	14,52	estar_participio	57	55	0,07	bastar_con
2	10440	12,51	deber_infinitivo	58	50	0,06	salir_participio
3	9534	11,43	estar_gerundio	59	43	0,05	ir_participio
4	9151	10,97	ir_a_infinitivo	60	39	0,05	continuar_siendo
5	5166	6,19	tener_que_infinitivo	61	37	0,04	empenarse_en
6	4340	5,20	seguir_gerundio	62	37	0,04	comenzar_gerundio
7	3293	3,95	volver_a_infinitivo	63	37	0,04	acabar_gerundio
8	2500	3,00	empezar_a_infinitivo	64	35	0,04	quedar_por_infinitivo
9	2484	2,98	haber_que_infinitivo	65	33	0,04	llevar_participio
10	2340	2,80	comenzar_a	66	33	0,04	estar_cansado_de
11	1997	2,39	ir_gerundio	67	24	0,03	traer_participio
12	1909	2,29	lograr_infinitivo	68	24	0,03	largarse_a_infinitivo
13	1336	1,60	llegar_a_infinitivo	69	23	0,03	quedarse_participio
14	1230	1,47	venir_gerundio	70	22	0,03	terminar_siendo_participio
15	1213	1,45	dejar_de_infinitivo	71	22	0,03	lanzar_a
16	1134	1,36	soler_infinitivo	72	21	0,03	dejarse_de
17	1045	1,25	tener_participio	73	20	0,02	empezar_por
18	987	1,18	deber_estar_ser_participio	74	19	0,02	estar_empenado_en
19	981	1,18	terminar_gerundio	75	18	0,02	dejar_de_ser_participio
20	720	0,86	quedar_participio	76	17	0,02	estar_sin_infinitivo
21	599	0,72	dar_a_(luz,entender,conocer, préstamo)	77	16	0,02	deber_haber_sido_participio
22	578	0,69	verse_participio	78	16	0,02	darse_por_participio
23	533	0,64	llevar_a_cabo	79	16	0,02	dar_en_infinitivo
24	483	0,58	terminar_de_infinitivo	80	16	0,02	cesar_de
25	447	0,54	seguir_participio	81	15	0,02	quedar_en_infinitivo
26	420	0,50	conseguir_infinitivo	82	15	0,02	estar_harto_de
27	408	0,49	estar_a_punto_de	83	14	0,02	volver_de_infinitivo
28	357	0,43	venir_a_infinitivo	84	14	0,02	estar_al_infinitivo
29	356	0,43	continuar_gerundio	85	13	0,02	quedarse_sin_infinitivo
30	339	0,41	hace_poco	86	12	0,01	estar_que_verbo_conjugado
31	317	0,38	estar_de_acuerdo	87	11	0,01	tener_por_infinitivo
32	283	0,34	hace_mucho	88	10	0,01	volver_gerundio
33	283	0,34	alcanzar_a	89	10	0,01	estar_en_vias_de
34	262	0,31	venir_de_infinitivo	90	10	0,01	echar_a

35	257	0,31	dar_por_adjetivo_participio	91	10	0,01	deber_de
36	241	0,29	llevar_a_infinitivo	92	9	0,01	comenzar_por
37	220	0,26	deber_haber_participio	93	6	0,01	se_le_dar_por
38	205	0,25	dejar_de_lado	94	6	0,01	acertar_a
39	174	0,21	parar_de	95	6	0,01	acabar_participio
40	169	0,20	salir_gerundio	96	5	0,01	falta_por_infinitivo
41	152	0,18	dejar_participio	97	4	0,00	no_pasar_de_ser
42	140	0,17	seguir_sin_infinitivo	98	4	0,00	no_estar_de_mas
43	127	0,15	entrar_a_infinitivo	99	4	0,00	dista_mucho_de
44	125	0,15	estar_para_infinitivo	100	3	0,00	tener_sin_infinitivo
45	118	0,14	venir_participio	101	3	0,00	echar_de_menos
46	112	0,13	poner_a_infinitivo	102	3	0,00	acabar_siendo_participio
47	108	0,13	acostumbrar_a	103	2	0,00	romper_a_infinitivo
48	102	0,12	andar_gerundio	104	2	0,00	no_hay_mas_que_infinitivo
49	98	0,12	terminar_por_infinitivo	105	2	0,00	gustar_de_infinitivo
50	97	0,12	haber_de_infinitivo	106	1	0,00	ver_de_infinitivo
51	93	0,11	cansarse_de	107	1	0,00	no_hace_mucho_que
52	90	0,11	disponerse_a	108	1	0,00	meterse_a_infinitivo
53	87	0,10	no_hacer_mas_que_infinitivo	109	1	0,00	estar_hastiado_de
54	80	0,10	tardar_en_infinitivo		83445	100,00	TOTAL
55	74	0,09	pasar_a_infinitivo				
56	66	0,08	andar_participio				
57	55	0,07	quedarse_gerundio				

Las perífrasis verbales están concentradas principalmente en los primeros cuarenta tipos, que representan la gran mayoría de las perífrasis en total.

Para la clasificación verbal para su respectiva subcategorización, requiere conocer la frecuencia por lema, en infinitivo, ya que en el texto se encuentran con sus respectivas variaciones morfológicas. Para ello, de los verbos en sus formas simples, 1059633, que representan más del 67 % de las formas verbales en total, se eligieron los 2000 verbos más frecuentes, cuyas formas alcanzan a 880.000 ocurrencias en el corpus. Dichos verbos fueron convertidos a sus lemas correspondientes, obteniéndose 615 lemas básicos, que cubren aproximadamente más del 75% de los verbos del corpus total. En la Tabla 3 se muestran los 100 lemas verbales más frecuentes, ordenados en forma decreciente.

La frecuencia de aparición de los lemas muestra como los primeros 10 lemas más frecuentes poseen aproximadamente más de la mitad de todas las ocurrencias. Basta señalar que los verbos más estudiados en términos de subcategorización por la literatura, son aquellos de mayor riqueza y variedad de argumentos. En este caso, el verbo ser representa una gran variedad de formas a identificar, por ejemplo, conjuntamente con los adjetivos que suelen acompañar al verbo: “ser preciso”, “ser probable”, “ser conciente de”, “ser necesario”, “ser urgente”. De la misma forma, el

verbo “tener” posee una gran variedad de realizaciones. También es necesario indicar una limitación, que observando las formas simples más frecuentes encontramos verbos como ser, estar, ir, llegar, querer y haber, podemos inferir que existe una variedad de perífrasis no detectadas. Sería necesario realizar una investigación más exhaustiva sobre las posibles perífrasis. Esto llevaría a que los datos finales contendrían un porcentaje mayor de perífrasis en desmedro de las formas simples.

Tabla 3. Frecuencia de aparición de los 99 lemas más frecuentes del corpus.

ORDEN	VERBO	CANTIDAD	%	ORDEN	VERBO	CANTIDAD	%
1	ser	142625	16,47	50	trabajar	2726	0,31
2	tener	37269	4,30	51	mantener	2701	0,31
3	ir	35415	4,09	52	ocurrir	2700	0,31
4	estar	32888	3,80	53	afirmar	2691	0,31
5	haber	31077	3,59	54	sumar	2669	0,31
6	hacer	28825	3,33	55	existir	2627	0,30
7	decir	27948	3,23	56	intentar	2615	0,30
8	ser-ir	26369	3,04	57	llamar	2605	0,30
9	querer	10596	1,22	58	pagar	2563	0,30
10	llegar	10401	1,20	59	sostener	2562	0,30
11	dar	10183	1,18	60	pensar	2489	0,29
12	quedar	7371	0,85	61	producir	2407	0,28
13	pasar	6758	0,78	62	caer	2389	0,28
14	jugar	6654	0,77	63	anunciar	2354	0,27
15	parecer	6353	0,73	64	considerar	2331	0,27
16	llevar	5517	0,64	65	conocer	2318	0,27
17	saber	5505	0,64	66	morir	2314	0,27
18	contar	5217	0,60	67	informar	2288	0,26
19	recibir	5131	0,59	68	incluir	2273	0,26
20	asegurar	5002	0,58	69	resultar	2253	0,26
21	encontrar	4935	0,57	70	cumplir	2207	0,25
22	explicar	4878	0,56	71	confirmar	2134	0,25
23	ver	4850	0,56	72	creer	2121	0,24
24	pedir	4841	0,56	73	empezar	2114	0,24
25	ganar	4782	0,55	74	faltar	2039	0,24
26	poner	4687	0,54	75	indicar	2036	0,24
27	seguir	4682	0,54	76	ofrecer	1996	0,23
28	tratar	4613	0,53	77	sufrir	1980	0,23
29	crear	4239	0,49	78	entrar	1961	0,23
30	deber	4207	0,49	79	necesitar	1940	0,22
31	volver	4092	0,47	80	convertir	1907	0,22
32	hablar	4046	0,47	81	reunir	1863	0,22
33	dejar	3973	0,46	82	abrir	1823	0,21
34	esperar	3724	0,43	83	viajar	1805	0,21
35	salir	3710	0,43	84	participar	1780	0,21
36	comenzar	3593	0,41	85	lograr	1773	0,20
37	agregar	3578	0,41	86	declarar	1715	0,20
38	presentar	3534	0,41	87	cobrar	1638	0,19
39	señalar	3457	0,40	88	reconocer	1608	0,19
40	vivir	3400	0,39	89	admitir	1596	0,18
41	terminar	3363	0,39	90	enfrentar	1589	0,18
42	realizar	3314	0,38	91	alcanzar	1587	0,18
43	venir	3306	0,38	92	recordar	1553	0,18
44	mostrar	3256	0,38	93	advertir	1532	0,18

45	buscar	3072	0,35	94	sentir	1530	0,18
46	permitir	3058	0,35	95	responder	1512	0,17
47	tomar	2972	0,34	96	vencer	1508	0,17
48	aparecer	2957	0,34	97	cambiar	1501	0,17
49	decidir	2808	0,32	98	bajar	1499	0,17
50	perder	2804	0,32	99	apuntar	1493	0,17

4. Clasificación automática de los verbos por Clustering

Una forma de autoclasificar palabras tanto por sus similaridades semánticas como por sus características sintácticas y morfológicas, es hacerlo por medio de algoritmos que evalúan la pertenencia de una palabra a una determinada clase o cluster de acuerdo a su contexto inmediamente posterior y anterior. Los algoritmos asignan la mayor probabilidad de una palabra de pertenecer a una clase o a otra.

Para facilitar la tarea de etiquetado y posterior clasificación y agrupamiento, se realizan operaciones sobre el corpus utilizadas en procesamiento del lenguaje natural Goodman, 2001 [8]. La primera consiste en reemplazar todos los valores numéricos por una clase, por ejemplo, [NUMEROS], como es usual escribir las clases en modelos estadísticos del lenguaje. El segundo, consiste en mapear todas los nombres propios a la clase [NOMBRES]. El fin es reducir la variación introducida por palabras justamente que dan gran variación al corpus, de manera que el corpus quede más compacto.

Algunos autores consideran que a los fines del reconocimiento automático de habla la autoclasificación de palabras posee mejores resultados que la clasificación manual. Sin embargo, esto suele ser difícil de establecer.

Se utilizó el software Cluster, perteneciente al conjunto de herramientas HTK [7] de gran difusión en la comunidad de reconocimiento automático de habla. Se asignó un conjunto de 1000 clases para la clasificación de todas las palabras en 10 iteraciones. Las 64.000 palabras más frecuentes del corpus fueron utilizadas para realizar la autoclasificación. Las clases obtenidas reflejan en gran medida las características morfológicas, semánticas y sintácticas comunes a las clases o clusters.

Por ejemplo, la clase número 8, presenta en su inicio la siguiente lista de palabras, correspondiente a participios y/o adjetivos plurales, masculinos en su mayoría, que reflejan proximidad en un contínuo semántico: abarrotados, abrazados, abrochados, acertados, acomodados, admirados, afligidos, afónicos, agigantados, agotados, agradecidos, agrandados, ahogadas, ahogados, aislados, ajustados, alarmados, alertas, aliviados, alterados, amargados, amarrados, amenazados, amontonados, amordazados, anegados, angustiados, ansiosos, apagados, apretados, apurados, apáticos, arreglados, articuladas, asfixiados, […] .

Las clases que recibieron más atención son aquellas constituidas por verbos. A continuación se exponen algunos ejemplos, y se ensayan algunas intrepretaciones sobre los agrupamientos en clases.

La clase 24 posee los infinitivos: adicionar, ahondar, anotarse, beberse, debitar, desparramar, divisar, encajar, erigir, esparcir, extirpar, guardarse, imprimir, incluir, incluirlas, inhalar, insertar, invertir, leerlos, palpar, poner, ponerle, ponerles, ponerse, programarlo, radicar, radiografiar, reinstalarse, rendirle, reparar, sembrarse, suministrarse, traducirlo. Dichos verbos, de acuerdo a la interpretación, tienen en común poseer como argumento un objeto material.

Algunas clases poseen pocos verbos y otras una gran cantidad, siendo difícil intuir a simple vista cuáles características morfológicas, sintácticas o semánticas podrían estar actuando. Un ejemplo es la clase 460, a la cual pertenecen 2400 verbos. Por ejemplo: abandona, abandonaba, abandonara, abandonará, abandone, abandonó, abarataría, abarate, abarató, ablandó, aboga, abona, abonaron, abonaría, abonó, aborda, abordaban, abordaron, abordará, abordó, aborrece, abra, abre, abren, abriera, abriga, abrimos, abrirá, abrirán, abrió, abrí, absorba, absorbe, absorbieron, absorbió, absorbía, abundó, aburrieron, acabara, acabe, acabó, acapara, acaparó, acaricia, acarició, acató, acciona, accionó, aceitó, acelera, aceleraba, aceleran, aceleraron, acelerará, acelerarán, aceleraría, acelere, aceleró, acentúe, acentúo, acepte, aceptó. Diversas variables subyacentes podrían estar detrás de esta clasificación, seguramente más de una. Podemos ver que la mayoría los verbos tienen por lo menos tienen una acepción en los cuales son transitivos.

Otras clases tienen una proximidad evidente, como la clase 790 en las cuales encontramos 264 verbos relacionados con decir, posiblemente seguidos por una completiva. Por ejemplo:

CLASS790 790 264 IN,
aclara, aclaraba, aclaró, aconsejando, acota, acotaba, acotó, adelantó, admite, admitiera, admitió, admitían, aduce, adujo, advertirán, advertía, advierte, advirtiendo, advirtiera, advirtió, afirma, afirmaba, afirmó, agrega, agregaba, agregó, alegó, alertó, amanecieron, amplía, anticiparon, anticipó, anunciaran, anunciaron, anunció, apresó, apuntó, arbitrará, arengó, argumenta, argumentaba, argumentara, argumentó, asegura, aseguraba, asegurándole, aseguró, asevera, aseveró, atestigua, auguró, aventuró, averiguó, avisa, avisaba, avisamos, añade, añadió, balbuceó, bramó, bromea, bromeó, bufó, calculó, clama, clamó, clarificaba, comenta, comentaba, comentó, comprendió, concluye, concluyó, confesó, confiesa, confirmaba, confirmando, confirmaron, confirmó, confió, conjeturó, conmemoraron, consideró, contesta, contestó, contó, corroboró, decimos, declaró, decía, decíamos, dedujo, delatan, deleita, demostramos, demostrará, demostré, demostró, demuestra, denunció, deploró, desconociera, desmintieron, desmintió, destacando, destacó, detalló, devito, diagnosticó, dice, diciéndole, dictaminaron, dictaminó, dije, dijera, dijimos, dijiste, dijo, dirá, diría, dispararan, ejemplifica, ejemplificó, elegís, enfatiza, enfatizó, entendió, entonan, especificó, estimaban, estimo, estimó, evaluó, evitamos, exclamó, explica, [...].

La clase 473 posee 18 elementos, los cuales pertenecen claramente a verbos que constituyen perífrasis verbales:

CLASS473 473 18 IN
acostumbran, deban, deben, deberán, deberían, debieran, debieron, debían, parecerían, podrian, podrían, pretendan, pudieran, pudiesen, puedan, simulaban, solían, suelen

La clase 539 posee 69 verbos que indican algún tipo de actividad refexiva con respecto a lo dicho, que permitirían distinguirlos de la clase 790, por ejemplo.

CLASS539 539 69 IN
acreditó, adivinaba, advierta, afirmara, alega, argumentará, calcula, calculaba, comprenderá, comprobaba, comprobara, comprobó, comprueba, compruebe, concibe, confirmo, considera, consideraba, considere, consignaba, constató, cree, creyera, creyó, creía, deduce, demostraría, descarta, descartaba, descontaba, descubre, descubrió, descuenta, deslizaba, destaca, destacaba, destacará, determinó, especifica, estableció, estima, estimaba, estipuló, evitará, imaginaba, imagine, intuye, opine, percibió, predijo, presume, presumía, presupone, previó, procuraba, pronostica, pronosticaba, recordara, rumorea, rumoreó, sospechaba, sospeche, sostenía, supone, suponía, supuso, teme, temía, verificara,

Por último, la clase 145 presenta 16 verbos que expresan el resultado o conclusión de una evaluación:

CLASS145 145 16 IN
considerarla, considerarlas, considerarlo, denominarse, parecerle, quedarle, resultar, resultarle, sentirnos, ser, serle,,,

No siempre es fácil determinar en todos los casos el sentido del agrupamiento de los verbos, el cual obedece a varias dimensiones ocultas. Una forma complementaria y que en este sentido puede considerarse superior para clasificar los verbos, puede encontrarse en SOM -Self-organizing Maps- como los estudios realizados por Ritter y Kohonen [9] y [10], para un estudio sobre los cuentos de los hermanos Grimm o, más específicamente para la clasificación de los verbos en roles temáticos con redes neuronales, como en Miikkulainen [11]. Para una comparación entre los métodos de clustering y self-organizing maps puede verse Siivola [12]. Dichos autores trabajan con métodos que permitirían, conversacionalmente, representar la distribución de los verbos como un contínuo semántico, en los cuales puede determinarse tanto la proximidad como la distancia, y no como clases discretas, estancas, las cuales agrupan verbos próximos pero tienen dificultad para expresar la gradación semántica y posibles particiones dentro de la misma clase. Las técnicas de clustering, muy utilizadas por ejemplo en diversas disciplinas científicas, como por ejemplo las neurociencias, o en tecnologías del habla, no han sido debidamente evaluadas en términos de la lingüística computacional. Las técnicas, que tienen en principio resultados apreciables, deben seguir siendo investigadas, buscando variantes y optimizando resultados. Por ejemplo, una variación interesante sería clasificar las palabras con frases verbales. Otra variante, reduciendo la variación morfológica por la utilización de lemas. Esta última variación, por ejemplo, se encuentra en trabajos como en [10].

5. CONCLUSIONES

La frecuencia de los eventos lingüísticos no siempre ha jugado un rol relevante en los estudios lingüísticos en español. La frecuencia de los verbos y paráfrasis permite establecer un mapa desde el cual realizar decisiones teóricas, tanto para proyectos de investigación como para su utilización en tecnologías del habla. El grado de cobertura de las herramientas lingüísticas es uno de los puntos críticos para las aplicaciones reales. El estudio exploratorio permite poseer un conocimiento sobre las limitaciones y dificultades que pueden encontrarse en un dominio. En este caso, el presente estudio se propone como un mapa para el estudio de la sub-categorización verbal.

El estudio de la frecuencia de los verbos en un corpus grande reveló cuáles serían los verbos de mayor relevancia en corpus periodísticos para alcanzar un grado aceptable de adecuación empírica, al mismo tiempo que señala carencias de conocimiento, por ejemplo, en cuanto a determinar un conjunto de perífrasis verbales, que resultaron ser, en apariencia, más que aquellas extraídas de gramáticas y de la literatura en general.

El agrupamiento automático para la clasificación de verbos mostró ser un método promisorio para la investigación lingüística. El método ya ha sido probado con éxito en el campo de la construcción de modelos de lenguaje para el reconocimiento automático de habla. Diversas variantes del método permitirían poseer cada vez mejores clasificaciones basadas en grandes corpus, los cuales desde el punto de vista de la investigación lingüística, han resultado muchas veces indomables hasta el presente. Los métodos de auto-clasificación basados en redes neuronales, tienen antecedentes que permiten suponer que pueden ser importantes en este sentido. Más allá de los proyectos de grandes

dimensiones que se están realizando para obtener datos lingüísticos, es de igual o superior importancia experimentar con las técnicas y algoritmos que permitan acceder a dicho conocimiento.

Agradecimientos

Se reconoce el apoyo financiero para la realización de este trabajo, al proyecto 3147/06, de la FINANCIADORA DE ESTUDIOS E PROJETOS - FINEP www.finep.gov.br – y del Ministério de Ciência e Tecnologia do Governo Brasileiro.

Referencias

[1] Roland, D., Dick, Federic y Elman, Jeffrey (2007). Frequency of basic English grammatical structures: A corpus analysis. Journal of Memory and Language 57, páginas 348–379.

[2] Miller, G. (1995) WordNet: A lexical database for English. Communications of the ACM, 38 (11), pág. 39-41.

[3] Paul Kingsbury, Martha Palmer y Mitch Marcus. (2002) Adding Semantic Annotation to the Penn TreeBank. HTL.

[4] Proyecto FrameNet en inglés: framenet.icsi.berkeley.edu/

[5] FrameNet en español. gemini.uab.es/SFN

[6] Diario Clarín, versión digital. www.clarin.com.ar/

[6] OpenNLP. Open Natural Language Processing Toolkit. opennlp.sourceforge.net/index.html

[7] HTK. Hidden Markov Model Toolkit (HTK). University of Cambridge. htk.eng.cam.ac.uk/

[8] Goodman, J. T. (2000). A bit of progress in language modeling, extended version. Technical Report MSR-TR-2001-72, Microsoft Research.

[9] Riiter, H. and Kohonen, T. (1989). Self-organizing semantic maps. Biological Cybernetics, 61(4): 241-254.

[10] Honkela, T., Pulkki, V., and Kohonen, T. (1995). Contextual relations of words in Grimm tales analyzed by self-organizing maps. En Proceedings of the International Conference on Artificial Networks (ICANN), pages 3-7.

[11] Miikulainen, Risto.(1993) Subsymbolic Natural Language Processing. An Integrated Model of Scrips, Lexicon and Memory. MIT Press.

[12] Siivola, Vesa. (2007) Language Models for Automatic Speech Recognition: Construction and Complexity Control. Helsinki University of Technology. Dissertations in Computer and Information Science.