Francisco José Sánchez García
(Universidad de Granada)

EL ÍNDICE DE RIQUEZA LÉXICA EN EL NIVEL DE ESTUDIOS BAJO DEL CORPUS PRESEEA GRANADA[1]

Fecha de recepción: 07.02.2018 **Fecha de aceptación**: 04.05.2018

Resumen: En los últimos años, la proliferación de estudios de lingüística cuantitativa y estadística léxica viene arrojando resultados muy relevantes, especialmente gracias a la riqueza léxica y el léxico disponible, si bien la mayoría de estos trabajos se ha limitado a estudiar el vocabulario de los jóvenes preuniversitarios. En el caso de Granada se llevó a cabo el estudio del léxico disponible de la provincia, pero aún carecemos de otros indicadores lexicométricos sobre la riqueza del lexicón mental de los granadinos. Con este trabajo aportamos una primera aproximación a la riqueza léxica del español coloquial hablado en Granada, centrada en el nivel de estudios bajo el corpus PRESEEA Granada (mediante una muestra compuesta por 18 informantes), a fin de obtener el promedio de vocablos, intervalo de aparición de palabras de contenido nocional y, finalmente, el promedio de hápax a partir de las variables "sexo" y "edad".

Palabras clave: riqueza léxica, PRESEEA Granada, hápax, estadística léxica, sociolingüística

Title: Lexical Richness in the Low Level of Studies of the Corpus PRESEEA Granada

Abstract: In recent years, studies of quantitative linguistics and lexical statistics have proliferated, providing very relevant results, especially thanks to the lexical richness and the available lexis, although most of these works have been limited to the study of the vocabulary of young pre-college students. In the case of Granada, the study of the available lexicon of the province has been published, but we still lack other lexicometric indicators. With this work, we provide a preliminary approximation to the lexical richness of colloquial Spanish spoken in Granada, focused on the low level of studies in PRESEEA Granada corpus (using a sample composed of 18 informants), in order to obtain the average of words, interval of appearance of words of notional content and, finally, the average of hapax from the variables 'gender' and 'age'.

Key words: Lexical richness, PRESEEA Granada, hapax, lexical statistics, sociolinguistics

[1] Este estudio se ha realizado en el marco del Proyecto ECOPASOS-Granada (Estudio complementario de los patrones sociolingüísticos del español de Granada), financiado por el Ministerio de Ciencia e Innovación (Ref. FFI2015-68171-C5-2-P) e integrado en el proyecto PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y América).

 DOI: 10.23825/ITINERARIOS.28.2018.05

INTRODUCCIÓN

En las últimas décadas, son muchos y valiosos los trabajos sobre enseñanza del léxico. En el ámbito hispánico, la mayoría de estos estudios gravita en torno a las inestimables aportaciones de López Morales, gran impulsor de los estudios sobre la disponibilidad léxica. Como sabemos, el léxico disponible

> es el conjunto de unidades léxicas que los hablantes conocen y, potencialmente, están en condiciones de usar –incluyendo léxico pasivo–, aunque su actualización esté condicionada por el tema concreto que se aborde en cada situación comunicativa. Como es natural, el número medio de estas voces dependerá del grado de formación y cultura de los hablantes. Así, como nos dice H. López Morales (1999), una persona culta maneja entre 4.000 y 5.000 palabras, mientras que una persona común, con una formación académica básica, no alcanza más de las 2.000. (Pastor Milán y Sánchez García 2008: 13)

Por otra parte, es imprescindible diferenciar entre *palabras frecuentes* y *palabras disponibles*:

> Las palabras frecuentes son, como las denomina Michéa (1950), palabras 'atemáticas', unidades gramaticales (artículos, preposiciones y conjunciones y en orden decreciente, verbos, adjetivos y sustantivos, que se actualizan en cualquier situación comunicativa, mientras que las palabras disponibles –o 'temáticas'– representan el léxico potencial que se presenta en situaciones concretas y condicionadas por un tema que sirva de estímulo. Según Humberto López Morales, el estudio de ambos grupos de palabras es necesariamente complementario si lo que queremos es ofrecer una visión panorámica de la realidad léxica de la comunidad que estudiamos (1986: 63). De este modo, la suma de ambos léxicos constituye el vocabulario fundamental de una comunidad, el eje vertebrador del idioma. (Pastor Milán y Sánchez García 2008: 13)

Indudablemente, el estudio del léxico disponible es especialmente relevante para conocer el lexicón mental de una comunidad de hablantes (se ha revelado especialmente productivo en el estudio de los estudiantes preuniversitarios), pero también presenta ciertas carencias. En primer lugar, para determinarlo, es imprescindible aplicar una encuesta sobre un listado cerrado de temas, conocidos como "centros de interés"[2]. Además, como acabamos de exponer, solo se toman en consideración las palabras nocionales y, en particular, sustantivos y adjetivos. Si lo que queremos es determinar con preci-

[2] En el Proyecto panhispánico de disponibilidad léxica se consideran dieciséis centros de interés: 1) Las partes del cuerpo; 2) La ropa; 3) La casa; 4) Los muebles de la casa; 5) Los alimentos; 6) Los objetos situados en la mesa para las comidas; 7) La cocina: muebles y utensilios; 8) La escuela: muebles y material escolar; 9) La iluminación y el aire acondicionado; 10) La ciudad; 11) El campo; 12) Los medios de transporte; 13) Los trabajos del campo y del jardín; 14) Los animales; 15) Los juegos y diversiones; 16) Las profesiones y oficios. Adicionalmente, en el estudio del léxico disponible de Granada, se incluyeron dos centros de interés suplementarios: 17) El mar y 18) La religión. *Cf.* Pastor Milán y Sánchez García (2008).

sión el grado de conocimiento global del léxico, es preciso atender también al llamado "índice de riqueza léxica".

Los estudios sobre riqueza léxica surgen hace más de seis décadas gracias a Giraud (1960), preocupado por correlacionar el número de palabras y vocablos de un texto como punto de partida para la obtención de un índice válido que permitiera cuantificar el potencial lingüístico de los hablantes. Este autor fue el primero en acuñar la diferencia entre *palabra* (unidad del texto: material gráfico comprendido entre dos espacios en blanco) y *vocablo* (unidad del léxico: palabras diferentes que podemos encontrar en un texto), sentando con ello las bases de la larga y prolífica estela de estudios sobre léxico-estadística de las décadas siguientes, en las que debemos destacar algunos hitos. Por su parte, Müller (1968) profundizó en el análisis cuantitativo de la frecuencia de vocablos en un texto atendiendo a su naturaleza gramatical. Además, Müller concedió especial importancia a aquellas unidades que aparecen en el texto una sola vez: el *hápax*, que podemos obtener dividiendo el número total de vocablos por aquellos que tienen frecuencia 1.

La fórmula de Giraud para obtener el índice de riqueza léxica no era complicada. Como venimos diciendo, el autor distinguía entre las palabras nocionales o de contenido semántico (sustantivos, adjetivos calificativos, verbos y adverbios) y gramaticales (artículos, preposiciones, conjunciones, pronombres y adjetivos determinativos).

Así, tenemos:

$$R = \frac{V}{N} \qquad\qquad R = \frac{V}{2N}$$

En la primera fórmula, consideraríamos todos los vocablos en V; en la segunda, únicamente las voces nocionales. En ese caso, el número total de palabras del texto (N) se multiplica por 2, ya que Giraud entendía que las palabras nocionales habían de representar la mitad del texto.

La información que nos proporciona este índice y el método para calcularlo se han mantenido, sin grandes novedades, hasta hace relativamente pocos años. Así, por ejemplo, Těšitelová (1992) propone considerar la repetición de palabras de un texto, la fuerza de la zona de palabras de baja frecuencia (comprendida entre 1-10) y los fenómenos de dispersión y concentración del vocabulario.

Por su parte, Ávila Sánchez propone un método para medir la riqueza léxica, para lo que se sirve de

> tres procedimientos comparativos: el número de vocablos recogidos en el total de textos de cada subconjunto de niños, la densidad léxica promedio por cien palabras, el número de vocablos acumulados por deciles de acuerdo con su frecuencia descendiente. (1986: 511)

Entendemos que esta aportación es especialmente relevante: de entrada, hay que tener en cuenta únicamente las primeras 100 palabras, lo que simplifica enormemente la tarea, al mismo tiempo que, según Haché de Yunén (1991), Ham (1979) y el propio Ávila Sánchez (1986), a partir de la centena de unidades obtenidas, el promedio de vocablos deja

de aumentar de manera significativa. De este modo, podemos calcular el coeficiente de densidad léxica si dividimos el número de tipos léxicos (T) que aparece en un fragmento del texto de una determinada longitud entre el número de palabras del segmento (N). Así, podemos analizar nuestra muestra atendiendo a textos individuales.

$$D = \frac{T}{N}$$

Por último, hay que cuantificar las frecuencias acumuladas por deciles[3], y ello resulta especialmente útil para comparar la riqueza léxica de textos que no tienen la misma extensión.

Sobre esa base, López Morales (2011) introduce una fórmula para determinar el porcentaje de vocablos. Como vemos, se trata de dividir el número total de vocablos entre el total de unidades léxicas para luego multiplicar el resultado por 100:

$$PV = \frac{V \times 100}{N}$$

Como él mismo señala, este es un "indicador grueso" que debe complementarse con el intervalo de aparición de palabras de contenido nocional (IAT):

$$IAT = \frac{N}{PN}$$

De esta fórmula se desprende que, a mayor número de palabras nocionales, menor será el intervalo, lo que redunda en un índice de riqueza léxica más favorable. Se trata de una fórmula especialmente interesante para cotejar entre individuos particulares, a fin de estudiar la relación de un sujeto con el resto de la muestra o un determinado grupo de informantes de esta.

En nuestro estudio, prestamos especial atención a las palabras nocionales. Como nos recuerda López Morales:

> Las palabras nocionales (PN) son aquellas unidades léxicas con contenido semántico, es decir, sustantivos, verbos, adjetivos y adverbios, aunque con estos dos últimos se requiere de algunas especificaciones. La riqueza léxica se obtiene aquí al considerar la cantidad de vocablos o unidades léxicas diferentes y el total de palabras de contenido nocional (PN). El primer cálculo que se realiza es el que determina el porcentaje de vocablos (PV). El procedimiento requiere que se divida el total de vocablos (V) entre el total de las unidades léxicas comprendidas en el texto (N) y luego se multipliquen por 100. (2011: 20)

[3] "El *decil* es una medida de localización o posición no central. Los deciles son los nueve puntos que dividen la distribución en diez puntos de forma tal que, dentro de cada una, está incluido el 10 % de los datos. Entonces, un *decil* es un valor que representa la décima parte de un conjunto de información" (INEI 2006).

METODOLOGÍA

El propósito de este trabajo es examinar el índice de la riqueza léxica en el español coloquial hablado en Granada, sirviéndonos de los tres indicadores más operativos que se vienen manejando en los últimos años: la frecuencia de palabras nocionales, el intervalo de aparición con respecto al total de palabras y el índice de hápax.

El estudio de la riqueza léxica ha sido aplicado con gran acierto a la enseñanza/aprendizaje de la lengua materna en Secundaria y Bachillerato[4] (y también de español como L2), pero hasta ahora son escasos los estudios dedicados a analizar cómo puede funcionar este índice en la conversación coloquial de hablantes ya formados, pertenecientes a diferentes grupos de edad, sexo y nivel de instrucción.

Por ello, consideramos especialmente interesante aprovechar los materiales del corpus PRESEEA de Granada, un corpus oral formado por entrevistas obtenidas mediante un muestreo por cuotas de afijación uniforme, atendiendo a las variables sociales antes mencionadas. Con este trabajo, nos adentramos por vez primera en la dimensión léxica del corpus, ya que, hasta la fecha, la mayoría de estudios se han centrado en la investigación sobre la fonética o la gramática.

La muestra de hablantes está compuesta por un total de 54 informantes[5], distribuidos en tres niveles de instrucción (nivel de estudios primarios, secundarios y universitarios) y tres tramos de edad (jóvenes –entre 19 y 34 años–, adultos –entre 35 y 54– y mayores de 55), además de la variable *sexo*.

Cuadro 1 Distribución de los hablantes del corpus PRESEEA (Granada)

Nº informantes: 54		Nivel de estudios					
		Bajo		Medio		Alto	
		Mujer	Hombre	Mujer	Hombre	Mujer	Hombre
Edad	19-34	3	3	3	3	3	3
	35-54	3	3	3	3	3	3
	> 55	3	3	3	3	3	3

Como esta primera aproximación se limita al nivel de estudios primarios, nos ceñiremos a estudiar la riqueza léxica de los primeros 18 informantes, tal y como resaltamos

[4] Aparte del análisis de la riqueza léxica de los estudiantes de Bachillerato de Tenerife (Reyes Díaz 2007), destacamos los trabajos de Haché de Yunén (1991) sobre la riqueza léxica de los escolares dominicanos de enseñanza primaria; Echeverría, Valencia *et al.* (1992) sobre los estudiantes chilenos de enseñanza media; Portela (1992) sobre los estudiantes dominicanos y Cintrón Serrano (1992) sobre el español hablado en Puerto Rico. En el ámbito del español peninsular nos interesan especialmente las investigaciones de Andrés Pérez (1997), realizadas en la Universidad de Alcalá de Henares sobre alumnos de EGB; Torres González (1999, 2003) sobre alumnos tinerfeños, o García Rosas sobre el español como lengua extranjera (*apud* Reyes Díaz 2007: 151-152). A propósito de la metodología sobre el estudio lexicométrico, remitimos a los trabajos de Baayen y Tweedie (1998), Ávila Muñoz (2014) y Capsada Blanch y Torruella Casañas (2017).

[5] La codificación de la muestra se basa en las indicaciones generales aplicadas por el PRESEEA. Para más información sobre el sistema de codificación de los informantes, remitimos a Moya Corral (2009: 25).

 DOI: 10.23825/ITINERARIOS.28.2018.05

en el cuadro 1. Así, atenderemos básicamente a las variables sociales *sexo* y *edad*. Del mismo modo que en otros trabajos realizados previamente sobre la expresión del sujeto pronominal en el mismo corpus (Manjón-Cabeza, Pose Furest y Sánchez García 2017), hemos optado por analizar fragmentos de 100 palabras de cada uno de los informantes del nivel bajo, teniendo en cuenta que el promedio no se altera de manera significativa analizando un texto más extenso. Con todo, nuestro propósito no es otro que la obtención de resultados significativos y fiables, que puedan ser contrastados con otros del resto de niveles, o incluso de otras áreas geográficas estudiadas en el entorno PRESEEA.

En primer lugar, hemos seleccionado un fragmento relativamente extenso de cada una de las entrevistas, considerando únicamente las 100 primeras palabras, de las cuales se han eliminado las onomatopeyas, nombres propios, interjecciones y anacolutos. Como es habitual en este tipo de trabajos, nos apoyamos en el concepto de "unidad léxica", entendiendo de esta manera la

> palabra o el conjunto de palabras que tienen un solo significado, esto es, las formas verbales compuestas, las locuciones prepositivas, adverbiales o conjuntivas [que] son consideradas y contabilizadas como una única unidad léxica. (Torres González 2003: 441)

Como podrá verse, hemos obtenido el listado de frecuencia de uso de las palabras de cada informante, utilizando para ello la aplicación sobre variedad léxica desarrollada por el Instituto de Ingeniería del Conocimiento de la Universidad Autónoma de Madrid[6]. Seguidamente, hemos procedido a la lematización de las unidades (desechando las formas antes mencionadas), delimitando las formas gramaticales de las nocionales, que clasificaremos atendiendo a su categoría morfológica (sustantivos, adjetivos, verbos y adverbios), para comparar el promedio de uso para cada una de las variables estudiadas. Una vez contabilizados los porcentajes de cada uno de los indicadores mencionados, nos servimos del paquete estadístico SPSS para su procesamiento final, que consistirá en el cotejo de los resultados por tramos de edad y por sexo, determinando el total de palabras y, más específicamente, el índice relativo a las formas nocionales.

Anteriormente describíamos las fórmulas necesarias para calcular los diferentes índices que nos permiten conocer la riqueza léxica. Puede entenderse mejor si analizamos detalladamente un fragmento de la entrevista a uno de nuestros informantes [GRAN-H11-37], escogido aleatoriamente hasta obtener un total de 100 palabras:

> I: pasa algo a lo mejor allí en el barrio pues // no roban // no revientan coches ni revientan retrovisores //allí el barrio está protegido / mm bueno protegido // que lo tenemos controlado / que sabemos que si viene alguien vemos alguien sospechoso pues ya sabemos que ése va a hacer una jangada // (tiempo = 24:59) pues entonces lo seguimos y efectivamente vemos cómo se apoya en el coche/ intenta forzar el coche

[6] Se trata de una aplicación interactiva, de libre acceso, que facilita el listado de las palabras más frecuentes de un texto dado. Puede manejarse en el enlace siguiente: <http://innova.iic.uam.es/acl/>.

(simultáneo: E= no me lo digas) // ya en el barrio los tenemos ya muy / muy guipados a la gente / ya sabemos si viene uno y //sabemos si va a lo que va o / o va de/ de pasada /// si pasa por ahí //.

Siguiendo el método propuesto por López Morales (2011), en primer lugar debemos determinar cuáles de esos vocablos se repiten: *ya* (4), *va* (4), *sabemos* (4), *barrio* (3), *viene* (2), *vemos* (2), *tenemos* (2), *revientan* (2), *protegido* (2), *pasa* (2), *coche* (2), *allí* (2), etc.

A continuación, consideraremos aparte las palabras sin contenido semántico (formas no plenas): *que* (5), *el* (5), *si* (4), *a* (4), *pues* (3), *en* (3), *y* (2), *o* (2), *de* (2), *por* (1), *ni* (1), *los* (1), *la* (1), *cómo* (1), etc.

Normalmente, las 10 o 15 primeras palabras del listado de frecuencias suelen ser precisamente las gramaticales aunque, en este caso, la distribución entre unas y otras ha quedado bastante equilibrada.

Atendiendo al conjunto de unidades consideradas (*tokens*), es preciso conocer el número total de palabras distintas del informante (*types*) por intervalos regulares (habitualmente, se considera más ilustrativo presentarlo por deciles, esto es, de 10 en 10 palabras), que en el caso que nos ocupa arroja como resultado un 54 %. Por tanto, de estas 54 palabras diferentes, hay que separar las nocionales de las gramaticales, lo que nos permite obtener un total de 39 vocablos.

Cuadro 2 Distribución por deciles del número de palabras distintas

Nº palabras	Nº palabras distintas
10	10
20	16
30	22
40	29
50	35
60	41
70	45
80	48
90	49
100	54

Recapitulando, del análisis del fragmento estudiado, obtenemos la siguiente información:

a) Total de palabras: 100
b) Total de palabras de contenido semántico: 60
c) Total de palabras de contenido semántico repetidas en el texto: 15 (suman 39 registros entre todas)
d) Número de vocablos (palabras de contenido semántico) diferentes: 39
e) Resto (palabras gramaticales): 41

 DOI: 10.23825/ITINERARIOS.28.2018.05

Con estos datos ya podemos aplicar la fórmula de López Morales para examinar la proporción entre nocionales y el resto (nocionales repetidas y no nocionales):

PV=39x100/ 100 T: 39

Normalmente, se considera que el índice de riqueza léxica es positivo por encima del 50 %, de modo que este primer cálculo arroja un resultado significativamente bajo.

Una vez conocido este índice, es preciso obtener el intervalo de aparición de palabras nocionales, es decir, a partir de qué palabra del texto aparecerá una nueva palabra de contenido semántico nocional. Dicho índice se obtiene dividiendo el total de registros entre el número de vocablos.

IAT= 100/39: T: 2,56

De modo que es necesario esperar de media a 2,56 palabras para encontrarnos con una unidad nocional. No hace falta explicar que, cuanto mayor sea este intervalo de aparición de estas palabras, menor será la riqueza léxica. Nuevamente, en este caso, nos encontramos ante un intervalo de aparición de palabras nocionales por debajo de los estándares que consideraríamos positivos. Si lo comparamos con los resultados ofrecidos por López Morales (2011) sobre un corpus de estudiantes de secundaria, que, por ejemplo, pueden encontrarse en torno al 0,5-1, queda claro que se trata de un indicador de pobreza léxica en cuanto al empleo de palabras nocionales.

Finalmente, nos interesa conocer el índice de hápax (palabras de una sola ocurrencia en el texto), que resulta de dividir el número total de vocablos entre aquellos que tienen frecuencia 1 (V/V_1):

$$\text{Hápax} = V/V_1 = 38/23 = 1{,}65.$$

También en este caso nos encontramos ante un resultado relativamente pobre que, a priori, encaja bien con el nivel sociocultural del informante. Sabemos que la riqueza léxica disminuye a medida que el índice va aumentando; así, un promedio que hubiera revelado una mayor riqueza léxica normalmente estaría más cerca de 1.

ANÁLISIS DE LOS RESULTADOS

Una vez obtenido el recuento de palabras totales (*tokens*) y diferentes (*types*) de cada uno de los informantes, se han procesado sus datos a fin de determinar el número de vocablos diferentes (de tipo semántico), así como el intervalo y el índice de hápax, como podemos ver en el cuadro 3.

Cuadro 3 Resultados globales de los informantes del corpus

	Palabras contenido semántico	Resto de palabras	Vocablos	Intervalo	Hápax
GRAN-H11-37	60	41	39	2,56	1,69
GRAN-H11-38	48	52	36	2,77	1,5
GRAN-H11-39	54	46	41	2,43	1,36
GRAN-M11-40	57	43	35	2,85	1
GRAN-M11-41	50	50	38	2,63	1,31
GRAN-M11-42	46	54	38	2,63	1,15
GRAN-H21-43	48	52	35	2,85	1,45
GRAN-H21-44	47	53	44	2,27	1,1
GRAN-H21-45	46	54	40	2,5	1,14
GRAN-M21-46	58	42	47	2,12	1,17
GRAN-M21-47	57	43	47	2,12	1,23
GRAN-M21-48	47	53	43	2,32	1,1
GRAN-H31-49	57	43	39	2,56	1,3
GRAN-H31-50	50	50	39	2,56	1,39
GRAN-H31-51	47	53	39	2,56	1,21
GRAN-M31-52	55	45	37	2,7	1,6
GRAN-M31-53	49	51	37	2,7	1,15
GRAN-M31-54	53	47	38	2,63	1,26

Fijémonos ahora en los promedios de cada uno de los indicadores, atendiendo a las variables analizadas. En primer lugar, vamos a examinar los resultados que ofrece la variable edad:

Cuadro 4 Promedio de vocablos, intervalo de aparición de palabras nocionales e índice de hápax por tramos de edad

	Vocablos	Intervalo	Hápax
19-34	38,15	2,64	1,33
35-54	42,6	2,36	1,19
> 55	38,15	2,61	1,31
	39,63	2,53	1,27

En sociolingüística, el factor social "edad" tradicionalmente se ha considerado clave para determinar los usos lingüísticos de una comunidad de hablantes (Mitkova 2007). En este trabajo, podemos comprobar que, efectivamente, los tramos de edad arrojan resultados interesantes, si bien ninguno de los tres grupos (como era previsible) evidencia un índice de producción léxica elevada, más bien al contrario.

Para empezar, llama la atención el resultado de los informantes adultos (comprendidos entre 35 y 54 años) en el promedio de vocablos (42,6 % frente al 38,15 % de los jóvenes

 DOI: 10.23825/ITINERARIOS.28.2018.05

y el 39,63 % de los mayores de 55) y en el intervalo de aparición (2,36 frente a 2,64 de los primeros y 2,53 de los últimos). No obstante, no podemos decir que ese intervalo de 2,36 sea excepcional: si nos fijamos, por ejemplo, en las investigaciones desarrolladas sobre la riqueza léxica de estudiantes de secundaria nos damos cuenta enseguida de que esos promedios suelen ser más positivos incluso en estudiantes de nivel preuniversitario.

También hallamos una diferencia significativa en el índice de hápax de este grupo (1,19), con un resultado ostensiblemente menor que los jóvenes y los mayores (que precisamente evidencian un resultado casi idéntico: 1,33 y 1,27 respectivamente).

Como ha señalado, entre otros, Torres González (1999, 2003), la variable sexo no suele resultar operativa en los estudios contrastivos de riqueza léxica.

Cuadro 5 Promedio de vocablos, intervalo de aparición de palabras nocionales e índice de hápax por sexo y tramos de edad

	Vocablos		Intervalo		Hápax	
	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino
19-34	39,3	37	2,58	2,70	1,51	1,15
35-54	39,6	45,6	2,54	2,18	1,23	1,16
> 55	39	37,3	2,56	2,67	1,3	1,33
	39,3	39,9	2,56	2,51	1,34	1,21

En el cuadro podemos observar unos resultados casi idénticos en el promedio de vocablos y el de intervalo, si bien el índice de hápax es ligeramente más bajo en el total de las mujeres. En cambio, si profundizamos un poco más y cruzamos la variable "sexo" con la de "edad", sí que afloran algunos datos curiosos: son las mujeres del primer y segundo tramo de edad las que nos aportan los mejores promedios de hápax del corpus de estudio, con resultados cercanos a 1: las mujeres jóvenes con un 1,15 y las de edad comprendida entre 35 y 54, un 1,16.

Por último, fijémonos en la distribución de las palabras nocionales. Como era de esperar, son más frecuentes los sustantivos y los verbos:

Cuadro 6 Promedio de palabras nocionales por tramos de edad

	Sustantivos	Adjetivos	Verbos	Adverbios	Total
19-34	20,45	1,99	20,66	9,33	52,43
35-54	24,5	1,33	18,49	6,16	50,48
> 55	21,3	3,83	19,66	6,99	51,78
	22,08	2,38	19,60	7,49	51,55

Únicamente destaca un mayor uso de sustantivos entre los informantes de edad intermedia (24,5 frente al 20,4 de los jóvenes y el 21,3 de los mayores). También es llamativo el mayor uso de adjetivos entre los mayores de 55 años, que prácticamente duplica a los otros dos grupos (1,99 para los jóvenes y 1,33 para los adultos).

Cuadro 7 Promedio de palabras nocionales por sexo y tramos de edad

	Sustantivos		Adjetivos		Verbos		Adverbios	
	M	F	M	F	M	F	M	F
19-34	24,3	16,6	2,33	1,66	19,33	22	8	10,66
35-54	23	26	1,33	1,33	15,33	21,66	7,33	5
> 55	21	21,6	3	4,66	19,66	19,66	7,66	6,33
	22,76	21,4	2,22	2,55	18,10	21,10	7,66	7,33

Por sexos, apreciamos varios datos de interés: las mujeres jóvenes utilizan más verbos y más adverbios que los hombres, pero menos sustantivos y menos adjetivos. Entre los adultos (35-54), el resultado está bastante equilibrado por sexos, salvo en el uso de verbos, también mayor para las mujeres. Por último, los informantes mayores de 55 años arrojan unos promedios muy similares, salvo en lo tocante al uso de los adjetivos, con un 4,66 en las mujeres frente a 3 en los hombres. Esto es más llamativo todavía si lo comparamos con el uso de adjetivos de las mujeres de otros tramos de edad, que aquí es prácticamente 4 veces superior, evidenciando la predilección de las mujeres mayores de 55 por calificar y describir con mayor detalle que los hombres de su tramo de edad y, en general, del conjunto de los informantes del corpus.

RECAPITULACIÓN

Hemos querido presentar con este trabajo una primera aproximación al estudio de la riqueza léxica en un corpus de español coloquial, a fin de ir calibrando las variables estadísticamente más significativas en el nivel de instrucción básico.

En líneas generales, los resultados obtenidos revelan un nivel de riqueza léxica muy bajo en todos los grupos y tramos de edad estudiados, y ello puede explicarse quizá por el abandono temprano de la escolarización y rápida entrada en el mercado laboral de los informantes del nivel sociocultural bajo, si bien la diferencia observada en el grupo de edad de los adultos apunta a un resultado ligeramente más favorable. En cualquier caso, un índice de riqueza léxica general de 39,63 es paupérrimo si lo comparamos con los resultados apuntados ya como bajos por López Morales (2011: 24), para quien, por ejemplo, un índice de 46 en estudiantes de secundaria sería manifiestamente mejorable, a la vista de los estándares establecidos por Ávila Sánchez (1986), Haché (1988) o Cintrón (1992), entre otros.

Por tanto, considerando la diferencia de medias de las variables examinadas, únicamente ha resultado estadísticamente significativa la variable "edad", tanto en el cómputo global como en el desglose según el tipo de palabra nocional utilizada.

Con todo, los datos que acabamos de analizar hacen necesario un estudio comparativo con los niveles de instrucción medio y alto del corpus PRESEEA Granada, que esperamos poder llevar a cabo en próximos trabajos.

 DOI: 10.23825/ITINERARIOS.28.2018.05

BIBLIOGRAFÍA

Andrés Pérez, Bernardo (1997) *Riqueza léxica en textos escritos de tres niveles de EGB*. Memoria de Licenciatura. Alcalá de Henares, Universidad de Alcalá.

Ávila Muñoz, Antonio Manuel (2014) "Patrones sociolingüísticos de la riqueza léxica. Estudio basado en una propuesta original para el cálculo del índice de la densidad léxica virtual de los hablantes". *Lingüística Española Actual*. 36: 249-272.

Ávila Sánchez, Francisco Raúl (1986) "Léxico infantil de México: Palabras, tipos, vocablos". En: *Actas del II Congreso Internacional sobre el «Español de América»*. México, UNAM: 510-517.

Baayen, R. Harald y Tweedie, Fiona J. (1998) "How variables may a constant be? Measures in lexical richness in perspective". *Computers and the Humanities*. 32: 323-352.

Capsada Blanch, Ramón y Torruella Casañas, Joan (2017) "Métodos para medir la riqueza léxica de los textos: Revisión y propuesta". *Verba: Anuario Galego de Filoloxia*. 44: 347-408.

Cintrón Serrano, Filomena (1992) *Índices de riqueza léxica en escolares de Barranquitas*. Tesis de maestría. San Juan, Universidad de Puerto Rico.

Echeverría, Max Sergio, Valencia, Alba, Ávila, Emilio, Muñoz Rigollet, Gloria, Núñez, Nicolás y Véliz, Mónica (1992) "Evaluación de la riqueza léxica de estudiantes de último año de enseñanza media". *Estudios Filológicos*. 27: 59-71.

Giraud, Pierre (1960) *Problèmes et méthodes de statistiquelinguistique*. Paris, Presses Universitaires.

Haché de Yunén, Ana Margarita (1991) "Aportes de las pruebas de riqueza léxica a la enseñanza de la lengua materna". En: Humberto López Morales (ed.) *La enseñanza del español como lengua materna*. Río Piedras, Universidad de Puerto Rico: 47-60.

Ham, Roberto (1979) "Del 1 al 100 en lexicografía". En: Luis Fernando Lara (ed.) *Investigaciones lingüísticas en lexicografía*. México, El Colegio de México: 41-83.

López Morales, Humberto (2011) "Los índices de riqueza léxica y la enseñanza de lenguas". En: Javier de Santiago Guervós, Hanne Bongaerts, Jorge J. Sánchez y Marta Seseña (coords.) *Del texto a la lengua: la aplicación de los textos a la enseñanza-aprendizaje del español L2-LE. ASELE, Asociación para la enseñanza del español como lengua extranjera*. 1: 15-28.

Manjón-Cabeza, Antonio, Pose Furest, Francisca y Sánchez García, Francisco José (2017) "El factor social edad y la expresión del sujeto pronominal en el español hablado de Granada". *Lingüística Española Actual*. 39 (1): 31-51.

Mitkova, Adriana (2007) "El léxico juvenil por áreas temáticas". *Tonos digital*. 14. https://www.um.es/tonosdigital/znum14/secciones/estudios-17-lexicojuvenil.htm

Moya Corral, Juan Antonio, coord. (2007, 2008, 2009) *El español hablado en Granada. Corpus para su estudio sociolingüístico. Nivel de estudios bajo, medio y alto*. Granada, Universidad de Granada.

Müller, Charles (1968) *Estadística lingüística*. Madrid, Gredos.

Pastor Milán, Mª Ángeles y Sánchez García, Francisco José (2008) *El léxico disponible de Granada y su provincia*. Granada, Universidad de Granada.

Portela, Clara (1992) *Índices de riqueza léxica en estudiantes de primer año universitario*. Tesis de maestría en Lingüística. Santiago de los Caballeros, Pontificia Universidad Católica Madre y Maestra.

Reyes Díaz, María Josefa (2007) "Apuntes para la enseñanza del vocabulario". *Revista de Filología*. 25: 529-538.

Těšitelová, Marie (1992) *The main areas of quantitative linguistics*. New York, Planum Press.

Torres González, Antonia Nelsi (2003) "Riqueza léxica en textos narrativos escritos por estudiantes de Tenerife". En: Francisco Moreno Fernández, Francisco Gimeno-Menéndez, José Antonio Samper, María Luz Gutiérrez, María Vaquero y César Hernández Alonso (coords.) *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*. Madrid, Arco Libros: 435-449.

----- (1999) "Incidencia de las variables sociales en los índices de producción léxica de estudiantes del último curso de la enseñanza no universitaria". En: Julián de la Cuevas y Dalila Fasla (eds.) *Contribuciones al estudio de la Lingüística Aplicada*. Castellón, Asociación Española de Lingüística Aplicada: 393-401.

SITOGRAFÍA

Instituto Nacional de Estadística Informática (2006) *Glosario básico de términos estadísticos*. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib0900/Libro.pdf [30.01.2018]