ARTÍCULOS

Círculo de Lingüística Aplicada a la Comunicación
ISSN: 1576-4737

EDICIONES COMPLUTENSE

https://dx.doi.org/10.5209/clac.81206

Métodos de lexicometría sociolingüística: análisis del corpus oral contemporáneo PRESEEA-Santander

Inmaculada Martínez[1]; Hiroto Ueda[2]

Recibido: 28 de marzo de 2022 / Aceptado: 5 de mayo de 2022

Resumen. La lexicometría es un método que nos permite identificar unidades temáticas derivadas de la extracción automática de patrones de conocimiento en datos de naturaleza textual (Romero, Alarcón y García, 2018). De su aplicación emergen las tendencias léxicas de un corpus a través de la cuantificación de la ocurrencia de las palabras. Los distintos estilos léxicos sociolingüísticos se han estudiado en amplias variedades de las lenguas del mundo, incluida la lengua española. Sin embargo, no existen, en los estudios llegados a nuestro alcance hasta el momento, suficientes análisis cuantitativos del léxico de un corpus sociolingüístico oral contemporáneo. El objetivo general de este artículo es detectar las preferencias de uso del vocabulario de la lengua española hablada en el marco de la lexicometría sociolingüística. Para ello, se analizó una muestra representativa de un corpus estratificado en torno a tres variables (sexo, edad, nivel educativo). Dicha muestra pertenece al corpus PRESEEA-Santander, enmarcado en el *Proyecto para el Estudio Sociolingüístico del Español de España y América* (Moreno Fernández, 2021). En el análisis se empleó el sistema LYNEAL (*Letras y Números en Análisis Lingüísticos*) (Ueda, 2021), así como el software estadístico en código abierto R. Los resultados apuntan a que el sexo se revela como una variable importante en el proceso de variación léxica al detectarse, entre otros hallazgos, el uso del estilo nominal sobre el verbal y el empleo preferente de adverbios en *-mente* por parte del hombre; con respecto a la edad, se advierte la tendencia al empleo del truncamiento léxico en la generación de jóvenes y en el género mujer; por último, se aprecia la concentración de uso de *muchísimo* en mujer, joven, de nivel primario de instrucción.
Palabras clave: lexicometría, variación sociolingüística, corpus oral, léxico del español actual.

[en] Methods of sociolinguistic lexicometry: analysis of the contemporary oral corpus PRESEEA-Santander

Abstract. Lexicometry is a method that allows us to identify thematic units derived from the automatic extraction of knowledge patterns in data of a textual nature (Romero, Alarcón and García, 2018). From its application, the lexical tendencies of a corpus emerge through the quantification of the occurrence of words. The different sociolinguistic lexical styles have been studied in wide varieties of the world's languages, including the Spanish language. However, in the studies available to us to date, there are not enough quantitative analyzes of the lexicon of a contemporary oral sociolinguistic corpus. The general objective of this article is to detect the preferences for the use of the vocabulary of the spoken Spanish language within the framework of sociolinguistic lexicometry. To do this, a representative sample of a corpus with stratification in three variables (sex, age, educational level) was analyzed. This sample belongs to the PRESEEA-Santander corpus, framed in the *Project for the Sociolinguistic Study of Spanish in Spain and America* (Moreno Fernández, 2021). The LYNEAL system (*Letters and Numbers in Linguistic Analysis*) (Ueda, 2021) was used in the analysis, as well as the open-source statistical software R. The results indicate that gender is revealed as an important variable in the process of lexical variation, detecting, among other findings, the use of nominal over verbal style and the preferential use of adverbs in *-mente* by men; with respect to age, the tendency to use lexical truncation in the younger generation and in the female gender is noted; finally, the concentration of use of *muchísimo* in women, young people, with a primary education level, is appreciated.
Keywords: lexicometry, sociolinguistic variation, oral corpus, current Spanish lexicon.

[1] Centro Internacional de Estudios Superiores del Español (CIESE-Comillas) (España)
Correo electrónico: martinezi@fundacioncomillas.es
N.º ORCID: https://orcid.org/0000-0003-4760-0903

[2] Universidad de Tokio (Japón)
Correo electrónico: uedahiroto@jcom.home.ne.jp
N.º ORCID: https://orcid.org/0000-0003-3204-609X

adverbio *muchísimo*. 4.2.2. Adverbios en *-mente*. 4.2.3. Interjecciones. 4.3. Visión de conjunto. 5. Conclusiones. Agradecimientos. Contribución de autoría CREdIT. Referencias bibliográficas.

Cómo citar: 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://dx.doi.org/10.5209/clac.81206

1. Introducción

Parece constatable la escasez de estudios variacionistas en niveles de análisis lingüístico superiores al fonético-fonológico, especialmente en el nivel léxico (Escoriza Morera, 2012). Este hecho nos acerca, por otro lado, a una realidad en la investigación, centrada más en intuiciones e impresiones subjetivas que en datos probados científicamente y apoyados en evidencias. La lexicometría sociolingüística que planteamos actualmente viene a cubrir este vacío, al proporcionarnos las herramientas necesarias para llevar a cabo el análisis de los estilos léxicos desde la objetividad y la fiabilidad que la cuantificación garantiza. Desde este ángulo diferente abordamos en este estudio la realidad del léxico oral contemporáneo del español. De esta manera, el marco teórico establecido en nuestro estudio es la combinación de la lexicometría y la sociolingüística de corpus. Para el abordaje de la primera, nos centraremos en esta ocasión en las palabras particulares y dejaremos a un lado la riqueza léxica y el campo semántico. Los tres -palabras particulares, riqueza léxica y campo semántico- constituyen los componentes que incluye Kock (1983) en la disciplina de la lexicometría.

A continuación, se expondrán los fundamentos teóricos de la investigación para, más adelante, abordar los presupuestos metodológicos que permitieron realizar el estudio. Posteriormente, se procede al análisis de los datos, que incluye dos subsecciones: la primera aborda los resultados en torno a las variables sociolingüísticas (edad, sexo y nivel educativo) y la segunda, el léxico de alta frecuencia acotado en el uso de intensificadores, adverbios en *-mente* e interjecciones. Se incluye una visión de conjunto y se ofrecen las conclusiones obtenidas y las líneas futuras de investigación que emergen del estudio.

2. Marco teórico

2.1. La lexicometría sociolingüística

La lexicometría constituye el estudio del uso léxico partiendo de su cuantificación. La lexicometría sociolingüística aborda esta cuantificación como criterio identificador de tendencias de uso en el léxico, a partir de variables de carácter sociolingüístico, como la edad, el sexo y el nivel de instrucción. Tomar, como terreno de análisis, un corpus de lengua oral como PRESEEA implica abordar la investigación sobre la realidad de uso de la lengua española contemporánea.

Queda, por tanto, delimitada la lexicometría a partir de los métodos estadísticos que se aplican a los análisis léxicos de los textos. En lo que respecta a las herramientas propias de la estadística para dichos análisis, se ofrecen, por un lado, las básicas de recuento de frecuencia absoluta, relativa o normalizada; por otro lado, se abordan también las técnicas de análisis más avanzadas como las multivariantes (análisis de conglomerado o *cluster*, componentes principales o de correspondencia y análisis factorial, entre otras); por último, se incluyen aquellos métodos fundamentados en distintos cálculos de probabilidad (Ueda y Moreno, 2017). En el apartado 3.2. se concretarán los procedimientos analíticos y los criterios cuantitativos concretos que se han elegido para este estudio.

Con respecto a los análisis léxicos, la investigación suele centrarse, no solo en formas y lemas como objeto de estudio; también, en constituyentes de formas (prefijos, raíces o sufijos), colocaciones, tipos léxicos (de contenido y función), lemas y categorías gramaticales. Para este estudio nos centraremos en el análisis de formas.

En general, los objetivos principales de la lexicometría como disciplina abarcan, desde la caracterización léxica de textos objeto hasta la correlación entre unidades léxicas y parámetros extralingüísticos (factores sociales, geográficos, históricos, etc.), pasando por la determinación de autoría de un texto histórico o contemporáneo. En este marco de líneas de trabajo, constituyen análisis de importancia fundamental los focalizados en la observación de las palabras particulares, la riqueza léxica y el campo semántico. El primero de estos ejes de la lexicometría será abordado en este estudio; los dos restantes constituirán nuestras líneas futuras de investigación.

2.2. El corpus sociolingüístico

Los respectivos desarrollos que la sociolingüística y la lingüística de corpus han ido trazando a lo largo de los años han generado entre ambas disciplinas una especial afinidad. Ello ha llevado a sostener, por parte de ciertos autores, que la sociolingüística es lingüística de corpus, al menos con respecto a una prominente rama de esta disciplina dedicada al estudio de la lengua hablada y escrita en contexto (Romaine, 2008). En íntima reciprocidad, todo corpus sería sociolingüístico por definición, dado que cualquier producto lingüístico natural se localiza en un punto determinado del espectro geográfico y social en que una lengua se manifiesta (Moreno Fernández, 2022).

Una de las metas de este tipo de sociolingüística es la compilación de corpus de datos apropiados para el análisis cuantitativo de las variables sociales y lingüísticas, tales como la clase social, el género, el estilo o la edad, entre otros (Romaine, 2008). En la actualidad, los corpus electrónicos se usan cada vez con mayor frecuencia para conocer el alcance de la variación lingüística en torno a estos factores y los métodos estadísticos que se aplican (véanse los análisis multivariantes como los aquí realizados) se apoyan de manera clara en datos disponibles computacionalmente (Lüdeling y Kyto, 2008).

Un último dato que nos interesa abordar en estos momentos para completar este acercamiento a la caracterización de los corpus sociolingüísticos es el de la representatividad. En este sentido, la representatividad sociolingüística de un corpus parece garantizada por un adecuado tratamiento de los factores citados y de los grupos sociales que permiten la identificación de los hablantes.

Uno de los corpus con mayor relevancia para el estudio de la variación geográfica y social que reúne los requisitos arriba trazados para el establecimiento de un corpus sociolingüístico es el corpus PRESEEA (Moreno Fernández, 2022). Sistemáticamente compilado conforme a criterios dados que se abordarán a continuación, se ha construido con un fin específicamente lingüístico y los materiales en él reunidos aspiran a la representatividad, por lo que las conclusiones que se obtengan de su estudio deberían ser generalizables al todo que es la lengua española a la que representan.

3. Método

3.1. Contextualización del corpus

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 –que se detallará más adelante- 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).

La muestra 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. Entre ellos se destaca, en primer lugar, el hecho de que los núcleos de población en los que se recogen los datos deben corresponderse con núcleos de población asentada: Santander posee conciencia de comunidad de habla acreditada, en torno a la cual surge el hablante santanderino. Autores como Peña Arce (2021: 444) señalan que este español "presenta una serie de características que, sin ser privativas de este dominio, dibujan unos contornos lingüísticos que lo personalizan". En segundo lugar, 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: 13). Las variables acotadas son tres: sexo, edad y nivel de instrucción (Blas Arroyo, 2005; Cestero *et al.* 2006). 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 del cuadro 1:

	Generación 1		**Generación 2**		**Generación 3**	
	H	**M**	**H**	**M**	**H**	**M**
Nivel educativo 1	1	1	1	1	1	1
Nivel educativo 2	1	1	1	1	1	1
Nivel educativo 3	1	1	1	1	1	1

Cuadro 1. Muestra-tipo por cuotas con número mínimo de informantes (adaptación de Moreno Fernández 2021: 14)

Por último, cabe realizar una breve caracterización social y lingüística del núcleo urbano en el que se obtienen los datos. Santander, como capital de la comunidad autónoma uniprovincial de Cantabria, cuenta con 173 375 habitantes (2020) y es la urbe más poblada de la comunidad autónoma (56,1 % del total de habitantes de Cantabria). Las características de esta estructura demográfica son, principalmente: la escasa proporción de población correspondiente a grupos de edad jóvenes (0-14 años), unida al marcado peso de la población perteneciente al grupo de personas mayores de 65 años; ello provoca un índice de envejecimiento municipal levemente superior a la media regio-

nal y claramente superior a la media del país (Ayuntamiento de Santander, 2020: 221). El relevo generacional, por tanto, corre un serio peligro.

Todos estos rasgos parecen presentar una semblanza lingüística de la ciudad de Santander determinada, en primer lugar, por una fuerte conciencia de habla santanderina, así como por la homogeneidad y la ausencia de mezclas lingüísticas. Su situación geográfica en el punto central de la provincia y mirando al mar ha provocado el hecho de que no se produzcan trasvases lingüísticos significativos desde las comunidades limítrofes de Asturias y País Vasco (Martínez y Ueda, 2021).

3.2. Procedimientos analíticos

La investigación aborda el estudio del léxico empleado por 18 informantes que viven en la ciudad de Santander, estratificados en torno a las variables ya mencionadas de sexo, edad y nivel educativo. El estudio ahonda en los patrones de comportamiento del léxico en el geolecto de esta ciudad española, en tres sociolectos (alto, medio y bajo) de una muestra oral extraída de corpus de población preestratificada. Se aspira a abordar, con métodos sociolingüísticos (McEnery y Hardie, 2012; Moreno Fernández, 2021) y lexicométricos (Koch, 1983; Higuchi, 2017), los rasgos léxicos destacados en las conversaciones transcritas y semidirigidas entre el informante y el encuestador. La metodología empleada es cuantitativa, con aplicaciones estadísticas de probabilidad y análisis multivariantes.

El procesamiento de los datos se ha llevado a cabo gracias al sistema LYNEAL (*Letras y Números en Análisis Lingüísticos*), que permite "buscar e identificar formas, efectuar análisis cuantitativos, descriptivos y multivariantes, y obtener resultados en cifras -absolutas, relativas y normalizadas-, en gráficos y en mapas" (Ueda, 2021: 5). Previamente al análisis del corpus, se ha llevado a cabo un proceso de lematización automático que ha sido revisado por un experto nativo con el fin de depurar incorrecciones (Martínez y Ueda, 2021).

En torno a las variables se distribuyen los lemas de manera estructural y estratificada, con lo que se evita, por tanto, el riesgo de una distribución sesgada. Los datos fueron obtenidos en entrevista oral semidirigida con una duración que oscila entre los 40 y los 50 minutos. Presentan un cómputo variado de palabras que aparece detallado en la tabla 1 y que debe tenerse en cuenta a la hora de calcular la frecuencia normalizada (por cada 100 000 palabras) que presentamos en este estudio.

Sexo:	**Hombre**			**Mujer**			**Total**
Edad	**E1**	**E2**	**E3**	**E1**	**E2**	**E3**	
Nivel-1	6380	6395	8160	6347	6971	5141	39394
Nivel-2	4250	8319	7633	3337	10278	5459	39276
Nivel-3	7699	5437	5888	4343	3710	9352	36429
Total	18329	20151	21681	14027	20959	19952	115099

Tabla 1. Constitución del corpus en número de palabras

Los 115 099 lemas obtenidos en total, clasificados en torno a las variables mencionadas y la frecuencia de aparición, quedan distribuidos tal y como aparece en la tabla 2:

Lema_Cat.	**Sexo**	**Edad**	**Nivel**	**Frec.**
abajo_adv.	H	E1	N2	2
abajo_adv.	H	E2	N2	3
abajo_adv.	H	E3	N1	2
abajo_adv.	H	E3	N2	1
abajo_adv.	M	E1	N3	1
abajo_adv.	M	E2	N1	3
abajo_adv.	M	E2	N2	3
abajo_adv.	M	E3	N2	1
abastecer_vb.	H	E3	N1	1
abastecer_vb.	H	E3	N2	1
…	…	…	…	…

Tabla 2. Distribución de los lemas y su frecuencia de aparición

A partir de estos datos, se indaga en el léxico que caracteriza a los distintos sociolectos que integran el corpus. Para destacar los lemas que caractericen cada categoría de sexo, edad y nivel educativo, se han establecido cuatro criterios cuantitativos:

(1) Frecuencia relativa. Para incluir los lemas preferidos de una determinada categoría -por ejemplo, del hombre con respecto a la mujer-, hemos determinado la frecuencia relativa mayor de 0.7. El mismo criterio lo hemos aplicado a la mujer con respecto al hombre. De la misma forma se ha aplicado en otros factores, como la edad y el nivel educativo. Para representar la Edad-1 (20-34 años), el lema debe abarcar más de 0.7 en su frecuencia relativa dentro de las tres categorías (E1, E2, E3) del factor Edad.
(2) Suma. Para evitar la posible accidentalidad de la alta frecuencia relativa, causada por la baja frecuencia absoluta, hemos establecido un umbral mayor de 9. De esta manera, excluimos los lemas de reducida frecuencia absoluta, desde 1 hasta 9.
(3) Número en persona. Una cierta frecuencia relativamente alta de un vocablo puede corresponder solo a unas pocas personas, lo que no representaría la categoría en cuestión. Sobre la base de la observación de los datos, hemos fijado el número mínimo necesario para tratar los lemas en cuestión a partir de 5, cifra que surge como resultado de aplicar distintos controles. De este modo, se excluyen los vocablos que utilizan solo una, dos, tres y cuatro personas.
(4) Probabilidad binomial. Nos interesa la significatividad estadística de la frecuencia relativa con respecto a la ratio en su totalidad. Para conocerla, recurrimos al método estadístico de la prueba binomial a partir del cual hemos elaborado la función de probabilidad binomial, Binom:

$$p = Binom(x, s, r) = IF(x = 0, 1, 1 - BINOMDIST(x-1, s, r, 1))$$

donde x = frecuencia, s = suma, r = número total de categoría / totalidad, BINOMDIST = función Excel. El criterio para determinar la significatividad, lo hemos establecido en $p < 0.01$.

4. Resultados y discusión

Procedemos, a continuación, al análisis de los datos, el cual dividiremos en tres apartados. En primer lugar, se abordará el estudio cuantitativo del vocabulario en el corpus PRESEEA-Santander en función del sexo, edad y nivel de instrucción. En segundo lugar, se procederá al análisis del léxico que presenta una alta frecuencia de aparición (mayor de 10) y que se corresponde con las categorías gramaticales de adjetivo, adverbio e interjección. Por último, se ofrecerá una visión de conjunto en distribución diagonalizada y esparcimiento biplot.

4.1. Variables sociolingüísticas

4.1.1. Sexo

Se advierte un marcado estilo nominal en el sexo. Del total de 51 palabras seleccionadas con una frecuencia alta de aparición, 28 son sustantivos y solo 6 adjetivos; es decir, más de la mitad de las palabras en este sexo se corresponden con nombres (*plaza*, *regalo*, *salsa*, *diferencia*, *perdón*, *edificio*, *vivienda*, entre otras) y menos de un cuarto, con adjetivos (*absoluto*, *exacto* o *curioso*). Además, de entre las diez primeras palabras, 7 son los sustantivos arriba mencionados.

En el caso del sexo Mujer, aunque el estilo también es mayoritariamente nominal, la frecuencia de uso de sustantivos es menor: son 12 los sustantivos encontrados entre las 28 palabras más frecuentes usadas por este sexo: no se llega a completar la mitad del total de formas mayoritariamente utilizadas, de entre las cuales destacan los 7 adjetivos empleados (*precioso*, *interesante*, *social*, *diferente*, *raro*, *bastante*, *distinto*). En ambos sexos, el estilo nominal destaca sobre el verbal, pues tan solo se aprecia el uso de un verbo, entre las diez palabras más frecuentes, en Hombre (*existir*) y el uso de un verbo en Mujer (*durar*).

Por otro lado, destaca en el sexo Hombre el uso del adverbio en *-mente*, hasta en tres ocasiones (*exactamente*, *normalmente* y *simplemente*) y un uso bastante reducido del adverbio en grado superlativo. Este último hecho contrasta con una alta frecuencia en el sexo Mujer (*muchísimo*), tal y como trataremos más adelante, en la segunda parte de este epígrafe; se confirma, por tanto, el "empleo femenino de los intensivos", que menciona García Mouton (1999: 73) con ejemplos de *monísimo, muy muy mono, tan simpático* y los prefijos *super-, híper-*, utilizados más por las mujeres que por los hombres. El mismo uso se observa en distintas lenguas de Europa, como es el caso del inglés, danés, francés, alemán y ruso (Jerpersen, 1922: 250).

Según Lastra (1992: 305), "las mujeres producen más formas [fonéticas] estándar" y "prefieren las formas de prestigio, ya sea porque tienen valor en la movilidad social o porque evitan las formas estigmatizadas". Continúa señalando la autora que hay grupos donde las interjecciones varían según el sexo y añade que "se han hecho estudios que indican que en Suecia, Estados Unidos y Brasil, los hombres usan más malas palabras que las mujeres".

Este hecho ha podido constatarse en nuestro corpus a través de la variación con respecto al sexo que se manifiesta en torno a la preferencia de uso de las interjecciones en español. En el caso del Hombre, la interjección *joder* está situada entre las 23 primeras palabras más frecuentes. Destacan, asimismo, los dos primeros lugares que ocupan

jolín y *jo* dentro de las palabras más frecuentes empleadas por las mujeres; con la misma funcionalidad, el sexo Hombre prefiere la utilización de *joder*, que además aparece bastante más atrás, en el lugar 23 entre las palabras más usadas. La diferencia de uso entre *joder*, preferido por los hombres, y *jolín* y *jo*, por las mujeres, puede deberse al hecho de que "la mujer autocorrige su forma de hablar, evitando lo que está mal considerado" (García Mouton, 2003: 110-118). En contraste, "la exclusividad de las palabras groseras" se atribuye tradicionalmente al lenguaje masculino (García Mouton, 1999: 52-53).

La frecuencia relativa de estos datos puede observarse en las figuras 1 y 2 que se presentan a continuación:

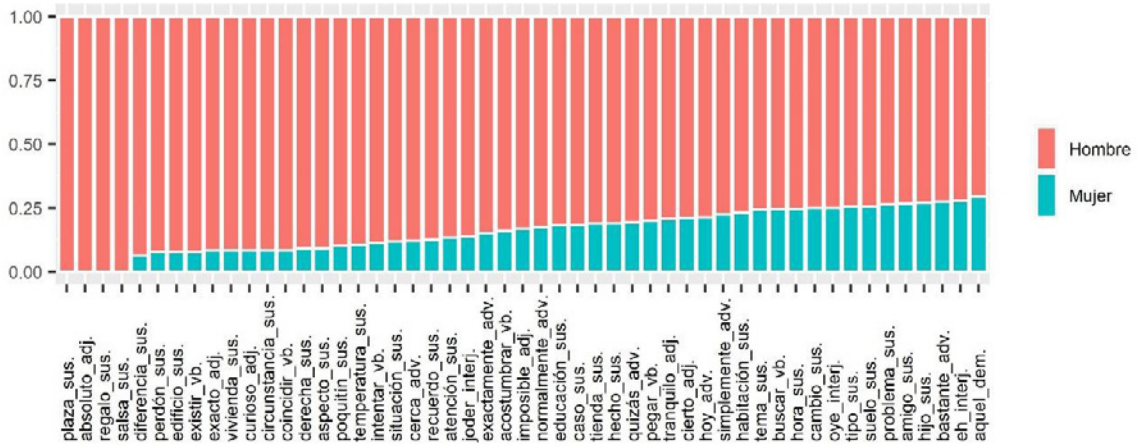

Fig. 1. Palabras preferidas en Hombre

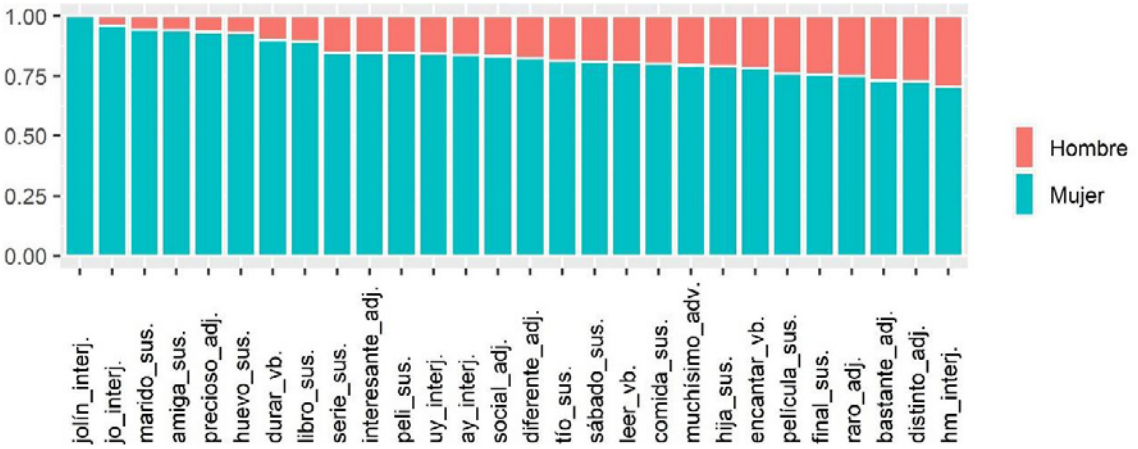

Fig. 2. Palabras preferidas en Mujer

4.1.2. Edad

El análisis basado en la división etaria del corpus PRESEEA-Santander muestra, en primer lugar, que las palabras preferidas en las tres horquillas de edad son sustantivos y que en los tres casos pueden reflejar los intereses particulares de cada segmento de población. En concreto, en la primera generación de jóvenes (20-34 años), la arquitectura (*edificio*); en la segunda de adultos (35-54 años), la alimentación (*carne*); en la tercera de mayores (más de 55 años), la vida rural (*campo*).

En segundo lugar, fruto de este análisis lexicométrico, parecen aflorar de manera nítida, tanto en la primera como en la tercera generación, los campos semánticos de las preocupaciones que atañen a cada franja de edad. Así, en la correspondiente a los jóvenes se observan el uso frecuente de vocablos como *leer*, *libro*, *peli*, *película*, *carrera,* mientras los mayores muestran la lógica preferencia por palabras como *morir*, *nacer*, *nieto*, *jubilar*.

En tercer lugar, en el apartado de los adverbios y modificadores, parece sintomático el uso de *bastante* entre los jóvenes y *poquitín* entre los adultos, como claramente caracterizadores de ambas edades; la generación 3 se caracteriza, por su parte, por el uso frecuente de adverbios en *-mente*: *efectivamente*, *perfectamente*.

Por último, destaca la apócope o acortamiento *peli* en la primera generación, uso que parece constituir un icono léxico característico de esta edad; en contrapartida, no existe ningún acortamiento en el resto de las generaciones que ocupe los usos más frecuentes. Este mecanismo de truncamiento léxico aparece, además, con una alta frecuencia en el sexo mujer, tal y como se muestra en la figura 2 del epígrafe anterior. Es detectado ya por López García y Morant (1991) y se manifiesta en otros usos como *gordi, chuli, pelu, ilu, cari*.

Estos datos aparecen listados en las figuras 3, 4 y 5 que se muestran a continuación:

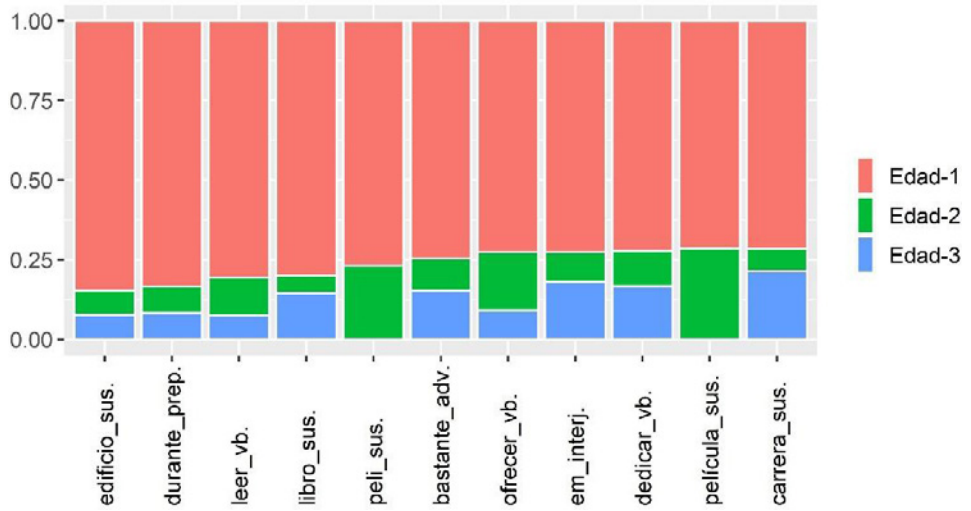

Fig. 3. Palabras preferidas de la generación 1 (Edad-1)

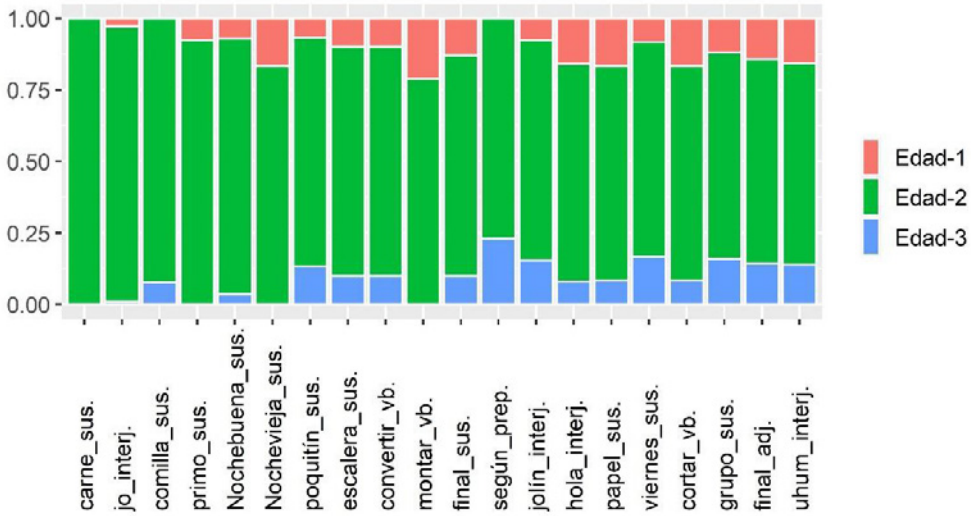

Fig. 4. Palabras preferidas de la generación 2 (Edad-2)

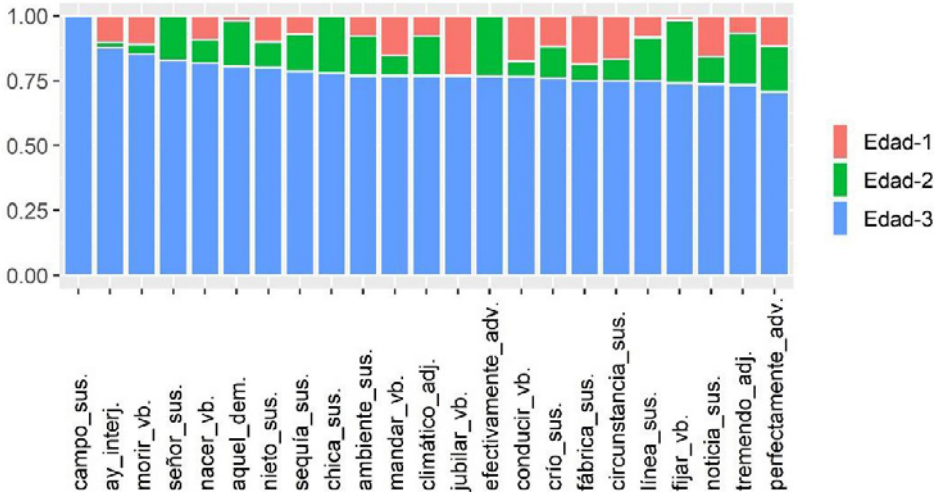

Fig. 5. Palabras preferidas de la generación 3 (Edad-3)

4.1.3. Nivel de instrucción

Al igual que en los parámetros anteriores, realizaremos aquí el análisis lexicométrico de los tres niveles de instrucción.

Advertimos en primer lugar, un cambio sustancial con respecto a la relación entre el estilo nominal y el verbal a lo largo de los tres niveles educativos. En el nivel bajo, se aprecia el desequilibrio entre ambos estilos, a favor del

estilo nominal: son 9 los nombres que aparecen entre las formas más frecuentes (*paso*, *crío*, *aspecto*, *pregunta*, *empresa*, *costumbre*, *cuestión*, *fábrica*, *circunstancia*), frente a una sola interjección (*joder*). En el nivel medio se mantiene la preferencia por el estilo nominal, pero aquí la distancia se acorta, pues son 7 los sustantivos empleados (*arroz*, *alegría*, *idea*, *tren*, *pasta*, *chico*, *sábado*) y 3, los verbos (*recoger*, *casar*, *aparcar*). En el nivel superior, es prácticamente idéntico el uso de sustantivos y verbos: 9 son los verbos empleados (*dedicar*, *procurar*, *utilizar*, *apetecer*, *evitar*, *coincidir*, *aportar*, *preguntar*, *pasear*) y 10, los sustantivos (*realidad*, *contacto*, *relación*, *iglesia*, *salsa*, *nota*, *carrera*, *línea*, *nivel*, *señor*). Todo ello parece mostrarnos que el empleo adecuado de la lengua -con un equilibrio entre los estilos nominal y verbal- parece estar relacionado con un mayor nivel de instrucción.

En segundo lugar, el empleo de formas pertenecientes al registro coloquial, como *joder* o *crío* aparece tan solo en el nivel de instrucción primario, mientras que este hecho no se advierte en el resto de niveles educativos. En el caso de emplear alguna interjección, el segundo nivel de instrucción prefiere las interjecciones atenuantes del tipo *jo*, *jolín* (frente a *joder*), mientras que no encontramos entre las palabras más frecuentes del nivel superior ninguna interjección.

Sorprende, asimismo, que en el primer nivel de instrucción se adviertan palabras como *aspecto* o *cuestión*, cuando en estos contextos quizá cabría esperar formas comodín como *tema*. En este mismo sentido, el vocablo *idea*, que aparece en quinto lugar de frecuencia en el nivel de instrucción 2, podría desvelar el uso comodín del sustantivo. O, incluso, el adjetivo *interesante* en el nivel 3, que también puede mostrar un uso general en el que se ha incurrido por el abuso de dicho término y que podría haberse sustituido por el empleo de otra forma más cargada léxicamente.

Por último, con respecto a la longitud de las palabras, que marca la mayor o menor legibilidad del discurso, todo parece apuntar a que en el nivel correspondiente a los estudios superiores el número de sílabas por palabra es mayor. En este nivel educativo, en comparación con el resto, parece que dominan las palabras de tres sílabas, hecho que no ocurre en los niveles 1 y 2. En este sentido, se atisba que el nivel 3 tiende a alargar las palabras en la falsa creencia de resultar más culta su habla, muy probablemente influido este hecho por el tipo de educación recibida.

Se muestran, a continuación, estos datos representados en las figuras 6, 7 y 8:

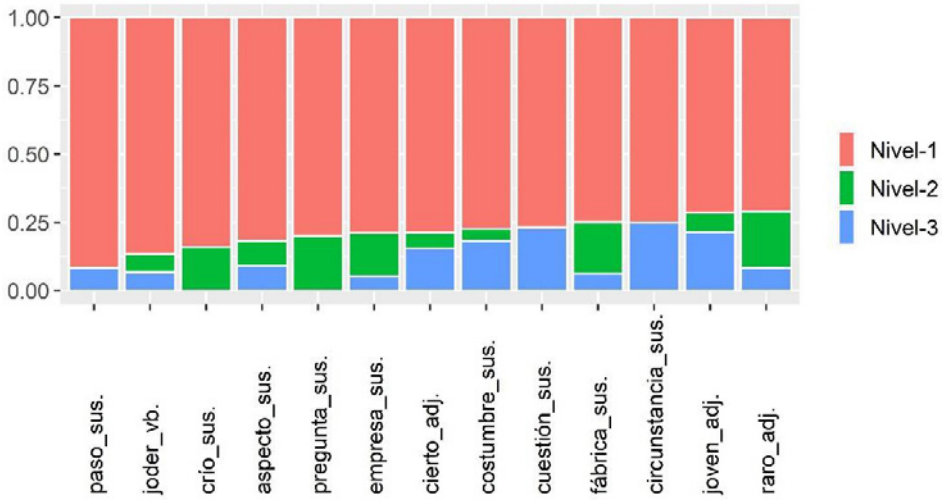

Fig. 6. Palabras preferidas del Nivel 1

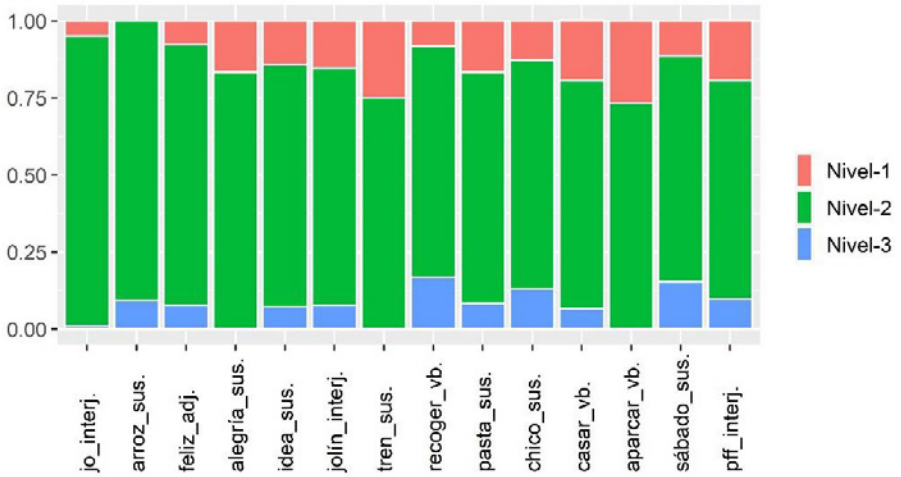

Fig. 7. Palabras preferidas del Nivel 2

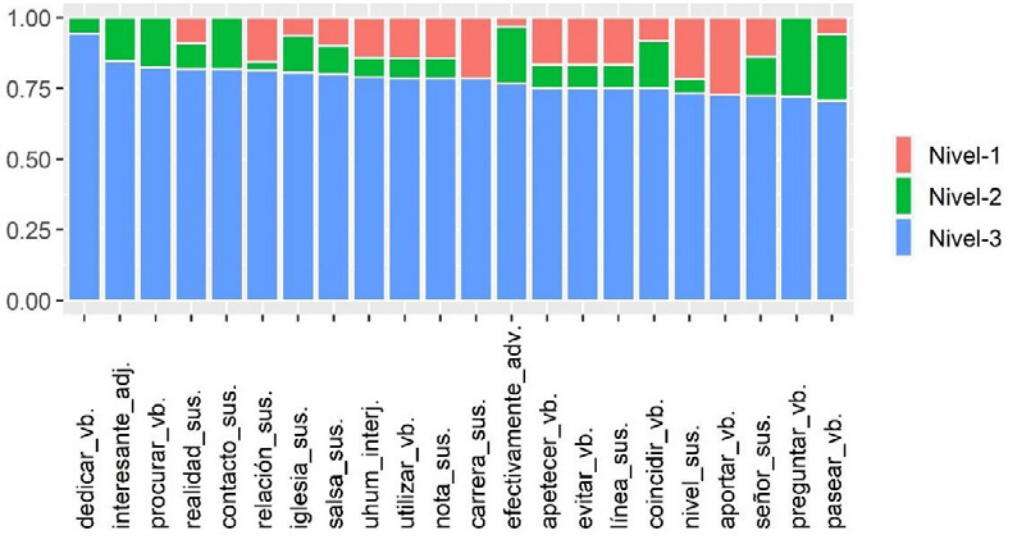

Fig. 8. Palabras preferidas del Nivel 3

4.2. Léxico de alta frecuencia

Llegados a este punto, abordamos ahora el léxico de alta frecuencia (mayor de 10) determinado por factores sociolingüísticos. Nos centraremos, para ello, en el análisis de tres categorías gramaticales: en concreto, adjetivo, adverbio e interjección, por presentar peculiares rasgos sociolingüísticos.

4.2.1. Adjetivo y adverbio muchísimo

En la sección 4.1.1. hemos señalado la preferencia femenina de *muchísimo* (como adjetivo y adverbio), cuyo detalle se observa en la figura 2 y en la tabla 3. Jespersern (1922), Lakoff (1977) y García Mouton (1999) coinciden en destacar el uso femenino de intensificadores tanto en inglés (*so, such, divine, gorgeous*) como en español (*monísimo, muy, muy mono*). El análisis de frecuencia que aquí se muestra corrobora empíricamente dicha teoría.

En cuanto a la variable Edad, se destaca el mayor uso de esta forma en los adultos (generación 2) y su menor preferencia en el nivel educativo superior, tal y como se advierte en la tablas 4 y 5; así, la línea representativa del nivel educativo muestra en la figura 9 una clara línea descendente en su uso, desde el nivel inferior (N1) hasta el superior (N3), pasando por el nivel medio (N2).

	Frec. absol.		**Suma Frec.absol.**	**Frec. normaliz.**	
Sexo	**H**	**M**		**H**	**M**
muchísimo	19	76	95	0.316	1.383
Palabras	60161	54938	115099	1000.000	1000.000

Tabla 3. Adjetivo y adverbio *muchísimo* (Sexo)

Edad	**E1.f.a.**	**E2.f.a.**	**E3.f.a.**	**Suma.f.a.**	**E1.f.n.**	**E2.f.n.**	**E3.f.n.**
muchísimo	20	52	23	95	0.618	1.265	0.552
Palabras	32356	41110	41633	115099	1000.000	1000.000	1000.000

Tabla 4. Adjetivo y adverbio *muchísimo* (Edad)

Nivel	**N1.f.a.**	**N2.f.a.**	**N3.f.a.**	**Suma.f.a.**	**N1.f.n.**	**N2.f.n.**	**N3.f.n.**
muchísimo	41	32	22	95	1.041	0.815	0.604
Palabras	39394	39276	36429	115099	1000.000	1000.000	1000.000

Tabla 5. Adjetivo y adverbio *muchísimo* (Nivel educativo)

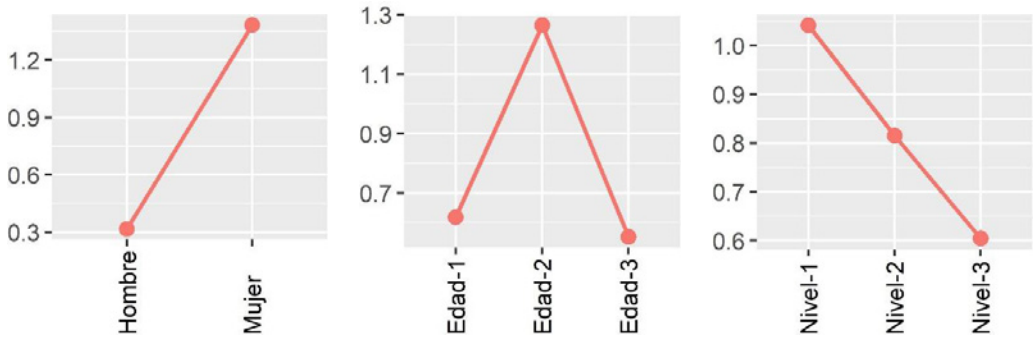

Fig. 9. Adjetivo y adverbio *muchísimo* (Sexo, Edad, Nivel educativo)

Al observar estas distribuciones de frecuencia, advertimos una cierta concentración de uso en los parámetros Mujer-Generación 2-Nivel de instrucción 1. Podría pensarse que representa la frecuencia destacada de una única informante: si fuera así, se trataría de una particularidad individual (e incluso puntual, localizada en el momento en que lleva a cabo su encuesta), y no de una marcada preferencia de uso de *muchísimo* por Mujer en general. Para confirmar o refutar esta hipótesis, hemos averiguado el modo de concentración numéricamente sesgada de este vocablo y hemos realizado el recuento en los siguientes idiolectos con los resultados que se muestran a continuación en la tabla 6:

Idiolecto	Frecuencia
M-E2-N1	**28**
M-E2-N2	13
M-E3-N3	7
M-E3-N2	7
M-E2-N3	7
H-E3-N2	7
M-E1-N1	6
M-E1-N3	6
H-E1-N1	3
H-E3-N1	2
H-E2-N1	2
M-E1-N2	2
H-E1-N2	2
H-E1-N3	1
H-E2-N3	1
H-E2-N2	1

Tabla 6. Frecuencia del vocablo *muchísimo* en idiolectos

Se puede apreciar que destaca el mayor uso en M-E2-N1 (28) dentro de los 95 usos en total. Sin embargo, el porcentaje de concentración (28 / 95 = 29.5%) no llega a la mitad, ni al tercio, lo que garantiza la caracterización general, y no necesariamente individual, de su uso femenino. De la misma manera, los siete casos de relativa mayor frecuencia apoyan nuestra hipótesis.

Al observar la frecuencia de otras formas del superlativo absoluto en la tabla 7, la concentración femenina es, igualmente, absoluta:

Superlativo	Hombre	Mujer
baratísimo	0	1
buenísimo	0	2
clarísimo	0	5

Superlativo	Hombre	Mujer
convencidísimo	0	1
crudísimo	1	0
enganchadísimo	0	1
facilísimo	0	1
grandísimo	0	1
importantísimo	0	1
limpísimo	0	1
malísimo	0	2
queridísimo	1	0
segurísimo	0	1
supuestísimo	0	1
tantísima	0	1
tontísimo	0	2
tremendísimo	0	1

Tabla 7. Frecuencia de uso de superlativos absolutos

Estos datos parecen apuntar a una cierta tendencia intensificadora de la expresión femenina a través de este morfema, sobre todo en el estilo sociolingüístico de las mujeres adultas de nivel educativo bajo (M-E2-N1). Este hecho necesitaría ser confirmado, obviamente, con análisis futuros que tengan como base una muestra mayor. Se exponen, a continuación, algunos ejemplos:

(1) […] es que aquí aparece *muchísima* más gente que aparece en el libro […] (M, E1, N1).
(2) […] se hizo *muchísima* propaganda / se gastaron *muchísimo* dinero en hacer publicidad // y quedó en nada […] (M, E1, N3).
(3) […] y entonces valoramos menos lo que es la amistad, entonces *muchísimas* veces es mi amiga es mi amiga […] (M, E2, N1).

4.2.2. Adverbios en -mente

En esta sección, analizaremos las frecuencias normalizadas de los adverbios en *-mente*, que hemos mencionado con anterioridad (cf. 4.1.2.) e indicaremos la preeminencia numérica en Hombre y en Edad-3. Para argumentarlo, se muestran en las figuras 10-12 los siguientes gráficos de proporción:

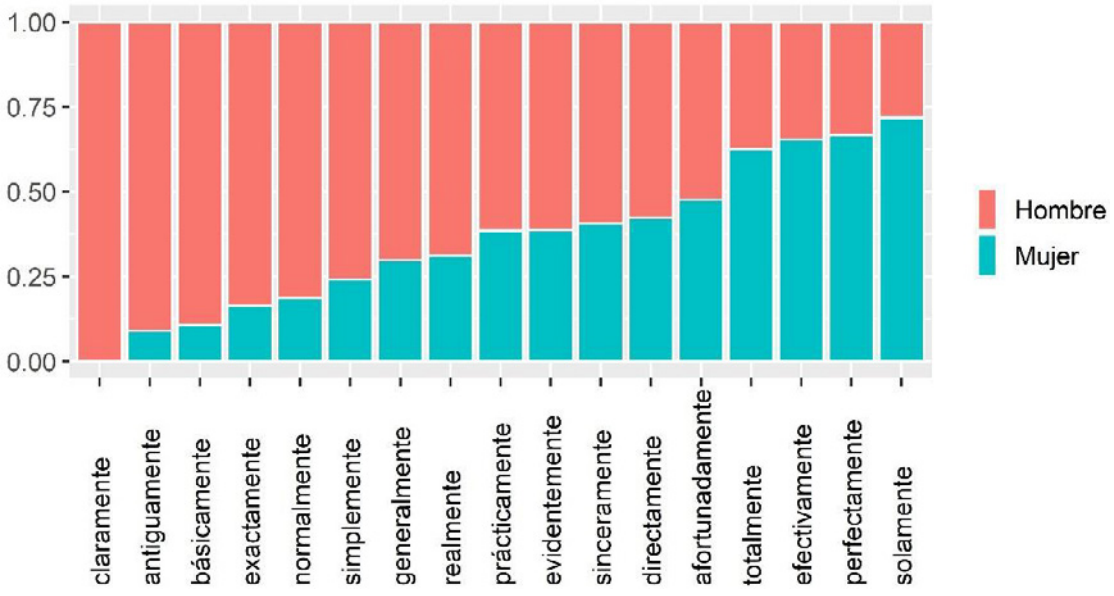

Fig. 10. Adverbios en *-mente*. Sexo (Hombre, Mujer).

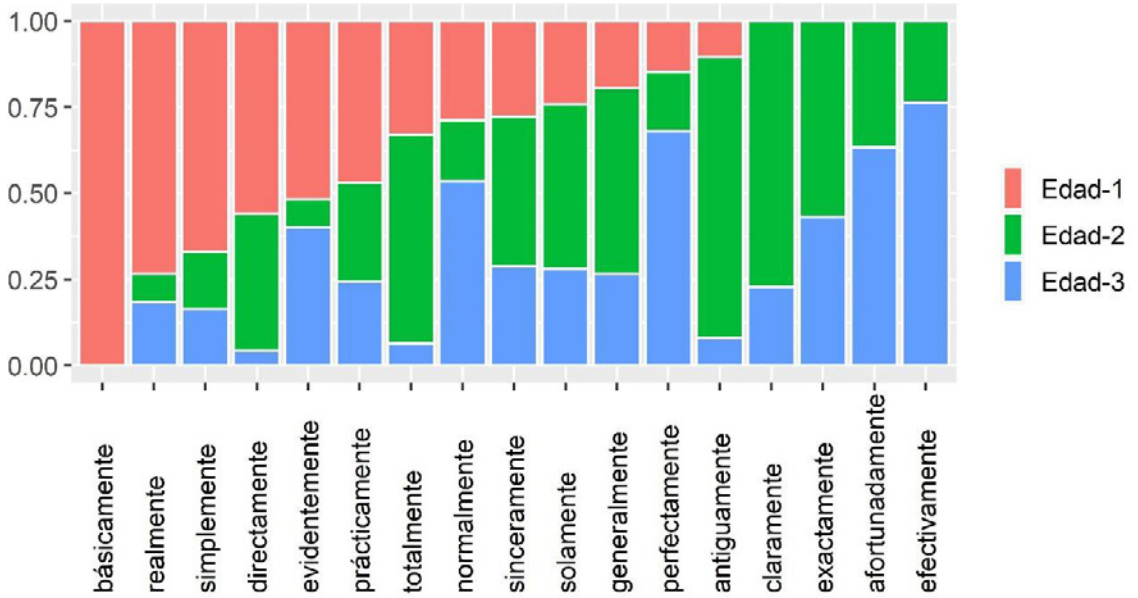

Fig. 11. Adverbios en *-mente*. Edad

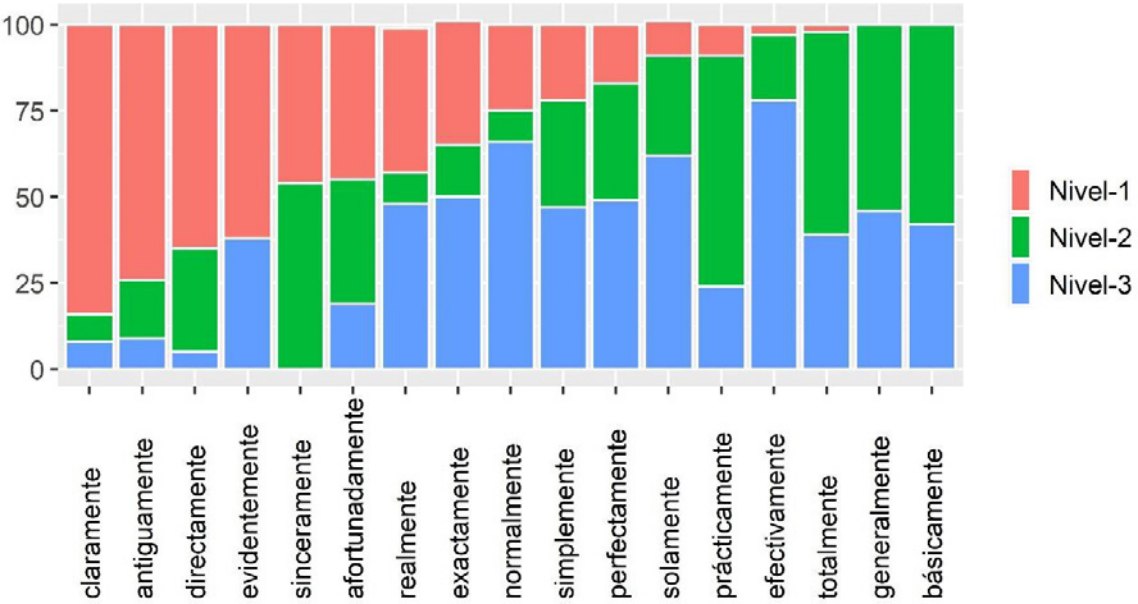

Fig. 12. Adverbios en *-mente*. Nivel educativo.

A la luz de estos datos, se advierte la predominancia general de estos adverbios en Hombre con respecto a Mujer. Esta tendencia se observa en la figura 10, donde se aprecia la diferencia notable de frecuencia normalizada (Hombre. 6.782; Mujer. 3.841). En cuanto a la Edad, la diferencia etaria no es notable en la figura 11 (E1. 5.409 / E2. 5.206 / E3. 5.524), pero sí en el nivel educativo, donde el uso de este tipo de adverbios es mayoritario en el nivel superior, como aparece reflejado en la figura 12 (N1. 4.138 / N2. 4.685 / N3, 7.467), con algunas excepciones según las palabras.

A nuestro modo de ver, los hombres utilizan los adverbios en *-mente* más que las mujeres en general, a excepción de los siguientes vocablos: *totalmente*, *efectivamente*, *perfectamente*, *solamente*, donde el uso por las mujeres es mayor. La causa de la diferencia podemos suponerla en una cierta connotación intensificadora del grado, que suele caracterizar el habla femenina, ejemplificada por García Mouton (2003: 84) con *divinamente*, *enormemente*, *fatalmente*, *fenomenalmente*, tal y como se observa en los siguientes ejemplos de nuestro corpus santanderino:

(4) […] nos conformábamos con cosas más simples $ sí / *totalmente* / *totalmente* / *totalmente* $ era más fácil // (M, E2, N2).
(5) […] ¡claro! tenemos caracteres también *totalmente* diferentes // mi hermana es una niña ¡jo! supereducadita es decir yo qué sé es diferente a mí (M, E2, N2).
(6) […] pero *efectivamente* yo que soy una vaga y que no voy al gimnasio […] (M, E2, N2)
(7) […] no se lo vas a decir a tu amiga que sabes *perfectamente* que va a estar detrás de ti diciéndote tal tal tal […] (M, E2, N1).

4.2.3. Interjecciones

Con respecto a la categoría gramatical de la interjección, se advierte una diferencia numérica entre sexos. El contraste mayor se encuentra entre *joder* en el habla masculina, por una parte, y *jolín* y *jo* en la femenina por otra, como se señaló líneas más arriba. En menor grado se advierte, a través de la figura 13, la preferencia de *oye*, *eh*, *hola*, *uf* en la masculina y *uy*, *ay*, *vamos*, *madre mía*, en la femenina:

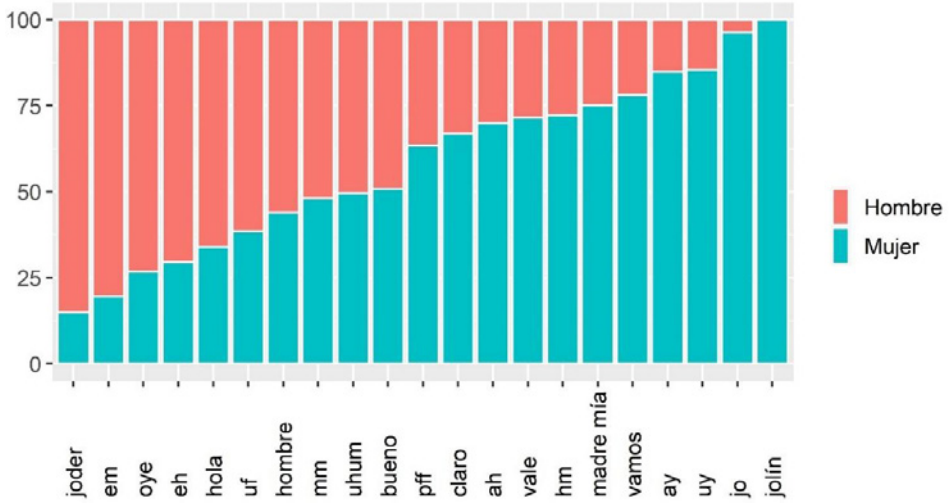

Fig. 13. Uso de las interjecciones en función del Sexo

Ejemplos:

(8) […] y dices tú *joder* parece que he vendido lo de toda la semana […] (H, E1, N1).
(9) […] mmm volvemos a lo mismo / porque me vas a decir / *joder* / tu hijo va a acabar de ti […] (H, E2, N2)
(10) […] de Obermaier es que eran los bocetos de las cuevas de Altamira // o sea hoy lo miras desde esa tal y dices / desde esa perspectiva y dices ¡joder! ¡si es que teníamos aquí una joya! […] (H, E3, N1)
(11) […] claro y // y la la lo que la gente la la chocó mucho digo yo / ¡jo! tampoco era para tanto dice ¿tú qué has leído? y yo todo lo que pasaba por mis manos (M, E1, N1).
(12) […] mi plato / favorito / bueno / a ver / es que *jolín* / parece que son preguntas que no nos las hacemos ¿eh (M, E2, N1)
(13) […] y creo que he sido feliz de la vida / jugando de pequeña encima que *jolín* yo era muy imaginativa […] (M, E2, N2)

En cuanto a la edad, se encuentran formas preferidas para cada generación (E1, E2, E3), que enumeramos en orden descendente con porcentaje mayor de 50%: E1: *em* (77.4%), *vamos* (70.1%); E2: *jo* (95.3%), *jolín* (75.5%), *hola* (72.9%), uf (67.7%), *uhum* (67.4%), *vale* (55.9%), *oye* (53.8%), *madre mía* (53.6%); E3: *ay* (85.2%), *ah* (57.7%), *claro* (52.2%), *uy* (50.9%). El mayor detalle se aprecia en la figura 14:

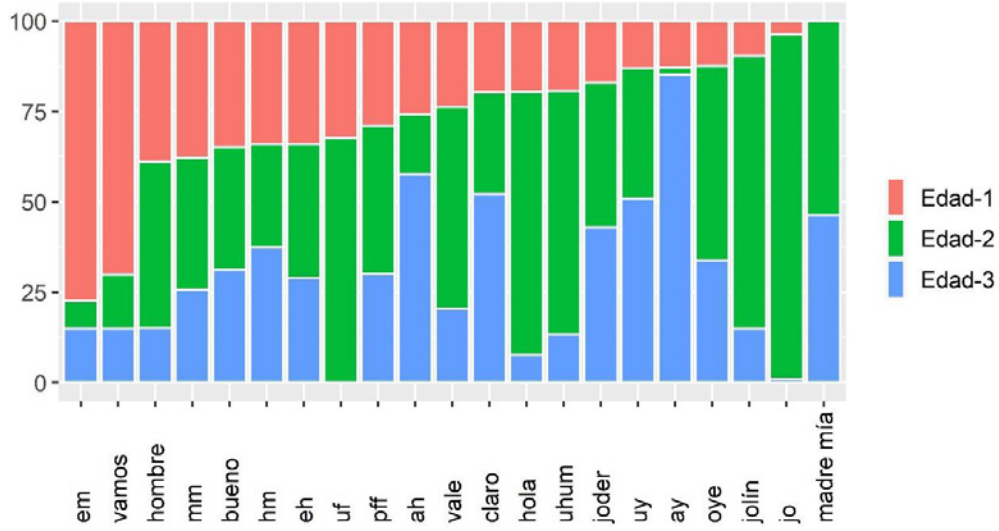

Fig. 14. Uso de las interjecciones en función de la Edad

Por último, en el nivel educativo, no se encuentran formas preferidas con porcentaje mayor de 50% en el nivel más bajo de instrucción, como observamos en la figura 15. En el nivel medio se han detectado las siguientes formas: *jo* (94.1%), *jolín* (76.5%), *pff* (70.5%), *vale* (66.6%), *ay* (64.4%), *vamos* (63.6%), *uf* (62.3%), *madre mía* (58.5%), *hola* (55.5%), *joder* (55.1%), *ah* (52.6%), *uy* (52.5%); en el nivel superior se advierten las interjecciones *uhum* (80.2%) y *em* (56.4%).

Ejemplos:

(14) […] yo creo que para todo hay un tratamiento ahora de usted y de tú depende de las circunstancias y la situación / pero *vamos* en un momento que estás entablando una conversación […] (H, E1, N1)

(15) […] él quería ir a Italia // porque es un enamorado de Florencia // pero *claro* //con dos mil euros que les pagan […] (H, E3, N1)

(16) […] y era / *pff* / a ver / era la habitación de mis padres / el salón era muy grande […] (H, E2, N2)

(17) […] ahora / estoy aquí he estado *uhum* en algunas ciudades trabajando […] (H, E1, N3)

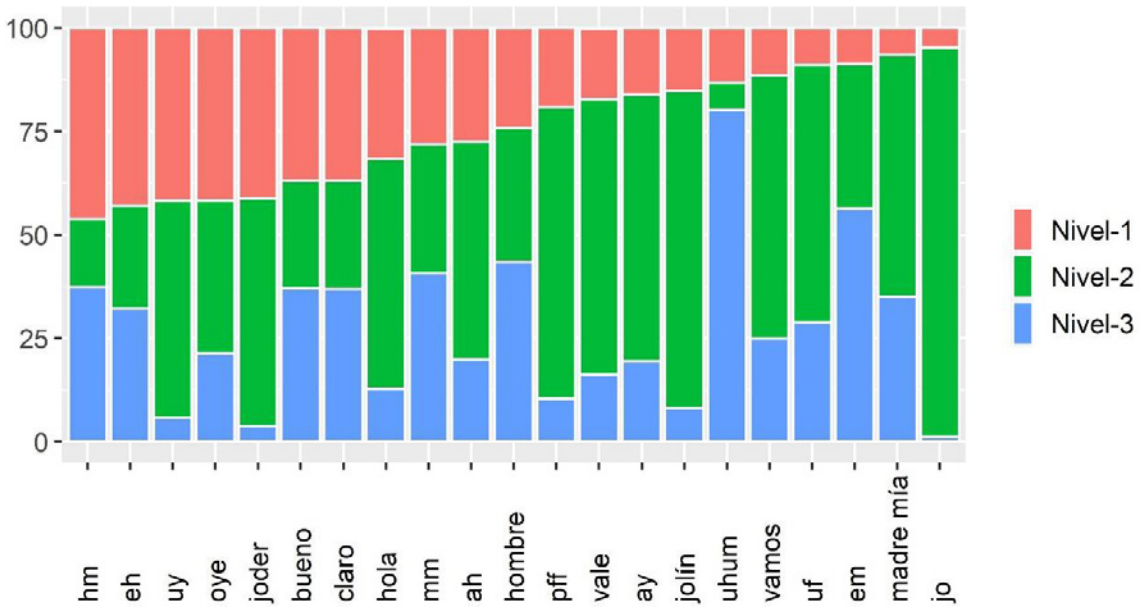

Fig. 15. Uso de las interjecciones en función del Nivel educativo

4.3. Visión de conjunto

Hasta este punto hemos venido analizando los usos léxicos dentro de cada variable sociolingüística: sexo, edad, nivel educativo. Conviene, asimismo, abordar la totalidad de la distribución sesgada en el mismo plano bidimensional constituido por el léxico y las variables mencionadas. Para ello, se presenta a continuación la tabla de frecuencia relativa horizontal (porcentaje) de los vocablos relevantes más frecuentes, con el fin de observar la preferencia relativa de cada vocablo. Utilizamos la plataforma R, con los paquetes ggplot y MASS. El resultado se muestra en la figura 16:

Esta figura muestra, gracias a la distribución diagonalizada, la concentración de altas frecuencias en la zona diagonal, desde la parte superior izquierda hasta la inferior derecha, con las clasificaciones en dos grupos de las variables y en cuatro grupos del léxico, obtenidas por el análisis de conglomerado (*cluster*, con distancia euclidiana y agrupamiento completo). A partir de este gráfico, podemos observar, tanto la relación existente entre variables y léxico por separado como la existente entre variables y léxico al mismo tiempo.

De este modo, confirmamos los factores ordenados desde H (Hombre) hasta M (Mujer) pasando por N1, N3, E1, E3, E2, cuyo orden es significativo, puesto que refleja precisamente la distribución diagonal. Lo mismo puede decirse de las palabras, desde *claramente, realmente, normalmente*, …, hasta …, *madre mía, ay, jolín, jo*, cuyo ordenamiento también está garantizado por la misma distribución sesgada diagonal de frecuencias. Reiteramos que se destacan los dos extremos de la variable Sexo: Mujer (M) y Hombre (H). En cambio, los rasgos de Edad (E1, E2, E3) se reúnen en la parte central, lo que manifiesta la relativamente poca distinción lexicométrica. Lo mismo puede decirse de las palabras que se sitúan en la parte central del eje vertical: *mm, hm, generalmente, sinceramente, efectivamente, hola, claro*, etc.

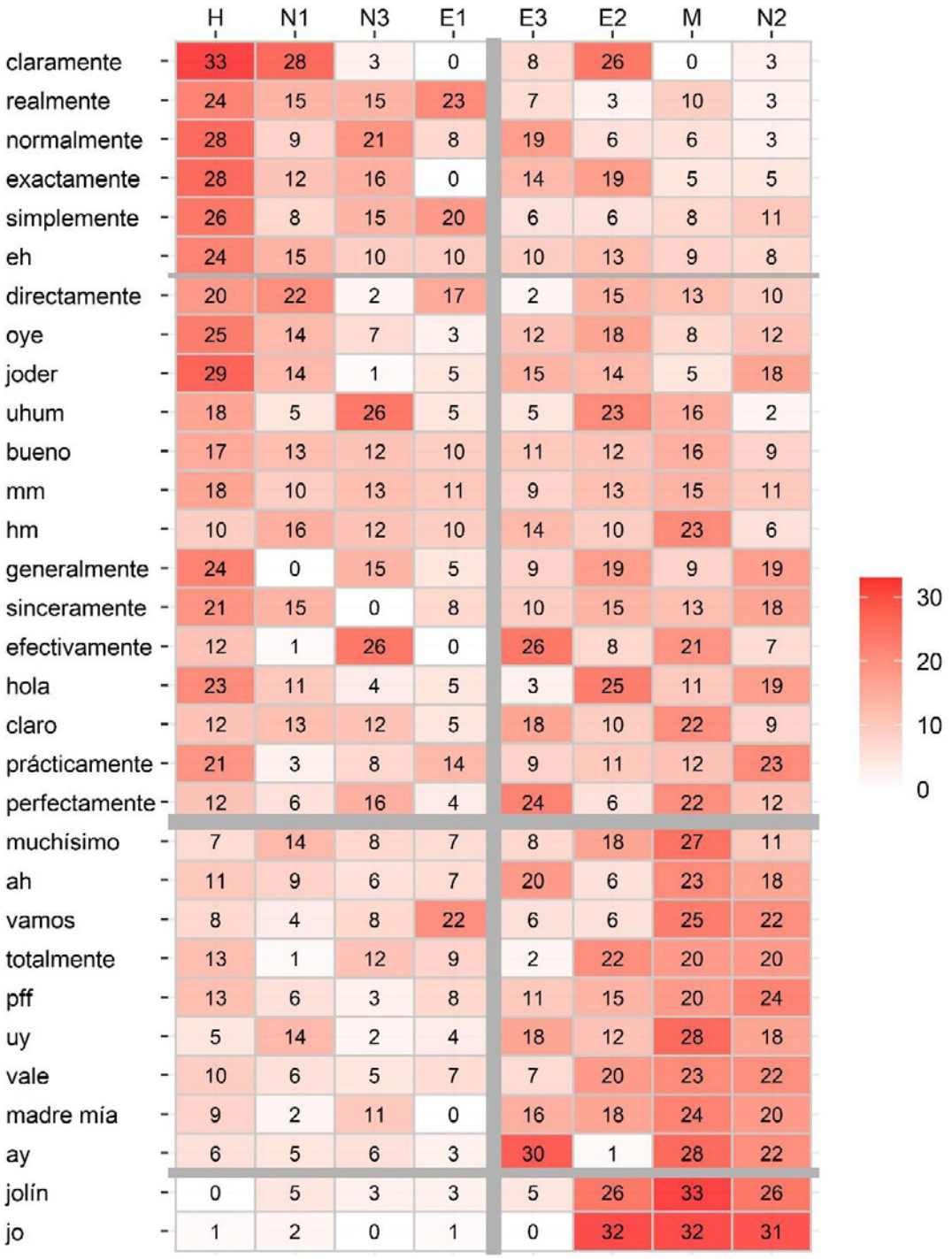

Fig. 16. Distribución diagonalizada de palabras y factores

Seguidamente, presentamos otra visión de conjunto, ahora en el gráfico de esparcimiento tanto de vocablos como de variables, situados en el mismo espacio bidimensional con el eje horizontal correspondiente a los primeros valores de carga y el vertical de los segundos valores de carga, ambos conseguidos en el análisis multivariante de correspondencia (Benzécri, 1976; Peña, 2002; Hair *et al.* 2007; Higuchi 2016). Este gráfico denominado *biplot* posee el mérito de representar las distintas direcciones de flechas pertenecientes a las variables y los puntos situados de vocablos y nos permite interpretarlos al mismo tiempo de la siguiente manera (Vicente Villardón 2021):

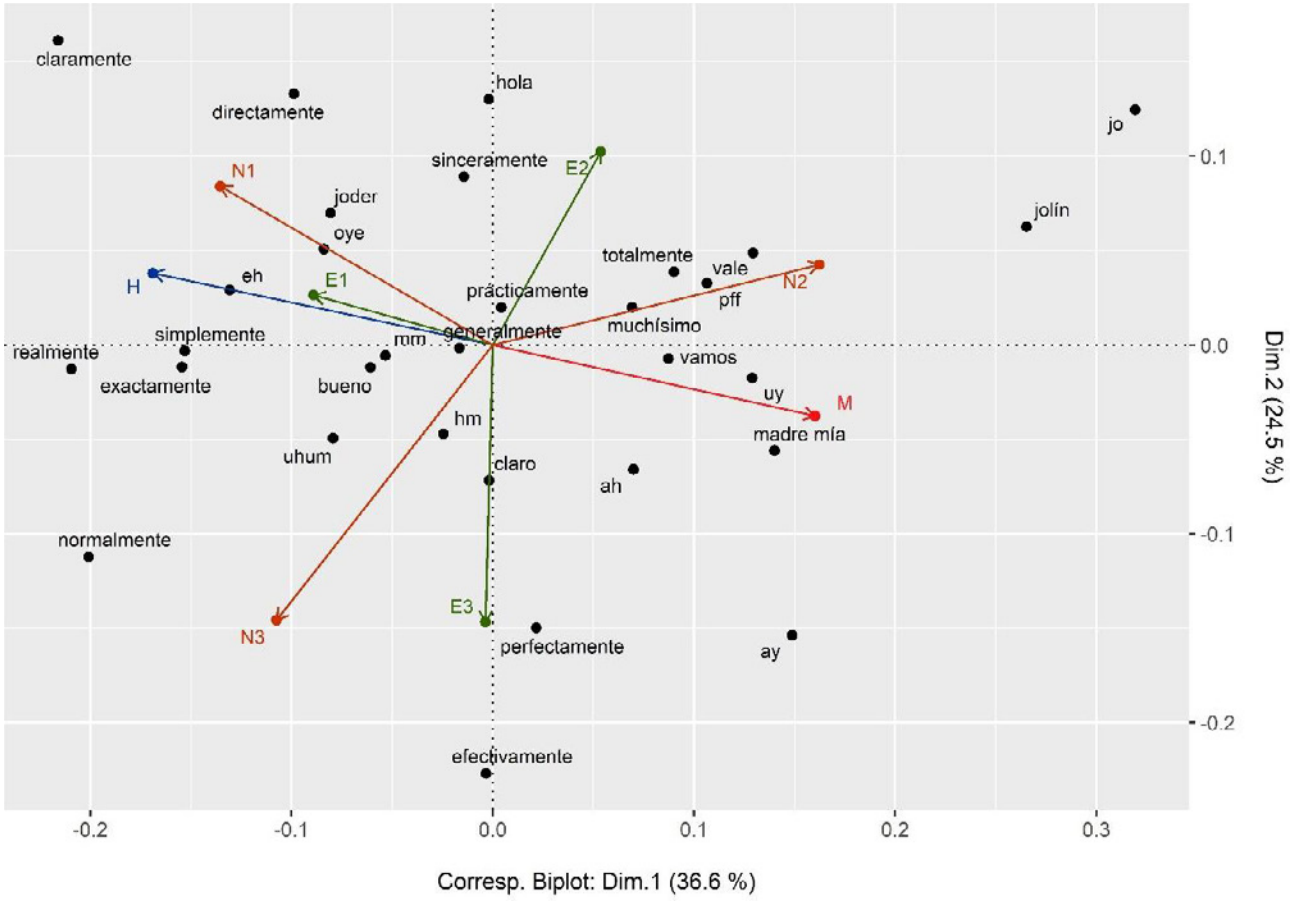

Fig. 17. Biplot

Veamos, en primer lugar, la dirección horizontal de variables. En nuestra observación la misma dirección parece representar la oposición entre H (hombre) y M (mujer), con distintos adverbios de '-mente', *claramente, realmente, normalmente, exactamente,* etc., a la izquierda (H), y los vocablos de *jo, jolín, vale, uy, ay, madre mía,* a la derecha (M), respectivamente. Las palabras situadas en la zona central, *hola, prácticamente, mm, claro, perfectamente, efectivamente*, son neutrales con respecto a la oposición Hombre - Mujer, lo mismo que en la figura 16.

El eje vertical ofrece menor amplitud y parece presentar el orden de Edad: Edad-2 (E2) con *hola* y *sinceramente*, etc., Edad-1 (E1) con *eh, oye, joder* y, finalmente, Edad-3 (E3) con *perfectamente* y *efectivamente*. El porcentaje que presenta el eje horizontal (36.6%) es más alto que el del segundo eje (24.5%), lo que manifiesta su relativa mayor importancia.

Nos llama la atención la coincidencia relativa de dirección de flechas de H (hombre), N1 (nivel educativo bajo) y E1 (edad joven), lo que manifiesta cierta correlación de estos factores sociales. Por otra parte, también es interesante observar la parecida dirección de N3 (nivel educativo alto) y E3 (edad avanzada), lo que es lógicamente natural. La tendencia parecida de M (mujer), N2 (nivel educativo medio) y E2 (edad mediana) es más dispersa que la de los dos grupos anteriores y abarca más espacio y vocablos. Esta clasificación tripartita se comprueba también en la distribución diagonalizada que se advierte en el gráfico de la figura 16.

5. Conclusiones

Las *palabras preferidas* en este estudio se refieren a las palabras que han presentado las frecuencias relativas significativas en el corpus, lo que no implica necesariamente que los sujetos de la categoría correspondiente las utilicen siempre con la misma preferencia, ni tampoco con exclusividad discriminatoria (Smith, 1987). Se trata de indicios estadísticos aparecidos en los datos del corpus santanderino, lo que no permite una generalización excesiva de preferencias.

A pesar de esta limitación interpretativa, hemos advertido que existen ciertas tendencias del uso lingüístico, lo que constituye un factor de primer orden a la hora de analizar las distintas hablas sociolingüísticamente caracterizadas. Al respecto, García Mouton (2003: 23) precisa:

> […] no todas las mujeres responden en su forma de hablar a unos esquemas de los que se consideran femeninos, ni todos los hombres se apartan de ellos. […] no hablan lenguajes diferentes, pero que sí tienen unas preferencias claras al usar el lenguaje que no siempre coinciden.

Efectivamente, Irizarry (1997: 63), tras investigar cuantitativamente los textos de Octavio Paz y Rosario Castellanos, se pregunta "si estas [diferencias significativas] son una manifestación de idiolecto individual más que de sexo abierto a interpretación", y se limita a anotar que "el grado en que mis resultados caracterizarán a otros ensayistas masculinos y femeninos en idioma español o en contextos transculturales queda por investigar".

Por nuestra parte, en la presente investigación se han podido documentar algunos patrones de comportamiento léxico conforme a las variables de sexo, edad y nivel educativo. Se sintetizan, a continuación, las tendencias más generales halladas en la muestra santanderina:

a. El sexo, que se revela como una variable fundamental en el uso de otros fenómenos como la atenuación (Cestero y Albelda, 2020), también en el léxico presenta su variación: así, en el hombre domina el estilo nominal sobre el verbal, el uso de adverbios en *-mente* y el no uso del adjetivo en grado superlativo propio de la mujer, tal y como ocurre en otras lenguas. También el uso masculino de determinadas interjecciones como *joder*, en contraste con *jo*, *jolín* de la mujer.
b. Con respecto a la edad, advertimos la tendencia al uso de determinados modificadores: *bastante*, entre los jóvenes y *poquitín* entre los adultos; el uso de adverbios en *-mente* parece corresponderse con los mayores y el truncamiento léxico, por último, es estrategia léxica propia de la primera generación y del sexo mujer.
c. Se observa una tendencia progresiva al equilibrio entre los estilos nominal y verbal, al mayor alargamiento de las palabras y al no uso de las interjecciones a medida que se avanza en el nivel educativo.
d. Por último, se aprecian estilos léxicos diferentes con respecto al uso de *muchísimo* en cuanto al sexo, edad y nivel de instrucción. La concentración aquí se da en mujer, nivel bajo de instrucción y generación de adultos. También en el uso de los adverbios en *-mente,* los estilos sociolingüísticos marcan una notable diferencia entre hombres y mujeres. Las interjecciones señalan un uso divergente entre sexos y generaciones, no así en el nivel educativo, donde no se advierten formas preferidas.

Finalmente, en el análisis lexicométrico es preciso distinguir entre los aspectos lingüísticos generales y los temas particulares de conversación. En definitiva, ciertas formas en cuestión pueden indicar las preferencias generales sociolingüísticas y/o las selecciones particulares concernientes al asunto que se trata en la conversación. Cabe precisar, por tanto, que hay casos de temas escogidos en el momento de la conversación, aparte del uso lingüístico general del habla. Creemos que es conveniente distinguir entre estos dos aspectos fundamentales -la vida y la lengua-, que están unidos de manera estrecha.

En este punto conviene recordar la crítica de Lastra (1992: 308) sobre Lakoff (1973):

> Casi ninguna de las sugerencias de Lakoff (1973) sobre las diferencias entre el habla masculina y femenina resultaron ser verdaderas, de acuerdo con las investigaciones que sobre el tema se hicieron. Hay estereotipos que se atribuyen a los hombres o a las mujeres y tal es el caso de lo que pasó con lo que había firmado Lakoff.

A nuestro modo ver, podemos evitar la creación de estereotipos exentos de evidencias concretas mediante el análisis cuantitativo de los datos sociolingüísticos que nos proporciona la lexicometría; también, mediante el estudio específico de vocablos individuales, sin caer en la generalización excesiva de categorías gramaticales y semánticas. Hay que analizar los casos individuales y alejarse de los prejuicios impresionistas, tal y como se ha pretendido acometer en este estudio.

Como bien se reconoce, la lengua no es solo para la comunicación, aunque esta sea, sin duda, su función principal; también sirve para confirmar la pertenencia al grupo social (Trudgill, 1983); de ahí que se mantengan ciertas características sociolingüísticas de grupo. Creemos haber encontrado, a través de este análisis lexicométrico del corpus oral PRESEEA-Santander, ciertos rasgos preferentes en el uso del léxico de las variables sociolingüísticas que ejercen la función identificadora del grupo social.

Este estudio de lexicometría sociolingüística deberá, necesariamente, completarse con futuros análisis que abarquen una muestra mayor y estudios coordinados de cotejo con otras áreas urbanas. Por otro lado, tal y como nos recuerda Cabré (1977), un estudio completo de un texto no puede acabar en el análisis de las palabras aisladas, sino que debe continuar analizando unidades más amplias mediante la lematización de las unidades. Por ello, será preciso abordar estudios que nos permitan, a partir de la lematización, medir la extensión real del vocabulario de estos textos orales. Con ello, se podrá establecer el coeficiente de gramaticalidad o funcionalidad, así como el de lexicalidad y el grado de redundancia lexical de este (Cabré, 1977). Por último, esta investigación habrá de constituir el primer fundamento de un futuro estudio más amplio dedicado a la estilometría sociolingüística, en el que el abordaje de la riqueza léxica y el campo semántico resulten decisivos.

Agradecimientos

Este trabajo se inscribe dentro de las actividades científicas del proyecto de I+D+i *Agenda 2050. El español del centro-norte de España: procesos de variación y cambio espaciales y sociales* (PID2019-104982GB-C51), financiado por el Ministerio de Ciencia e Innovación-Agencia Estatal de Investigación/10.13039/501100011033. El proceso de transcripción, necesario para poner a disposición de la comunidad científica internacional las muestras orales del corpus, ha podido llevarse a cabo gracias a la financiación recibida de la Fundación Comillas. Nuestro agradecimiento es también para Francisco Moreno Fernández por el apoyo prestado en este estudio.

Contribución de autoría CREdIT

Inmaculada Martínez (IM) y Hiroto Ueda (HU).
Las aportaciones realizadas por cada uno de los autores al artículo son las siguientes:

- Conceptualización – (IM y HU)
- Curación de datos – No procede.
- Análisis formal – (HU)
- Adquisición de fondos – (IM)
- Investigación – (IM y HU)
- Metodología – (IM y HU)
- Administración del proyecto – (IM y HU)
- Recursos – (IM y HU)
- Software – (HU)
- Supervisión – (IM)
- Validación – (HU)
- Visualización – (IM)
- Redacción – borrador original – (IM y HU)
- Redacción – revisión y edición – (IM y HU)

Referencias bibliográficas

Ayuntamiento de Santander. (2020). *Plan estratégico Santander 2010-2020*. Consultado de http://www.planestrategicosantander.com/documentos.php.

Benzècri, Jean Paul (1976). *L'Analise des données*. II. *L'Analyse des correspondences.* París: Dubod.

Blas Arroyo, José Luis (2005). *Sociolingüística del español. Desarrollos y perspectivas en el estudio de la lengua española*. Madrid: Cátedra.

Cabré, Maria Teresa (1978). La lexicometría como método de localización de rasgos ideológicos. *Revista Española de Lingüística* 8 (2), 335-344. Consultado https://dialnet.unirioja.es/servlet/articulo?codigo=41045

Cestero Mancera, Ana María, Molina Martos, Isabel, Paredes García, Florentino (2006). *Estudios sociolingüísticos del español de España y América*. Madrid: Arco Libros.

Cestero Mancera, Ana María y Albelda Marco, Marta (2020). Estudio de variación en el uso de atenuación I: Hacia una descripción de patrones dialectales y sociolectales de la atenuación en español. *Signos* 53(104), 935-961. Consultado http://dx.doi.org/10.4067/S0718-09342020000300935 .

Escoriza Morera, Luis (2012). La variación de expresión en el plano léxico. dificultades y perspectivas. *Lingüística* 28, 247-273.

García Mouton, Pilar (1999). *Cómo hablan las mujeres*. Madrid: Arco Libros.

García Mouton, Pilar (2003). *Así hablan las mujeres*. *Curiosidades y tópicos del uso femenino del lenguaje.* Madrid: La Esfera de los Libros.

Hair, Joseph, Anderson, Rolph, Tatham, Ronald. y Black, William (2007). *Análisis multivariante* (5ª ed.). Madrid: Pearson Prentice Hall.

Higuchi, Koichi (2017). A two-step approach to quantitative content analysis: kh coder tutorial using *Anne of Green Gables* (Part I, II). *Ritsumeikan Social Science Review*, 52, 77-91; 53: 137-147.

Irizarry, Estelle (1992). A computer-assisted investigation of gender-related idiolect in Octavio Paz and Rosario Castellanos. *Computers and the Humanities,* 26, 103-117. En Irizarry, Estelle (1997) *Informática y literatura. Análisis de textos hispánicos.* Traducción española. 41-69. Barcelona: Proyecto A Ediciones.

Jespersen, Otto (1922). *Language. Its nature, development and origin*. Londres: George Allen & Unwin.

Koch, Josse de (1983). *Elementos para una estilística computacional.* I, II. Madrid: Editorial Coloquio.

Lastra, Yolanda (1992). *Sociolingüística para hispanoamericanos. Una introducción*. México: Colegio de México.

Lakoff, Robin (1973). Language and woman's place. *Language in Society* 2, 45-80.

Lakoff, Robin (1977). Women's language. *Language and style* 10, 222-247.

López García, Ángel y Morant, Ricardo (1991). *Gramática Femenina*. Madrid: Cátedra.

Lüdeling, Anke y Kytö, Merja (2008). *Corpus Linguistics: An International Handbook*. Vol. 1. Berlin: De Gruyter.

Martínez, Inmaculada y Ueda, Hiroto. 2020. *Inventario léxico del corpus PRESEEA-Santander*. en: https://lecture.ecc.u-tokyo.ac.jp/~cueda/kenkyúrekisi/santander.pdf [12/09/2021]

Martínez, Inmaculada y Ueda, Hiroto. 2021. Aspectos estadísticos del español oral. Análisis del corpus sociolingüístico de Santander (España). Comunicación oral en X Congreso Asiático de Hispanistas, Seúl: Universidad Hankuk de Estudios Extranjeros.

McEnery, Tony y Hardie, Andrew (2012). *Corpus linguistics. Methods, theory and practice.* Cambridge: Cambridge University Press.

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

Moreno Fernández, Francisco (2022). La variación geográfica y social en los corpus hispánicos. En *Lingüística de corpus en español: The Routledge Handbook of Spanish Corpus Linguistics*, eds. G. Parodi, P. Cantos-Gómez y C. Howe, (en prensa). Nueva York: Routledge. Consultado de: https://doi.org/10.4324/9780429329296

Moreno Fernández, Francisco, Paredes García, Florentino, Molina Martos, Isabel y Cestero Mancera, Ana María (2000). La sociolingüística de Alcalá de Henares en el Proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA). *Oralia* 3, 149-168.

Peña, Daniel (2002). *Análisis de datos multivariantes.* Madrid: McGraw Hill.

Peña Arce, Jaime (2021). La complejidad dialectal de Cantabria. Diacronía y sincronía del yeísmo regional. *Zeitschrift für romanische Philologie* 137, no. 2, 426-450. https://doi.org/10.1515/zrp-2021-0017

Romaine, Suzanne (2008). Corpus linguistics and sociolinguistics. En *Corpus Linguistics: An International Handbook*, eds. A. Lüdeling y M. Kytö, 96-111. Berlin: De Gruyter.

Romero Pérez, Ivón, Alarcón Vásquez, Yolima y García Jiménez, Rafael (2018). *Lexicometría: enfoque aplicado a la redefinición de conceptos e identificación de unidades temáticas*. Biblios 71. DOI: 10.5195/biblios.2018.466

Smith, Philip (1987). *Language, the sexes and society* (trad. Inoue, Kazuko; Kono, Takeshi y Mineko, Masamune), Oxford: Basil Blackwell.

Trudgill, Peter (1983). *Sociolinguistics. An introduction to language and society.* Londres: Penguin Books.

Ueda, Hiroto. 2021. *Guía para el uso de LYNEAL con materiales de PRESEEA*. en https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/doc/guia.pdf

Ueda, Hiroto y Moreno Sandoval, Antonio. 2017. *Análisis de datos cuantitativos para estudios lingüísticos.* en: https://h-ueda.sakura.ne.jp/gengo/4-numeros/doc/numeros-es.pdf

Vicente Villardón, José Luis (2021). *Los métodos biplot (Teoría)*. Consultado de: http://biplot.usal.es/multbiplot/documentatio-ñnotas-sobre-biplot-clasico-.pdf