REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2025 PUCV, Chile 58(117) 121-143

La Novedad léxica como medida de riqueza léxica en un corpus oral contemporáneo: el corpus PRESEEA-Santander

The lexical novelty as lexical richness measure in a contemporary oral corpus: the PRESEEA-Santander corpus

Inmaculada Martínez Martínez
UNIVERSIDAD DE CANTABRIA
ESPAÑA
inmaculada.martinez@unican.es

Hiroto Ueda
UNIVERSIDAD DE TOKIO
JAPÓN
hiroto.ueda.tokio@gmail.com

Recibido: 4/1/2023 / **Aceptado:** 12/4/2024
DOI: 10.4151/S0718-09342025011701076

Resumen

Los estudios cuantitativos del léxico más generalizados son, entre otros, los léxicos básicos, los disponibles y los que miden la riqueza léxica (Ávila Muñoz, 2014). En esta última dirección se ubica el presente estudio. Tradicionalmente, la fórmula empleada a partir de la cual se obtiene la riqueza media de cada texto es TTR (*Type Tokens Ratio*), consistente en dividir el número de vocablos diferentes entre el total de frecuencias de palabras (Capsada & Torruella, 2017). Esta fórmula es válida cuando se comparan corpus de igual tamaño, pero no resulta fiable al cotejar textos de distinta dimensión. Junto a la Densidad (Ávila Muñoz, 2014), otros índices para medir la riqueza léxica —como la Diversidad (Baayen, 2001), la Información (Shannon & Weaver, 1963) o lo que se conoce como Peculiaridad (Baayen, 2001, 2008)— tampoco están exentos de problemas. El objetivo final de este trabajo es comprobar la eficacia de nuestra propuesta, la Novedad léxica, como un nuevo índice de riqueza léxica. Para medir dicha eficacia, se ofrece la fórmula específica y se compara con los otros índices, tomando como base del análisis el corpus PRESEEA-Santander en dos direcciones: los parámetros sociolingüísticos (sexo, edad y nivel educativo) y la categoría gramatical. Los resultados obtenidos muestran una diferencia notable de Novedad léxica en función de la categoría gramatical y escasa en lo que atañe a los criterios sociolingüísticos.

Palabras clave: riqueza léxica, Novedad léxica, variación léxica, corpus sociolingüístico, análisis cuantitativo del léxico

Abstract

The most common quantitative lexical studies are, among others, the basic lexicon, the available lexicon and those that measure lexical richness (Ávila Muñoz, 2014). The present study is framed within the latter category. Traditionally, the formula used to obtain the average richness of each text is TTR (Type Tokens Ratio), which consists of dividing the number of different words by the total frequency of words (Capsada & Torruella, 2017). This formula is valid when comparing corpora of equal size, but it is not reliable when comparing texts of different dimensions. Along with Density (Ávila Muñoz, 2014), other indices to measure lexical richness —such as Diversity (Baayen, 2001), Information (Shannon & Weaver, 1963) or what is known as Peculiarity (Baayen, 2001, 2008)— are also not without problems. The final objective of this work is to verify the effectiveness of our proposed Lexical Novelty as a new index of lexical richness. To measure this effectiveness, the specific formula is offered and compared with the other indices, taking the PRESEEA-Santander corpus as the basis of the analysis in two directions: the sociolinguistic parameters (sex, age and educational level) and the grammatical category. The results obtained show a notable difference in Lexical Novelty as a function of grammatical category but a little difference in terms of sociolinguistic criteria.

Keywords: lexical richness, Lexical Novelty, lexical variation, sociolinguistic corpus, quantitative analysis of lexicon

INTRODUCCIÓN

En el ámbito de la sociolingüística han sido varios los autores que se han acercado al análisis del léxico y de su riqueza mediante la consideración de diversos parámetros: la 'densidad léxica' (Ávila Muñoz, 2014), la 'diversidad léxica' (Baayen, 2001), la 'información léxica' (Shannon & Weaver, 1963) o lo que se denomina 'peculiaridad léxica' (Baayen, 2001, 2008). De forma correlativa, son varias las fórmulas que se han previsto para medir la riqueza léxica de los textos, escritos u orales. Estos intentos han supuesto avances —sin duda trascendentales— en el complejo proceso de delimitar la capacidad que tienen los individuos de producir textos léxicamente variados y, por lo tanto, enriquecidos.

Frente a los progresos alcanzados, estos índices no han sido capaces de precisar con la máxima eficacia el grado de riqueza léxica de los textos. Así, la densidad léxica, equivalente en castellano de la fórmula TTR (Baker et al., 2006), se calcula solo a partir de dos valores, las formas diferentes o 'vocabulario' (*types*) y el total de formas o 'extensión' (*tokens*), sin considerar la frecuencia de cada lema; la diversidad léxica necesita realizar ajustes en la fórmula que emplea y con ello deja muestras de su precariedad; por su parte, la información léxica parece un índice más coherente que la diversidad, pero apunta a una probabilidad cuya ocurrencia no es del todo segura; por último, la peculiaridad léxica parte del método que permite situar en contraste los hápax (voces que se registran una sola vez en un texto) y los *types*. Deja así de lado las palabras que aparecen solo dos veces (*dis legomena*) y las que aparecen tres veces (*tris*

legomena) y que también son importantes, aunque lo sean en menor grado de contribución a la riqueza léxica.

El método que aquí se propone para calcular la riqueza léxica es la Novedad léxica (Nlex), entendida como la medida representativa de la Novedad léxica. Consideramos que el texto que manifiesta más Novedad léxica posee un alto grado de riqueza léxica, mientras que el texto en donde los vocablos se repiten mucho contiene poca novedad y manifiesta, por tanto, pobreza léxica. Al igual que en el caso de la peculiaridad léxica, se parte del método que permite situar en contraste los hápax y los *types*. Sin embargo, la Novedad léxica amplía el ámbito de frecuencias tratadas a las palabras que aparecen dos y tres veces, y así sucesivamente, hasta llegar a la última palabra de mayor frecuencia, aunque en este caso su aportación pueda considerarse casi nula.

Los objetivos principales del presente estudio son, por un lado, fijar con precisión las limitaciones de las distintas medidas de riqueza léxica empleadas hasta el momento desde la perspectiva sociolingüística y, por otro lado, proponer la Novedad léxica como índice alternativo en los análisis de corpus orales, para lo cual se ofrecerá, junto a la definición del constructo, la fórmula correspondiente, así como los beneficios que conlleva su aplicación. Las preguntas de investigación planteadas en relación con los objetivos son las siguientes:

1. ¿Son eficientes los métodos conocidos hasta el momento para determinar la riqueza léxica de un texto?

2. ¿Es la Novedad léxica el índice más eficaz si se contrasta con los otros índices con base en el análisis de un corpus de lengua oral?

En la primera parte del estudio, se expondrá el estado de la cuestión sobre los índices de riqueza léxica y los avances conseguidos en investigaciones reseñables. En la sección 2, se presentará el apartado metodológico, con la propuesta de la fórmula para la Novedad léxica, así como el proceso de estandarización que conlleva. En la sección 3, se procederá al análisis del corpus PRESEEA-Santander[1] con el empleo de la citada fórmula y la comparación con las medidas anteriores a partir de dos tipos de variables: la categoría gramatical y los parámetros sociolingüísticos (edad, sexo y nivel educativo). Por último, el artículo se cerrará con las principales conclusiones y las líneas futuras de la investigación derivadas del análisis.

1. Medidas de riqueza léxica

En los estudios de la lingüística de corpus (McEnery & Hardie, 2012; Rojo, 2021), el primero de los índices que aquí analizamos, la densidad léxica (Ávila Muñoz, 2014), aborda la riqueza léxica dentro de un determinado texto a partir de la cantidad de palabras diferentes (*types*) —o vocabulario— dividida por la frecuencia total de palabras (*tokens*) —o extensión—. Esta proporción que representan las formas

diferentes en el conjunto de palabras del texto nos hace deducir que, en buena lógica, cuantos más vocablos diferentes se empleen, tanto mayor será su densidad léxica y más se evaluará, por tanto, su riqueza léxica.

En la bibliografía sobre la fórmula TTR, se ha discutido el problema que se presenta al aplicarla a textos de distintos tamaños (McEnery & Hardie, 2012; Rojo, 2021) y, por tanto, con diferente número de palabras. El problema reside en que la TTR no es constante sino descendente en textos cada vez más grandes: cuanto más grande sea el texto, menos formas nuevas se encontrarán.

Si tomamos como referencia estudios previos, la razón de este comportamiento peculiar de la fórmula TTR estriba para autores como Ishikawa (2012) en la limitación del inventario léxico del hablante. Así, en el inicio del texto se presentan los lemas nuevos con una frecuencia relativamente alta; sin embargo, al aumentar la magnitud del texto, la ocurrencia de nuevos lemas es menos probable, debido a una lógica restricción del vocabulario; de este modo, suelen repetirse las mismas palabras y, por tanto, la densidad léxica disminuye. Otros autores como Stubbs (2006) señalan que los nuevos términos suelen ser lemas de poca probabilidad en la parte posterior del texto, porque se han integrado ya en la lista de lemas diferentes al comienzo de dicho texto. Por todo ello, si los nuevos lemas son de mínima probabilidad en las fases finales del texto, la TTR se vuelve cada vez menor. Esta limitación será subsanada por el método de estandarización que trataremos más adelante (ver apartado 2.2).

En resumen, la densidad léxica se calcula solo a partir de dos valores, el vocabulario (*types*) y la extensión (*tokens*), sin considerar la frecuencia de cada forma, información fundamental a la hora de calcular la riqueza léxica y que sí consideran otros métodos que trataremos a continuación.

Para abordar la siguiente medida de riqueza, la diversidad léxica, consideramos la fórmula de K de Yule (1944), que, aunque revisada años más tarde por Simpson (1949), no consigue ser mejorada (Baayen, 2001)[2]. La tomamos en cuenta puesto que la riqueza léxica se considera como un concepto contrario al de repetición léxica; es decir, cuanto menos se repite el vocabulario, tanto mayor será la diversidad léxica y, por tanto, la riqueza léxica de un determinado texto.

Según Tweedie y Baayen (1998), el valor K de Yule (1944) (K = 10^4 * (s - t) / t^2 = 10^4 * (10000 - 100) / 100^2 = 9900) mantiene una constancia relativa a pesar de la diferente magnitud del texto. Para el cálculo de K, se prepara una lista de distribución de frecuencia, denominada 'espectro de frecuencia' (Baayen, 2008) representada por F:C, donde F es la frecuencia y C es su cómputo o 'frecuencia de frecuencia' (Baayen, 2008), es decir, el número de veces que ocurre la frecuencia en cuestión. El espectro de frecuencia de lemas, por ejemplo, (a, b, b, c, c, d, e), se explicaría de la siguiente manera: F:C (a, b, b, c, c, d, e) = 1:3, 2:2, lo cual quiere decir que la frecuencia 1 se encuentra en 3 lemas (a, d, e) y la frecuencia 2 en 2 lemas (b, c).

Kin (2009: 56-57) parte, para explicar la fórmula, de un caso extremo del valor K, aquel en el que un texto aparece construido con 100 palabras iguales, (a, a, …, a), lo cual representa la repetición total:

F:C (a, a, …, a) = 100:1 (frecuencia 100, 1 vez)
s = 100^2 * 1 = 10000
t = 100
K = 10^4 * (s - t) / t^2 = 10^4 * (10000 - 100) / 100^2 = 9900

Este valor de K=9900 es difícil de evaluar de manera inmediata y es aquí donde reside su limitación y, por tanto, la necesidad de modificar esta fórmula para obtener un valor normalizado con el rango de [0, 1]. Para ello, en primer lugar, es preciso eliminar el primer término de la fórmula, 10^4 (=10000), que no sirve sino para aumentar la dimensión por una cifra grande (Maekawa, 1995: 25), de modo que llegamos a K1:

K1 = (s - t) / t^2 = (10000 - 100) / 100^2 = 9900/10000 = 0.99

Por consiguiente, hemos obtenido un valor bastante próximo a 1. Para llegar al valor deseado 1, dividimos K1 por (1-1/t), es decir, 1-1/100 = 0.99, y llegamos a K2:

K2 = (s - t) / t^2 / (1-1/t)
= (10000 - 100) / 100^2 / (1-1/100) = (1-1/100) / (1-1/100) = 1

En el cálculo de K (y K1, K2) se utilizan valores de números cuadrados, por lo que ofrece una variación de mayor escala (Herdan, 1956: 33). Por ello, conviene subsanarlo con la función de la raíz cuadrada (√) en K3:

K3 = √ (s - t) / t^2 / (1-1/t))

Como el rango de K3 es [0, 1], igual o menor de 1, su raíz cuadrada aumenta la dimensión, manteniendo el valor mínimo (0) y el máximo (1). El valor K y sus modificaciones (K1, K2, K3) manifiestan el grado de repetición, por lo que se evalúa que tanto más rico es el vocabulario cuanto menor es su K (y K1, K2, K3). Por lo tanto, al aprovechar el rango de K3 [0, 1], conviene utilizar la cifra complementaria de K3 con respecto a 1, para llegar al indicador de la diversidad o la riqueza léxica. Con las modificaciones mencionadas, proponemos denominarlo 'K normalizado' (Kn):

Kn = 1 - √((s - t) / t^2 / (1-1/t))

La necesidad de estas modificaciones explica por sí sola la precariedad del índice de la diversidad léxica. El mayor problema que presentan las cifras de K y Kn es su falta de coherencia con la densidad léxica.

Otra característica estadística que parece superar las limitaciones arriba trazadas es la información léxica. La entropía informativa propuesta por Shannon y Weaver

(1963) es concebida como uno de los indicadores de la riqueza léxica. Su fórmula es la siguiente (E):

E(X) = -Σ[i=1:n] p(i) * log2 p(i),
donde n = *type*, p(i) = probabilidad(i) = frecuencia(i) / total de frecuencia de i=1:n

Si analizamos un texto de 8 palabras con frecuencia de (4, 2, 1, 1), sus probabilidades correspondientes son: (4/8, 2/8, 1/8, 1/8). La entropía es la suma de todos los elementos (en R):

```
(4/8)*(-log2(4/8))+
(2/8)*(-log2(2/8))+
(1/8)*(-log2(1/8))+
(1/8)*(-log2(1/8)) # 1.75
```

Su valor máximo se presenta cuando todas las probabilidades son iguales, lo que representa la máxima información. La fórmula de la entropía máxima (E.max) es:

E.max = (n/n) (-log2(1/n)) = log2(n)

El valor mínimo (E.min) se presentaría cuando no existe más que una sola probabilidad 1:

E.min = 1/1 (-log2(n/n)) = -log2(1) = 0

Por consiguiente, sobre la entropía se aplica la estandarización o normalización para obtener una cifra que oscile en el rango [0, 1] y así llegar a la entropía normalizada (En), cuya fórmula es:

En(X) = E(X) / E.max = -Σ(i=1:n) p(i) * log2 p(i) / log2(n)

Aplicamos el siguiente código para la entropía normalizada (en R):

```
Ent=function(L, s=0){ # L: Lemas; s=1: normalizar
F=as.data.frame(table(L)) # F: Frecuencia
d=nrow(F); if(d==1) return (0) # d: Type
P=F[,2]/sum(F[,2]) # P: Probabilidad
r=sum(-P*log2(P))  # r: entropía
ifelse(s==0, r, (r/log2(d))) # devolver

}
```

Los resultados son los siguientes:

```
> s=0 # Sin normalizar
> Ent(c('a','a','a','a','a'),s) # 0
> Ent(c('a','a','a','a','b'),s) # 0.7219281
> Ent(c('a','a','a','b','b'),s) # 0.9709506
```

```
> Ent(c('a','b','b','c','c'),s) # 1.521928
> Ent(c('a','b','b','c','d'),s) # 1.921928
> Ent(c('a','b','c','d','e'),s) # 2.321928
> Ent(L,s) # 7.780323
> s=1 # Normalizar
> Ent(c('a','a','a','a','a'),s) # 0
> Ent(c('a','a','a','a','b'),s) # 0.7219281
> Ent(c('a','a','a','b','b'),s) # 0.9709506
> Ent(c('a','b','b','c','c'),s) # 0.9602297
> Ent(c('a','b','b','c','d'),s) # 0.960964
> Ent(c('a','b','c','d','e'),s) # 1
> Ent(L,s) # 0.6408732
```

Advertimos que la entropía normalizada puede superar a la fórmula de densidad léxica ya analizada TTR, al ofrecer valores diferentes en los ejemplos segundo y tercero, lo que es imposible en TTR. Podemos entender intuitivamente que la entropía de un texto homogéneo como el primero ('a', 'a', 'a', 'a', 'a') no conlleva ninguna información, puesto que la repetición total es sinónimo de información nula. En cambio, la información que contiene el último ejemplo es máxima ('a', 'b', 'c', 'd', 'e'), en el sentido de que no sabemos qué elemento aparece con seguridad por presentar su máxima variabilidad. En el segundo ejemplo, sabemos que la mayoría de las veces ocurre 'a', lo que presenta bastante poca información. En el penúltimo ejemplo, la probabilidad de 'b' es el doble de otros elementos, pero no es muy segura su ocurrencia. Por ello, la información es relativamente alta. Como veremos más adelante el índice de la información léxica resulta más coherente con la diversidad léxica que otros índices de la riqueza léxica.

Como un avance importante en este análisis de la riqueza, el índice de la peculiaridad léxica aparece representado por la relación entre hápax y *tokens*, puesto que los hápax son vocablos sumamente peculiares que caracterizan al texto léxicamente.

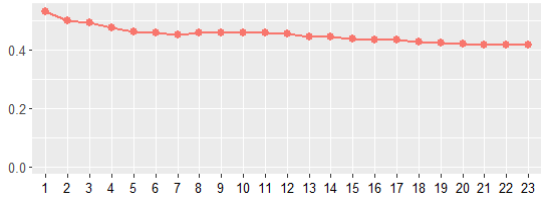
Los hápax son importantes tanto cuantitativa como cualitativamente, puesto que ocupan casi el 40% del texto y muestran una alta capacidad informativa, como ya se señaló líneas más arriba, al comienzo de esta sección. Efectivamente, los lingüistas han venido tomándolos en consideración en sus estudios lexicométricos, especialmente en forma de la ratio hápax/tokens (Baayen, 2001, 2008).

Rojo (2002) expone un dato interesante sobre esta relación, al señalar que mantiene una proporción casi constante entre corpus de distintas magnitudes. Es decir, la ratio de hápax/*types* sería estable y no se vería afectada por el tamaño del texto, a diferencia de lo que ocurría en la fórmula TTR. Así lo demuestran los grandes datos del Corpus de Referencia del Español Actual (CREA), en los que se mantiene casi la misma ratio,

entre 39,7% y 42,9%. Por otra parte, Capsada y Torruella (2017) mencionan que, al tratarse de obras de menor escala, ciertamente la ratio de hápax/ *types* disminuye de acuerdo con el aumento de *tokens*. Consideramos que ambas opiniones, aparentemente opuestas, resultan complementarias, puesto que Rojo (2002) habla de grandes corpus, mientras que Capsada y Torruella (2017), de textos reducidos.

Observamos la proporción hápax/*type*s en nuestro corpus a través de la Figura 1.

Figura 1
División acumulativa hápax/types.



Token: *5000

Como la cantidad de *tokens* en el corpus PRESEEA-Santander es reducida (115 136), se observará la tendencia descendente, aunque leve (coeficiente de variación = .0621). También se observa la configuración de la figura descendente, cada vez más plana. La razón de estas dos tendencias parece estar en el comportamiento numérico de los hápax, que son los nuevos lemas en cada división.

El mayor inconveniente de la proporción hápax/*types* se manifiesta en textos que, o bien no presentan ningún hápax o bien muestran muy pocos casos, lo que puede ocurrir en textos breves. Así la proporción hápax/*types* del tercer ejemplo, ('a', 'a', 'a', 'b', 'b'), es igual a cero (hápax = 0), a pesar de que TTR = 2/5 = 0.4.

A modo de valoración, el índice de la peculiaridad léxica parece demasiado sencillo, puesto que toma en consideración solo el número de hápax con respecto a *types;* de esta manera, se ignoran los lemas que ocurren dos o tres veces, que son también importantes, aunque, como ya se ha mencionado, en menor grado que los casos de hápax.

2. Método

2.1 La Novedad léxica

Como señalábamos al comienzo, proponemos la Novedad léxica como una medida representativa de la riqueza léxica. El texto que manifiesta más Novedad léxica poseerá un alto grado de riqueza léxica, mientras que el texto en donde los vocablos se repiten mucho contendrá poca novedad y manifestará, por tanto, pobreza léxica. El modelo teórico subyacente del método que aquí presentamos reside en el método de hápax con ampliaciones de uso de todos los casos en lugar de solo uno, el hápax.

Para llegar al concepto de Novedad léxica, partimos del método de hápax/*types*, que acabamos de presentar, puesto que tomar en consideración el número de palabras que aparecen solo una vez no parece muy representativo. Ciertamente los hápax son altamente informativos en comparación con las palabras repetitivas. Sin embargo, se propone ampliar aquí el ámbito de frecuencias tratadas, tal y como se adelantó al comienzo del estudio.

Para obtener el grado de contribución a la Novedad léxica, utilizamos la información numérica que poseemos de cada palabra, su frecuencia. El grado de contribución a la Novedad léxica de un hápax debe ser el máximo dentro del rango [0, 1], es decir, 1. El de las palabras que aparecen dos veces es 1 dividido por 2 (1/2), y el de las palabras de frecuencia tres (1/3), y así sucesivamente. De este modo, la cifra que representa la riqueza léxica la calculamos por la siguiente fórmula de la Novedad léxica y la Novedad léxica normalizada (Nlex.n):

$$Nlex. = \Sigma[i=1:d]\ 1/f(i)$$
$$Nlex.n = Nlex/d$$

donde *f(i)* es la frecuencia del lema correspondiente y *d* es el número de lemas, es decir, *types*. De esta manera, utilizamos el número de lemas (*d*) como denominador de la cifra normalizada, puesto que el valor máximo teórico se presenta cuando todas las palabras son hápax (por ejemplo, una lista de estudiantes); en este caso, el grado de información será máximo. Su suma se iguala, entonces, al número de lemas diferentes (*types*), por tanto, Nlex.n = Nlex/d = d/d = 1.

El siguiente código de la función (Nlex), al recibir la lista de lemas (L), devuelve la cifra de la Novedad léxica (en R):

```
Nlex=function(L,sel=0){ # L: Lemas; sel: seleccionar
  F=as.data.frame(table(L)) # Frecuencia
  d=nrow(F); if(d==1) return (0)
  s=sum(1/F[,2]) # sum(1/Frecuencia)
  ifelse(sel==0, s, s/d) # devolver
}
```

A continuación, presentamos los resultados de la ejecución de la fórmula propuesta de la Novedad léxica a partir de datos sencillos (en R):

```
> s=0
> Nlex(c('a','a','a','a','a'),s) # 0
> Nlex(c('a','a','a','a','b'),s) # 1.25
> Nlex(c('a','a','a','b','b'),s) # 0.8333333
> Nlex(c('a','b','b','c','c'),s) # 2
> Nlex(c('a','b','b','c','d'),s) # 3.5
> Nlex(c('a','b','c','d','e'),s) # 5
```

```
> Nlex(L,s) # 2572.31
> s=1
> Nlex(c('a','a','a','a','a'),s) # 0
> Nlex(c('a','a','a','a','b'),s) # 0.625
> Nlex(c('a','a','a','b','b'),s) # 0.4166667
> Nlex(c('a','b','b','c','c'),s) # 0.6666667
> Nlex(c('a','b','b','c','d'),s) # 0.875
> Nlex(c('a','b','c','d','e'),s) # 1
> Nlex(L,s) # 0.5698515
```

Observamos que el segundo ejemplo devuelve una cifra de Novedad léxica normalizada (Nlex.n) (=.625) mayor que el tercero (=.417), por contener un caso de hápax ('b') que da la máxima cifra 1. Como el tercer ejemplo no contiene ningún hápax, arroja una cifra menor. Recordamos que en este ejemplo el índice de hápax/*types* devuelve cero (0) por no presentar ningún caso de hápax. En realidad, en los textos grandes esto no causa problema, puesto que el número de hápax es tan grande que ocupa casi el 40%. Sin embargo, nuestra Nlex y Nlex.n ofrecen la cifra representativa de la riqueza léxica con mayor precisión, puesto que incluyen no solamente los casos de hápax sino todos los casos (lemas).

2.2 Estandarización

Para eliminar el efecto de la magnitud de *tokens* que ejerce en el cálculo de la fórmula TTR y sus variantes, presentamos un método consistente, no en el cambio de la fórmula matemática, sino en el modo de aplicarla. Este método se denomina estandarización y será sometido a evaluación (ver el apartado 2.3) a partir de la comparación con el resto de las fórmulas abordadas en el marco teórico.

El método que se llama TTR estandarizada (*standardised type-tokens ratio* o *mean type-tokens ratio*) es simple: fijar de antemano el tamaño de intervalo del corpus, calcular la TTR de cada intervalo y finalmente derivar el valor medio de TTR de todos los intervalos (Baker et al., 2006; McEnery & Hardie, 2012). A nuestro modo de ver, el método de estandarización es aplicable no solamente a TTR, sino a todos los índices mencionados. Vamos a realizarlo con el siguiente código a partir del intervalo de 1000 formas:

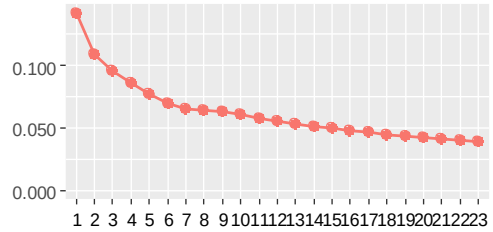
```
# Riqueza léxica
RL=function(L, s=0){ # L: Lemas, sel: Selección
  if(s==0) return(TTR(L,5)) # 0:Tokens
  if(s==1) return(TTR(L,4)) # 1:Type
  if(s==2) return(TTR(L,0)) # 2:TTR (Type/Tokens)
  if(s==6) return(Hp(L,1))  # 6:Hápax/Type
  if(s==7) return(Ent(L,0)) # 7:Entropía
  if(s==8) return(Ent(L,1)) # 8:Entropía normalizada
```

```
  if(s==9) return(Nlex(L,0))  # 9:Novedad léxica
  if(s==10) return(Nlex(L,1)) # 10:Novedad léxica normalizada
  if(s==11) return(Kn(L,0)) # 11:Yule-K
  if(s==12) return(Kn(L,1)) # 12:Yule-K normalizado
}
# Estandarización
Est=function(L, ti=1000, s=1){ # L: Lemas, s: seleccionar, ti: tamaño de intervalo
  ni=floor(length(L)/ti); t=0 # ni: número de intervalos, t: TTR
  for(i in 1:ni){ # Repetir
    It=L[((i-1)*ti+1):(i*ti)] # Intervalo
    t=t+RL(It, s) # s==0: TTR
  }; t/ni # Valor medio
}
Est(L,1000,9) # 9: Novedad léxica: 162.108
```

Procedemos a su aplicación en divisiones acumulativas (*tokens*: 5000, 10000, 15000, …) para ver el efecto de *tokens* (intervalo=1000). Los datos se extraen del corpus oral contemporáneo PRESEEA-Santander. La Figura 2 corresponde a TTR sin estandarización (TTR) y la Figura 3 a TTR con estandarización (TTR.est):

Figura 2

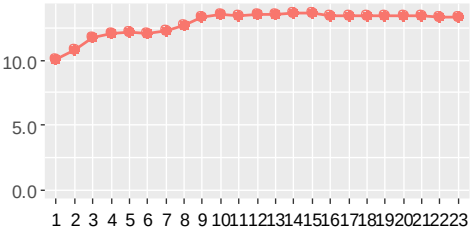
TTR. División acumulada.



Token: *5000

Figura 3

TTR.est. División acumulada.



Token: *5000

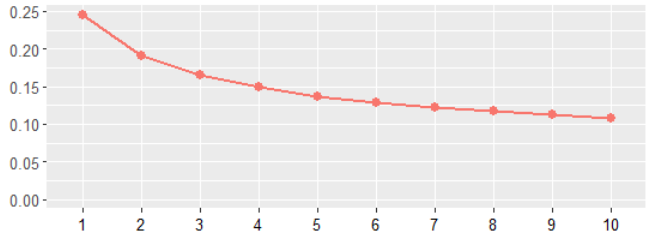
Efectivamente, la eficacia de la estandarización en TTR.est se observa mayoritariamente, aunque no perfectamente en forma plana. Observamos que el comportamiento numérico de TTR.est. por la división acumulada es bastante estable,

menos en las primeras divisiones donde las cantidades de lemas son relativamente reducidas: 5000, 10000, …, lo que demuestra que su estabilidad depende de la cantidad total de lemas.

En la Figura 4 observamos el comportamiento de acuerdo con el aumento del tamaño del intervalo, 1000, 2000, …, 10000.

Figura 4

TTR estandarizada según tamaño de intervalo.



Intervalo *1000

Es comprensible la curva trazada, puesto que estamos observando los valores de TTR, que suelen presentar la cifra cada vez menor a medida que aumenta la totalidad (*tokens*), como hemos visto anteriormente.

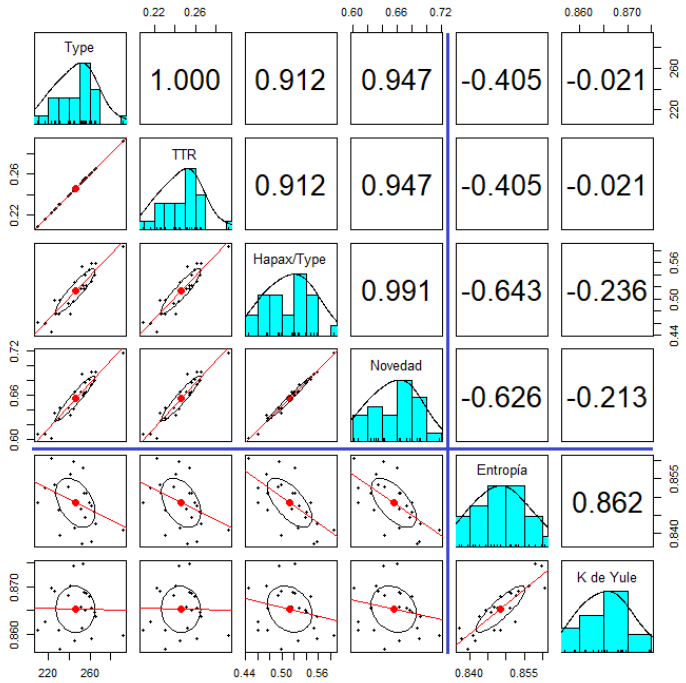
2.3 Comparación de los índices de riqueza

Para observar la coherencia entre los índices tratados, hemos recurrido al coeficiente de correlación de Pearson entre todos los pares posibles, siguiendo el método de Capsada y Torruella (2017). La metodología a partir de la cual se lleva a cabo la estadística léxica emplea el paquete R de *ggplot* y *psych*.

La Figura 5 muestra el resultado de aplicar los datos derivados de los valores estandarizados que se han presentado en distintas divisiones separadas[3].

Figura 5

Correlación entre índices de riqueza.



La Figura 5 muestra con nitidez que los índices se dividen en dos grupos: Type, TTR, Hápax/Type y Novedad léxica, por una parte, y Entropía normalizada y K de Yule normalizado, por otra. Esto es así, puesto que dentro de cada grupo se presentan las mayores cifras de coeficiente de correlación y el gráfico de esparcimiento correlacionado, mientras que entre los grupos se presentan los coeficientes de correlación escasos o a la inversa. De este modo, los índices del segundo grupo no son coherentes con los del primero: los últimos dos índices (Entropía normalizada y K de Yule) parecen indicar no precisamente la riqueza léxica, sino más bien la diversidad léxica.

2.4 El corpus PRESEEA-Santander

El Proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA) se concreta en la creación de corpus sociolingüísticos sincrónicos de lengua española hablada, representativos del mundo hispánico en su variedad geográfica y social. En la actualidad, agrupa a cerca de 50 equipos de investigación sociolingüística coordinados en torno a una misma metodología y a idénticos principios teóricos de carácter sociolingüístico: la concepción del dialecto como propiedad de una comunidad de habla, la variabilidad como rasgo caracterizador de la

lengua, la cuantificación como método analítico y la representatividad de las muestras de habla (Moreno Fernández et al., 2000).

El corpus para el estudio sociolingüístico de la ciudad de Santander, situada en el norte de España, se ha conformado siguiendo los requisitos mínimos contemplados en la metodología común del proyecto internacional (Martínez & Ueda, 2023). Entre ellos se destaca que el muestreo debe ser representativo del universo que sirve de base al estudio sociolingüístico. Se trabaja con muestras por cuotas con afijación uniforme, consistentes en "dividir el universo relativo en subpoblaciones, estratos o cuotas - atendiendo a unas variables sociales determinadas- y en asignar igual número de informantes a cada una de esas cuotas" (Moreno Fernández, 2021, p.13). Las variables acotadas son tres: sexo, edad y nivel de instrucción. Por sexo, la población queda agrupada en Hombres (H) y Mujeres (M); en la estratificación por edad los informantes se distribuyen en Generación 1 (de 20 a 34 años), Generación 2 (de 35 a 54 años) y Generación 3 (de 55 años en adelante). La estratificación por grado de instrucción se delimita en tres niveles: Nivel 1 (educación básica, hasta la edad de 10 años, aproximadamente); Nivel 2 (educación secundaria hasta la edad de 16-18) y Nivel 3 (educación superior, hasta la edad de 21-22).

La muestra queda representada a través de la Tabla 1, con un informante por cuota o estrato, lo que da lugar a los 18 informantes que la integran.

Tabla 1
Muestra-tipo por cuotas con número mínimo de informantes (adaptación de Moreno Fernández, 2021, p. 14).

Sexo	Hombre			Mujer		
Nivel / Edad	Nivel 1	Nivel 2	Nivel 3	Nivel 1	Nivel 2	Nivel 3
Edad 1	1	1	1	1	1	1
Edad 2	1	1	1	1	1	1
Edad 3	1	1	1	1	1	1

3. Análisis de datos y discusión

En esta sección, analizamos la riqueza léxica del corpus PRESEEA-Santander utilizando el índice de Novedad léxica. Como ya señalamos en la introducción, nuestra perspectiva es doble: por un lado, los parámetros sociolingüísticos (sexo, edad y nivel educativo) y, por otro lado, la categoría gramatical. En este orden, por tanto, serán analizados los resultados del análisis en las cifras de Nlex.

3.1 Parámetros sociolingüísticos

En esta subsección, nos dedicamos a analizar el corpus entero sin dividir por categorías gramaticales sino por los tres parámetros en torno a los cuales se estructura el corpus. Los lemas del corpus se clasifican de la siguiente manera, tal y como se muestra en la Tabla 2.

Tabla 2

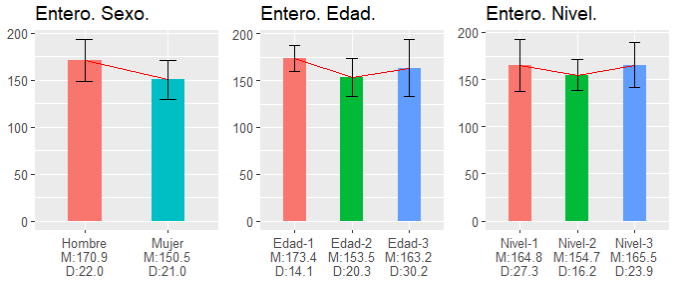
Frecuencia de lemas en función de los parámetros sociolingüísticos.

Sexo	Hombre			Mujer		
Nivel / Edad	Nivel 1	Nivel 2	Nivel 3	Nivel 1	Nivel 2	Nivel 3
Edad 1	6380	4263	7699	6348	3337	4343
Edad 2	6395	8319	5437	6992	10278	3710
Edad 3	8161	7634	5888	5141	5459	9352

Las cifras del índice de Novedad léxica se distribuyen como muestran las Figuras 6, 7 y 8.

Figuras 6, 7 y 8

Índice N del corpus en función del sexo, edad y nivel de instrucción.



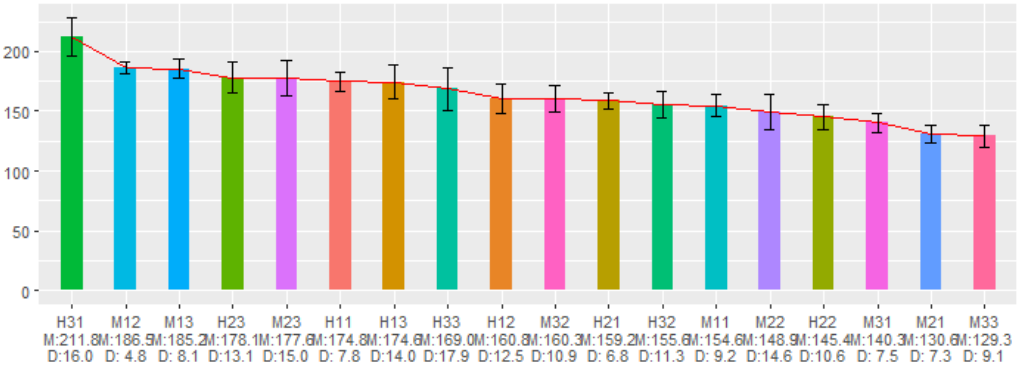
Las tendencias numéricas que observamos en el corpus entero son muy parecidas a las de las palabras de contenido (sustantivos, adjetivos, adverbios y verbos). Ello significa que la riqueza léxica del corpus se representa en la propia de las palabras de contenido; es decir, cuanto más elevada sea la cifra de Nlex en las palabras de contenido, se considera que el texto es más rico.

El valor Hombre, dentro del parámetro Sexo, manifiesta un índice ligeramente mayor de Nlex con respecto al de Mujer, en general. En el parámetro de edad y en el de nivel, nos sigue llamando la atención la curva en forma de V, con el descenso en edad-2 y nivel-2. Para indagar la causa de esta tendencia, vamos a ver cada caso particular dentro de los 18 casos que constituyen la muestra representativa antes descrita.

La Figura 9 muestra los valores de Novedad léxica en orden descendente.

Figura 9

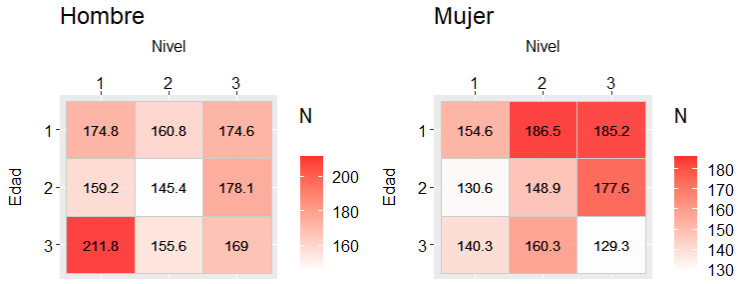
Índice de N del corpus en función de los parámetros sociolingüísticos en orden descendente.



Como resulta difícil interpretar este orden descendente, recurrimos a los mapas de calor clasificados por parámetros representados en las Figuras 10 y 11.

Figuras 10 y 11

Mapas de calor clasificados por parámetros.



En estas figuras tampoco encontramos tendencias generales. La situación es tan compleja que exige un análisis más detallado. Anteriormente hemos explicado que Nlex en Hombre es mayor que en Mujer, en general. En particular, sin embargo, la situación no es tan sencilla como demuestra un análisis comparativo enfocado. Ahora bien, no se trata de una comparación entre Hombre y Mujer de manera conjunta, sino enfocada en pares distintivos que se diferencian solo en un parámetro único, en este caso, el del sexo. Por ejemplo, se compara el caso de H-E1-N1 y M-E1-N1, donde se encuentra la diferencia solo en Sexo: H y M. De esta manera comparamos 9 pares y marcamos el valor más grande con un subrayado. Veamos el resultado a través de la Tabla 3.

Tabla 3

Comparación enfocada en pares distintivos por sexo.

Sexo	**Hombre**			**Mujer**		
Nivel / Edad	Nivel 1	Nivel 2	Nivel 3	Nivel 1	Nivel 2	Nivel 3
Edad 1	174.8	160.8	174.6	154.6	<186.5>	185.2
Edad 2	159.2	(145.4)	178.1	130.6	148.9	177.6
Edad 3	<211.8>	155.6	169.0	140.3	160.3	(129.3)

Efectivamente, en el nivel educativo bajo el índice para Hombre es mayor que el de Mujer. En cambio, en el nivel educativo medio ocurre lo contrario, pues el índice de N para Mujer es más alto. En el nivel superior se diversifica entre edad-1, donde Mujer manifiesta un índice mayor y edad-2 y 3, donde aquí el índice mayor lo refleja Hombre. Como se ha indicado antes, esto representa una descripción de la muestra sin pretensión de generalizar.

Al realizar la misma comparación de tríos distintivos en edad y nivel educativo, no hemos podido encontrar ninguna tendencia general ni particular, tal y como se muestra en las Tablas 4 y 5.

Tabla 4
Comparación enfocada en pares distintivos por Edad.

Sexo	Hombre			Mujer		
Nivel / Edad	Nivel 1	Nivel 2	Nivel 3	Nivel 1	Nivel 2	Nivel 3
Edad 1	174.8	160.8	174.6	(154.6)	**<186.5>**	185.2
Edad 2	159.2	145.4	**<178.1>**	(130.6)	148.9	177.6
Edad 3	<211.8>	155.6	169.0	140.3	160.3	(129.3)

Tabla 5
Comparación enfocada en pares distintivos por Nivel educativo.

Sexo	Hombre			Mujer		
Nivel / Edad	Nivel 1	Nivel 2	Nivel 3	Nivel 1	Nivel 2	Nivel 3
Edad 1	174.8	160.8	174.6	154.6	**<186.5>**	<185.2>
Edad 2	159.2	(145.4)	178.1	(130.6)	148.9	177.6
Edad 3	<211.8>	155.6	169.0	140.3	160.3	(129.3)

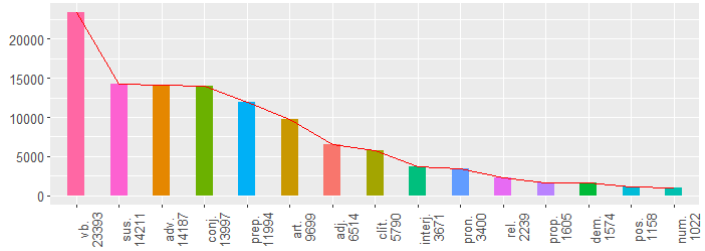
Hemos marcado las cifras más altas entre paréntesis angulares <…> y las dos más bajas entre paréntesis normales (…) dentro de cada conjunto de comparaciones. Estas cifras destacadas ejercen el efecto positivo y el negativo donde aparecen como miembro del par o del trío en comparación. Como estos valores pertenecen a la muestra antes descrita (18 sujetos), no pueden ser representativos de ninguna manera. En este sentido se desea llevar a cabo un estudio con más datos de personas pertenecientes a cada estrato de los distintos sociolectos que conforman el corpus. Será quizá en este momento cuando podamos resolver el hallazgo de la forma en V antes mencionado.

3.2. Categorías gramaticales

Nos interesa observar ahora las cifras de la Novedad léxica en las categorías gramaticales más frecuentes. Abordaremos, en primer lugar, un análisis contrastivo para pasar, a continuación, a un análisis por categoría.

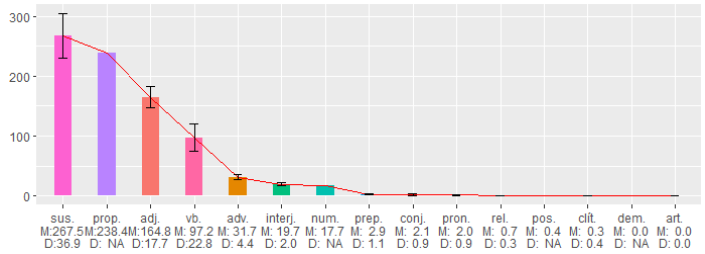
La Figura 12 muestra las frecuencias de las 11 categorías analizadas en orden descendente[4].

Figura 12
Categorías gramaticales más frecuentes.



Estas diferencias numéricas (de *tokens*) no afectan directamente al comportamiento numérico de la Novedad léxica, como muestra la Figura 13, puesto que se trata de valores estandarizados en los que se aborda la frecuencia de manera inversa: 1/f. Veamos, pues, en esta Figura 13 el orden descendente del valor medio.

Figura 13
Categoría gramatical. Novedad léxica (M: Media, D: Desv.est.).



Es importante la diferencia de cifras que se encuentra entre las siete primeras categorías gramaticales (sustantivo, nombre propio, adjetivo, verbo, adverbio, interjección y numeral), por una parte, y las restantes (preposición, conjunción, pronombre, relativo, posesivo, clítico y artículo). Naturalmente, el primer grupo pertenece a las palabras de contenido, las que conllevan más información léxica y el segundo grupo, a las palabras funcionales, carentes de significado léxico. De esta manera, la diferencia gramatical se manifiesta estadísticamente a las claras.

También comprobamos las diferencias de Novedad léxica dentro del primer grupo: sustantivo, nombre propio, adjetivo y verbo frente a adverbio, interjección y numeral. Las cuatro primeras categorías (sustantivo, nombre propio, adjetivo, verbo) son más informativas que las tres últimas (interjección, adverbio y numeral), como muestran sus cifras de Nlex.n. La categoría 'sustantivo' ocupa el primer lugar, lo que resulta convincente en el sentido de que se utiliza para transmitir primordialmente la información nueva, y por ello se corresponde con su alta Novedad léxica.

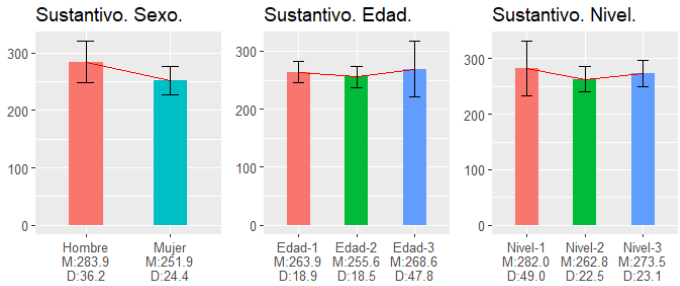
La frecuencia del nombre propio es reducida (=1605) y precisamente por esta razón hay bastantes casos de una sola ocurrencia o de dos ocurrencias, y así sucesivamente, siempre de reducida ocurrencia, lo que contribuye a una cifra alta de

Novedad léxica. A continuación, sigue el adjetivo y ya después, a cierta distancia, el verbo. El último lugar lo ocupan los artículos, palabras funcionales por excelencia. Las diferencias entre las primeras categorías son significativas, por no presentar coincidencia de los ámbitos de variación (media ± desviación estándar).

Estas características informativas manifestadas en la Novedad léxica se reconocen en nuestro uso lingüístico cotidiano. Al utilizar estos términos con frecuencia, se establecen en la memoria y se convierten en términos normales, no especiales, es decir, poco informativos. Ahora bien, la cuestión del marcaje lingüístico (marcado vs. no marcado) no se reduce a una dicotomía, sino a una gradación cuantificable de Novedad léxica cuantificable por la Novedad léxica.

En el análisis pormenorizado por categorías, comenzamos abordando el sustantivo. Las Figuras 14, 15 y 16 presentan en contraste las cifras de la Novedad léxica de los sustantivos del corpus en función de los parámetros sociolingüísticos: sexo, edad y nivel educativo. Nuestro propósito no es presentar la significación de la diferencia de frecuencia, lo que se lleva a cabo en la estadística inferencial, sino presentar grados de riqueza léxica en forma de Novedad léxica, donde no se aplica el cálculo de la significatividad.

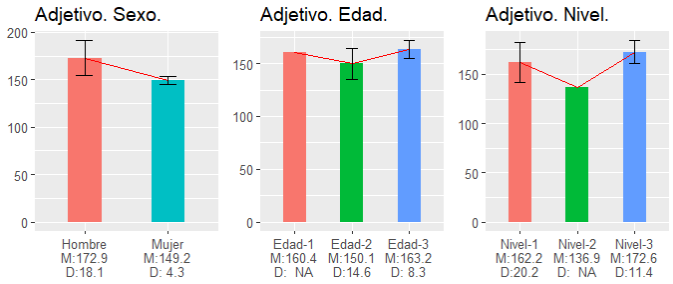
Figuras 14, 15 y 16
Novedad léxica del sustantivo en función del sexo, edad y nivel de instrucción.



Como veremos de aquí en adelante, el mayor valor de Novedad léxica en Hombre con respecto a Mujer es constante. En cambio, tanto en edad como en nivel educativo, las tendencias de la curva varían según el caso. Especialmente, nos llama la atención la forma de V en Edad: Edad-1 alto, Edad-2 bajo, Edad-3 alto; y en Nivel educativo: Nivel-1 alto, Nivel-2 bajo, Nivel-3 alto. Este patrón de alto - bajo - alto se encuentra en edad y nivel educativo en distintas categorías gramaticales.

Con respecto al adjetivo, la situación del parámetro edad cambia en esta categoría, tal y como se aprecia en las Figuras 17, 18 y 19, nuevamente en contraste.

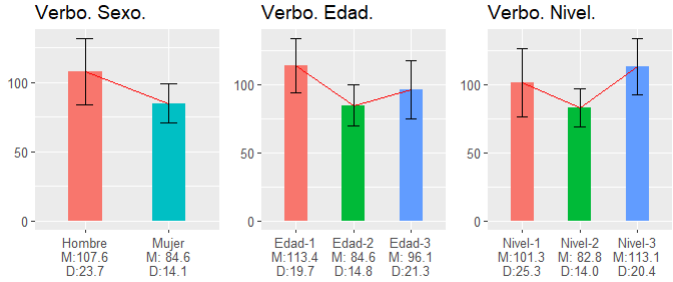
Figuras 17, 18 y 19
Novedad léxica del adjetivo en función del sexo, edad y nivel de instrucción.



En adjetivos, tanto la edad como el nivel educativo presentan de nuevo la línea en forma de V, pero esta vez de manera más marcada que en los sustantivos. En el parámetro sexo, el índice de hombre es mayor que el de mujer en lo que respecta a la Novedad léxica del adjetivo.

Con respecto al verbo, es interesante observar la tendencia de esta categoría en el parámetro de edad en las Figuras 20, 21 y 22.

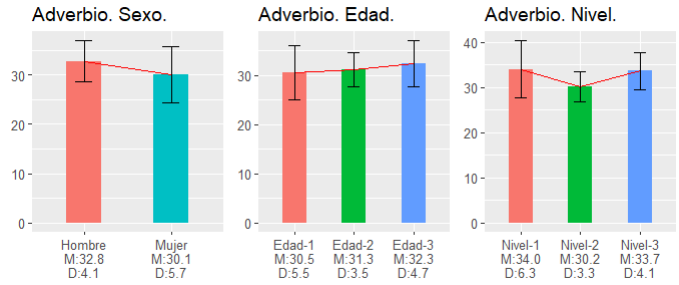
Figuras 20, 21 y 22
Novedad léxica del verbo en función del sexo, edad y nivel de instrucción.



El valor para hombre sigue siendo mayor que el de mujer, esta vez con gran diferencia. La forma en V se repite en edad y nivel educativo, esta vez con una diferencia aún más marcada.

Por último, la situación del adverbio es diferente en el parámetro de edad, como se advierte en las Figuras 23, 24 y 25 que se muestran a continuación.

Figuras 23, 24 y 25
Novedad léxica del adverbio en función del sexo, edad y nivel de instrucción.



Es palpable a partir del índice Nlex que la variedad de adverbios avanza de acuerdo con la edad.

CONCLUSIONES

En este estudio hemos revisado los distintos índices que existen hasta el momento para analizar la riqueza léxica y hemos propuesto nuestro propio índice: la Novedad léxica. Consideramos que dicho concepto demuestra el grado de Novedad léxica, en la medida en que creemos conveniente y útil conocer la constitución de frecuencia de palabras en un determinado texto. Para observar su aplicabilidad, hemos analizado con ella nuestro corpus PRESEEA-Santander en sus aspectos sociolingüísticos y gramaticales.

En la sección 3.2 dedicada a las categorías gramaticales, hemos advertido diferencias notables del índice de Novedad léxica entre dichas categorías. En cambio, las diferencias entre parámetros sociales (sexo, edad, nivel educativo) son leves, lo que demuestra cierta homogeneidad en la Novedad léxica.

Consideramos que las observaciones de categorías gramaticales son válidas, puesto que se basan en la muestra de 18 personas sociolingüísticamente no sesgadas con una distribución equitativa de sexo, edad y nivel educativo. Sus datos en conjunto son, por tanto, representativos. En las Figuras 6-8 y 14-25, no hemos descubierto las tendencias generales de las cifras. La única tendencia constante ha sido el sexo, donde siempre el índice Nlex. es mayor en Hombre que en Mujer. Sin embargo, existe una diferencia social, aunque esta sea leve. Para detectarla, es imprescindible buscar los índices de mayor sensibilidad.

Creemos haber demostrado la validez de nuestra decisión de seleccionar la Novedad léxica en lo anteriormente expuesto. También creemos necesario realizar un análisis general con el conjunto de datos y análisis particulares con los datos separados para formar pares distintivos. Por otro lado, quedan para otro estudio la comparación de los distintos índices mencionados a partir del análisis del corpus PRESEEA-Santander: Type, T (TTR), Tm (TTR modificada), Th (TTR de Herdan), H (Hápax),

HT (Hápax/Type), E (Entropía), En (Entropía normalizada), Nlex (Novedad léxica), Nlex.n (Novedad léxica normalizada), K (K de Yule), Kn (K de Yule normalizado), así como sus correspondientes valores estandarizados.

Junto a la necesidad de ampliar análisis en las líneas trazadas, cabe proponer como líneas futuras de la investigación el abordaje de los datos en cotejo con el resto de las ciudades españolas y de las americanas.

REFERENCIAS BIBLIOGRÁFICAS

Ávila Muñoz, A. (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, 171-194.

Baayen, H. (2001). *Word frequency distributions*. Kluwer Academic Publishers.

Baayen, H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge University Press.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh University Press.

Capsada, R., & Torruella, J. (2017). Métodos para medir la riqueza léxica de los textos. Revisión y propuesta. *Verba,* 44, 347-408.

Herdan, G. (1956). *Language as choice and change*. Groningen.

Ishikawa, S. (2012). *A Basic Guide to Corpus Linguistics*. Hitsuzishobo.

Kin, M. (2009). *Introduction to statistic science of text data*. Iwanamishoten.

Maekawa, M. (1995). *Analyze text with science*. Iwanamishoten.

Martínez, I., & Ueda, H. (2021). *Inventario léxico de PRESEEA-Santander. Proyecto para el estudio sociolingüístico del español de España y América*. https://h-ueda.sakura.ne.jp/lyneal/preseea.htm

Martínez, I., & Ueda, H. (2023). Métodos de lexicometría sociolingüística: análisis del corpus oral contemporáneo PRESEEA-Santander. *Círculo de Lingüística Aplicada a la Comunicación, 94*, 227-245. https://doi.org/10.5209/clac.81206.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics*. Cambridge University Press.

Moreno Fernández, F. (2021). Metodología del proyecto para el estudio sociolingüístico del español de España y América (PRESEEA). Universidad de Alcalá de Henares. https://preseea.linguas.net/Metodolog%C3%ADa.aspx

Rojo, G. (2002). *Sobre la lingüística basada en el análisis de corpus* [Ponencia plenaria en las *Jornadas sobre Corpus Lingüísticos,* San Sebastián].

Rojo, G. (2021). *Introducción a la lingüística de corpus en español.* Routledge.

Shannon, C. E., & Weaver, W. (1963). *The mathematical theory of communication.* University of Illinois Press.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*: 688. https://www.nature.com/articles/163688a0.pdf

Stubbs, M. (2006). *Words and phrases. Corpus studies of lexical semantics.* Blackwell Publishing.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*, 323-352.

Yule, G. U. (1944). *Statistical study of literary vocabulary*. Cambridge University Press.

NOTAS

[1] Véase Martínez y Ueda (2021) https://h-ueda.sakura.ne.jp/lyneal/preseea.htm

[2] El método de estandarización es aplicable incluso a K de Yule para unificar la magnitud del texto objeto de investigación, a pesar de que K no la necesita en principio.

[3] Leyenda de abreviaturas: vb: verbo; sus: sustantivo; adv: adverbio; conj: conjunción; prep.: preposición; art.: artículo; adj.: adjetivo; clit.: clítico; interj.: interjección; pron.: pronombre, rel: relativo, prop: nombre propio, dem: demostrativo, pos: posesivo, num: numeral.

[4] M: valor medio, D: Desviación estándar. En los gráficos de barras hemos agregado líneas horizontales de error, una línea superior que corresponde al valor medio + desviación estándar, y otra inferior que corresponde al valor medio - desviación estándar. Estas líneas de error (vacilación) sirven para ver las posibilidades de coincidencia y discrepancia dentro de la posible variación. Al coincidir la mayor parte entre las barras, la diferencia se vuelve poco significativa, lo que ha ocurrido en los gráficos de parámetros sociales.