Capítulo X
Estudio lexicométrico del español hablado en Granada. El índice de riqueza léxica del corpus PRESEEA

Francisco José Sánchez García
Universidad de Granada

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 no disponíamos aún de otros indicadores lexicométricos. Este trabajo está dedicado al examen de la riqueza léxica del español coloquial hablado en Granada, y viene a completar uno anterior centrado en el nivel de estudios bajo del corpus PRESEEA Granada (Sánchez García 2018). Apoyándonos en la metodología de López Morales (2011), presentamos aquí el promedio de vocablos, intervalo de aparición de palabras de contenido nocional y, finalmente, el promedio de hápax a partir de las variables "nivel de instrucción", "sexo" y "edad". Se constata que el nivel de instrucción es una variable significativa, toda vez que los resultados obtenidos para los informantes con estudios secundarios y universitarios son superiores a los del nivel bajo en ambos sexos y todos los tramos de edad.
Palabras clave: riqueza léxica, PRESEEA Granada, hápax, estadística léxica, sociolingüística

Lexicometric study of the Spanish spoken in Granada. The Lexical Richness Index of the PRESEEA corpus

Abstract: In recent years, studies of quantitative linguistics and lexical statistics have proliferated, providing 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 precollege 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 an approximation to the lexical richness of colloquial Spanish spoken in Granada, which completes a previous one focused on the low level of studies in PRESEEA- Granada (Sánchez García 2018). Based on the methodology of López Morales (2011), we analyze here the average of words, interval of appearance of words of notional content and, finally, the average of hapax from the variables 'level of studies", 'gender' and 'age'. It is found that the variable 'level of education' is significant, since the results obtained for the informants with secondary and university studies are higher than those of the low level in both genders and all age groups.
Keywords: Lexical richness, PRESEEA Granada, hapax, lexical statistics, sociolinguistics

Sánchez García, F. J. (2021). "Estudio lexicométrico del español hablado en Granada. El índice de riqueza léxica del corpus PRESEEA". En *El español de Granada. Estudio sociolingüístico,* 279-300. Peter Lang.

1. Introducción

Esta investigación presenta los resultados globales de nuestro análisis lexicométrico del corpus PRESEEA-Granada. En un trabajo previo, realizamos una primera aproximación dedicada al estudio la riqueza léxica del nivel bajo (Sánchez García 2018), que completamos aquí con los niveles de instrucción medio y alto, a fin de ofrecer un examen exhaustivo de indicadores tales como el porcentaje de vocablos, intervalo de aparición de términos e índice de hápax, además del desglose de las palabras nocionales por categorías gramaticales. La hipótesis de partida de este trabajo pasa por evaluar cuantitativamente la calidad del léxico coloquial de los granadinos atendiendo a las variables sociales 'sexo', 'edad' y 'nivel de instrucción', en el supuesto de que, a mayor nivel de estudios, mayor será el índice de riqueza léxica de los informantes.

1.1. *Los estudios de riqueza léxica*

Los trabajos sobre el léxico y su enseñanza se han acrecentado en los últimos años. Si ponemos el foco en el ámbito hispánico, el investigador esencial que ha coordinado estos estudios es Humberto López Morales, principal impulsor de los estudios sobre la disponibilidad léxica. En primer lugar, conviene abordar la definición de léxico disponible, entendido como

> 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)

Sobre esta base, hemos de tener clara la diferencia entre palabras frecuentes y palabras disponibles; las palabras frecuentes (en denominación de Michéa, 1950) hacen referencia a las palabras 'atemáticas', es decir, unidades gramaticales (artículos, preposiciones y conjunciones) e, igualmente, en orden decreciente a verbos, adjetivos y sustantivos, que se actualizan en cualquier situación comunicativa. Las palabras disponibles (o palabras temáticas), tal y como abordábamos en Pastor Milán y Sánchez García (2008: 13) "representan el léxico potencial que se presenta en situaciones concretas y condicionadas por un tema que sirva de estímulo". Atendiendo a esto, la unión de ambos léxicos conforma el vocabulario esencial de una comunidad, el pilar sobre el que se asienta el idioma.

Expuesto lo anterior, es evidente que los estudios sobre léxico disponible son altamente significativos para comprender el lexicón mental de una comunidad de hablantes, singularmente en la etapa preuniversitaria, pero también presenta algunas carencias asociadas al modelo de encuesta (basado en los "centros de interés", una selección cerrada de temas que deben servir para que los informantes evoquen las palabras relacionadas con esos campos semánticos) del que únicamente se valoran los ya citados vocablos de carácter nocional, y, en concreto, sustantivos y adjetivos. Para un estudio que determine con rigor el grado conocimiento global del léxico no nos basta conocer el léxico disponible: hemos de ocuparnos también del índice de "riqueza léxica".

El concepto fue acuñado por Guiraud (1954), cuyo interés se centraba en la correlación del número de palabras y vocablos de un texto a fin de obtener un índice válido que permitiera cuantificar el potencial lingüístico de los hablantes de cualquier

idioma. El semiólogo instituyó la diferencia entre palabra (considerada como unidad del texto; es decir, material gráfico comprendido entre dos espacios en blanco) y vocablo (entendida como unidad del léxico para referirse a las palabras diferentes que podemos encontrar en un texto). Con ello se inauguran los estudios léxico-estadísticos que progresivamente y, en las siguientes décadas, han ido ganando interés e impacto en nuestro campo de estudio. Hay que relievar igualmente las aportaciones de Müller (1968), quien, centrándose a su carácter gramatical, se ocupó del análisis cuantitativo de la frecuencia de vocablos en un texto. De ahí se derivó otro hito significativo: acuñar el término hápax para referirse a aquellos vocablos que solo aparecen en un texto una sola vez y que se calcula dividiendo el número total de vocablos por aquellos que tienen frecuencia 1.

A fin de obtener el índice de riqueza léxica, Guiraud plantea una fórmula partiendo de esa distinción antes aludida entre palabras nocionales o de contenido semántico (sustantivos, adjetivos calificativos, verbos y adverbios) y gramaticales (artículos, preposiciones, conjunciones, pronombres y adjetivos determinativos).
De esta forma, 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.

El método, que se ha revelado como eficaz, apenas ha experimentado variaciones a lo largo de los años; tan solo pequeñas modificaciones como la de Těšitelová (1992), quien sugiere valorar la repetición de palabras de un texto, el valor de la zona de palabras de baja frecuencia (comprendida entre 1-10), y los fenómenos de dispersión y concentración del vocabulario. Algunos años antes, el mexicano experto en sociosemántica Raúl Ávila había planteado un método para calcular la riqueza léxica atendiendo a "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). Desde ese planteamiento, Haché (1991), Ham (1979) y el propio Ávila (1986), consideran que, a partir de la centena de unidades obtenidas, el promedio de vocablos deja de crecer de manera relevante. Atendiendo a ello se calcula el coeficiente de densidad léxica si dividimos el número de tipos léxicos (T) que aparecen 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, cuantificar las frecuencias acumuladas por deciles resulta especialmente útil para comparar la riqueza léxica de textos que no tienen la misma extensión

Sobre esa base, Humberto López Morales introduce una fórmula para determinar el porcentaje de vocablos. Como vemos, el proceso consiste en dividir el número total de vocablos entre el total de unidades léxicas para, a continuación, multiplicar el resultado por 100:

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

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

El resultado es lo que él define como un "indicador grueso" que por sí solo no tendría la utilidad requerida si no se complementara con el intervalo de aparición de palabras de contenido nocional (IAT):

Procede ahora definir las palabras nocionales que poseen una gran relevancia dentro del corpus que se examinará: se trata de aquellas que, como anticipábamos, poseen un contenido semántico (sustantivos, adjetivos, verbos y adverbios) a diferencia de las formas gramaticales. Atendiendo a lo expuesto y nuevamente según López Morales, "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)". (López Morales 2011: 20).

Atendiendo a los resultados del IAT se concluye que, cuanto mayor resulte el número de palabras nocionales el intervalo será menor lo que supone que será más alto el índice de riqueza léxica. Se trata de una fórmula especialmente interesante para cotejar entre individuos particulares[1], a fin de estudiar la relación de un sujeto con el resto de la muestra o un determinado grupo de informantes.

Hemos de advertir, no obstante, que nuestro trabajo se apoya en un corpus del español coloquial, de modo que los resultados no son directamente equiparables a los obtenidos en estudios como el de Reyes Díaz (2017-2018), que analiza la riqueza léxica de los estudiantes preuniversitarios a partir de textos escritos; evidentemente, la riqueza léxica de una redacción siempre arrojará un resultado más alto que el discurso oral.

2. 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 (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 los 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, distribuidos en tres niveles de instrucción (nivel de estudios primarios, secundarios y universitarios)

[1] Para tal fin, resulta también muy prometedora la propuesta de Ávila Muñoz (2017), que ha desarrollado un algoritmo para calcular el "léxico virtual", entendido como la capacidad léxica de una selección de individuos, basado en la estimación del vocabulario de que disponen en su lexicón mental. Su modelo ha sido probado con una muestra de hablantes de la ciudad de Málaga.

y tres tramos de edad (jóvenes —entre 19 y 34 años—, adultos —entre 35 y 54— y mayores de 55), además de la variable de sexo. La distribución de informantes queda como sigue:

Tabla 1

Distribución de los informantes del corpus PRESEEA-Granada

Edad	E. Primarios		E. secundarios		E. Universitarios	
	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre
19-34	3	3	3	3	3	3
35-54	3	3	3	3	3	3
Mayor de 55	3	3	3	3	3	3

En la misma línea que investigaciones previas como la de Manjón, Pose y Sánchez (2017) sobre la expresión del sujeto pronominal, se ha decidido analizar fragmentos de cien palabras de cada informante, habida cuenta que el promedio no se alteraría de manera significativa atendiendo a una mayor amplitud textual. Con todo, nuestro propósito no es otro que la obtención de resultados significativos, fiables y verificables, que puedan ser cotejados con los de otras áreas geográficas estudiadas en el entorno PRESEEA.

En primer lugar, hemos escogido un fragmento suficientemente extenso de cada una de las entrevistas, sin solapamientos con el/la entrevistador/a; para cada texto, nos hemos servido únicamente de las cien primeras palabras, de las que se han descartado las onomatopeyas, nombres propios, interjecciones y anacolutos. Lógicamente, como corresponde a una investigación de estas características, nos apoyamos en el concepto de "unidad léxica", entendiendo por tal 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 el software AntConc (Anthony 2019). 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 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-H11-037), escogido aleatoriamente hasta obtener un total de 100 palabras:

> (1) 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 (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 [GRAN-H11-037].

Siguiendo el método propuesto por López Morales, 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 diez o quince 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 diez en diez 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.

Tabla 2

Distribución por deciles del número de palabras diferentes

Nº de palabras	Nº de 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:

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

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 acuerdo con este indicador, 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 Humberto López Morales sobre un corpus de estudiantes de secundaria, que por ejemplo pueden encontrarse en torno al 1,5-2, 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):

Hápax= V/V_1= 39/23 =1,69.

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.

3. Análisis de los resultados

3.1. *El nivel de estudios bajo*

Seguidamente, rescatamos los resultados obtenidos en el estudio de los informantes del nivel de estudios primarios ya adelantados en nuestra primera toma de contacto con el léxico del corpus PRESEEA-Granada (Sánchez García 2018), para su cotejo posterior con los dos niveles restantes. 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 la tabla siguiente.

Tabla 3

Recuento global y resultados del IAT y Hápax por informante (nivel de estudios bajo)

	Palabras nocionales	Resto de palabras	Vocablos	Intervalo	Hápax
GRAN-H11-037	60	41	39	2,56	1,69
GRAN-H11-038	48	52	36	2,77	1,5
GRAN-H11-039	54	46	41	2,43	1,36
GRAN-M11-040	57	43	35	2,85	1
GRAN-M11-041	50	50	38	2,63	1,31
GRAN-M11-042	46	54	38	2,63	1,15
GRAN-H21-043	48	52	35	2,85	1,45
GRAN-H21-044	47	53	44	2,27	1,1
GRAN-H21-045	46	54	40	2,5	1,14
GRAN-M21-046	58	42	47	2,12	1,17
GRAN-M21-047	57	43	47	2,12	1,23
GRAN-M21-048	47	53	43	2,32	1,1
GRAN-H31-049	57	43	39	2,56	1,3
GRAN-H31-050	50	50	39	2,56	1,39
GRAN-H31-051	47	53	39	2,56	1,21

GRAN-M31-052	55	45	37	2,7	1,6
GRAN-M31-053	49	51	37	2,7	1,15
GRAN-M31-054	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’.

Tabla 4

Porcentaje de vocablos, IAT e índice de hápax por tramos de edad (nivel de estudios bajo)

	Vocablos	Intervalo	Hápax
19-34	38,15	2,64	1,33
35-54	42,6	2,36	1,19
Mayor de 55	38,15	2,61	1,31
Media	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 nuestro corpus, los tramos de edad arrojan resultados interesantes, aunque ninguno de los tres grupos etarios evidencia un índice de producción léxica elevado, 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 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.

Tabla 5

Porcentaje de vocablos, IAT e índice de hápax por tramos de edad (nivel de estudios bajo)

	Vocablos		Intervalo		Hápax	
Edad	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
Mayor de 55	39	37,3	2,56	2,67	1,3	1,33
Media	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 más próximos 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:

Tabla 6

Porcentaje de palabras nocionales por tramos de edad (nivel de estudios bajo)

	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
Mayor de 55	21,3	3,83	19,66	6,99	51,78
Media	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 duplican a los otros dos grupos (1,99 para los jóvenes y 1,33 para los adultos).

Tabla 7

Porcentaje de palabras nocionales por sexo y tramos de edad (nivel de estudios bajo)

	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
Mayor de 55	21	21,6	3	4,66	19,66	19,66	7,66	6,33
Media	22,76	21,4	2,22	2,55	18,10	21,10	7,66	7,33

Por sexos sí apreciamos varios datos de interés: las mujeres jóvenes utilizan más verbos y más adverbios que los hombres, pero menos sustantivos y 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.

3.2. *El nivel de estudios medio*

Los datos globales para cada informante del nivel de estudios medio ya revelan un incremento significativo en el total de palabras de contenido semántico, al igual que en el resto de los indicadores.

Tabla 8

Recuento global y resultados del IAT y Hápax por informante (nivel de estudios bajo)

	Palabras nocionales	Resto de palabras	Vocablos	Intervalo	Hápax
GRAN-H12-019	71	29	42	2,38	1,61
GRAN-H12-020	54	46	39	2,56	1,44
GRAN-H12-021	66	34	52	1,92	1,23
GRAN-M12-022	61	39	49	2,04	1,63
GRAN-M12-023	62	38	37	2,7	1,48
GRAN-M12-024	58	42	49	2,04	1,08
GRAN-H22-025	66	34	53	1,88	1,76
GRAN-H22-026	63	37	54	1,85	1,54
GRAN-H22-027	64	36	43	2,32	1,89
GRAN-M22-028	62	38	40	2,5	1,29
GRAN-M22-029	62	38	50	2,0	1,31
GRAN-M22-030	64	36	58	1,72	1,11
GRAN-H32-031	58	42	49	2,04	1,16
GRAN-H32-032	53	47	44	2,27	1,06
GRAN-H32-033	66	34	57	1,75	1,16
GRAN-M32-034	57	43	45	2,22	1,28
GRAN-M31-034	57	43	46	2,17	1,24
GRAN-M32-035	66	34	55	1,81	1,14

Analizaremos ahora los tres indicadores de riqueza léxica por tramos de ‘edad’.

Tabla 9

Porcentaje de vocablos, IAT e índice de hápax por tramos de edad (nivel de estudios bajo)

	Vocablos	Intervalo	Hápax
19-34	44,66	2,27	1,40
35-54	49,66	2,04	1,48
Mayor de 55	49,33	2,04	1,17
Media	47,88	2,11	1,35

Como vemos en la tabla 8, los informantes jóvenes son quienes aportan un porcentaje de vocablos ligeramente inferior al resto, aunque el promedio de este nivel de estudios secundarios es bastante aceptable (47,8). Tanto los adultos como los mayores de 55 años presentan un IAT de 2,04, que se sitúa en los estándares estudiados por López Morales.

En esta ocasión, la variable ‘sexo’ tampoco arroja resultados relevantes.

Tabla 10

Porcentaje de vocablos, IAT e índice de hápax por tramos de edad (nivel de estudios bajo)

	Vocablos		Intervalo		Hápax	
Edad	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino
19-34	44,33	45	2,28	2,27	1,42	1,39
35-54	50	49,33	2,01	2,07	1,73	1,23
Mayor de 55	50	52	2,02	2,06	1,12	1,22
Media	48,11	48,77	2,10	2,13	1,42	1,28

La semejanza entre los informantes de ambos sexos por cada tramo de edad es evidente en los tres indicadores, a excepción del índice de hápax de los varones adultos (entre 35-54 años), que muestran un porcentaje ligeramente desfavorable con respecto a las mujeres.

Si nos centramos en la distribución morfológica de las palabras nocionales, encontramos una clara preferencia por los verbos en todos los tramos de edad, aunque

este resultado está casi equilibrado con el obtenido para los sustantivos, que destacan especialmente entre los informantes del grupo de edad intermedio.

Tabla 11

Porcentaje de palabras nocionales por tramos de edad (nivel de estudios bajo)

	Sustantivos	Adjetivos	Verbos	Adverbios	Total
19-34	18,83	7,2	20,3	12,6	58,93
35-54	20,4	4,95	21,9	13,15	60,4
Mayor de 55	18,5	3,3	24,6	7,99	54,39
Media	19,24	5,15	22,26	11,24	57,9

Tabla 12

Porcentaje de palabras nocionales por sexo y tramos de edad (nivel de estudios bajo)

	Sustantivos		Adjetivos		Verbos		Adverbios	
	M	F	M	F	M	F	M	F
19-34	19,3	18,3	8,8	5,6	20,3	20,3	11,3	14
35-54	23,6	17,3	4,3	5,6	19,6	24,3	12	14,3
Mayor de 55	18	19	3	3,6	23,6	25,6	9,3	9,6
Media	20,3	18,2	5,3	4,9	21,16	23,4	10,88	12,6

Si cruzamos las variables ‘sexo’ y edad’, observamos una preferencia clara por el verbo entre las mujeres de edad intermedia (24,3% del total de registros), mientras que los varones de ese tramo etario precisamente destacan por el uso del sustantivo (23,6%). El resto de los resultados obtenidos está bastante equilibrado en ambos sexos, lo que confirma nuevamente que se trata de una variable poco representativa.

3.3. *El nivel de estudios alto*

Repasemos ahora los datos globales para cada uno de los informantes del nivel de estudios alto (estudios universitarios).

Tabla 13

Recuento global y resultados del IAT y Hápax por informante (nivel de estudios bajo)

	Palabras nocionales	Resto	Vocablos	Intervalo	Hápax
GRAN-H31-001	61	39	54	1,85	1,08
GRAN-H31-002	57	43	46	2,17	1,17
GRAN-H31-003	58	42	43	2,32	1,26
GRAN-M31-004	56	44	41	2,43	1,24
GRAN-M31-005	60	40	52	1,92	1,13
GRAN-M31-006	54	46	42	2,38	1,35
GRAN-H32-007	57	43	48	2,08	1,2
GRAN-H32-008	59	41	49	2,04	1,13
GRAN-H32-009	56	44	44	2,27	1,18
GRAN-M32-010	51	49	44	2,27	1,15
GRAN-M32-011	60	40	49	2,04	1,16
GRAN-M32-012	57	43	48	2,08	1,2
GRAN-H33-013	57	43	48	2,08	1,18
GRAN-H33-014	56	44	50	2	1,13
GRAN-H33-015	66	34	52	1,92	1,23
GRAN-M33-016	58	42	47	2,12	1,14
GRAN-M33-017	59	41	44	2,27	1,33
GRAN-M33-018	59	41	52	1,92	1,23

Como puede comprobarse en la tabla 12, el promedio de los datos de los informantes es semejante a los del nivel de estudios medio. Hemos obtenido un PV global de 47,3 y un intervalo muy positivo, de 2,11, siendo más alto entre los mayores, lo que ratifica la tendencia evidenciada en el nivel medio a propósito de la variable 'edad'.

Tabla 14

Porcentaje de vocablos, IAT e índice de hápax por tramos de edad (nivel de estudios bajo)

	Vocablos	Intervalo	Hápax
19-34	46,3	2,17	1,18
35-54	47	2,13	1,17
Mayor de 55	48,8	2,05	1,20
Media	47,3	2,11	1,18

En efecto, se observa una clara correlación entre la edad y la riqueza léxica también en el nivel alto, toda vez que los resultados obtenidos son progresivamente más favorables a medida que aumenta el tramo etario de los informantes. También aquí se alcanza un promedio de entorno a 50% de porcentaje de vocablos, y un intervalo bastante aceptable, de 2,05 en los mayores de 55.

Tabla 15

Porcentaje de vocablos, IAT e índice de hápax por tramos de edad (nivel de estudios bajo)

	Vocablos		Intervalo		Hápax	
Edad	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino
19-34	47,6	45	2,11	2,24	1,17	1,20
35-54	47	47	2,13	2,13	1,17	1,17
Mayor de 55	50	47,6	2	2,1	1,18	1,23
Media	48,2	46,53	2,08	2,15	1,17	1,20

Esta curva ascendente se confirma también si examinamos por separado los informantes de ambos sexos: a excepción del grupo de varones jóvenes, el resto de los valores obtenidos evidencian una mejora clara tanto en hombres como en mujeres adultos y mayores de 55 años.

Por último, nos fijaremos en la distribución morfológica de las palabras nocionales para el nivel alto.

Tabla 16

Porcentaje de palabras nocionales por tramos de edad (nivel de estudios bajo)

	Sustantivos	Adjetivos	Verbos	Adverbios	Total
19-34	20,45	4,8	19,15	11,45	55,85
35-54	24,3	4,95	18,3	8	55,55
Mayor de 55	23,3	3,95	18,15	11,95	57,35
	22,68	4,56	18,53	10,46	56,23

En este caso, la categoría más destacada es el sustantivo (seguido del verbo) en todos los tramos de edad, especialmente entre los adultos y los mayores.

Tabla 17

Porcentaje de palabras nocionales por sexo y tramos de edad (nivel de estudios bajo)

	Sustantivos		Adjetivos		Verbos		Adverbios	
	M	F	M	F	M	F	M	F

19-34	20,6	20,3	4,6	5	20	18,3	12,3	10,6
35-54	22,6	26	5,6	4,3	19	17,6	9	7
Mayor de 55	25,3	21,3	3,3	4,6	19	17,3	10,3	13,6
	22,8	22,5	4,5	4,6	19,3	17,7	10,5	10,4

Los resultados del nivel de estudios alto por sexo y tramos de edad también muestran un equilibrio entre los informantes de cada grupo etario, a excepción del grupo de edad entre 35-54 (mayor uso de sustantivos en las mujeres: 26/22,6) y el de los adultos (aquí la preferencia se invierte, 25,3 para los varones/21,3 para las mujeres). En cualquier caso, para la variable sexo no apreciamos divergencias con relación al promedio general.

4. Conclusiones

El análisis de los datos obtenidos nos permite confirmar la hipótesis planteada al inicio, toda vez que ha quedado demostrada la correlación entre el nivel de instrucción y la riqueza léxica de los informantes. Los registros de los niveles de educación secundaria y universitaria revelan un significativo incremento de los índices examinados (porcentaje de vocablos e intervalo de aparición de términos) en contraposición con los resultados del nivel de estudios primarios, en el que hallamos un nivel de riqueza léxica muy pobre en todos los grupos y tramos de edad estudiados, lo que podría explicarse 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 ínfimo si lo comparamos con los resultados apuntados ya como bajos por Humberto 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 (1986), Haché (1988) o Cintrón (1993), entre otros. El único indicador que ha roto con la tendencia general es el índice de hápax, para el que hemos documentado un registro más desfavorable en el nivel medio de instrucción (1,35 global, con 1,42 para los varones y 1,28 para las féminas) frente a 1,27 del nivel bajo. Sí que obtenemos un dato positivo en el nivel alto (1,18).

Por tanto, considerando los datos globales, 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. Precisamente, uno de los resultados más interesantes se explica por el factor edad. Claramente, el promedio de vocablos, así como el intervalo de aparición de términos se incrementan gradualmente a medida que aumenta la edad de los informantes, especialmente en los niveles medio y alto. Este dato puede atribuirse a dos causas posibles: el factor experiencia y las carencias del sistema educativo. En primer lugar, la experiencia y el bagaje cultural que dan los años repercutirían positivamente en la riqueza léxica de los hablantes, independientemente del nivel de instrucción (la tendencia es más favorable para los mayores en todos, pero especialmente en los niveles medio y alto). En segundo lugar, este resultado constataría el fracaso del sistema educativo español en la última década, que, tal y como refleja el último informe PISA (2018), sigue evidenciando serias carencias en la competencia lectoescritora de los estudiantes preuniversitarios españoles, y especialmente de los andaluces. Evidentemente, dicha competencia se halla ligada al lexicón mental o caudal de vocabulario de los hablantes, o lo que es lo mismo: su riqueza léxica.

Referencias bibliográficas

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. *LEA. Lingüística Española Actual, 36,* 249-272.

Ávila Muñoz, Antonio Manuel (2016). Can speakers' virtual lexical richness be calculated? Individual and social determining factors. *Spanish in Context, 13(2),* 285-307. DOI: dx.doi.org/10.1075/sic.13.2.06avi

Ávila, Raúl (1986). Léxico infantil de México: Palabras, tipos, vocablos. En *Actas del II Congreso Internacional sobre el español de América*, (pp. 510-517). México D.F.: Universidad Nacional Autónoma de México.

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, 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. DOI: https://doi.org/10.15304/verba.44.3155

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

Echeverría, Max; Valencia, Alba, *et alii*. (1992). Evaluación de la riqueza léxica de estudiantes de último año de enseñanza media. *Estudios Filológicos, 27,* 59-71.

Guiraud, Pierre (1960). *Problèmes et méthodes de statistique linguistique*. 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,* (pp. 47-60). Río Piedras: Universidad de Puerto Rico.

López Morales, Humberto (2011). Los índices de riqueza léxica y la enseñanza de lenguas. En Javier de Santiago Guervós, Hanne Bongaerts, Jorge Juan Sánchez Iglesias y Marta Seseña Gómez (coords.) *Del texto a la lengua: la aplicación de los textos a la enseñanza-aprendizaje del español L2-LE,* vol. I, (pp. 15-28). Madrid: ASELE.

López Morales, Humberto (1984). *La enseñanza de la lengua materna. Lingüística para maestros de español*. Madrid: Playor.

Manjón-Cabeza, Cruz, Antonio; Pose Furest, Francisca y Sánchez García, Francisco José. (2017). Factores determinantes en la expresión el sujeto pronominal en el corpus PRESEEA Granada. *Boletín de Filología, 51 (2),* 181-207.

Moya Corral, Juan Antonio, coord. (2007). *El Español hablado en Granada. Corpus oral para su estudio sociolingüístico. I Nivel de estudios alto*. Granada: Editorial Universidad de Granada.

Moya Corral, Juan Antonio, coord. (2008). *El Español hablado en Granada II. Corpus oral para su estudio sociolingüístico. Nivel de estudios medio*. Granada: Editorial Universidad de Granada.

Moya Corral, Juan Antonio, coord. (2009). *El Español hablado en Granada III. Corpus oral para su estudio sociolingüístico. Nivel de estudios bajo*. Granada: Editorial 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: Editorial 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, Mª Josefa (2007). Apuntes para la enseñanza del vocabulario. *Revista de Filología*, *25*, 529-538.

Reyes Díaz, Mª Josefa (2007-2008). Riqueza léxica de textos redactados por alumnos de Bachillerato de Las Palmas de Gran Canaria. *Anuario de Lingüística Hispánica*, *23-24*, 147-163.

Sánchez García, Francisco José (2018). El índice de riqueza léxica en el nivel de estudios bajo del corpus PRESEEA-Granada. *Itinerarios. Revista de Estudios Lingüísticos, 28*, 95-107. DOI:10.23825/ITINERARIOS.28.2018.05

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, Mª Luz Gutiérrez Araus, María Vaquero y César Hernández Alonso (coords.), *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*, (pp. 435-449). Madrid: Arco Libros.

Torres González, Antonia Nelsi (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*, (pp. 393-401). Castellón: Asociación Española de Lingüística Aplicada.