Boletín de Filología, Tomo LX
Número 2 (2025): 391-414

La ratio de categoría gramatical en español. Rasgos estilísticos y sociolingüísticos

Inmaculada Martínez Martínez[1]
Universidad de Cantabria, España

Hiroto Ueda[2]
Universidad de Tokio, Japón

Resumen

La Ratio de Categoría Gramatical (RCG) apenas ha sido estudiada en el ámbito de la sociolingüística, principalmente en relación con factores extralingüísticos como el sexo, la edad o los antecedentes educativos. Este hecho se debe, probablemente, a que la RCG muestra escasa variación sociolingüística: si este es el caso, la constancia sociolingüística de la RCG indicará la esencia de la lengua española. En esta investigación presentamos el análisis de la RCG en las entrevistas realizadas en Santander a 54 hablantes clasificados por sexo, edad y nivel educativo correspondientes al proyecto PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y América). Examinamos las causas de esta constancia de acuerdo con el "Sistema" y "Norma" del lenguaje (Coseriu 1973) y la "Ley de los grandes números" de la estadística teórica (Corbalán y Sanz

[1] Para correspondencia, dirigirse a: Inmaculada Martínez Martínez (inmaculada.martinez@unican.es), Departamento de Filología. Universidad de Cantabria. Avda. de los Castros s/n, 39005 Santander, Cantabria, España. ORCID iD: 0000-0003-4760-0903).

[2] Para correspondencia, dirigirse a: Hiroto Ueda (hiroto.ueda.tokio@gmail.com), Universidasd de Tokio, Profesor emérito, ORCID iD: 0000-0003-3204-609X.

2011). Dentro de este estándar de uso, encontramos ciertos sesgos hacia los sustantivos en hombres, hacia los verbos en mujeres, hacia los adverbios en personas mayores y hacia sustantivos y adjetivos en personas con un alto nivel educativo. Estos sesgos que constituyen la norma lingüística son fiables, pues el coeficiente de variación de la RCG se volvió lo suficientemente pequeño debido a la cantidad expandida del número de materiales.

Palabras clave: categoría gramatical; ratio de categoría gramatical; variación sociolingüística; corpus PRESEEA-Santander

The Grammatical Category Ratio in Spanish. Stylistic and sociolinguistic features

Abstract

The Grammatical Category Ratio (GCR) has barely been studied in sociolinguistics, mainly in relation to extralinguistic factors such as sex, age, or educational background. This is probably due to the fact that the GCR shows little sociolinguistic variation: if this is the case, the sociolinguistic constancy of the GCR will indicate the essence of the Spanish language. In this research, we present the analysis of the GCR in interviews conducted in Santander with 54 speakers classified by sex, age, and educational level, corresponding to the PRESEEA project (Project for the Sociolinguistic Study of Spanish in Spain and the Americas). We examine the causes of this constancy according to the "System" and "Norm" of language (Coseriu 1973) and the "Law of Large Numbers" of theoretical statistics (Corbalán and Sanz 2011). Within this pattern of usage, we found certain biases toward nouns in men, toward verbs in women, toward adverbs in older people, and toward nouns and adjectives in highly educated people. These biases, which constitute the linguistic norm, are reliable, as the coefficient of variation of the GCR became sufficiently small due to the expanded number of materials.

Key Words: grammatical category; grammatical category ratio; sociolinguistic variation; PRESEEA-Santander corpus

Recibido: 15/07/2025 Aceptado: 02/10/2025

1. INTRODUCCIÓN

La Ratio de Categoría Gramatical (RCG) contempla numerosos ámbitos de aplicación: desde la estilística literaria general hasta el análisis de autor, obra o género en obras concretas, pasando por el estudio de los cambios históricos de las palabras en español. En cambio, en el campo de la sociolingüística no hemos encontrado apenas bibliografía sobre la relación entre la RCG y factores extralingüísticos como el sexo, la edad o el nivel educativo, probablemente debido al hecho de que la RCG no muestra, a priori, mucha variación sociolingüística. Si este es el caso, la constancia sociolingüística de la RCG indicará la esencia fundamental de la lengua española en general.

Confirmamos que la RCG de palabras de contenido varía considerablemente entre las tipologías presentes en análisis efectuados con anterioridad, como el Diccionario de frecuencias de Juilland y Chang-Rodríguez (1964). En concreto, este diccionario deja constancia de que en obras teatrales hay más verbos y menos sustantivos y adjetivos, con diferencias destacadas. En ensayos, prensa y textos científicos, la proporción se invierte, mientras las novelas exhiben un estado intermedio entre ambos. Con respecto al análisis que se lleva a cabo para este estudio, correspondiente a la lengua oral presente en el corpus PRESEEA de Santander (Proyecto para el Estudio Sociolingüístico del Español de España y América) (Moreno Fernández 1993), cabe señalar aquí una menor variabilidad que la encontrada en el diccionario de frecuencias citado. La RCG casi constante observada en las entrevistas no puede ser resultado del azar, puesto que se trata de los hábitos lingüísticos que cada hablante ha adquirido de forma independiente en diferentes entornos lingüísticos.

Se hará necesario, por tanto, examinar las causas de la constancia de la RCG en el lenguaje hablado desde la perspectiva de constructos como los de *sistema* y *norma* del lenguaje (Coseriu 1973) y la *ley de los grandes números* establecida por la estadística teórica (Corbalán y Sanz 2011). Desde la perspectiva del sistema lingüístico, es posible utilizar una categoría gramatical tantas veces como se desee, pero existe un cierto estándar de uso como norma. Por ejemplo, es habitual que la RCG de verbos en el lenguaje hablado se aproxime al 20% según nuestro cálculo. Este estándar presenta una gran variabilidad en muestras pequeñas. Sin embargo, a medida que el material aumenta, la amplitud de la variación se vuelve más pequeña por la mencionada ley de los grandes números.

Dentro de este estándar de uso, encontramos ciertas tendencias: hacia los sustantivos en hombres; hacia los verbos en mujeres; hacia los adverbios en personas mayores; hacia los sustantivos y adjetivos en personas con un alto

nivel educativo. Estos sesgos que constituyen la norma lingüística son fiables, puesto que el coeficiente de variación de la RCG se volvió lo suficientemente pequeño debido a la cantidad expandida del número de materiales.

En la investigación que aquí se presenta, se aborda, en primer lugar, un estado de la cuestión de los estudios previos sobre la RCG (§2); en segundo lugar, se presentan los métodos estadísticos con los que hemos realizado la investigación (§3); en tercer lugar, se analizan, bajo el mismo prisma de la RCG las 54 entrevistas clasificadas por sexo, edad y nivel educativo correspondientes al proyecto PRESEEA (§4) y se finaliza con las oportunas conclusiones (§5).

2. ESTUDIOS ANTERIORES

La Ratio de Categoría Gramatical (RCG) ha sido tratada en distintos campos de la lingüística y la literatura. En lo que sigue, presentaremos el estado de la cuestión de aquellos que consideramos importantes para nuestra investigación de la RCG: la Lingüística de corpus (§2.1) y las variables sociolingüísticas (§2.2). Es preciso señalar que incluimos los estudios no solamente sobre RCG, sino también sobre frecuencias de categoría gramatical en general para dotar a la investigación de un mayor rigor y profundidad.

2.1. Lingüística de corpus

En el ámbito de la metodología lingüística basada en el trabajo con corpus, un nutrido grupo de autores nos han acercado, a través de análisis empíricos, a los principales patrones lingüísticos que se observan en las categorías gramaticales. Así, Geens (1978) es uno de los primeros autores que abre el camino al comparar las frecuencias de palabras funcionales del *Brown Corpus* y del *Theatre Corpus*, utilizando el valor *z* para conocer la significatividad de la diferencia. Conforme a lo esperado, este autor consigue valores significativamente más altos de formas deícticas y exclamaciones en el corpus de obras teatrales y de formas conjuntivas en el *Brown Corpus*.

Biber (1988: 76-78, 246-269), uno de los investigadores representativos en esta área, presenta los valores estadísticos de media, mínimo, máximo, rango (recorrido) y desviación estándar. Los aplica en múltiples categorías y subcategorías gramaticales, como tiempo, persona y número del verbo,

o pronombre, demostrativo, indefinido, nombre, y preposición, entre otros, de todos y cada uno de los corpus del inglés hablado y escrito. Este estudio representa el primer acercamiento al análisis estadístico de las categorías gramaticales desde la Lingüística de corpus: de ahí su valor.

Koskela (1998), por su parte, determinó que, aunque el uso de frases nominales puede encontrarse en todo tipo de textos, estas se usan con mayor proporción en textos de especialidad. Su gran aportación descansa en la cuantificación, no solo de las frases nominales, sino también de la categoría verbal en ensayos de carácter filosófico de la lengua finesa. El objetivo de su estudio es el de descubrir cómo se relacionan estos fenómenos entre sí y qué podrían decirnos sobre el ensayo como género y sobre el autor como escritor. Los resultados apoyan la hipótesis de que el ensayo es un género híbrido entre la escritura académica propiamente dicha y la literatura. Por un lado, demuestra que hay muchos verbos diferentes en los ensayos estudiados, pero que sus frecuencias individuales suelen ser relativamente bajas, lo cual parece indicar que existen expresiones subjetivas, concretas y dinámicas, propiamente literarias, también en este tipo de textos. Por otro lado, anota que la mayor frecuencia de los sintagmas nominales en estos textos atestigua el uso objetivo, abstracto y estático del lenguaje típico de la escritura académica.

Mair, Hundt, Leech y Smith (2003) comparan la RCG del corpus LOB (*Lancaster-Oslo-Bergen Corpus*) y de su actualización, el corpus Freiburg–LOB del inglés británico (FLOB). En torno a una comparabilidad de un millón de formas, estos autores señalan un aumento muy significativo en la frecuencia de sustantivos y adjetivos entre LOB (1961) y FLOB (1991). Los autores indican que este aumento se aplica tanto a los nombres comunes como a los propios, pero que es más significativo (un aumento del 11%) en el caso de los nombres propios.

Brezina (2018: 30-32) quiso indagar en la frecuencia con la que los académicos usan la categoría adjetival y la verificó con el *British National Corpus*. A través de la prueba *t* de *Student*, encuentra que los académicos usan los adjetivos con más frecuencia, pues advierte una diferencia significativa: 857 adjetivos por 10.000 palabras (ámbito académico) y 575 adjetivos por 10.000 palabras (ámbito de ficción).

Finalmente, Rojo (2021: 210), al reducir el número de palabras en todo el *Corpus del Español del siglo XXI* (CORPES) de la Real Academia Española, con la frecuencia normalizada (FN) >=0.005, FN>=0.05, FN>=0.1, FN>1, encuentra que la proporción que ocupan los verbos aumenta y la proporción de sustantivos, adjetivos y adverbios disminuye con el transcurso de los años.

Esta nutrida selección de estudios constituye el fundamento a partir del cual la investigación que aquí se presenta pretende avanzar en el ámbito de

la cuantificación en la lingüística de corpus, esta vez referida al área de la sociolingüística.

2.2. Variables sociolingüísticas

Alcanzamos el segundo de los constructos teóricos, el que aborda el análisis de la categoría gramatical desde el ámbito de la sociolingüística, en concreto, desde los parámetros sociolingüísticos de edad, sexo y nivel educativo. Se analizará también el registro (oral/escrito) en su vinculación con el concepto de categoría gramatical.

Hatano (1966: 22-47) solicitó a alumnos de Primaria (10-12 años), Secundaria (13-15) y Bachillerato (16-18) que escribieran las impresiones que tuvieran ante un dibujo concreto. Calculó la ratio de categoría gramatical a partir de la clasificación de sexo y curso, y midió el llamado 'coeficiente de acción' (verbo/adjetivo). Este análisis le permitió descubrir la siguiente progresión: este coeficiente es mayor en el hombre que en la mujer y el coeficiente se invierte en favor de la mujer; para la conjunción subordinante, de nuevo es el hombre quien presenta un coeficiente mayor y, finalmente, para la conjunción coordinante, la mujer presenta un mayor índice.

En el contexto hispánico, y ya en el siglo XXI, en lo que respecta a la edad, Terrádez Gurrea (2001) descubre que los adjetivos son mucho más utilizados por la tercera generación, mientras que los verbos presentan mayor uso en las generaciones primera y segunda, lo que puede interpretarse como la existencia de un mayor contenido elaborado y una menor tendencia a la implicación conversacional por parte de la generación longeva. En relación con el sexo, las mujeres utilizan más sustantivos, adjetivos y verbos que los hombres, lo que parece constatar que el género femenino muestra una mayor cohesión en el léxico que actualiza. Por último, en lo que atañe al nivel sociocultural, si este es bajo, utilizará menos adjetivos y sustantivos que los otros dos grupos (niveles medio y alto de instrucción), por lo que su contenido informacional y de elaboración será menor.

Por último, Moreno *et al.* (2005: 159-161) centran su análisis en el registro y comparan el estilo del diccionario de Juilland & Chang-Rodríguez (1964), arriba mencionado, y el estilo hablado de su propio corpus y llegan a las siguientes conclusiones: en primer lugar, se advierte una distribución inversa para las categorías léxicas (nombres, verbos y adjetivos, entre otras) y no léxicas (conjunciones y preposiciones, entre otras), pues las segundas son mucho más frecuentes en el registro oral, mientras que las léxicas están más presentes en los textos escritos; en segundo lugar, en el registro oral, la presencia de verbos y adverbios es mucho mayor que en la escritura,

mientras que los nombres y los adjetivos son mucho más frecuentes en la lengua escrita que en el registro oral.

Analizaremos, después de acercarnos a los métodos estadísticos empleados en este estudio, el comportamiento de estas variables sociolingüísticas en un corpus oral, urbano y contemporáneo como es el corpus PRESEEA-Santander. Hemos observado que existen conexiones entre los dos constructos aquí abordados, la lingüística de corpus (§2.1) y la sociolingüística variacionista (§2.2), en el sentido de que ambas disciplinas tratan los materiales lingüísticos, bien en forma de textos escritos, bien a través de textos hablados. De esta manera, las dos se coordinan de manera complementaria. Por nuestra parte, proponemos desarrollar la sociolingüística variacionista de corpus con transcripciones de múltiples hablas, lo que precisamente intentamos en el presente trabajo.

3. MÉTODOS ESTADÍSTICOS

En lo que sigue, se explicarán los métodos de carácter estadístico aplicados en esta investigación, como son la frecuencia ajustada (§3.1), la probabilidad estandarizada (§3.2) y la concentración diagonal (§3.3).

3.1. Frecuencia ajustada

En los datos sociolingüísticos, es normal que los parámetros posean cantidades totales diferentes. Por ejemplo, los sujetos masculinos y los femeninos no presentan una totalidad igualada con exactitud, aunque trabajemos con el mismo número de personas equitativamente distribuidas. Por ello, para comparar las frecuencias de cada categoría es necesario relativizar la frecuencia observada (absoluta) por la suma de cada parámetro.

En un ejemplo sencillo, mostrado en la Tabla 1[3], la columna-1 (c1) corresponde al hombre (H) y la columna-2 (c2), a la mujer (M). Observamos las sumas de las dos variables, 6 y 15, respectivamente. Es necesario dividir la matriz de datos (D) por las dos sumas de columnas y, para ello, hemos

[3] En esta sección, utilizamos el sistema R (*R Core Team* 2021) y el paquete *ggplot2*.

preparado la función que produce la frecuencia relativa (*RF: Relative frequency*), que se muestra en esta misma tabla:

```
> D=Matrix(1:6,3,b=F); AddSum(D)  # Matriz de datos
   c1  c2  Sh
r1  1   4   5
r2  2   5   7
r3  3   6   9
Sv  6  15  21
> R=RF(D,F); R # Frecuencia relativa
         c1        c2
r1 0.1666667 0.2666667
r2 0.3333333 0.3333333
r3 0.5000000 0.4000000
> sum(D) # 21
[1] 21
> sum(R) # 2
[1] 2
```

Tabla 1. Función que produce la frecuencia relativa

Esta frecuencia relativa (R) es comparable, puesto que está relativizada por la cantidad total. Sin embargo, esta frecuencia relativa no es conveniente para observar la realidad cuantitativa, por ofrecer una cantidad siempre muy reducida dentro del recorrido entre 0 y 1 ([0, 1]). Por ello, los investigadores utilizan el porcentaje, que es la cantidad de la frecuencia relativa multiplicada por 100: 16,7%, 33,3%, etc. De esta manera, en el porcentaje podemos observar la cantidad dentro de la escala de [0, 100]. Sin embargo, es preciso admitir que el porcentaje tampoco representa la cantidad real de frecuencia, puesto que tiene sentido solo dentro de 100.

Para obtener una cantidad comparable que sea realista en el mismo recorrido de los datos observados, proponemos utilizar la frecuencia relativa ajustada (*Adjusted frecuency: AF*), que definimos de la siguiente manera y que mostramos en la Tabla 2:

AF = R / sum(R) * sum(D)
donde R: frecuencia relativa, D: matriz de datos (frecuencia absoluta).

```
> R/sum(R)*sum(D) # Frecuencia ajustada          ...(2)
       c1  c2
r1 1.75  2.8
r2 3.50  3.5
r3 5.25  4.2
> A=AF(D,F); A
   c1   c2
r1 1.75  2.8
r2 3.50  3.5
r3 5.25  4.2
> sum(D) # 21
[1] 21
> sum(A) # 21
[1] 21
```

Tabla 2. Frecuencia relativa ajustada

De esta manera, comprobamos que la cantidad total de la frecuencia ajustada (sum(A)) coincide con la de frecuencia observada (sum(D)), lo que se demuestra fácilmente:

$\Sigma(A) = \Sigma(R/r*d)$	R: frecuencia relativa, $r=\Sigma(R)$, $d=\Sigma(D)$
$= \Sigma(R) / r * d$	r y f son valores constantes
$= r / r * d = d = \Sigma(D)$	$r=\Sigma(R)$, $f=\Sigma(D)$

La misma frecuencia ajustada representa, por tanto, un valor, ahora sí, realista con la misma escala de magnitud que la frecuencia observada.

Finalmente presentamos la matriz de datos (D) y calculamos la frecuencia ajustada (A) y la frecuencia relativa vertical (R) en la Tabla 3:

```
> A=AF(D,h=F,r=2); R=R3(RF(D,F)); Cbind(D,A,R)
  D c1 c2 A   c1  c2 R    c1    c2
r1 |  1  4 | 1.75 2.8 | 0.167 0.267
r2 |  2  5 | 3.50 3.5 | 0.333 0.333
r3 |  3  6 | 5.25 4.2 | 0.500 0.400
```

Tabla 3. Frecuencia ajustada y frecuencia relativa vertical

3.2. Probabilidad estandarizada

Al abordar el segundo de los métodos, conviene señalar que, para observar los datos en una escala estandarizada, se utiliza el método de estandarización (S), cuya fórmula es la siguiente:

$$S = (D - M) / SD, \qquad \text{(M:media, SD:desviación estándar)}$$

Como apreciamos en la Tabla 4, la media de la matriz estandarizada es cero (0) y su desviación estándar es 1. De esta manera, por medio de la matriz estandarizada podemos comparar distintas matrices de diferentes magnitudes dentro de la misma escala:

```
> D=Matrix(1:6,3,b=F); D  # Matriz de datos
   c1 c2
r1  1  4
r2  2  5
r3  3  6
> S=(D-mean(D))/sd(D); S
          c1        c2
r1 -1.3363062  0.2672612
r2 -0.8017837  0.8017837
r3 -0.2672612  1.3363062
> S=ST(D); S # Matriz estandarizada
          c1        c2
r1 -1.3363062  0.2672612
r2 -0.8017837  0.8017837
r3 -0.2672612  1.3363062
> mean(S) # media
[1] 0
> sd(S) # desviación estándar
[1] 1
```

Tabla 4. Matriz estandarizada

Por otra parte, la matriz estandarizada tiene el mérito de aproximarse a la distribución normal estándar con la media en 0 y la desviación estándar en 1. La probabilidad producida por medio de los valores estandarizados, la denominamos “probabilidad estandarizada”, cuyo mérito consiste en ofrecer la probabilidad constante no afectada por la magnitud de datos, tal y como se muestra en la Tabla 5:

```
> set.seed(5); V=round(sort(rnorm(10,mean=50, sd=10))); V
 [1] 37 42 44 44 45 47 51 51 64 67
> R3(ST(V)) # R3: redondear a tres decimales.
 [1] -1.278 -0.754 -0.545 -0.545 -0.440 -0.231 0.189 0.189 1.551 1.865
> R3(ST(V*100))
 [1] -1.278 -0.754 -0.545 -0.545 -0.440 -0.231 0.189 0.189 1.551 1.865
```

Tabla 5. Probabilidad estandarizada

Es bien sabido que la probabilidad binomial y la normal utilizadas en pruebas estadísticas presentan valores muy bajos y muy altos que son significativos en los datos grandes. Por su parte, la probabilidad estandarizada permanece ajena a tal efecto, como se muestra en la Tabla 6:

```
> set.seed(0); V=sort(rbinom(10,size=100, prob=0.5)); V
 [1] 46 47 48 49 50 50 51 52 52 57
> R3(pbinom(V,sum(V),0.1))
 [1] 0.296 0.350 0.407 0.466 0.526 0.526 0.584 0.640 0.640 0.861
> W=V*100; R3(pbinom(W,sum(W),0.1))
 [1] 0.000 0.000 0.001 0.037 0.387 0.387 0.884 0.996 0.996 1.000
```

Tabla 6. Probabilidad estandarizada en datos grandes

3.3. Concentración diagonal

Para conocer la tendencia general de distribución de frecuencias, es útil el método de concentración diagonal (Ueda y Moreno Sandoval 1998). Dicho método consiste en concentrar los valores altos en la zona diagonal de la matriz de datos, que parte del rincón superior izquierdo y llega al rincón inferior derecho de manera lineal, tal y como se muestra en la Tabla 7:

```
D=Matrix(c(1,3,2,5,0,5,1,4,6,1,9,3),3,b=T); D # Data matrix
   c1 c2 c3 c4
r1  1  3  2  5
r2  0  5  1  4
r3  6  1  9  3
> Dc=DC(D); Dc
   c3 c1 c2 c4
r3  9  6  1  3
r2  1  0  5  4
r1  2  1  3  5
```

Tabla 7. Método de concentración diagonal

En la realización anterior en R, podemos comparar las dos matrices: matriz de datos (*data matrix*: D) y matriz diagonalizada (DC(D)). En la matriz diagonalizada, podemos observar la reordenación tanto de filas (r1, r2, r3 en r3, r2, 1) como de columnas (c1, c2, c3, c4 en c3, c1, c2, c4); en ella, los valores altos están concentrados en la parte superior izquierda y en la inferior derecha. Por esta reordenación, interpretamos el movimiento de las frecuencias desde los dos puntos de vista, horizontal y vertical, al mismo tiempo. Si el orden de filas, r3 - r2 - r1, representa, por ejemplo, un cambio del estilo bajo al alto, los parámetros de columna, con los atributos de los sujetos, también se interpretan como indicadores del mismo cambio de estilo.

También podemos realizar el método de concentración unilateral, con uno de los dos ejes no reordenado, por alguna razón convincente, por ejemplo, en torno a las edades (Edad-1, Edad-2, Edad-3), tal y como observamos en la Tabla 8:

```
> D=Matrix(c(1,3,2,5,0,2,1,4,6,1,9,3),4,b=T); D # Data matrix      ..
   c1 c2 c3
r1  1  3  2
r2  5  0  2
r3  1  4  6
r4  1  9  3
> Dc=DC(D,d=1); Dc
   c1 c2 c3
r2  5  0  2
r1  1  3  2
r4  1  9  3
r3  1  4  6
```

Tabla 8. Método de concentración unilateral

Esta vez, se reordenan solo las filas, puesto que el orden de c1 - c2 - c3 representa la Edad-1, Edad-2, Edad-3. En este caso, el orden de r2 - r1 - r4 - r3 sigue el orden de la edad.

4. ANÁLISIS DE DATOS

En el análisis de datos se presentará, en primer lugar, la caracterización de los informantes del corpus, así como las categorías objeto de estudio (§4.1). En segundo lugar, se aplicarán los métodos estadísticos arriba desarrollados en las variables sociolingüísticas de sexo, edad y nivel educativo (§4.2). Se terminará con el abordaje de los datos en su totalidad (§4.3), para, de esta forma, conseguir un análisis más completo y riguroso.

4.1. Informantes y categorías

Los datos que aquí se analizan provienen de transcripciones de encuestas realizadas en la ciudad de Santander (España) a 54 personas, clasificadas

en dos sexos (H, M), tres edades (E1, E2, E3) y tres niveles de educación (N1, N2, N3). Cabe señalar que cada sociolecto agrupa a tres informantes. Así se muestra en la Tabla 9:

H (Hombres) M (Mujeres)
E1 (de 20 a 34 años) E2 (de 35 a 54 años) E3 (de 55 años en adelante)
N1 (educación básica, hasta la edad de 10 años, aproximadamente) N2 (educación secundaria hasta la edad de 16-18) N3 (educación superior, hasta la edad de 21-22)

Tabla 9. Caracterización de los informantes del corpus

Por último, nuestra clasificación de categorías gramaticales sigue el método tradicional consistente en las siguientes 13 categorías[4]: adjetivo (a), artículo (ar), adverbio (av), conjunción (c), interjección (ij), interrogativo (ir), nombre (n), numeral (nm), preposición (p), pronombre (pn), relativo (r), signo (sg) y verbo (vb).

4.2. VARIABLES SOCIOLINGÜÍSTICAS

4.2.1. Sexo

Para la variable de sexo, presentamos la siguiente tabla cruzada de la frecuencia observada de datos (D) (Figura 1), de su correspondiente frecuencia ajustada (A) (Figura 2) y de la frecuencia relativa horizontal (R) (Figura 3):

D	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
H	12,965	14,264	21,848	19,934	3,710	769	25,000	1,824	18,999	14,797	3,512	34,659
M	11,715	12,362	21,114	20,404	2,745	873	23,233	1,717	17,144	15,721	3,163	35,242

Figura 1. Distribución de frecuencia (D). Sexo

[4] Para ahondar en los detalles, véase Martínez-Martínez y Ueda (2025).

A	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
H	12,707	13,981	21,414	19,538	3,636	754	24,503	1,788	18,621	14,503	3,442	33,970
M	11,957	12,618	21,551	20,826	2,802	891	23,714	1,753	17,499	16,046	3,228	35,971

Figura 2. Frecuencia ajustada (A). Sexo

R	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
H	0.075	0.083	0.127	0.116	0.022	0.004	0.145	0.011	0.110	0.086	0.020	0.201
M	0.071	0.075	0.128	0.123	0.017	0.005	0.140	0.010	0.104	0.095	0.019	0.213

Figura 3. Frecuencia relativa horizontal (R). Sexo

La frecuencia ajustada es útil para conocer la tendencia cuantitativa general de la totalidad y de las diferencias reales entre las dos variables: Hombre (H) y Mujer (M). Sin embargo, a partir de estos datos, es un tanto difícil evaluar la diferencia relativa entre ambos parámetros. Para destacar las cifras menores y mayores, hemos utilizado las probabilidades correspondientes a los elementos de la matriz estandarizada y hemos destacado (en color rojo) en la Figura 4 las probabilidades superiores a 0.7:

P	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
H	0.760	0.760	0.240	0.240	0.760	0.240	0.760	0.760	0.760	0.240	0.760	0.240
M	0.240	0.240	0.760	0.760	0.240	0.760	0.240	0.240	0.240	0.760	0.240	0.760

Figura 4. Probabilidad estandarizada (P). Sexo

Cuando se trata de los datos con dos variables (H, M), las probabilidades estandarizadas siempre presentan 0.760 (correspondiente a la frecuencia mayor de la media) y 0.240 (correspondiente a la frecuencia menor de la media). En este caso, nos interesa solo la diferencia de las dos, mayor y menor. Seguidamente, vamos a observar a través de la Figura 5 el resultado de la concentración diagonal:

C	a	ar	ij	n	nm	p	r	av	c	ir	pn	v
H	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.240	0.240	0.240	0.240	0.240
M	0.240	0.240	0.240	0.240	0.240	0.240	0.240	0.760	0.760	0.760	0.760	0.760

Figura 5. Probabilidad concentrada (C). Sexo

Según el resultado de la concentración diagonal, los hombres (H) utilizan el adjetivo, el artículo, la interjección, el nombre, el numeral, el pronombre y el relativo más que las mujeres, lo que se comprueba en la frecuencia ajustada, que hemos visto anteriormente (Figura 3). Sin embargo, en el caso de dos variables, el análisis de concentración diagonal aplicado a la

probabilidad estandarizada no nos ofrece mucha información. Por ello, en su lugar, presentamos el resultado de la concentración diagonal de frecuencia relativa horizontal a través de la Figura 6:

Rc	n	p	ar	a	ij	r	nm	ir	av	pn	c	v
H	0.145	0.110	0.083	0.075	0.022	0.020	0.011	0.004	0.127	0.086	0.116	0.201
M	0.140	0.104	0.075	0.071	0.017	0.019	0.010	0.005	0.128	0.095	0.123	0.213

Figura 6. Frecuencia relativa concentrada (P). Sexo.

En la matriz concentrada (Figura 6), podemos observar el siguiente movimiento de categorías gramaticales:

n - p - ar - a - ij - r - nm - ir - av - pn - c - v

En este orden, de acuerdo con los dos parámetros de sexo. El hallazgo estriba en que las primeras categorías (n, p, ar, ...) están más concentradas en los hombres (H), mientras que las últimas (v, c, pn, ...), en las mujeres (M). La gradación, por su parte, es continua, de manera que las probabilidades se presentan de mayor a menor sin interrupción.

Edad

En cuanto a la variable Edad, se presentan las siguientes tablas de frecuencia observada (D) (Figura 7), ajustada (A) (Figura 8) y relativa (R) (Figura 9):

D	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
E1	6,846	7,024	11,203	11,393	1,323	373	13,167	950	9,777	7,803	1,746	18,645
E2	8,856	9,386	14,962	14,387	2,624	560	17,127	1,274	13,060	10,694	2,325	25,249
E3	8,978	10,216	16,797	14,558	2,508	709	17,939	1,317	13,306	12,021	2,604	26,007

Figura 7. Distribución de frecuencia (D). Edad

A	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
E1	8,539	8,761	13,974	14,211	1,650	465	16,424	1,185	12,195	9,733	2,178	23,256
E2	8,273	8,768	13,977	13,440	2,451	523	16,000	1,190	12,200	9,990	2,172	23,587
E3	7,961	9,058	14,893	12,908	2,224	629	15,906	1,168	11,798	10,659	2,309	23,060

Figura 8. Frecuencia ajustada (A). Edad

R	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
E1	0.076	0.078	0.124	0.126	0.015	0.004	0.146	0.011	0.108	0.086	0.019	0.207
E2	0.073	0.078	0.124	0.119	0.022	0.005	0.142	0.011	0.108	0.089	0.019	0.210
E3	0.071	0.080	0.132	0.115	0.020	0.006	0.141	0.010	0.105	0.095	0.021	0.205

Figura 9. Frecuencia relativa horizontal (R). Edad

Observemos ahora la tendencia general del movimiento en la matriz diagonalmente concentrada de la Figura 10:

A	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
E1	0.872	0.854	0.835	0.714	0.636	0.433	0.133	0.187	0.205	0.295	0.281	0.275
E2	0.345	0.452	0.521	0.722	0.782	0.859	0.797	0.424	0.387	0.269	0.283	0.289
E3	0.230	0.175	0.153	0.124	0.130	0.183	0.610	0.860	0.867	0.876	0.876	0.876

Figura 10. Probabilidad concentrada (C). Edad

El orden de categorías gramaticales observado en la línea de edad, E1 - E2 -E3, es:

a - ar - av - c - ij - ir - n - nm - p - pn - r - v

Parece que las categorías de adjetivo (a), artículo (ar) y adverbio (av) se utilizan más en las personas jóvenes (E1), mientras que los verbos (v), relativos (r), y pronombres (pn), entre otros, son más bien propios de las personas mayores (3). Los adultos (E2), por su parte, coinciden con E1 y E3 en el uso relativo de categorías.

4.2.3. Nivel de educación

El último parámetro sociolingüístico tratado en este estudio es el nivel de educación. Presentamos los mismos análisis de distribución de frecuencia, frecuencia ajustada y frecuencia relativa horizontal a través de las Figuras 11, 12 y 13, respectivamente:

D	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
N1	7,618	8,314	13,051	13,391	1,634	589	15,074	1,243	11,690	10,255	2,104	23,011
N2	8,091	8,655	15,330	13,655	2,187	581	16,079	1,232	11,953	10,915	2,147	24,257
N3	8,971	9,657	14,581	13,292	2,634	472	17,080	1,066	12,500	9,348	2,424	22,633

Figura 11. Distribución de frecuencia (D). Nivel de educación

A	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
N1	7,942	8,668	13,607	13,961	1,704	614	15,716	1,296	12,188	10,692	2,194	23,991
N2	7,914	8,466	14,996	13,357	2,139	568	15,728	1,205	11,692	10,677	2,100	23,728
N3	8,808	9,481	14,316	13,050	2,586	463	16,769	1,047	12,273	9,178	2,380	22,221

Figura 12. Frecuencia ajustada (A). Nivel de educación

R	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
N1	0.071	0.077	0.121	0.124	0.015	0.005	0.140	0.012	0.108	0.095	0.019	0.213
N2	0.070	0.075	0.133	0.119	0.019	0.005	0.140	0.011	0.104	0.095	0.019	0.211
N3	0.078	0.084	0.127	0.116	0.023	0.004	0.149	0.009	0.109	0.082	0.021	0.197

Figura 13. Frecuencia relativa horizontal (R). Nivel de educación

La matriz concentrada de las categorías gramaticales en los tres niveles educativos muestra la no continuidad en la distribución, tal y como se advierte en la Figura 14:

*	c	nm	ir	v	pn	av	p	r	ij	ar	a	n
N1	0.862	0.816	0.802	0.761	0.721	0.157	0.669	0.415	0.160	0.352	0.291	0.279
N2	0.415	0.570	0.600	0.668	0.715	0.840	0.126	0.191	0.496	0.225	0.273	0.285
N3	0.191	0.141	0.135	0.126	0.124	0.506	0.760	0.862	0.842	0.872	0.876	0.876

Figura 14. Probabilidad concentrada (C). Nivel de educación

Según estos datos, las categorías se dividen en dos grupos: conjunción, numeral, interrogativo, verbo, pronombre y adverbio pertenecen a los dos primeros niveles: N1 (primaria) y N2 (secundaria); en contraste, nombre, adjetivo, artículo, interjección), relativo), y preposición destacan en N3, la educación superior.

4.2.4. Totalidad

En último lugar, nos acercamos a los datos en su conjunto, en forma de probabilidades correspondientes a los valores estandarizados, de los datos tratados en las secciones anteriores en los parámetros sociolingüísticos, pero en forma combinada, por ejemplo, H-E1-N1. El primer abordaje lo realizamos a través de la Figura 15:

P	a	ar	av	c	ij	ir	n	nm	p	pn	r	v
H-E1-N1	0.757	0.812	0.079	0.859	0.234	0.078	0.497	0.905	0.865	0.440	0.624	0.406
H-E1-N2	0.775	0.379	0.701	0.500	0.287	0.192	0.739	0.123	0.905	0.206	0.356	0.227
H-E1-N3	0.948	0.718	0.427	0.831	0.470	0.084	0.513	0.248	0.373	0.102	0.856	0.308
H-E2-N1	0.790	0.852	0.119	0.376	0.580	0.755	0.830	0.543	0.874	0.197	0.974	0.217
H-E2-N2	0.320	0.658	0.648	0.048	0.298	0.455	0.574	0.434	0.621	0.697	0.252	0.747
H-E2-N3	0.792	0.330	0.961	0.157	0.999	0.122	0.285	0.597	0.258	0.137	0.132	0.039
H-E3-N1	0.377	0.686	0.510	0.246	0.277	0.489	0.567	0.873	0.816	0.478	0.413	0.531
H-E3-N2	0.180	0.562	0.761	0.359	0.600	0.804	0.081	0.816	0.142	0.876	0.582	0.569
H-E3-N3	0.463	0.979	0.099	0.125	0.873	0.621	0.980	0.091	0.927	0.102	0.887	0.150
M-E1-N1	0.072	0.131	0.115	0.888	0.377	0.857	0.335	0.935	0.204	0.959	0.019	0.954
M-E1-N2	0.282	0.050	0.930	0.427	0.302	0.621	0.677	0.516	0.342	0.482	0.380	0.547
M-E1-N3	0.661	0.770	0.360	0.780	0.196	0.297	0.843	0.330	0.533	0.162	0.582	0.446
M-E2-N1	0.292	0.147	0.386	0.780	0.323	0.522	0.146	0.237	0.568	0.753	0.396	0.892
M-E2-N2	0.045	0.127	0.252	0.972	0.488	0.837	0.101	0.734	0.102	0.807	0.061	0.983
M-E2-N3	0.949	0.808	0.152	0.355	0.446	0.098	0.895	0.530	0.900	0.121	0.841	0.247
M-E3-N1	0.114	0.131	0.821	0.809	0.264	0.986	0.034	0.833	0.062	0.922	0.380	0.810
M-E3-N2	0.807	0.195	0.764	0.272	0.901	0.357	0.410	0.434	0.131	0.711	0.566	0.198
M-E3-N3	0.366	0.684	0.862	0.109	0.254	0.755	0.511	0.012	0.294	0.782	0.777	0.418

Figura 15. Probabilidad estandarizada (P). Totalidad

A partir de esta matriz de datos, llegamos al estado de la siguiente matriz concentrada (Figura 16). Esta vez, reordenamos no solamente filas horizontales, sino también columnas verticales, puesto que el orden inicial, tanto de filas como de columnas, no sabemos *a priori* si establece alguna gradación preconcebida:

C	a	ij	r	av	ar	n	p	nm	c	ir	pn	v
H-E2-N3	0.792	0.999	0.132	0.961	0.330	0.285	0.258	0.597	0.157	0.122	0.137	0.039
H-E1-N3	0.948	0.470	0.856	0.427	0.718	0.513	0.373	0.248	0.831	0.084	0.102	0.308
M-E2-N3	0.949	0.446	0.841	0.152	0.808	0.895	0.900	0.530	0.355	0.098	0.121	0.247
M-E3-N2	0.807	0.901	0.566	0.764	0.195	0.410	0.131	0.434	0.272	0.357	0.711	0.198
H-E2-N1	0.790	0.580	0.974	0.119	0.852	0.830	0.874	0.543	0.376	0.755	0.197	0.217
H-E3-N3	0.463	0.873	0.887	0.099	0.979	0.980	0.927	0.091	0.125	0.621	0.102	0.150
H-E1-N2	0.775	0.287	0.356	0.701	0.379	0.739	0.905	0.123	0.500	0.192	0.206	0.227
H-E1-N1	0.757	0.234	0.624	0.079	0.812	0.497	0.865	0.905	0.859	0.078	0.440	0.406
M-E1-N3	0.661	0.196	0.582	0.360	0.770	0.843	0.533	0.330	0.780	0.297	0.162	0.446
H-E3-N1	0.377	0.277	0.413	0.510	0.686	0.567	0.816	0.873	0.246	0.489	0.478	0.531
M-E1-N2	0.282	0.302	0.380	0.930	0.050	0.677	0.342	0.516	0.427	0.621	0.482	0.547
M-E3-N3	0.366	0.254	0.777	0.862	0.684	0.511	0.294	0.012	0.109	0.755	0.782	0.418
H-E3-N2	0.180	0.600	0.582	0.761	0.562	0.081	0.142	0.816	0.359	0.804	0.876	0.569
H-E2-N2	0.320	0.298	0.252	0.648	0.658	0.574	0.621	0.434	0.048	0.455	0.697	0.747
M-E2-N1	0.292	0.323	0.396	0.386	0.147	0.146	0.568	0.237	0.780	0.522	0.753	0.892
M-E3-N1	0.114	0.264	0.380	0.821	0.131	0.034	0.062	0.833	0.809	0.986	0.922	0.810
M-E2-N2	0.045	0.488	0.061	0.252	0.127	0.101	0.102	0.734	0.972	0.837	0.807	0.983
M-E1-N1	0.072	0.377	0.019	0.115	0.131	0.335	0.204	0.935	0.888	0.857	0.959	0.954

Figura 16. Probabilidad concentrada. Totalidad

Este resultado de la concentración diagonal es interesante, puesto que ofrece distintos modos de interpretación, concordantes unos con otros. Comenzamos por interpretar las filas de parámetros extralingüísticos:

H-E2-N3, H-E1-N3, ..., M-E1-N1.

Al dividir estos parámetros combinados en dos partes iguales, la primera parte (en negrita, en total, 9) contiene 6 hombres (H) y 3 mujeres (M), es decir, hay mayoría de hombres. En cuanto a la edad (E), observamos la siguiente distribución: E1 (4,) E2 (3) y E3 (2), lo que no muestra una característica determinada. En el nivel de educación (N): N1 (2), N2 (2) y N3 (5), se presenta la mayoría en N3. Por lo tanto, la primera parte se caracteriza mayoritariamente por H (hombre) y N3 (nivel educativo superior), que representa un estilo elevado formal, en contraste con la segunda parte correspondiente a la mujer (M) y nivel mediano o bajo de educación, que manifiestan un estilo llano coloquial. Es importante observar, así mismo, que los dos extremos de parámetros son de hombre y N3 y las cuatro extremos son de mujer y N1-N2.

Si es correcta esta interpretación, esta se aplica al orden de categorías gramaticales en torno a la siguiente secuencia:

a - ij - r - av - ar - n - p - nm - c - ir - pn – v

Según la misma interpretación, adjetivo, interjección, relativo y adverbio, entre otros, corresponden al hombre y al nivel superior de educación, mientras que verbo, pronombre, interrogativo o conjunción son propias de un estilo de mujer y nivel mediano o bajo de educación. Sin embargo, no se trata de una división tajante del estilo, sino de una tendencia de gradación continua, observada en la transición diagonal que empieza en la parte superior izquierda y termina en la parte inferior derecha.

5. CONCLUSIÓN

En este estudio, tras revisar las investigaciones anteriores sobre la ratio de categorías gramaticales, hemos intentado establecer nuestros métodos estadísticos en forma de la frecuencia ajustada (§3.1), la estandarización (§3.2) y la concentración diagonal (§3.3), antes de abordar el análisis de los datos sociolingüísticos del habla santanderina (PRESEEA-Santander). El propósito de establecer los métodos estadísticos ha sido determinar los criterios razonables para evitar decisiones o evaluaciones subjetivas. Una vez establecidos los criterios, podemos mantenerlos a lo largo del estudio sin fluctuar entre equivocaciones.

Al iniciar este estudio, partíamos de la hipótesis de trabajo según la cual las variaciones se producirían en función de los parámetros sociales. Sin embargo, hemos observado que las ratios de categorías gramaticales son casi constantes independientemente de los parámetros sociales. Pensamos que esta constancia es debida al mismo estilo coloquial, a pesar de las variaciones sociales.

No obstante, es cierto así mismo que existe cierta variación según los parámetros tratados por las distintas caracterizaciones estilísticas: estilos de hombre o mujer, de jóvenes y mayores, de individuos con un nivel educativo bajo, mediano, o alto, de distintos grados escolares, de teatro, novela o prensa. Creemos que podemos tratar estos estilos en una escala de elevación lingüística, donde intervienen distintas categorías gramaticales. Según nuestro análisis de datos, la escala de elevación empieza con el adjetivo (más elevado, formal) para llegar al verbo (más llano, coloquial) pasando por el adverbio (más cercano al adjetivo) y el nombre (más cercano al verbo) en un grado intermedio. El artículo es más elevado que el pronombre y la conjunción es más elevada que la preposición. Hemos trazado la dirección escalar de la siguiente manera:

> <más elevado>: a (adjetivo), ij (interjección), r (relativo), av (adverbio), ar (artículo), n (nombre), p (preposición), nm (numeral), c (conjunción), ir (interrogativo), pn (pronombre), v (verbo): <menos elevado>.

Naturalmente, esta gradación estilística puede variar según los datos tratados, puesto que se trata de una tendencia general, no definitiva. Sin embargo, creemos que se mantiene, *grosso modo*, la casi misma escalación, como acabamos de averiguar en distintos textos. Tampoco es recomendable establecer la escala definitiva antes de analizarla con más datos variables.

Una vez establecida la escala estándar apoyada por gran cantidad de datos, podremos analizar otros textos utilizando la misma escala, para caracterizar los textos en cuestión comparando las partes comunes y diferentes. No creemos que haya unas discrepancias enormes a la hora de rechazar esta escala definida a partir de números y datos con los métodos explicados en este trabajo. De esta manera, podremos aproximarnos a una realidad lingüística cada vez más fiable, más empírica y, en definitiva, más rigurosa.

FINANCIAMIENTO

Proyecto *ECOS-C/N, Estudio de los condicionantes sociales del español actual en el centro y norte de España: nuevas identidades, nuevos retos, nuevas soluciones*, (ref. PID2023-148371NB-C42). Ministerio de Ciencia, Innovación y Universidades

DECLARACIÓN DE AUTORÍA (ROLES CREDIT)

Autora 1: conceptualización; investigación; visualización; redacción – borrador original; redacción – revisión y edición.

Autor 2: investigación; metodología; visualización; redacción – borrador original; redacción – revisión y edición.

REFERENCIAS BIBLIOGRÁFICAS

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Brezina, V. 2018. *Statistics in Corpus Linguistics.* Cambridge University Press.

Corbalán, F. y G. Sanz. 2011. *La conquista del azar. La teoría de probabilidades*. RBA.

Coseriu, E. 1973. *Teoría del lenguaje y lingüística general. Cinco estudios.* Gredos.

Geens, D. 1978. On Measurement of Lexical Differences by Means of Frequency. *Glottometrika*, 1: 45-72.

Hatano, K. 1966. Psicología de estilo (en japonés). En *Introducción a la estilística*, pp. 22-47. Sanseido.

Juilland, A. y E. Chang-Rodríguez. 1964. *Frequency Dictionary of Spanish Words*. Mouton & Co.

Koskela, M. 1998. Verbs and Noun Phrases - Two Tendencies in Philosophical Essays. En M.-R. Lukka, S. Salla y H. Dufva (Eds.) *Puolin ja toisin. The Finnish Association for Applied Linguistics Yearbook 1998*, pp. 159-170. Publications of the Finnish Association for Applied Linguistics, No. 56. Jyväskylä.

Mair, C., M. Hundt, G. Leech y N. Smith. 2003. Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics* 7: 245-264.

Martínez Martínez, I. y H. Ueda. 2025. *Inventario léxico de PRESEEA-Santander. Proyecto LYNEAL*. En línea: https://h-ueda.sakura.ne.jp/lyneal/il/std/ [Consulta: 13/07/2025].

Moreno Fernández, F. 1993. Proyecto para el estudio sociolingüístico del español de España y América (PRESEEA). *Lingüística 5*: 268-271.

Moreno, A., G. De la Madrid, M. Alcántara, A. González, J. M. Guirao y R. De la Torre. 2005. The Spanish corpus. En E. Cresti y M. Moneglia (Eds.) *C-Oral-Rom. Integrated Reference Corpora for Spoken Romance Languages*, pp.135-161. John Benjamins Publishing Company.

R Core Team 2021. R: *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. https://www.R-project.org/ .

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

Terrádez Gurrea, M. 2001. *Frecuencias léxicas del español coloquial. Análisis cuntitativo y cualitativo*. Facultat de Filologia. Universitat València.

Ueda, H. y A. Moreno Sandoval. 1998. *Análisis de datos cuantitativos para estudios lingüísticos.* En línea: https://h-ueda.sakura.ne.jp/gengo/4-numeros/doc/numeros-es.pdf [Consulta: 23/06/2025]