ONOMÁZEIN
Revista de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS

Disponibilidad léxica y dominio de la ortografía: un estudio empírico basado en la influencia de los factores sociales

Lexical availability and spelling knowledge: an empirical study based on the influence of social factors

Ester Trigo Ibáñez
Universidad de Cádiz
España

Manuel Francisco Romero Oliva
Universidad de Cádiz
España

Inmaculada Clotilde Santos Díaz
Universidad de Málaga
España

ONOMÁZEIN 47 (marzo de 2020): 27-45
DOI: 10.7764/onomazein.47.02
ISSN: 0718-5758

Ester Trigo Ibáñez: Departamento de Didáctica de la Lengua y la Literatura, Universidad de Cádiz, España. | E-mail: ester.trigo@uca.es
Manuel Francisco Romero Oliva: Departamento de Didáctica de la Lengua y la Literatura, Universidad de Cádiz, España. | E-mail: manuelfrancisco.romero@uca.es
Inmaculada Clotilde Santos Díaz: Departamento de Didáctica de la Lengua, las Artes y el Deporte, Universidad de Málaga, España. | E-mail: santosdiaz@uma.es

Fecha de recepción: marzo de 2018
Fecha de aceptación: agosto de 2018

Resumen

Este artículo tiene como objetivo mostrar de forma empírica la influencia de los factores sociales en la ortografía tomando como referencia los planteamientos de la disponibilidad léxica. La muestra de estudio está formada por 400 estudiantes preuniversitarios de Sevilla que realizaron la encuesta siguiendo las directrices del Proyecto Panhispánico de Disponibilidad Léxica. Los análisis multivariantes muestran que las mujeres cometen menos errores ortográficos que los hombres en todos los segmentos analizados (tipo de centro, población y nivel sociocultural) y que la frecuencia de error de los hombres es mayor en centros privados que en centros públicos. Este estudio abre nuevas líneas de investigación en el plano sociolingüístico y de la lingüística aplicada a la enseñanza de idiomas ya que prueba el efecto de los factores sociales no solo en el léxico disponible sino también en la ortografía.

Palabras clave: disponibilidad léxica; ortografía; sociolingüística; didáctica de la lengua.

Abstract

The purpose of this article is to show empirically the influence of social factors on spelling, taking as reference the approaches of the lexical availability. The study sample consists of 400 preuniversity students from Seville who conducted the survey following the guidelines of the Pan-Hispanic Project of Lexical Availability. The multivariate analyses show that women commit fewer misspellings than men in all the segments analyzed (type of center, population and sociocultural level) and that the frequency of error of men is greater in private centers than in public centers. This study opens new lines of research in the sociolinguistic and linguistics applied to language teaching, as it tests the influence of social factors not only in the lexicon available but also in the spelling.

Keywords: lexical availability; spelling; sociolinguistics; language teaching.

1. Introducción

Todo modelo científico conforma también un sistema de ideas que, en principio, debe competir con los presupuestos aceptados en el momento en cuestión y, poco a poco, lograr transformar la configuración disciplinar (García Marcos, 2009). Bajo estas circunstancias surge la lingüística aplicada, en la década de los cincuenta del siglo pasado, disciplina que hoy cuenta con una aceptación innegable —visible en la tradición académica y científica— y una emergente proliferación desde perspectivas heterogéneas. En este marco referencial encontramos la disponibilidad léxica[1], cuyos estudios ven la luz en Francia (Gougenheim y otros, 1956, 1964) con el fin de determinar un vocabulario de base para la enseñanza del francés tanto a los habitantes de la antigua unión francesa como a los inmigrantes que llegaban a Francia. Para ello, se elabora un procedimiento de encuesta capaz de recopilar el léxico evocado por un hablante en determinadas situaciones comunicativas o centros de interés combinando frecuencia y orden de aparición (López Chavez y Strassburguer Frías, 1987).

La disponibilidad léxica se ha desarrollado especialmente en el ámbito hispánico en el marco del Proyecto Panhispánico de Disponibilidad Léxica (PPHDL), impulsado por Humberto López Morales. Este hecho ha permitido una unificación metodológica clave tanto para su avance como para el establecimiento de comparaciones sintópicas (Samper Padilla, 1998). Sin embargo, si bien el objetivo inicial de los trabajos de disponibilidad léxica no fue la detección de errores ortográficos[2], contamos con un amplio porcentaje de investigaciones que han manifestado la progresiva preocupación por este aspecto. En este sentido, centrados en la lengua materna de estudiantes preuniversitarios, encontramos los trabajos de Paredes (1999), Galloso (2003), Ortolano (2005), Fernández Smith y otros (2008), Saura (2008), Blanco (2011) y García Casero (2013). Enmarcados dentro de estudios universitarios de grado y postgrado, consignamos las investigaciones de Ávila (2007) y Santos (2015). Además, existen estudios que inciden en la adquisición de segundas lenguas, como los de Carcedo (1999), Frey (2007), Sánchez-Saus (2016), Hidalgo (2017) y Mariscal (2017).

No obstante, pese a la preocupación de los investigadores de disponibilidad léxica acerca de la disortografía consignada en las encuestas, a excepción del estudio exploratorio con estudiantes universitarios de Ávila (2007), tan solo se han realizado investigaciones descriptivas —independientemente de la perspectiva adoptada— para dar cuenta de esta situación. Por ello, nuestro trabajo intenta explorar con mayor profundidad la importancia que cobran los

1 López Morales (1995), Carcedo (1999) y Paredes (2012) ofrecen una amplia panorámica de cómo surgen y evolucionan estos estudios desde el punto de vista tanto metodológico como epistemológico.

2 Realmente se considera un aspecto secundario pues la propia metodología de la prueba invita al informante a escribir tantas palabras como vengan a su mente dado un estímulo o centro de interés en un tiempo máximo de dos minutos, sin prestar especial atención a los aspectos ortográficos.

factores sociolingüísticos en la frecuencia de error registrada; conocer el comportamiento de las distintas variables estudiadas proporcionará una válida información de cara a diseñar estrategias de enseñanza de lenguas desde la diversidad del hablante.

Las razones que nos llevan a pensar que los datos consignados en los repertorios de disponibilidad son de gran utilidad se centran en los siguientes aspectos: a) por un lado, desde el punto de vista ortográfico, el hecho de detectar los errores más frecuentes dentro de la esfera léxica usual de los informantes es muy importante de cara a diseñar materiales didácticos ajustados a las necesidades halladas, y, b) por otro lado, la determinación de variables de estudio dentro del PPHDL posibilitaría realizar análisis estadísticos univariantes y multivariantes que determinasen la influencia de los factores extralingüísticos para alcanzar el dominio ortográfico (Mairal y otros, 2010; Herrera y otros, 2011).

En consecuencia, nuestro objetivo principal se centra en evaluar qué factores inciden en la competencia ortográfica y, más precisamente, sobre una mayor o menor producción de errores ortográficos. Para ello, expondremos la relación entre el sexo, la clase social, el tipo de centro educativo y de población con el número total de disortografías registradas. Además, desde un planteamiento empírico, pretendemos contrastar la hipótesis de que el sociolecto utilizado por cada subgrupo social es diferente no solo en el vocabulario consignado —como demuestra el estudio de Ávila y Villena (2010)—, sino también en su domino ortográfico. Los hallazgos obtenidos evidenciaron unos resultados que pueden considerarse reveladores respecto a las creencias que se pudieran tener en torno al dominio ortográfico de algunos sociolectos como la variante sexo o el centro educativo. De esta forma, el conocimiento certero de estas diferencias debe servir de referencia en la toma de decisiones para una adecuada intervención didáctica.

2. Metodología

Los datos que componen el corpus de léxico disponible se conforman a partir de un test asociativo que explora los 16 centros de interés clásicos en este tipo de investigaciones[3]. Cada informante dispone de dos minutos para evocar el mayor número de palabras relacionadas con un centro de interés dado. Una vez recabada la información, se procesa con el programa *LexiDisp* (Moreno y otros, 1995) y se genera un listado de palabas que combina frecuencia y orden de aparición. De esta forma, la palabra más disponible es aquella que antes viene a la mente de los hablantes dado un centro de interés.

3 (01) Partes del cuerpo, (02) La ropa, (03) Partes de la casa (sin los muebles), (04) Los muebles de la casa, (05) Alimentos y bebidas, (06) Objetos colocados en la mesa para la comida, (07) La cocina y sus utensilios, (08) La escuela: muebles y materiales, (09) Iluminación, calefacción y medios para airear un recinto, (10) La ciudad, (11) El campo, (12) Medios de transporte, (13) Trabajos del campo y del jardín, (14) Los animales, (15) Juegos y distracciones, (16) Profesiones y oficios.

Para la realización de este trabajo, se confeccionó un corpus cacográfico durante el curso 2016-2017 a partir de un estudio de disponibilidad léxica realizado en la provincia de Sevilla (Trigo, 2011). La muestra de informantes está formada por 400 estudiantes de segundo curso de bachillerato, divididos en cuatro subgrupos formados por variables categóricas que estratifican la población (sexo: 159 hombres y 241 mujeres; tipo de centro: 298 públicos y 102 privados; población: 296 urbanos y 104 rurales, y nivel sociocultural: 65 alto, 104 medio y 231 bajo)[4]. Nuestro rango espacial se ciñe a la provincia de Sevilla y el rango temporal es el curso 2016/2017, momento en el que se compila el corpus atendiendo únicamente a los aspectos ortográficos de acuerdo con las pautas metodológicas —atendiendo a la clasificación de errores, catalogación y unificación— determinadas por Paredes (1999)[5].

El tamaño de la muestra formada por los 400 individuos arroja una precisión de muestreo de d=0,05, lo que indica que cualquier valoración real está con seguridad comprendida en p±5 %, siendo p la proporción estimada de la muestra. Aunque este dato de precisión hay que tomarlo con ciertas reservas al ser el procedimiento de recogida de datos de tipo accidental, el hecho de que los estratos de la muestra observada mantengan su proporcionalidad con los estratos poblacionales reales[6] permite confiar en la coherencia del dato suministrado.

4 Los informantes proceden de los siguientes centros escolares: público-rurales: IES Gerena (Gerena), IES Ostippo (Estepa), IES Lago Ligur (Isla Mayor), IES Torre de los Guzmanes (La Algaba); público-urbanos: IES V Centenario (Sevilla), IES Isbilya (Sevilla), IES Virgen del Castillo (Lebrija), IES Torreblanca (Sevilla), IES Macarena (Sevilla), IES Carmen Laffón (San José de la Rinconada), IES Albero (Alcalá de Guadaira), IES Ruiz Gijón (Utrera); privado-urbanos: Colegio San José (Sevilla), Colegio Aljarafe (Mairena del Aljarafe), Colegio Santa Ana (Sevilla).

5 Este autor establece cuatro niveles: (1) **acentuación y diéresis**: (1.1) *ausencia de tilde*; (1.1.1) aguda; (1.1.2) llana; (1.1.3) esdrújula; (1.1.4) hiato; (1.1.5) diacrítica; (1.2) *mala colocación de la tilde*; (1.2.1) porque la palabra no debe llevar; (1.2.2) porque no está bien colocada; (1.3) *diéresis*; (2) **letras**: (2.1) *error en la correspondencia fonema/letra o error de grafía*; (2.1.1) c/z/s; (2.1.2) g/j; (2.1.3) ll/y/hie-; (2.1.4) s/x; (2.1.5) h; (2.1.6) m/n; (2.1.7) b/v; (2.1.8) i/y; (2.1.9) gu/g, qu/q; (2.1.10) r/rr; (2.1.11) consonantes en posición implosiva; (2.1.12) sonidos en posición no implosiva; (2.1.13) metátesis; (2.2) *mayúsculas y minúsculas*; (2.2.1) minúsculas en acrónimos; (2.2.2) mayúsculas en nombres comunes; (2.2.3) mezcla de mayúsculas y minúsculas; (2.2.4) minúsculas en nombres propios; (2.3) *grafías dubitativas*; (2.4) *omisión de letras*; (2.5) *adición de letras*; (3) **morfosintaxis**: (3.1) *nombres compuestos*; (3.1.1) unión y separación de palabras; (3.1.2) empleo del guion; (3.2) *contracciones*; (3.2.1) del artículo y el nombre; (3.2.2) de las preposiciones y otro elemento; (3.3) *plural*; (3.4) *discordancias entre singular y plural*; (4) **léxico**; (4.1) *vulgarismos*; (4.2) *coloquialismos*; (4.3) *contaminación de lexemas*; (4.4) *extranjerismos*; (4.5) *marcas comerciales*; (4.6) *errores de concepto*; (4.7) *creaciones léxicas y palabras no preferidas por el DRAE*.

6 De los 14 153 estudiantes de segundo de bachillerato en la provincia de Sevilla, se consigna un 57,15 % de mujeres frente a un 42,84 % de hombres, un 79,8 % de centros públicos y un 20,2 % de centros privados, 74,5 % de centros situados en localidades urbanas —de más de 20.000 habitantes— frente a un 25,5 % de centros situados en localidades rurales y, por último, un 57,7 % de familias de nivel sociocultural bajo, un 25,8 % de nivel sociocultural medio y un 16,8 % de nivel sociocultural alto.

Además de las variables que estratifican la población, se han tenido en cuenta tres variables de interés para nuestro estudio o variables criterio: el *error*, el *número de errores* y la *tipología de error*. Por un lado, el *error* es una variable dicotómica 0/1 asociada a la palabra que determina la ausencia o presencia de error en cada vocablo del corpus. Será considerada dentro del enfoque bivariante que presentamos a continuación. Por otro lado, el *número de errores* (*NumErrores*) es una variable numérica que representa el número de errores cometidos por el informante. Esta será considerada dentro del contexto multivariante de la minería de datos, *data mining*, en el presente trabajo. Por último, la *tipología de error* es una variable categórica que refleja los cuatro grupos —acentuación, letras, morfosintaxis y léxico— previamente establecidos por Paredes (1999). Esta, por su alto grado de importancia, será estudiada de manera independiente en un estudio posterior prestando atención a la tipología de error consignada por cada subgrupo social. Todo el análisis que presentamos a continuación se realizó con el *software* estadístico SPSS v22 y el paquete R.

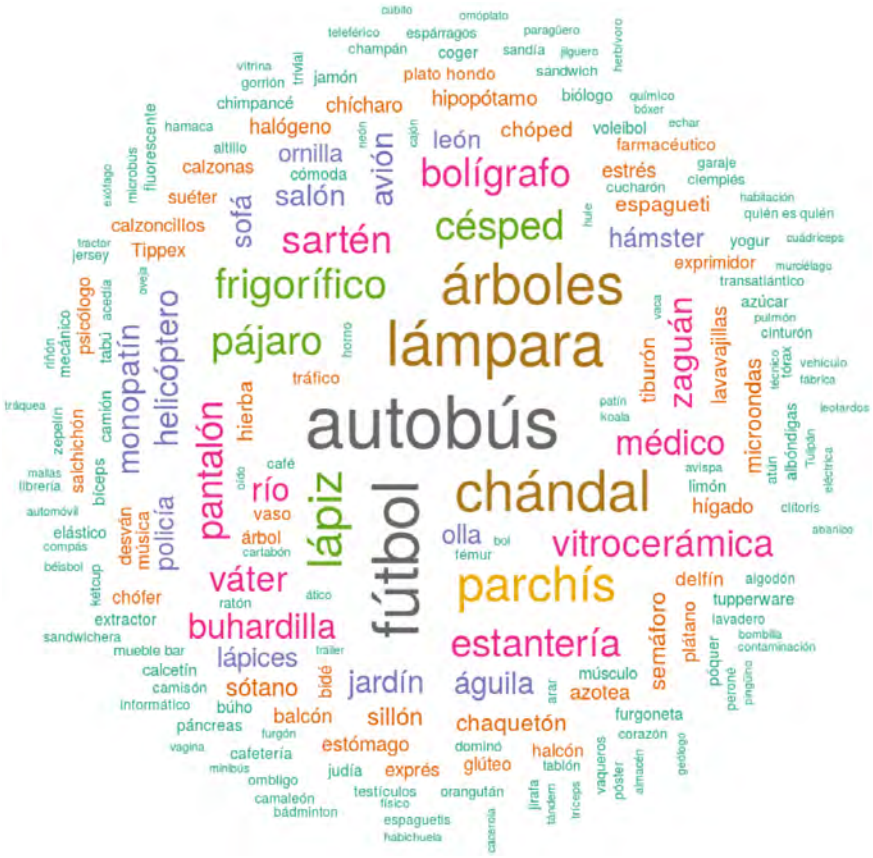
3. Resultados

Aunque el objetivo de nuestra investigación es analizar las características poblacionales de los individuos con mayor cantidad de errores ortográficos, se hace obligado en una primera aproximación reflejar aquellos vocablos que mayor incidencia de error poseen. Para este fin, aportaremos el gráfico de nube de palabras, construido mediante el paquete estadístico R, especializado en minería de datos —*vid.* figura 1—, en donde de una manera muy visual e intuitiva aparecen los vocablos con un tamaño (y color) que es función de la cantidad de errores que llevan asociados. Así, a medida que aumenta la frecuencia de los errores, las palabras aparecen más enfatizadas dentro de la nube.

Como se aprecia en la figura 1, la mayor frecuencia de error se registra en los errores de tipo 1 (acentuación). Sin embargo, también se reflejan errores de tipo 2 (letras), como *avanico* (*abanico*) o *amaca* (*hamaca*). Los errores de tipo 3 (morfosintaxis) y 4 (léxico), si bien se han estudiado en profundidad, como se refleja en Trigo y otros (2018), no aparecen reflejados en la figura 1 debido a su escasa frecuencia.

En cuanto a la lematización, dado que se ha procedido a hacer un vaciado de las disortografías consignadas en las encuestas, aparece, por ejemplo, el vocablo *autobus* (*autobús*). Este vocablo no ha sido lematizado siguiendo los criterios para elaborar los diccionarios de disponibilidad léxica (Samper Padilla, 1998) con la forma (*auto*)*bus*, puesto que lo que se pretendía era detectar el error ortográfico, en este caso, la tilde en palabra aguda —codificada, siguiendo a Paredes (1999), como 1.1.1—. Situación análoga ocurre con el hecho de consignar voces en plural y singular —en lugar de unificarlas en singular— como *arbol* o *lapiz* (*árbol* o *lápiz*) frente a *arboles* o *lapices* (*árboles* o *lápices*), puesto que las formas en singular registran un error de ausencia de tilde en palabra llana —codificada como 1.1.2— y las formas en plural consignan un error de ausencia de tilde en palabra esdrújula —codificada como 1.1.3—.

FIGURA 1
Vocablos según frecuencia de errores



3.1. Enfoque bivariante

Describiremos muy brevemente el concepto general de *contraste de hipótesis* y sus elementos asociados, así como el caso particular del *contraste chi-cuadrado* que será empleado dentro del enfoque bivariante.

Un *contraste de hipótesis* o *test de hipótesis* es una prueba estadística en la cual se decide con una cierta probabilidad fijada por el investigador si una proposición respecto de la población (hipótesis estadística) es cierta o no. Si rechazamos la hipótesis planteada en el contraste, queremos indicar que los datos de la muestra ofrecen cierta evidencia sobre su falsedad. Un contraste de hipótesis consiste, por tanto, en estudiar dos hipótesis: *Ho* (*hipótesis nula*) y *H1* (*hipótesis alternativa*), y verificar en función de los resultados muestrales cuál es la correcta.

El *nivel de significación alpha* o *error del contraste* se define como la probabilidad de rechazar la hipótesis nula en el supuesto de que esta sea cierta:

$\alpha = P\{rechazar\ H0/H0\ cierta\}$,

por lo que queda garantizado que si la hipótesis nula es cierta se acertará con una seguridad el $(1-\alpha)*100$ de las veces que se extraiga una muestra de la población. En general se establece $\alpha=0,05$, es decir, la fiabilidad del contraste será del 95 %.

El contraste que emplearemos para nuestro estudio se denomina *contraste chi-cuadrado* y en particular establece si dos variables categóricas son independientes (H0) o si, por el contrario, existe algún tipo de relación entre ellas (H1). Dicho contraste está basado en una función denominada *estadístico del contraste —vid.* tabla 1—, que cuantifica el error que comete la prueba usando las diferencias entre las frecuencias observadas o_{ij} de la muestra y las frecuencias teóricas e_{ij} que tendría si la hipótesis nula fuese cierta: si el resultado es un valor que supera un determinado umbral denominado *región crítica*, la hipótesis H0, de independencia, sería rechazada (véase la tabla 1). Equivalentemente se suele usar un criterio denominado *criterio del p-valor*, esto es: si la probabilidad del estadístico del contraste está por debajo del nivel de significación establecido para el contraste $(\alpha=0,05)$, se rechaza la hipótesis H0.

TABLA 1
Contraste chi-cuadrado

HIPÓTESIS DEL CONTRASTE	REPRESENTACIÓN	ESTADÍSTICO DEL CONTRASTE	REGIÓN CRÍTICA O DE RECHAZO DE H0
H0: X e Y son independientes H1: X e Y están relacionadas	Tabla de frecuencias r x c	$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$	p-valor < 0,05

3.1.1. Aplicación y resultados del contraste chi-cuadrado

Usando el contraste descrito anteriormente se pretende dilucidar si existen relaciones de dependencia entre las frecuencias de los errores cometidos por la población y las que estratifican dicha población de individuos: sexo, población, tipo de centro y nivel sociocultural. La tabla 2 muestra los resultados estadísticos obtenidos.

3.1.2. Distribución observada de los errores

Las tablas 3, 4, 5 y 6 representan las distribuciones bidimensionales de frecuencias de los *errores / no errores* cruzadas con cada variable de estratificación y sus porcentajes por filas, expresados con la finalidad de poder comparar entre y obtener en una primera aproximación las características poblacionales de los individuos donde el error se da en mayor medida. El hecho de que las tablas presenten totales distintos se debe a la pérdida de datos o ausencia de respuesta.

TABLA 2

Resultados del contraste chi-cuadrado

VARIABLES	CONTRASTE	FRECUENCIAS OBSERVADAS	ESTADÍSTICO DEL CONTRASTE Y SU P-VALOR
X: Error / No error en cada palabra Y: Sexo	Ho. El error es independiente del sexo. H1: El error es dependiente del sexo	Tabla 3	68,231 p = 0,000 < 0,05 Aceptada H1
X: Error / No error en cada palabra Y: Tipo de centro	Ho. El error es independiente del tipo de colegio. H1: El error es dependiente del tipo de colegio	Tabla 4	5,519 p = 0,019 < 0,05 Aceptada H1
X: Error / No error en cada palabra Y: Población	Ho. El error es independiente de la población. H1: El error es dependiente de la población	Tabla 5	9,491 p = 0,002 < 0,05 Aceptada H1
X: Error / No error en cada palabra Y: Nivel Sociocultural	Ho. El error es independiente del nivel social. H1: El error es dependiente del nivel sociocultural.	Tabla 6	72,559 p = 0,000 < 0,05 Aceptada H1

TABLA 3

Error *versus* sexo

			ERRORES1		TOTAL
			NO ERROR	ERROR	
sexo	Hombre	Recuento % dentro de sexo	45808 94,4 %	2726 5,6 %	48534 100,0 %
	Mujer	Recuento % dentro de sexo	71685 95,4 %	3433 4,6 %	75118 100,0 %
Total		Recuento % dentro de sexo	117493 95,0 %	6159 5,0 %	123652 100,0 %

TABLA 4

Error *versus* tipo de centro

			ERRORES2		TOTAL
			NO ERROR	ERROR	
centro	Público	Recuento % dentro de centro	85731 94,9 %	4582 5,1 %	90313 100,0 %
	Privado	Recuento % dentro de centro	32354 95,4 %	1577 4,6 %	33931 100,0 %
Total		Recuento % dentro de centro	118085 95,0 %	6159 5,0 %	124244 100,0 %

TABLA 5

Error *versus* tipo de población

			ERRORES3		TOTAL
			NO ERROR	ERROR	
población	Urbana	Recuento % dentro de la población	89040 95,1 %	4582 4,9 %	93622 100,0 %
	Rural	Recuento % dentro de la población	29045 94,8 %	1603 5,2 %	30648 100,0 %
Total		Recuento % dentro de la población	118085 95,0 %	6185 5,0 %	124270 100,0 %

TABLA 6

Error *versus* nivel sociocultural

			ERRORES4		TOTAL
			NO ERROR	ERROR	
Nivel Sociocultural	Bajo	Recuento % dentro de NSociocultural	65282 94,6 %	3735 5,4 %	69017 100,0 %
	Medio	Recuento % dentro de NSociocultural	30231 95,4 %	1454 4,6 %	31685 100,0 %
	Alto	Recuento % dentro de NSociocultural	22462 95,9 %	970 4,1 %	23432 100,0 %
Total		Recuento % dentro de NSociocultural	117975 95,0 %	6159 5,0 %	124134 100,0 %

3.1.2.1. Hallazgos más relevantes

Como resultado de las inferencias de la tabla 2 y de las tabulaciones presentadas en las tablas 3, 4, 5 y 6, podemos observar que todas las características analizadas en la población tienen dependencia significativa con respecto al número de errores ortográficos cometidos. Los estratos de mayor riesgo de error son los que siguen:

- Los hombres con un 5,6 % de error frente al 4,6 % de error de las mujeres,
- los centros públicos con un 5,1 % de error frente al 4,6 % de los privados,
- la población rural con un 5,2 % frente al 4,9 % de la población urbana y
- el nivel sociocultural bajo con un 5,4 % de error frente al 4,6 % y 4,1 % de los niveles medio y alto respectivamente.

Desde el punto de vista sociolingüístico, estos resultados demuestran que el comportamiento lingüístico de los subgrupos sociales es diferente, siendo más vulnerables los *hombres*, los

informantes de centros públicos, los *ubicados en núcleos rurales* y los de *nivel sociocultural bajo*. Esta cuestión corrobora nuestra hipótesis de partida pues las diferencias entre los distintos sociolectos se manifiestan claramente en el dominio ortográfico.

Conocer el comportamiento de cada subgrupo social es clave para plantear futuras investigaciones enfocadas en el ámbito de la enseñanza de lenguas que ayuden a paliar las diferencias encontradas. En este sentido, sería interesante compilar repertorios cacográficos (Gómez Camacho, 2006) para diseñar artefactos digitales que atiendan a la diversidad del hablante, respetando su propio ritmo de aprendizaje (Valdés y Romero, 2017; Romero y otros, 2018).

Además, constatando que son los informantes de centros públicos y de núcleos rurales los que consignan mayor número de error, será necesario, principalmente en estos contextos, hacer visibles las estrategias que la Administración Pública andaluza ofrece a los centros sostenidos con fondos públicos para la mejora de la competencia lingüística, como es el caso de los Proyectos Lingüísticos de Centro (Romero y Trigo, 2015, 2018), y, así, proporcionar una formación a estos centros educativos que les permita trabajar la ortografía desde perspectivas integradoras que garanticen los principios de igualdad y equidad (Trujillo y Rubio, 2014).

3.2. Enfoque multivariante y minería de datos

El estudio anterior muestra que la probabilidad de que una palabra se escriba erróneamente depende de los estratos específicos de sexo, tipo de centro, población y nivel sociocultural. Sin embargo, estamos interesados en centrarnos en la investigación dando un paso más, enfoncándonos en la submuestra de los informantes que cometen error, con el fin de caracterizar qué segmentos acumulan las mayores y menores tasas de error. Para ello nos centraremos en la variable criterio *NumErrores* ortográficos de cada individuo.

Entendemos por *segmento poblacional* cualquier combinación de categorías formada por una o varias de las variables de estratificación (sexo, tipo de centro, población y nivel sociocultural). La búsqueda de tales segmentos discriminantes implica una alta dificultad algorítmica de búsqueda, ya que el número de contrastes de hipótesis realizables para distinguir y seleccionar entre los segmentos candidatos es muy elevado.

A tal fin trasladaremos nuestro estudio al contexto multivariante empleando para ello una herramienta encuadrada dentro de la minería de datos o del aprendizaje automático cuyas características presentamos y describimos a continuación.

3.2.1. Minería de datos y el planteamiento CHAID

La minería de datos es una tecnología emergente que contiene métodos estadísticos y de inteligencia artificial orientados a explorar grandes bases o conjuntos de datos, de manera

automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

En particular, dentro de la minería de datos se encuentra el *método CHAID* (*Chi-Squared Automatic Interaction Detector*), el cual resuelve el problema de segmentación más eficientemente que con el planteamiento clásico de búsqueda exhaustiva, evitando así el problema de la alta complejidad y volumen de la información que se ha de analizar.

Entre las ventajas que aporta cabe mencionar:

- Es de fácil aplicación y su algoritmo se incluye en la mayor parte de los paquetes estadísticos actuales: SPSS, R, Statistica, etc.
- Presenta resultados interpretables e intuitivos sin necesidad de conocimientos estadísticos avanzados.

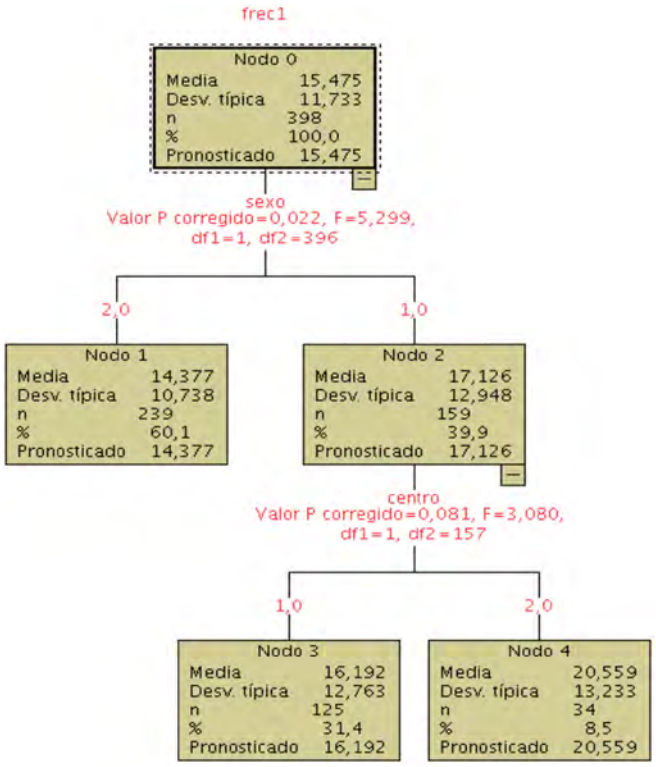
Teóricamente, *CHAID* se basa en la construcción automática de un árbol donde cada nodo representa un segmento formado por un conjunto de características de los informantes construido por un proceso interno de optimización, de forma que de manera inteligente maximiza el impacto en la variable *NumErrores* en cada segmento.

La optimización es un proceso recursivo con el que se deciden los nodos del árbol y consta de 2 fases: *fusión* y *división*. La *fusión* usa el *contraste F* para comparar las medias de error en los estratos poblacionales. Si el contraste distingue diferencias entre una categoría y el resto, la aísla y forma un nuevo nodo del árbol. En caso contrario, la categoría resulta irrelevante y la agrupa con el resto de categorías. Si todas las categorías así obtenidas son irrelevantes, la fusión crea un nodo terminal del árbol y para el cómputo sobre dicho nodo. En caso contrario, entra en juego la *división*, seleccionando aquella variable que presente mayor diferencia significativa (*p-valor* del *contraste F*) y repite el proceso.

3.2.1.1. Principales hallazgos

La aplicación del *árbol de segmentación CHAID* —gráfico 1— muestra la aparición de los segmentos poblacionales óptimos según la lógica del método y las estadísticas de cada uno de los segmentos. En este sentido, al estudiar la frecuencia de error ortográfico, nos encontramos con dos segmentos, uno referido al *sexo*, que muestra que las mujeres, independientemente del contexto analizado, consignan una media menor de error por persona que los hombres (14,37 frente a 17,13), y otro, centrado en los hombres, referido al *tipo de centro*, que refleja que los hombres asistentes a centros privados cometen un promedio de error más elevado que los asistentes a centros públicos (20,55 frente a 16,19).

GRÁFICO 1
Análisis CHAID



Puede observarse la ausencia del nivel sociocultural, y de la población, lo cual puede ser debido a que la parte de información discriminante que aportan quede parcialmente solapada por las variables presentes en el modelo descrito. El hecho de que en el gráfico 1 aparezca un total de 398 informantes (n=398) en lugar del total de 400 expuesto en la metodología se debe a la pérdida de datos o ausencia de respuesta, cuestión que no afecta a la fiabilidad de los resultados obtenidos.

4. Conclusión

Una de las líneas prospectivas de la disponibilidad léxica es el análisis pormenorizado de los errores ortográficos que consignan las encuestas realizadas. Por ello, desde el estudio inicial de Paredes (1999) son varios los autores mencionados que se han preocupado por los aspectos ortotipográficos. Estas investigaciones, si bien han constituido un valioso intento de avance disciplinar, tan solo aportan una visión descriptiva de la situación.

El presente estudio ha pretendido avanzar y profundizar en este tipo de investigaciones: partiendo del análisis de una primera discriminación de informantes en relación a la distri-

bución de palabras con y sin error registradas por informante —como ocurría en las investigaciones precedentes sobre disponibilidad léxica y ortografía (Paredes, 1999; Ortolano, 2005; Ávila, 2007, y Santos, 2015)—, se consignó un escaso número de encuestas sin ningún error (n=8), seis mujeres y dos hombres. Es por ello que fue necesario realizar un análisis estadístico bivariante, para el contraste de hipótesis y sus elementos asociados al *contraste chi-cuadrado*, y otro más avanzado multivariante y minería de datos que se centraría en caracterizar los segmentos de estudios que fueron caracterizados desde las submuestras de informantes.

Estos análisis nos han permitido desentrañar el comportamiento de cada sociolecto y corroborar la hipótesis acerca de que las mujeres escriben más y mejor que los hombres, independientemente del contexto analizado y del enfoque adoptado tanto en lengua materna (Valencia, 1997; Reyes, 1999; Gómez Devís, 2004; Lagüéns, 2008; Trigo y González, 2011; Sandu, 2012; Lugones, 2015, y Pacheco y otros, 2017) como en lengua extranjera (Carcedo, 1998, 2000; Samper Hernández, 2002; Jiménez y Ojeda, 2009; Agustín y Fernández Fontecha, 2014).

Además, hemos podido desmontar prejuicios altamente arraigados en nuestra sociedad, como el hecho de que en las escuelas privadas se aprende más y mejor que en las públicas (Fernández Llera y Muñiz, 2012), ya que nuestros resultados reflejan que los estudiantes hombres matriculados en centros privados, aun contando con condiciones socioculturales más favorables, consignan un mayor promedio de errores ortográficos que los matriculados en centros públicos.

El análisis pormenorizado de cada subgrupo social nos permitirá tomar decisiones de intervención didáctica centradas en la diversidad del hablante y nos abrirá una serie de líneas futuras de investigación desde diferentes disciplinas. En el ámbito de la didáctica de la lengua, el estudio detenido de la tipología de error consignada permite la realización de inventarios cacográficos con las palabras que componen la esfera léxica usual de los informantes (Trigo y otros, 2018) que sirvan de base en el diseño de aplicaciones digitales para trabajar la ortografía desde procesos inductivos y la intervención, desde las administraciones educativas, en la implantación de proyectos lingüísticos de centro que permitan mejorar el comportamiento lingüístico de los estudiantes en general y los aspectos ortotipográficos en particular. Desde la psicolingüística, a través del estudio de los subcentros y *clusters* detectados se podrá no solo conocer cuáles son las relaciones de palabras más frecuentes (Ferreira y Echeverría, 2010, y Santos, 2017), sino también qué tipo de fenómeno asociativo ha llevado a la mente una grafía u otra.

Además, esta investigación supone una nueva línea de avance metodológico en los estudios de disponibilidad léxica pues, hasta la fecha, no se ha trabajado con los estadísticos anteriormente detallados. Sería muy interesante, dada la existencia del PPHDL, que otros investigadores continuaran con esta línea de investigación. Esta cuestión permitiría la realización de comparaciones ulteriores y el afianzamiento o rechazo de las hipótesis de partida.

Paredes (2012) daba cuenta de los desarrollos teóricos y metodológicos de la disponibilidad léxica y la situaba como una disciplina ampliamente consolidada y con una ingente proliferación de enfoques. Concluía augurando un futuro muy prometedor a este tipo de estudios. En este sentido, consideramos que la presente investigación contribuye a dar veracidad al mencionado pronóstico.

5. Bibliografía citada

Agustín, María del Pilar, y Almudena Fernández Fontecha, 2014: "Lexical Variation in Learners' Responses to Cue Words: The Effect of Gender" en Rosa María Jiménez (ed.): *Lexical Availability in English and Spanish as a Second Language*, Dordrech: Springer Netherlands, 69-81.

Ávila, Antonio Manuel, 2007: "Léxico disponible y ortografía. Condicionantes sociales y hábitos culturales de influencia" en Juan Antonio Moya y Marcin Sosisnky (eds.): *Las hablas andaluzas y la enseñanza de la lengua. Actas de las XII Jornadas sobre la enseñanza de la lengua española*, Granada: Universidad de Granada, 25-47.

Ávila, Antonio Manuel, y Juan Andrés Villena (eds.), 2010: *Variación social del léxico disponible en la ciudad de Málaga*, Málaga: Editorial Sarriá.

Blanco, Marta, 2011: "La ortografía en el léxico disponible del español de Galicia" en Belén López Meirama (ed.): *Estudios sobre disponibilidad léxica en el español de Galicia*, Santiago de Compostela: Universidad de Santiago de Compostela, 189-216.

Carcedo, Alberto, 1998: "Sobre las pruebas de disponibilidad léxica para estudiantes de español LE", *RILCE* 142, 205-224.

Carcedo, Alberto, 1999: "Análisis de errores léxicos del español en la interlingua de los finlandeses" en Tomás Jiménez, María del Carmen Losada y José Francisco Márquez (eds.): *Español como lengua extranjera: enfoque comunicativo y gramática. Actas del IX Congreso Internacional de ASELE*, Santiago de Compostela: Universidad de Santiago, 465-472.

Carcedo, Alberto, 2000. *Disponibilidad léxica en español como lengua extranjera: el caso finlandés (estudio de nivel preuniversitario, y cotejo con tres fases de adquisición)*, Turku: Turun Yliopisto.

Fernández Llera, Roberto, y Manuel Muñiz, 2012: "Colegios concertados y selección de escuela en España: un círculo vicioso", *Presupuesto y Gasto Público* 67, 97-112. [https://goo.gl/7kH6Zd, fecha de consulta: 28 de febrero de 2018].

Fernández Smith, Gérard, Ana María Rico y María José Molina, 2008: *Léxico disponible de Melilla: estudio sociolingüístico y repertorios léxicos*, Madrid: Arco/Libros.

Ferreira, Roberto, y Max S. Echeverría, 2010: "Redes semánticas en el léxico disponible de inglés L1 e inglés LE", *Onomázein* 21, 133-153.

Frey, María Luisa Helen, 2007: "Disponibilidad léxica y escritura del español como lengua extranjera: propuesta de comparación de dos corpus", *Interlingüística* 17, 366-373.

Galloso, María Victoria, 2003: *El léxico disponible de Ávila, Salamanca y Zamora*, Burgos: Fundación Instituto Castellano y Leonés de la Lengua.

García Casero, María José, 2013: *El léxico disponible en estudiantes de 4.º de Educación Secundaria Obligatoria en Santander*. Tesis de doctorado, Universidad de Santander.

García Marcos, Francisco, 2009: *Aspectos de historia social de la lingüística I. De Mesopotamia al Siglo XXI*, Barcelona: Octaedro.

Gómez Camacho, Alejandro, 2006: "Los inventarios cacográficos en la enseñanza de la ortografía", *Escuela Abierta* 9, 63-74.

Gómez Devís, María Begoña, 2004: *La disponibilidad léxica de los estudiantes preuniversitarios valencianos: reflexión metodológica, análisis sociolingüístico y aplicaciones*. Tesis de doctorado, Universidad de Valencia.

Gougenheim, George, René Michea, Paul Rivenc y Aurélien Sauvageot, 1956: *L'élaboration du français élémentaire (I Degrée). Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris: Didier.

Gougenheim, George, René Michea, Paul Rivenc y Aurélien Sauvageot, 1964: *L'élaboration du français fondamental (I Degrée). Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris: Didier.

Herrera, Honesto, Rosario Martínez y Marian Amengual, 2011: *Estadística aplicada a la investigación lingüística*, Madrid: EOS.

Hidalgo, Matías, 2017: *La disponibilidad léxica como método de detección del vocabulario y de su selección en manuales: Aplicación en una muestra de estudiantes sinohablantes de ELE*. Tesis doctoral inédita, Universidad de Jaén.

Jiménez, Rosa María, y Julieta Ojeda, 2009: "Girls' and Boys' lexical availability in English as a foreign language", *ITL International Journal of Applied Linguistics* 158, 57-76.

Lagüéns, Vicente, 2008: "La variable sexo en el léxico disponible de los jóvenes aragoneses" en María Luisa Arnal (ed.): *Estudios sobre la disponibilidad léxica de los jóvenes aragoneses*, Zaragoza: Institución Fernando El Católico, 103-162.

López Chávez, Juan, y Carlos Strassburger Frías, 1987: "Otro cálculo del índice de disponibilidad léxica. Presente y perspectivas de la investigación computacional en México" en *Actas del IV Simposio de la Asociación Mexicana de Lingüística Aplicada*, México: Universidad Nacional Autónoma de México.

López Morales, Humberto, 1995: "Los estudios de disponibilidad léxica. Pasado y presente", *Boletín de Filología de la Universidad de Chile (homenaje a Rodolfo Oroz)* 35, 245-259.

Lugones, Ana, 2015: *El léxico disponible de los alumnos de secundaria bilingüe (español e inglés) en Salamanca*. Tesis de doctorado, Universidad de Salamanca.

Mairal, Ricardo, Sandra Peña, Francisco José Cortés y Francisco José Ruiz de Mendoza, 2010: *Teoría lingüística. Métodos, herramientas y paradigmas*, Madrid: Editorial Universitaria Ramón Areces.

Mariscal, Alicia María, 2017: *Análisis de errores ortográficos (inglés/español) en estudiantes de educación secundaria en una zona de contacto lingüístico*. Tesis de doctorado, Universidad de Cádiz.

Moreno, Francisco, José Enrique Moreno y Antonio García, 1995: "Cálculo de disponibilidad léxica. El programa LexiDisp", *Lingüística* 7, 243-249.

Ortolano, Bárbara, 2005: "Estudios de disponibilidad léxica sobre una muestra de alumnos de Ayamonte (Huelva)", *Tonos. Revista electrónica de Estudios Filológicos* IX [https://goo.gl/gsAaAJ, fecha de consulta: 24 de enero de 2018].

Pacheco, Carmen Rosa, Juan Silvio Cabrera e Iselys González, 2017: "Incidencia de la variable 'sexo' en la disponibilidad léxica de estudiantes de preuniversitario en Pinar del Río, Cuba", *Íkala. Revista de lenguaje y cultura* 22 (2), 237-253.

Paredes, Florentino, 1999: "La ortografía en las encuestas de disponibilidad léxica", *Revista Estudio Adquisición de la Lengua Española (REALE)* 11, 75-98.

Paredes, Florentino, 2012: "Desarrollos teóricos y metodológicos recientes de los estudios de disponibilidad léxica", *Revista Nebrija de Lingüística Aplicada* 11 [https://goo.gl/s7yFih, fecha de consulta: 28 de febrero de 2018].

Reyes, María Josefa, 1999: "Acerca de la relación entre la variable sexo y el aprendizaje léxico" en Julián de las Cuevas y Dalila Fasla (eds.): *Contribuciones al estudio de la Lingüística Aplicada, Logroño: Asociación Española de Lingüística Aplicada*, 387-391.

Romero, Manuel Francisco, y Ester Trigo, 2015: "Herramientas para el éxito", *Cuadernos de Pedagogía* 458, 16-21.

Romero, Manuel Francisco, y Ester Trigo, 2018: "Los proyectos lingüísticos de centro. Desarrollar la comprensión lectora en áreas no lingüísticas", *Textos. Didáctica de la Lengua y la Literatura* 79, 51-59.

Romero, Manuel Francisco, Ignacio Valdés y José Ramón Romero, 2018: "Aplicaciones móviles y ortografía. Innovando desde la tradición", *Aula de Secundaria* 25, 37-40.

Samper Padilla, José Antonio, 1998: "Criterios de edición del léxico disponible: sugerencias", *Lingüística* 10, 311-333.

Samper Hernández, Marta, 2002: *Disponibilidad léxica en alumnos de español como lengua extranjera*, Colección Monografías 4, Málaga: ASELE.

Sánchez-Saus, Marta, 2016: *Léxico disponible de los estudiantes de español como lengua extranjera en las universidades andaluzas*, Sevilla: Universidad de Sevilla.

Sandu, Blanca, 2012: "La disponibilidad léxica en alumnos rumanos de ELE: incidencia de la variable 'sexo/género' y su correlación con el 'nivel escolar'", *Lingua Americana,* año XVI 31, 61-85 [https://goo.gl/BrPCq6, fecha de consulta: 28 de febrero de 2018].

Santos, Inmaculada Clotilde, 2015: *Evaluación de la competencia léxica bilingüe en estudiantes del Máster Universitario en Profesorado. Análisis de pruebas de disponibilidad léxica y de identificación de tecnicismos en español, inglés y francés*. Tesis de doctorado, Universidad de Málaga.

Santos, Inmaculada Clotilde, 2017: "Organización de las palabras en la mente en lengua materna y lengua extranjera (inglés y francés)", *Pragmalingüística* 25, 603-617.

Saura, José Antonio, 2008: "La ortografía en las encuestas aragonesas de disponibilidad léxica" en María Luisa Arnal (ed.): *Estudios sobre disponibilidad léxica en jóvenes aragoneses*, Zaragoza: Institución Fernando el Católico, 196-206.

Trigo, Ester, 2011: *Dialectología y cultura. El léxico disponible de los preuniversitarios sevillanos*, Valencia: Aduana vieja.

Trigo, Ester, y Adolfo Emilio González, 2011: "Estudio del comportamiento de la variable sexo en el léxico disponible de los preuniversitarios sevillanos", *Diálogo de la Lengua* 3, 28-41.

Trigo, Ester, Manuel Francisco Romero e Inmaculada Clotilde Santos, 2018: "Elaboración de un corpus cacográfico desde la disponibilidad léxica en estudiantes sevillanos. Un análisis para la enseñanza de la lengua", *Revista de Lingüística y Lenguas Aplicadas* 13, 119-131.

Trujillo, Fernando, y Raúl Rubio, 2014: "El PLC como respuesta sistémica al reto de la competencia comunicativa en entornos educativos formales: propuestas de análisis de casos", *Lenguaje y Textos* 39, 29-38.

Valdés, Ignacio, y Manuel Francisco Romero, 2017: "Apostando por metodologías activas. Artefactos digitales para la enseñanza de la ortografía", *Publicaciones Didácticas* 85, 134-139.

Valencia, Alba, 1997: "Disponibilidad léxica. Muestreo y estadísticos", *Onomázein* 2, 197-226.