LOS ÍNDICES DE 'RIQUEZA LÉXICA' Y LA ENSEÑANZA DE LENGUAS

Humberto López Morales
Secretario General de la Asociación de Academias de la Lengua Española

1. INTRODUCCIÓN

La llamada 'calidad de la escritura' está integrada por una serie de factores, entre los que destacan sin duda: la riqueza léxica, la madurez sintáctica, los esquemas de cohesión y la coherencia discursiva. La amplitud y variedad del vocabulario está muy apoyado en la disponibilidad léxica del hablante, la madurez sintáctica, en su grado de entrenamiento combinatorial de oraciones simples en el discurso, los esquemas de cohesión y la coherencia discusiva dependen esencialmente del 'orden' que quiera dársele conscientemente a los elementos constitutivos del discurso.

Todo esto se aprende naturalmente desde los primeros años de enseñanza, si hablamos de lengua materna, y desde el primer curso si nos referimos a segundas lenguas. Pero para adquirir estas destrezas comunicativas se necesita un programa de enseñanza moderno y actualizado, que vaya mucho más allá de peticiones como subrayar en rojo el sujeto de una oración dada y en azul el predicado. Confundir la enseñanza de la gramática de una lengua, con la enseñanza-adquisición de esa lengua es un grave error que, por fortuna, va desapareciendo aunque lentamente[1].

1 Hace ya casi 70 años el germano-chileno Rudolf Lenz escribió lo siguiente: "Querer aprender una lengua por el estudio de su gramática es como aprender a tocar el violín leyendo tratados de música y métodos de violín sin tocar el instrumento, sin ejercitar los dedos", y una década más tarde, Américo Castro repetía la idea, aunque acudiendo a otras comparaciones: "Una primera confusión que conviene remover es la idea absurda de que el idioma se enseña enseñando gramática […] La gramática no sirve para enseñar a hablar y a escribir correctamente la lengua propia (podría añadirse que tampoco las extranjeras) lo mismo que el estudio de la fisiología o de la acústica no enseñan a bailar o que la mecánica no enseña a montar en bicicleta"

Un ejemplo será suficiente para ilustrarlo. Se trata del texto de un alumno de cuarto grado de primaria de escuela pública de Puerto Rico.

> La playa es muy bonita. La playa es azul. El agua de la playa es salada. En el mar hay peces, tiburones pez espada, pez martillo. Mi papá, mi mamá, mis hermanos y yo fuimos a la playa. Nos divertimos mucho. Nos bañamos mucho. Brincamos y saltamos mucho. Nos gusta ir a la playa con nuestros familiares. En la playa es bueno estar. Hace mucho fresco. Es bueno el ambiente. Mis hermanos juegan mucho. Y no también. A mí me gusta ir a la playa con mis primos, mis tíos, mis hermanos y mis abuelos.

Aunque se puede ver de inmediato que estamos ante un texto muy deficiente y no solo en cuanto a riqueza léxica, es asunto que será analizado con detalle más adelante. Desde el punto de vista de la madurez sintáctica es también muy pobre pues ninguna de las 15 oraciones simples de que consta ha recibido el menor tratamiento combinatorial del tipo 'La playa es muy bonita, azul y de agua salada', 'Nos divertimos, nos bañamos y brincamos y saltamos mucho' o 'En la playa es bueno estar porque hace mucho fresco y el ambiente es bueno'. Aquí se trata de inhabilidad y falta de entrenamiento, no de intención expresa de simplificar el discurso como sello estilístico, como sí es el caso de Azorín frente a Unamuno, por ejemplo, según han demostrado María Antonieta Andión y Ana María Ruiz (1996) con estadísticas en la mano.

Tampoco es un caso ejemplar de cohesión discursiva. La cadena 'playa', la más importante de este texto consigue índices bajísimos, pues la palabra 'playa' está presente en 7 de las 15 oraciones:

1, 2, 3, 0, **5,** 0, 0, 0, 0, **9, 10**, 0, 0, 0, **15** = 46,6%

Ante este resultado se estará casi convencido sin necesidad de efectuar ningún análisis de que la coherencia del texto queda puesta en entredicho.

2. LA RIQUEZA LÉXICA

El estudio del léxico, como es sabido, cuenta con diferentes vías de aproximación; las cualitativas y las cuantitativas. Dentro de las primeras caen los estudios de frecuencias, los referenciales, muy unido a los campos semánticos, y los relacionales, que se

y añadiría "Eso es de tal vulgaridad, que avergüenza tener que escribirlo una y otra vez" Sin embargo, estos deslindes entre lo teórico, representado por las gramáticas-reflexión, y el desarrollo de las destrezas comunicativas, no han llegado a su fin, como cabría imaginar; todavía se hace necesario seguir insistiendo en que la enseñanza de la gramática teórica tiene unos objetivos muy precisos que, desde luego, no son la base para la adquisición de estrategias y habilidades que permitan expresarse satisfactoriamente y comunicarnos con éxito. Vid. López Morales (1984).

ocupan de hiponimias, sinonimias, antonimias y homonimias. En el ámbito de las cuantitativas encontramos, entre otros, lo relativo a los léxicos básicos, los 'disponibles', y la 'riqueza léxica'.

Los primeros estudios de riqueza léxica aparecen en 1954 de la mano de Giraud. A pesar de que con posterioridad estas investigaciones se han ido refinando cada vez más, los postulados iniciales básicos se mantienen: así la relación existente entre el número de palabras y de vocablos de un determinado texto como elemento básico del análisis. Desde entonces quedó claro la diferencia entre *palabra* y *vocablo*: la primera, el material gráfico comprendido entre dos espacios en blanco de un texto, y el segundo, palabras diferentes que aparecen en un texto, sin contar las repeticiones.

La léxico-estadística ha sido el primer peldaño en la constitución de la estadística lingüística; la estadística léxica o léxico-estadística abarca el conjunto de operaciones, a veces sumamente complejas, que toman como unidades de trabajo las palabras y los vocablos; la palabra, unidad del texto, y el vocablo, unidad del léxico.

Müller (1968) pensaba que la estructura de un vocabulario comprende elementos cuantitativos simples que son el número de vocablos del texto y la frecuencia de cada uno de ellos; y elementos cualitativos que son la naturaleza gramatical de los vocablos y las relaciones de asociación (gramaticales o semánticas, paradigmáticas y sintagmáticas) que existen entre vocablos. Para llevar a cabo su conteo Müller define la norma lexicológica o norma de despojo como el conjunto de reglas o de convenciones que en el despojo cualitativo de un texto garantiza la constancia del tratamiento o de sus resultados. Para él, cuantificar vocabulario de un texto, es proceder a dos operaciones distintas que pueden ser sucesivas o simultáneas:

> A. El recuento de las palabras que componen el texto y cuyo número, representado por N, dará una medida de la extensión del texto, y
>
> B. El recuento de los vocablos empleados en el texto, y cuyo número, representado por V, mide la extensión del vocabulario.

La norma lexicológica debe dar reglas para delimitar la palabra y el vocablo. Por lo general se adopta la solución de sentido común que rige a los diccionarios, a pesar de las objeciones lingüísticas que se podrían formular. De esta manera se obtienen los índices de formas, que indican las ocurrencias de las palabras, es decir, las características estadísticas de las entradas, el lugar de aparición en el texto, el vocabulario común de una lengua y la jerarquización por grupos de mayor a menor presencia en la lengua.

Uno de los resultados más novedosos de la léxico-estadística consiste en confirmar que el vocabulario de los hablantes de una comunidad de habla es relativamente limitado; pues son muy pocas las palabras que presentan una frecuencia alta. Comúnmente el

individuo concentra sus necesidades de expresión en una cantidad relativamente reducida de entradas. Los estudios realizados en Francia hace casi cincuenta años mostraron que, independientemente de los términos de la especialidad laboral de cada cual, un hablante culto usa cuatro o cinco mil vocablos, mientras que el no culto puede manejarse con entre dos y tres mil.

3. LA MEDICIÓN DE LA RIQUEZA LÉXICA

Dentro del ámbito de la léxico-estadística el estudio de los índices de riqueza léxica ha sido un hito de mucha importancia. Las fórmulas, necesarias para su estudio, se han ido produciendo desde temprano: Giraud (1954), Ham (1979), López Morales (1984), Ávila (1986) y Tesitelová (1992).

Para calcular la riqueza léxica Giraud tomó en consideración dos tipos de palabras representadas en V: nocionales o palabras con contenido semántico (sutantivos, adjetivos calificativos, verbos y adverbios) y gramaticales o palabras funcionales del discurso (artículos, preposiciones, conjunciones, pronombres y adjetivos no calificativos). La extensión del texto se representa por N.

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

Cuando se toman en cuenta los dos grupos se utiliza la primera fórmula. En cambio el cómputo de las palabras nocionales exclusivamente requiere de la segunda fórmula en la que se duplica la extensión el texto (N), puesto que Giraud asumía que las palabas nocionales representaban la mitad del texto. Giraud a su vez añadió la fórmula de concentración de vocabulario para indicar la proporción de frecuencia total de las primas 50 palabras.

Tesitelová (1992), por su parte, elabora una interesante propuesta que recoge nuevos caracteres: repetición de palabras de un texto, la fuerza de la zona de palabras en baja frecuencia (1-10), la dispersión del vocabulario y la concentración del vocabulario.

Según Ham (1979), el principio que sustenta el empleo de la frecuencia de uso de vocablos para medidas de riqueza léxica es que no basta conocer el número de los existentes en la lengua o en una subordinación de esta, sino que también hay que tomar en cuenta la frecuencia con la que son utilizados. Aunque en una lengua haya muchos vocablos disponibles no podemos decir que realmente existe gran riqueza léxica si al momento de la producción de los textos la frecuencia de uso de estos se concentra en una minoría, haciéndose uso no significante de los demás. Por el contrario si en otra lengua que tuviera o no menos vocablos, aunque no sustancialmente menor, se hiciera un uso más disperso de los vocablos, tendríamos en realidad una situación de mayor riqueza

léxica manifestada en el hecho de mayor aprovechamiento del vocabulario disponible. Estos planteamientos ofrecen puntos de vista que amplían la noción clásica de riqueza léxica y que no deben olvidarse al evaluar el desarrollo del niño.

Ávila, por su parte, propone tres procedimientos comparativos para evaluar la riqueza léxica: el número de vocablos recogidos en el total de textos, la densidad promedio para 100 palabras y el número de vocablos acumulados por deciles.

La primera valoración de la riqueza léxica que establece Ávila pone en relación el número de palabras diferentes o vocablos (V) frente a la extensión del texto (N). Esta primera medida apunta hacia la diferencias en el vocabulario disponible. Para Ávila el léxico disponible es de baja frecuencia. El autor maneja este índice para comparar el número de vocablos que se obtienen en segmentos extensos de igual longitud.

El coeficiente de densidad léxica como segunda valoración, se obtiene al dividir el número de tipos léxicos (T) que parecen en un segmento del texto de una longitud determinada entre el número de palabras del segmento (N). La fórmula queda expresada del siguiente modo:

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

Este índice se basa en la evaluación individual de cada uno de los textos. Por lo tanto, es posible observar el comportamiento de la muestra en su conjunto con el apoyo de esta medida.

El tercer medio de valoración aplicado por Ávila –las frecuencias acumuladas por deciles- requiere ordenar los vocablos en frecuencia descendente de frecuencias en conjuntos de diez. Es decir, la lista de vocablos se inicia con el de frecuencia más alta y termina con los de más baja. La lista de vocablos ordenados por frecuencias descendente revela aquellos cuyas frecuencias acumuladas son suficientes para cubrir el 10%, 20% o 100% del total de frecuencias de un conjunto de textos. Ávila utiliza este índice para analizar textos de diferente longitud. Este procedimiento permite evaluar comparativamente el vocabulario de los conjuntos, ya que la extensión de cada uno de ellos, sus respectivas frecuencias no condiciona el número de vocablos que obtienen determinados deciles[2].

Para calcular la riqueza léxica, López Morales presenta dos fórmulas interrelacionadas que representan dos medidas obtenidas a base de cálculos empíricos: por una parte, toma en cuenta el porcentaje de vocablos (PV) del total de palabras de un texto

2 El procedimiento de evaluación de la riqueza léxica de los textos que hace Ávila sigue parcialmente a P. L. Baldi (1972). Una versión más detallada del pensamiento de estos autores, en Pastor, 1998:6-16.

(N), y por otra, mide el intervalo de aparición en el texto de palabras de contenido semántico nocional (IAT).

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.

La fórmula queda expresa de esta manera:

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

Este primer índice (PV) nos ofrece una visión de diversidad léxica y sirve como indicador grueso.

Para cuantificar el intervalo de aparición de palabras de contenido nocional (IAT), López Morales propone la siguiente fórmula:

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

El resultado de esta operación matemática refleja cifras relacionadas con la proporción de palabras nocionales en el texto. Esto es, a mayor número de palabras nocionales, menor es el intervalo, lo que se interpreta como mejor índice de riqueza léxica. Esta segunda medida resulta muy efectiva especialmente cuando se trabaja con textos de un grado académico particular y se pretende identificar la relación de un sujeto con el resto del grupo.

Sin ánimo de hacer aquí una revisión exhaustiva de estas propuestas, en particular de la Ávila, solo diré que los objetivos difieren en tal magnitud que prácticamente se trata de ejercicios diferentes e innecesariamente complejos para lo datos que quieren obtenerse. El objetivo debe ser siempre el texto y no un conjunto de ellos, de diferente extensión, además. Si con varias trabajos y sus evaluaciones en la mano, el investigador quiere sacar datos colectivos, las matemáticas actuales le brinda fórmulas de mucha menor complejidad.

Veamos, según el modelo de López Morales, el proceso de trabajo.

El texto que vimos al principio sobre 'La playa' necesita ser analizado para obtener los valores que nos permitan aplicar las fórmulas.

El total de palabras de que consta el texto es de 93.

De ellas, son vocablos: *playa, es bonita, azul, agua, salada, mar, hay, peces, tiburones*, *pez espada, pez martillo, papá, mamá, hermanos, fuimos, divertimos, mucho, bañamos, brincamos, saltamos, gusta, ir familiares, bueno, estar, hace, fresco, bueno, ambiente, hermanos, juegan, gusta, primos, tíos, abuelos*

Son vocablos, pero repetidos en el texto: *playa* (6), *es* (4), *mucho* (3), *hermanos* (2), *yo* (1)

Son palabras sin contenido semántico: *a* (4), *con* (2), *el* (3), *de* (1), *en* (2), *la* (7), *me* (1), *mi* (1), *mi* (2), mis (6), *muy* (1), *nos* (3), *nuestros* (1), *también* (1), *y* (4)

En resumen: Total de palabras de contenido semántico nocional: 55

Palabras repetidas	: 17
Palabras de contenido semántico nocional	: 38
Palabras sin contenido semántico nocional	: 38

La primera fórmula de López Morales, la que calcula la proporción entre las nocionales y las otras (nocionales pero repetidas y no nocionales) es así:

$$PV = \frac{38 \times 100}{93} \qquad T: 40{,}9$$

cantidad muy baja como se verá.

La segunda fórmula, la que mide el intervalo de aparición de palabras nocionales es:

$$IAT = \frac{93}{38} \qquad T: 2{,}4$$

cifras que indican que el índice de riqueza léxica es de solo 40,9, muy baja como se verá, y que en este texto es necesario esperar a 2,4 palabras para que haga su aparición una palabra de contenido semántico nocional.

Contrástese el texto anterior con este otro, producido por una alumna de segundo curso de escuela secundaria pública de Puerto Rico

> Quiero graduarme de maestra de educación primaria, luego inscribirme en una nivelación para poder dar preprimaria, sacar varios cursos como manualidades en papel, en globos, y luego hacer que la educación cambie, por lo menos en mi establecimiento. Pienso seguir en la universidad una maestría. También estudiar

> y terminar la carrera técnica de cultura de belleza para poder poner un salón de belleza y dar trabajo a mis compañeras. Para hacer todo eso cuento con el apoyo de mi familia, principalmente de mis padres y hermanos. Al tener una economía estable quiero irme a vivir a España algunos años pues.

Total de palabras: 100
Total de palabras de contenido semántico: 64
Total de palabras de contenido semántico repetidas en el discurso: 6
Palabras de contenido semántico: 58
Palabras sin contenido semántico: 41
Nombre propio: 1 (España)

El primer cálculo índica que el porcentaje de vocablos en el texto (PV) es de 58, y que el intervalo de aparición de vocablos diferentes es de 0.5 Si comparamos estos datos con los del texto anterior, se observan grandes diferencias: Mientras que la proporción del total del texto y las palabras de contenido semántico nocional fue allí de 40,9, aquí es de casi 15 puntos más; es por tanto este último, un texto que maneja un vocabulario más rico: Por otra parte mientras que allí el intervalo de aparición de palabras nocionales fue de 2,4, aquí esas palabras aparecen con una media de 0.5, una riqueza léxica casi cinco veces mayor. Obsérvese que esa segunda fórmula nos ayuda a afinar muchos los cálculos de riqueza léxica.

Un último ejemplo, esta vez producido por un periodista profesional, terminará de explicarnos las fórmulas propuestas y sus beneficios.

> Antes, mucho antes, cuando nuestros mayores hablaban de 'fatiga', lo hacían para expresar una debilidad estomacal, 'tengo hambre', 'tengo fatiga'. Pero cuando en este momento hablamos de 'fatiga política', cuando la gente del pueblo deja hacer y el mandatario y el partido se creen lo mejor del mundo, puede ocurrir cualquier cosa. Le sucedió a Rómulo, el novelista, y al partido AD entre aquellos años 45 y 48. Ahora el vocablo es interpretado como: agitación, cansancio, trabajo prolongado, ansia de vomitar, molestia causada por la pretensión de otro. Aburrir, vejar, agotar. Chávez está justificando al Generalísimo y es por ello que el cuento va para largo, con hampones incluido. Porque la fatiga también quiere decir "que hagan los que les venga en gana. Ignoremos al Gobierno".

Este texto está integrado por 122 palabras (sin contar la sigla AD, las cifras 45 y 48 y los nombres propios Rómulo [Betancourt] y [Hugo] Chávez), siempre entendiendo por 'palabra' cada una de las formas gráficas entre espacios en blanco, como suele procederse en un primer nivel de análisis. Como es habitual, es necesario que no se cuenten las palabras gráficas repetidas porque son el mismo vocablo (*antes* 2, *fatiga* 4, *tengo* 2 y *partido* 2) lo que deja un total de 110, cantidad muy cercana a las 100 sobre las que suelen hacerse estas operaciones. Se observará que aquí –como en todos los textos- hay vocablos de dos tipos: aquellos que significan algo (*pueblo*, *años*, *pretensión*, *viajar*, *cuen-*

to, etc.) y los que solo ejercen una determinada función gramatical (*cuando*, *de*, *lo*, *para*, *una*, *en*, etc.) que no hacen referencia a nada concreto o abstracto de mundo que nos rodea. Los primeros son nombres, adjetivos, verbos y adverbios, y los segundos, artículos, preposiciones, conjunciones, etc.

Como sabemos, la riqueza léxica se mide, en una primera fórmula, estableciendo una relación matemática entre el total de palabras que poseen contenido semántico (V) y el total de los vocablos del texto (N), salvo las excepciones señaladas.

$$PV = \frac{V \times 100}{N} \qquad \frac{65 \times 100}{110} \qquad T: 59{,}1$$

Por lo tanto, el 59,1 del total de palabras de este texto son de contenido semántico nocional no repetidas.

Un segundo índice trabaja con la fórmula:

$$IA = \frac{N}{PN} \qquad \frac{120}{65} \qquad T: 1.8$$

El resultado de esta otra operación matemática refleja cifras relacionadas con la porción de palabras nocionales en el texto (aquí sin importar las repetidas); esto es, a mayor número de palabras nocionales, menor es el intervalo, lo que se interpreta como mejor índice de riqueza léxica. En este caso, hay una palabra nocional cada 1,8 del total de palabras del texto.

Lo que significan estos números y cómo pueden interpretarse necesitan, desde luego, de elementos de comparación. Siempre está a la mano la comparación entre los individuos del grupo, pero eso nos permite hacer comparaciones modestas. Lo ideal es disponer de cifras más generales, que respondan, por ejemplo, a lo consignado en textos de escritores profesionales. Pero estos datos son muy escasos todavía, al menos para los índices de riqueza léxica, no así, por ejemplo, para las investigaciones sobre madurez sintáctica[3].

En efecto Raúl Ávila (2001), director desde el Colegio de México del proyecto de estudio del español en los medios de comunicación pública del Mundo Hispánico disponía de un importante conjunto de datos de riqueza léxica: de las radios locales, Radio Almería contaba con un índice de 66,6; de las nacionales, Radio Nacional de España 67,2, XEB de México, 67, las de Costa Rica, 67,1 y RCN de Colombia, 66,7. En cuanto a las televisiones nacionales, Televisión Española arrojó un índice de 68,1, XEWTV de México, 66, y Telenoticiera CM& de Colombia, de 68,6. Las de carácter internacional

3 El único ejemplo que conozco no está dedicado a la enseñanza sino a comprobar los altos índices de riqueza léxica en los medios de comunicación hispánicos de nuestros días.

CNN en español, 69,6 y ECO, 67,8. El promedio de estas cinco estaciones fue de 68,6. Por último, la prensa colombiana –*El Tiempo*– obtuvo unos índices de riqueza léxica de 68,4.

Ante las constantes críticas de muchos sobre el pobre español de los medios de todo el Mundo Hispánico, en los que no faltaban los reproches a la globalización que había conseguido empobrecer y uniformar esta lengua, Ávila analizó con cuidado un ensayo de Carlos Fuentes tomado al azar desde el punto de vista de la riqueza léxica, y el resultado fue de 69,7. ¡Una décima más que CNN en español! Y por si fuera poco esta comparación, acudió a los *corpora* de lengua hablada de la Ciudad de México y comprobó que entre los hablantes del nivel culto, el índice general de riqueza léxica fue de 68,5, dos décimas menos que el ensayo de Fuentes. Ante estos datos tan contundentes no cabe discusión alguna.

Pues, aunque el propósito que se perseguía era otro, de momento podemos aprovechar estos datos para establecer unos criterios provisionales. De manera que la chica de segundo curso de escuela secundaria cuyo texto hemos leído, que alcanza una puntuación de 46 en riqueza léxica, y el posterior del artículo de prensa venezolano, que consigue 59,1 son, en principio mejorables. Y digo en principio porque estos textos no son iguales en extensión, y sabemos que los datos se desvirtúan al pasar un texto de las 100 palabras. Tanto el ensayo de Carlos Fuentes como los textos de la norma culta tienen características que conspiran contra una comparación rígida: El ensayo tiene más de 100 palabras, y los textos de la norma culta son orales y transliterados posteriormente, transliteración sobre la que se llevó a cabo el conteo. Necesitamos, pues, para tener metas claras, exámenes de textos escritos que se consideren ilustrativos, de los que se pueden tomar varias calas pero nunca superiores a las 100 palabras.

4. EL ESTUDIO DE LA RIQUEZA LÉXICA EN EL MUNDO HISPÁNICO

Hasta el momento, el Mundo Hispánico no dispone de muchas investigaciones que midan la riqueza léxica en textos producidos por niños y adolescentes. Cabe mencionar las investigaciones realizadas en México (Ávila, 1986), Santo Domingo (Haché, 1988), Chile (Valencia *et allí*) y Puerto Rico (Cintrón, 1993)[4].

Ávila en su *Léxico infantil de México; palabra, tipos, vocablos*, aplicó estas medidas a una selección de 4,500 textos de un corpus léxico de textos de tema libre escritos por niños de tercero a sexto grado de primaria. Entre sus principales hallazgos se encuentran diferencias porcentuales de vocablos que favorecen a las niñas, y diferencias más acentuadas en relación con la variable socieconómica. Sus resultados revelan una relación

4 De nuevo remito a la tesis doctoral de Pastor (1998) para ampliar estas aquí breves consideraciones.

estrecha entre la riqueza léxica y el nivel sociocultural. En su estudio Ávila prueba que conforme se extiende el corpus se recogen proporcionalmente menos vocablos. Los hallazgos sobre la densidad resultaron equiparables a los porcentajes de vocablos; las niñas obtuvieron una densidad superior a los niños. La tercera medida usada por este investigador –la acumulación de frecuencias por deciles- demostró que la mayor o menor extensión de un corpus no condiciona el número de vocablos que se obtiene mediante estos procedimientos en los primeros 8 o 9 deciles, lo que lo lleva a concluir que las diferencias cuantitativas en la riqueza léxica son siempre menores en niños que en niñas y siempre relacionadas con el estrato sociocultural.

Ana Margarita Haché realizó un estudio *Aportes de la riqueza léxica a la enseñanza de la lengua materna,* en el que aplica los indicadores para medir riqueza léxica establecidos por López Morales. En el mismo trabajó sobre una población constituida por estudiantes de sexto y octavo grado de la escuela primaria de Santiago de los Caballeros en la República Dominicana. Las conclusiones de la investigación revelaron que la mayor extensión de los textos no parece tener influencia directa en el aumento del porcentaje de vocablos (PV). Al comparar los resultados de la primera medida (PV) en *corpora* de distintas extensiones notó porcentajes superiores de vocablos en textos menores. Es decir, que los índices parecían favorecer a aquellos sujetos que produjeron un menor número de palabras y penaliza a los más elocuentes y que después de las primeras 100 palabras el cálculo se distorsionaba. En este aspecto hubo coincidencia con los hallazgos ofrecidos anteriormente por Ávila. Sin embargo, en contraste con este, Haché encontró mayor riqueza léxica en niños que en niñas; la variable socioeconómica no discriminó adecuadamente a los sujetos debido a circunstancias particulares en las escuelas bajo estudio.

Entre los resultados asociados a la medida PV la investigadora concluyó que este índice no arrojó diferencias entre niños y niñas de la escuela pública y privada. No obstante informó que se incrementa el porcentaje de vocablos del sexto al noveno grado, lo que implicó que la edad incide en la riqueza léxica. En los resultados particulares sobre la segunda medida, Haché encontró que el intervalo disminuía con la edad y que este indicador resultó más sensible a la variable socio-económica.

Otro estudio que aplica las fórmulas de riqueza léxica de López Morales fue realizado por un equipo de investigadores de la Universidad de Chile y de la Universidad de Concepción. En esta ocasión, los investigadores hicieron ajustes a las fórmulas a tono con las necesidades específicas del proyecto; en particular, no utilizaron la fórmula de intervalo de aparición de palabras de contenido semántico. Sus análisis se elaboraron sobre los porcentajes de palabras nocionales (PN). Se llevó a cabo esta investigación sobre la muestra de 308 sujetos. Concluyó el estudio que no hay diferencias en el porcentaje de palabras nocionales entre niños y niñas. Los sujetos de nivel socio-económico y cultural alto presentaron un porcentaje levemente superior de palabras nocionales que los del nivel medio y bajo. En esta investigación, solo el régimen de estudio mantiene algún gra-

do de acción diferenciadora. Los investigadores observaron una relación proporcional entre el porcentaje de palabras nocionales y el total de unidades léxicas. Encontraron a su vez que el discurso narrativo permite una mayor riqueza léxica.

En Puerto Rico se llevó a cabo el estudio de riqueza léxica realizado por Filomena Cintrón, *Nuevos índices de riqueza léxica en escolares de Barranquitas*. En esta investigación, llevada a cabo sobre una muestra de estudiantes de los grados tercero, sexto, noveno y décimo, se encontró un crecimiento asistemático en el progreso léxico de los estudiantes según la variable grado de ambas medidas. Los datos sobre la primera de ellas y la variable sexo, mostraron una tendencia a favorecer a las niñas tanto en el tercer grado como en el sexto, hallazgo que coincide con las investigaciones de Ávila para el tercer grado. Para esta mediad y la variable socio-económica se encontró una relación asociativa que tendió a favorecer al nivel socio-económico alto, aunque muy ligeramente.

Los resultados sobre la segunda medida según la variable sexo y nivel socio-económico no registraron diferencias palpables. Sin embargo, resultó sorprendente que el intervalo aumentó a lo largo del nivel primario. Es decir, aparentemente la variable grado de escolaridad no resultó del todo relevante para la segunda medida.

5. LOS MATERIALES PARA LA PRODUCCIÓN DE LA RIQUEZA LÉXICA

El aspecto más importante, después de una especificación conceptual adecuada, es la creación del material de entrenamiento para los estudiantes. Se había creado en Puerto Rico una amplia gama de ellos, cuando la Universidad decidió cerrar sus cursos de español para extranjeros. Con respecto a la riqueza léxica, al menos una batería completa estaba terminada y probada. Ese conjunto de ejercicios para los primeros niveles atendía a las dos posibilidades existentes (aunque con múltiples variedades cada una de ellas) de trabajar la riqueza léxica: la ayuda que significaban los ejercicios para incrementar la madurez sintáctica y aquellos que se referían directamente a ella.

Los primeros partían de la base de revisar textos construidos *ad hoc*, de manera que la repetición léxica fuera uno de los aspectos más evidentes. Al realizar el ejercicio de eliminar estos elementos reiterados llevaba directamente a los alumnos a combinar oraciones; al realizar este ejercicio se eliminaban los elementos léxicos reiterados que tanta monotonía y falta de fluidez provocaban.

Un ejemplo de texto de debían modificar es el siguiente:

> El Ática es una región griega. El Ática es una región antigua. El Ática es una península triangular. El Ática tiene montañas. Los helenos del Ática extraían el mármol de las montañas. Esos mismos helenos del Ática construían monumentos fastuosos. Esos monumentos perduran hasta hoy.

El índice de riqueza léxica de este texto es 40 y el intervalo de aparición de palabras con contenido semántico, de 2,2. Cuando el alumno reescribe el texto:

> El Ática es una antigua región griega, una península triangular con montañas de donde los helenos extraían el mármol con el que construían monumentos fastuosos que perduran hasta hoy.

Sube el índice de riqueza léxica de 40 a 51 y baja el intervalo a 1,9.

En otros casos se presenta un texto con espacios en blanco y un conjunto de palabras que pudieran entrar en ellos. Todos los términos son sinónimos o cuasi sinónimos de los que pueden intercambiarse en el discurso. Así se consigue que los alumnos, ante un texto con mucha reiteración de términos, se empeñen en buscar sustitutos adecuados.

Véanse los siguientes ejemplos:

> Seleccione las palabras que crea adecuadas para cada espacio en blanco del siguiente texto:
>
> vimos contemplamos admiramos disfrutamos de
>
> El bosque nacional más importante del Caribe es el Yunque, y a él nos llevaron de excursión. Allí_____helechos gigantes y_____palmas, orquídeas de colores muy variados, enredaderas y musgos. _____también una gran cantidad de animales: halcones, búhos, palomas, lagartijas, ciempiés, tarántulas y hasta boas, y _____ sobre todo, la peculiar cotorra puertorriqueña.
>
> estupendo magnífico hermoso precioso bellísimo
>
> Hay varios grandes teatros en la Isla. El primero, construido por los españoles en la capital a finales del siglo XIX, fue el Tapia; es _____ . Después, también por esas fechas, pero en la ciudad sureña de Ponce, se levantó otro que puede describirse como un coliseo_____. Ya en el XX, otro_____ teatro fue terminado en el recinto universitario de Río Piedras, y más modernamente, el Gobierno de Puerto Rico construyó el Bellas Artes, que es un_____conglomerado de tres salas diferentes. Pero todos se caracterizan por su calidad; son_____.

6. FINAL

A finales de 1992 habían quedado en el Instituto Lingüística de Río Piedras, en su cátedra de Lingüística aplicada a la enseñanza de lenguas extrajeras dos baterías de 15 ejercicios cada una, algunos de ellos elaborados por los mismos alumnos. Aquel episodio terminó allí, pero el interés por el tema no desapareció. Debo y puedo decir que los resultados alcanzados entonces en cuanto a riqueza léxica fueron muy favorables. Vale la pena rescatar aquella experiencia, actualizar esos materiales y proseguir con la tarea. Ojalá que así sea.

BIBLIOGRAFÍA

Andión, M.ª A. y A. M.ª Ruiz (1996): "Azorín, Cela, Delibes y Unamuno. Análisis contrastivo de madurez sintáctica", *Revista de Estudios de Adquisición de la Lengua Española* (*REALE*) 6, 9-36.

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

— (2001) "Los medios de comunicación masiva y el español internacional", en *II Congreso Internacional de la lengua española*, [en línea]: <http://congresosdelalengua.es/valladolid/ponencias/unidad_diversidad_del_espanol/1_la_norma_hispanica/avila_r.htm>

Baldi, P. L (1972): "Fattori sociali dell'abilità lingüistica nella produzione di bambini de nove e dieci anni", *Studi Italiani di Lingüistica Teorica ed Applicata* I/3, 335-471.

Cintrón, F. (1993): *Nuevos índices de riqueza léxica en escolares de Barranquitas*, Río Piedras: Universidad de Puerto Ricio (Tesis de Maestría inédita).

Giraud, P. (1960): *Problemes et methodes de la statistique linguistique*, Paris: Presses Universitaires.

Haché, A. M. (1991): "Aportes de las pruebas de riqueza léxica a la enseñanza de la lengua materna", en H. López Morales (ed.), *La enseñanza del español como lengua materna*, Río Piedras: Universidad de Puerto Rico, 47-60.

Ham Chande, R. (1979): "Del 1 al 100 en lexicografía", en L. Fernando Lara (ed.), *Investigaciones lingüísticas en lexicografía*, México: El Colegio de México, 110-132.

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

Müller, C. (1968): *Estadística lingüística*, Madrid: Gredos

Pastor, M. (1998): *Descripción del léxico y de la sintaxis en textos producidos por niños del nivel primario del sistema público*, Río Piedras: Universidad de Puerto Rico (Tesis doctoral inédita).

Tesitelová, J. (1992): *The main areas of quantitative linguistics*, New York: Planum Press.

Valencia, A. *et al*. (1992): "Evaluación de la riqueza léxica en estudiantes de último año de enseñanza media", *Estudios Filológicos* 27, 59-72.