Estudios de Lingüística • Universidad de Alicante
ISSN 2171-6692

Núm. 43, 2025, pp. 161-181
https://doi.org/10.14198/ELUA.26993

Longitud léxica y frecuencia de uso en el español contemporáneo: un análisis estadístico de corpus

Lexical length and frequency of use in contemporary Spanish: a statistical corpus analysis

Resumen

En esta investigación se aborda la relación inversa que se produce en español entre la longitud léxica y la frecuencia de uso considerando los procesos de formación de palabras. Este objetivo central se aborda al tener en cuenta la estructura (derivación) de las palabras, para lo cual se analiza cuantitativamente la relación entre la longitud de las palabras, la frecuencia de uso, la formación de palabras y el estilo de escritura. Se revisan los análisis previos, basados exclusivamente en el cómputo de caracteres (Takefuta 1981; Yoshioka 1996), y también aquellos estudios que consideran la sílaba como unidad de medida de longitud (Herdan 1956; Gómez Guinovart 1999). En ambos casos se muestran sus carencias a través del análisis estadístico. Cuando la longitud de la palabra se mide utilizando el número de sílabas, el primer término de la distribución de frecuencia, las palabras de una sílaba, se convierten en un grupo que incluye palabras de distinto número de fonemas. De esta forma, el uso de unidades de medida aproximadas, como las sílabas, da como resultado observaciones aproximadas que no proporcionan una imagen precisa de la situación. Los resultados apuntan a que la longitud de las formas debe observarse desde una perspectiva no física, sino lingüística. Se considera en este estudio que la longitud lingüística de una palabra debe medirse desde la perspectiva de la morfología derivacional teórica y práctica, es decir, mediante prefijos y sufijos, que son unidades de morfología derivada. Lo que hace posible este tipo de análisis es, en definitiva, el estudio de frecuencia de los afijos. La metodología que se sigue es la correspondiente al análisis estadístico con medidas básicas como la distribución de frecuencia, la desviación estándar y otras fórmulas de creación propia en el programa R (*R Core Team* 2021). Asimismo, empleamos *ggplot2* (Wickham 2016) para crear gráficos. Los textos sometidos a análisis forman parte de un corpus del español contemporáneo hablado y escrito reunido *ad hoc* para el estudio. Esta selección pretende aportar una solución general a una variedad de materiales y no una visión particular de un único material, lo que da como resultado una conclusión no definitiva, pero sí más fiable.

Autoría

Inmaculada Martínez Martínez
Universidad de Cantabria, España
Inmaculada.martinez@unican.es
https://orcid.org/0000-0003-4760-0903

Hiroto Ueda
The University of Tokyo, Japan
hiroto.ueda.tokio@gmail.com
https://orcid.org/0000-0003-3204-609X

Para citar este artículo:
Martínez Martínez, I.; Ueda, H. (2025). Longitud léxica y frecuencia de uso en el español contemporáneo: un análisis estadístico de corpus, *ELUA*, 43, 161-181. https://doi.org/10.14198/ELUA.26993

Recibido: 05/02/2024
Aceptado: 16/04/2024

Conflicto de intereses: los autores declaran que no hay conflicto de intereses.

Palabras clave:
longitud léxica; frecuencia de uso; formación de palabras; análisis estadístico; español contemporáneo; longitud lingüística.

Abstract

This research addresses the inverse relationship that occurs in Spanish between lexical length and frequency of use, considering the word formation processes. This central objective is addressed by taking into account the structure (derivation) of words, for which the relationship between word length, frequency of use, word formation and writing style is quantitatively analyzed. Previous analyzes are reviewed, based exclusively on the counting of characters (Takefuta 1981; Yoshioka 1996) and also those studies that consider the syllable as a unit of length measurement (Herdan 1956; Gómez Guinovart 1999). In both cases their shortcomings are shown through statistical analysis. When word length is measured using the number of syllables, the first term of the frequency distribution, one-syllable words, becomes a group that includes words of various numbers of phonemes. In this way, the use of approximate units of measurement, such as syllables, results in approximate observations that do not provide an accurate picture of the situation. The results suggest that the length of the shapes should be observed from a linguistic but not a physical perspective. In this study, it is considered that the linguistic length of a word should be measured from the perspective of theoretical and practical derivational morphology, that is, through prefixes and suffixes, which are units of derivational morphology. What makes this type of analysis possible is, ultimately, the study of the frequency of affixes. The methodology followed is that corresponding to statistical analysis with basic measures such as frequency distribution, standard deviation, and other self-created formulas in the R program (R Core Team 2021), such as concentration analysis. Likewise, we used ggplot2 (Wickham 2016) to create graphs. The texts subjected to analysis are part of a corpus of contemporary spoken and written Spanish assembled ad hoc for the study. This selection is intended to provide a general solution to a variety of materials and not a particular vision of a single material, which results in a conclusion that is not definitive, but more reliable.

Keywords:
lexical length; frequency of use; formation processes; statistical analysis; contemporary Spanish; linguistic length.

1. INTRODUCCIÓN

Un aspecto que se aborda con frecuencia en la investigación lingüística cuantitativa es la longitud de las palabras. Generalmente, se dice que, cuanto más corta es una palabra, mayor es su frecuencia y cuanto más larga, menos frecuente es su uso. Aunque este parecería, *a priori*, un hecho evidente, se requiere una revisión desde una perspectiva científica, pues estamos ante un constructo esencial, también para la estilometría como herramienta de cuantificación estadística (Blasco Pascual y Ruiz Urbón 2022). Esta considera la longitud de las palabras como una de las características estilísticas de los textos, puesto que las palabras largas se consideran lingüística y estilísticamente marcadas si se usan con menos frecuencia.

Takefuta (1981, pp. 179-182) midió la longitud media de las palabras inglesas en el vocabulario utilizado (número de palabras expresadas, *token*)[1] en varios materiales, siempre conforme al número de caracteres. Sus datos establecían la media entre 4,0 y 5,6 caracteres. El autor añade en su estudio que este criterio no se debe ignorar, ya que las palabras con más letras pueden considerarse más difíciles que las palabras con menos letras. En un estudio realizado por Yoshioka (1996, p. 201) también para el inglés, los valores medios de letras en dos artículos

1 En cuanto al número de palabras, distinguimos entre *token* y *type*. Por ejemplo, el número de *token* en el conjunto de palabras {a, a, a, b, b, c} es 6, mientras que *type* es 3 {a, b, c}. En este estudio emplearemos *frecuencia de uso* como sinónimo de *token* y *número de palabras* o lemas referido a *type*.

periodísticos, dos novelas y dos libros de texto de Secundaria fueron 5,2, 5,0; 4,2, 4,3; 3,9, 4,4 letras, respectivamente.

En el caso del español, partimos de varios estudios (Juilland y Chang-Rodríguez 1964; Justicia 1995; Ueda 2021) para examinar el número promedio de caracteres en las palabras, junto al número promedio de caracteres en el vocabulario (número total de palabras). Los resultados fueron 7,4, 7,4 y 7,3 caracteres, respectivamente. Por tanto, es un dato contrastado el que las palabras en español tienen más caracteres (7 caracteres) que las palabras en inglés (4-5 caracteres).

Las razones de la mayor longitud de las palabras en español pueden estribar en el hecho de que se ha suprimido el acortamiento debido al debilitamiento fonológico y a que el proceso de formación de palabras es productivo. Más adelante señalamos (§2) que las palabras se acortan cuando se usan con demasiada frecuencia, pero esto no sucede frecuentemente en español. Ocurre, más bien, que el proceso de formación de palabras utiliza una rica variedad de prefijos y sufijos, crea nuevas palabras y el resultado son palabras relativamente largas que se usan de manera consistente y sin acortamientos.

Si continuamos con el análisis comparativo entre el inglés y el español, por ejemplo, *en.ton.ces* es una palabra de un solo morfema sin prefijo ni sufijo, y no hay palabras derivadas basadas en ella, por lo que es una palabra aislada. En inglés, *then* constituye una sílaba formada por cuatro letras, mientras *entonces* tiene tres sílabas y ocho letras. Aunque es una palabra de alta frecuencia, no hay señales de que se haya acortado en absoluto y la forma completa sigue intacta. Sin embargo, las palabras largas aisladas como *entonces* son excepcionales; generalmente, la palabra simple es corta y las palabras largas son formas derivadas por prefijos o sufijos, entre otros. Según el *Diccionario de la lengua española* de la Real Academia Española (DLE, 23.ª ed.), la palabra más larga en español es *electroencefalografista* (23 letras), pero se analiza como *electro.encefalo.graf.ista*. A partir de *encéfalo* 'cabeza, cerebro', muchas son las palabras derivadas (*cefalalgia, cefalitis, cefalópodo, hidrocefalia*, etc.) y, por tanto, se pueden aprender a partir de ella sin dificultad. También se pueden descomponer *anti.con.stitu.cion.al.idad* (22 caracteres) y *anti.norte.americ.an.ismo* (21 caracteres) que se sitúan en longitud detrás de *electroencefalografista*.

En el presente estudio abordamos el cálculo de la longitud de palabras en español a través del número de caracteres y de sílabas. Examinamos la relación inversa entre la longitud de la palabra y su frecuencia de uso, y la razón de la llamada "Ley de acortamiento" (Zipf 1936, 1949), según la cual la longitud de la palabra se acorta cuando la palabra se utiliza con más frecuencia. Se considera, asimismo, la estructura (derivación) de las palabras y se analiza cuantitativamente la relación entre la longitud de las palabras, la frecuencia de uso, la formación de palabras y el estilo de escritura. Se analizan con detalle, asimismo, los procesos de variación, para disponer de un análisis más completo. Los textos sometidos a análisis forman parte de un corpus del español contemporáneo hablado y escrito reunido *ad hoc* para el estudio que se detallará más adelante (§3). El análisis utiliza la herramienta R (R *Core Team* 2021). Los gráficos se crean a partir de *ggplot2* (Wickham 2016) y, por último, se emplean funciones propias que aparecen detalladas al final del estudio (§6).

2. ESTUDIOS ANTERIORES

Debido a las escasas referencias sobre la longitud léxica encontradas para el caso del español y por tratarse de un tema de lingüística general, mencionamos trabajos importantes sobre el inglés y el japonés, que conocemos de cerca.

Se pueden considerar los caracteres, las sílabas y los fonemas como las unidades utilizadas para medir la longitud de las palabras en español (Herdan 1956, pp. 176-197; Gómez Guinovart 1999, pp. 97-113). La longitud léxica medida a partir del número de letras no es exacta y plantea ciertos problemas: dos letras pueden representar un sonido (*ch*ico, *qu*e, *ll*orar, pe*rr*o), algunos caracteres son mudos (*h*ombre, *p*sicología, *m*nemotecnia) y, por último, una letra puede representar uno o dos sonidos (e*x*tremo, e*x*amen). Con el objetivo de solventar estos obstáculos, muchos estudios utilizan las

sílabas como unidades (Navarro Tomás 1966, pp. 54-55; Kabashima 1968, p. 16; Miller 1979, p. 108). Sin embargo, esta medición no está exenta de dificultades tampoco, pues *a, de, dos* y *tres* son palabras de una sílaba, y *ele, paso, vista, puesto, muestra* y *monstruo* son palabras de dos sílabas, pero la longitud de las palabras difiere sustancialmente. Por tanto, las unidades más precisas parecen ser los fonemas: *a* (1 fonema), *de* (2), *dos* (3), *tres* (4); *ele* (3), *paso* (4), *vista* (5), *puesto* (6), *muestra* (7), *monstruo* (8). En esta línea del análisis, se sitúan Saporta (1963, p. 69), quien midió el número de fonemas que compone un morfema; también Wierzbicka (1967 [2011], pp. 211-213), que calculó la longitud de las palabras por el número de fonemas; sin embargo, esta autora, en lugar de comparar la frecuencia de palabras individuales, dividió la frecuencia de uso en tres grupos discontinuos (I: rango 45-54, II: rango 295-304, III: rango 2495-2504), a la vez que realizaba comparaciones entre dichos grupos.

Algunos trabajos en torno a distintas lenguas analizan la relación entre la longitud de las palabras y la frecuencia de uso de estas. Coinciden en afirmar que la frecuencia de uso disminuye a medida que aumenta la longitud léxica (Zipf 1936, pp. 3-39; Zipf 1949, pp. 63-66; Whatmough 1956 [1960], p. 69; Wierzbicka 1967 [2011], pp. 211-226; Kabashima 1968, pp. 15-19; Martinet 1970 [1972], pp. 258-263; Miller 1951 [1979], pp. 107-109; Yasumoto 2009, p. 255; Tanaka 2021, pp. 168-174). Como factores causantes de este hecho, se abordan principios como la ley del mínimo esfuerzo, la economía y el coste.

En otras palabras, el hecho de que las palabras cortas aparezcan con mayor frecuencia podría deberse a la llamada economía de la lengua, que centraría el uso en el menor esfuerzo cognitivo. Según este principio, si una palabra frecuente es larga, esta se verá acortada, pues, de lo contrario, resultará más complejo su uso. Se trata de la *Ley de acortamiento* que Zipf (1949, p. 66) ejemplifica para el inglés (*telephone* > *phone, gasoline* > *gas, omnibus* > *bus)*, mientras Martinet (1970 [1972], p. 261) lo hace para el francés (*chemin de fer métropolitain* > *métro*). Por el contrario, es menos probable que se utilicen palabras largas, porque son más difíciles de procesar (Zipf 1949; Whatmough 1956; Martinet 1970). Kabashima (1968, pp. 15-19), que analizó materiales procedentes del informe del Instituto Nacional de Lengua y Lingüística Japonesa, *Terminología utilizada en noventa revistas modernas*, afirmaba que[2]:

> durante la larga historia del uso del lenguaje humano, se supone que existe la tendencia a la supervivencia de las palabras cortas para aligerar el esfuerzo en la actividad lingüística. (…) hay un efecto de selección en el que las palabras con formas cortas sobreviven como palabras de uso frecuente, mientras que nuevos conceptos, cosas, sistemas o conceptos complejos suelen ser expresados con palabras largas, puesto que los elementos se combinan para formar una palabra composicionalmente larga.

Otro autor japonés como Yasumoto (2009, p. 255) ahonda en este aspecto al afirmar: "En términos generales, en inglés, las palabras que se usan con frecuencia tienden a tener longitudes cortas, y las palabras con múltiples sílabas son términos técnicos o palabras difíciles". Toma, a modo de ejemplo, palabras con prefijos o sufijos que tienen una gran cantidad de sílabas, como *reconsideration*.

Por su parte, Bybee (2007, p. 260) señala que la "reducción del gesto articulatorio" que se observa en las palabras de alta frecuencia se debe a la "automatización normal de la actividad motora". Sin embargo, Tanaka (2021, p. 172) afirma rotundamente que "es cuestionable si la relación entre la longitud de las palabras y la frecuencia de uso representa la economía humana", e indica, más bien, que las palabras se hacen largas por razones lingüísticas, es decir por la estructura composicional de la palabra. Divjak (2019, pp. 33-36), por último, aborda el hecho de que muchos lingüistas han expresado dudas sobre el principio de economía y la ley del mínimo esfuerzo de Zipf (1936) arriba señalada.

El trabajo de Urrutibéheity (1972) sobre el léxico español es sugerente, en el sentido de que ha demostrado la relación que existe entre el uso (*usage* de Juilland y Chang-Rodríguez

2 La traducción española de textos japoneses e ingleses, de aquí en adelante en este estudio, es nuestra.

1964: frecuencia y dispersión) y el carácter funcional, la longitud léxica, la documentación antigua y la herencia latina. El autor nos anticipa con su análisis una posible relación de la longitud léxica con los factores restantes (Divjak 2019, p. 34).

Finalmente, no debemos olvidar la precisión de Kin (2018, p. 42) de que, en el caso del japonés, hay muy pocas palabras cortas. Con respecto a la longitud de las formas de las palabras japonesas, se dice que el promedio más común está en torno a 4 ± 1 moras, como unidad que mide el peso silábico, y se considera estable. Una o dos moras es una proporción demasiado pequeña para formar muchas palabras, y algo más de seis o siete moras resulta demasiado larga para soportar el coste cognitivo y la capacidad de retención, lo que hace que la frase sea larga.

En los desarrollos y análisis que se suceden a continuación, se toman en consideración estos estudios previos hasta aquí abordados.

3. EL CORPUS

El corpus constituido para el análisis del presente estudio aparece conformado por las siguientes obras españolas, detalladas en la Tabla 1:

Tabla 1. Composición de materiales

Datos	Frecuencia	Objetivo[3]
CORPES	352 899 385	500 000
Davies	17 326 289	250 000
GH	387 524	60 000
JCh	460 065	60 000
Moreno	685 905	60 000
Justicia	507 384	30 000
Santander	111 118	25 000
Manual	98 460	15 000
TOTAL	372 476 130	1 000 000

3 El "objetivo" en la Tabla 1 se refiere a la cantidad marcada como objetivo para constituir la proporción correspondiente de formas dentro de un millón. Por medio de estas cifras marcadas como objetivo, la totalidad original de palabras, más de 372 millones, se convierte en un millón.

A continuación, explicamos brevemente las características principales de cada material:

1. García Hoz (1953) (en adelante, GH) dividió en la década de 1940 la vida en cuatro ámbitos y recopiló 100 000 palabras de material lingüístico correspondientes a cada uno de estos ámbitos (GH, 20):

(a) Vida familiar: cartas privadas.

(b) Vida social indiferenciada: periódicos.

(c) Vida social regulada: documentos oficiales.

(d) Vida cultural: libros.

2. Juilland y Chang-Rodríguez (1964) recopilaron 100 000 palabras en cada uno de los cinco tipos de publicaciones que revisaron (obras dramáticas, novelas, ensayos, periódicos y revistas, textos científicos y técnicos) y que se publicaron en España en la primera mitad del siglo XX. Los autores ofrecen datos, tanto de la frecuencia de uso como de los lemas de las formas flexivas y no se incluyen palabras con una frecuencia de uso total de 4 o menos. De este corpus hemos extraído la frecuencia de los lemas y hemos calculado la distribución de frecuencia del vocabulario utilizado en cada publicación. El vocabulario utilizado en los cinco tipos de publicaciones muestra claramente las características de cada uno, convirtiéndolas en valiosos materiales estilísticos.

3. Ueda (1987) organizó a García Hoz (1953) y Juilland y Chang-Rodríguez (1964) y creó una tabla de frecuencias comparativas, y también añadió una tabla comparativa de frecuencias de palabras aparecidas en siete manuales de texto publicados en España, Francia y Estados Unidos (1957-1977). Se hizo una comparación con los datos de García Hoz (1953) y Juilland y Chang-Rodríguez (1964). De aquí usaremos las frecuencias de vocabulario de los libros de texto.

4. Justicia (1995) se conforma a partir de una lista de vocabulario estadístico de ensayos libres escritos en 20 minutos por 3402 alumnos de primaria (6-13 años) en Andalucía oriental (Almería, Granada, Jaén, Málaga). El vocabulario utilizado por los niños en cada ciclo de la educación primaria (A1:

6-7 años, A2: 8-10, A3: 11-13) se considera el material básico para estudiar el desarrollo del vocabulario, pero también es posible realizar una investigación estilística cuantitativa utilizando cada ciclo como variable.

5. Moreno *et al.* (2005) son grabaciones transcritas de discurso natural formal e informal de 500 hablantes. Además, se han agregado transcripciones de conferencias académicas y documentos de consultas de salud.

6. Davies (2006) muestra la frecuencia y difusión del vocabulario hablado, escrito, literario y periodístico para un total de 20 millones de palabras en España (43 %) y América Latina (57 %). Los materiales se limitan al período 1970-2000, siendo la mayoría de los materiales de la década de 1990.

7. Nuestro corpus PRESEEA-Santander (Martínez y Ueda 2021, 2023) contiene grabaciones de audio (tiempo de grabación superior a 45 minutos por entrevista) de 18 personas. La muestra está estratificada en torno a las variables de género (2: H, M), edad (3: E1, E2, E3) y nivel de educación (3: N1, N2, N3)[4]. La estructura general es la que se muestra en la Tabla 2 (Martínez y Ueda 2021).

Debido a la preestratificación del corpus, nos encontramos ante un material lingüístico ampliamente recogido que permite realizar comparaciones poco sesgadas. Usaremos este banco de datos para investigar la longitud y uso de las palabras del vocabulario utilizados por cada grupo.

8. Real Academia Española (2023). El *Corpus del español del siglo XXI* (CORPES) es el recurso más reciente (2023) y mayor (395 000 000 de palabras) que existe en la actualidad sobre el español escrito contemporáneo. Muestra los lemas, partes de la oración (categoría gramatical) y frecuencia del vocabulario utilizado en libros, periódicos, publicaciones periódicas e Internet publicados en España (30 %) y América Latina (70 %) en el siglo XXI (2001-2023).

Cabe señalar que se han empleado distintos materiales pertenecientes al registro oral y escrito, tal y como se hace constar ya en el resumen. La razón estriba en el hecho de que es nuestro deseo dar una solución general a una variedad de materiales y no una visión particular de solo un material. De esta manera, se obtiene una conclusión más fiable, aunque no definitiva, pues esto último resulta imposible.

4. LONGITUD Y FRECUENCIA

4.1. Análisis a partir del cómputo de sílabas

Como ya señalamos al comienzo (§1), el conjunto de los lemas constituye la lista completa de palabras disponibles (denominadas *types*), mientras que el uso real de las palabras difiere de esta lista (lo que se conoce como número total de palabras o *tokens*). Por ejemplo, al examinar el número

Tabla 2. Frecuencia de uso de palabras de la muestra PRESEEA-Santander

Sexo:	H: hombre			H.total	M: Mujer			M.total	Total
Edad: Nivel	E1	E2	E3		E1	E2	E3		
Nivel: N1	6380	6395	8160	20 935	6347	6971	5141	18 459	39 394
Nivel: N2	4250	8319	7633	20 202	3337	10 278	5459	19 074	39 276
Nivel: N3	7699	5437	5888	19 024	4343	3710	9352	17 405	36 429
Total	18 329	20 151	21 681	60 161	14 027	20 959	19 952	54 938	115 099

4 Sexo. H: Hombre / M: Mujer. Edad. E1 (20~34 años) / E2 (35~54 años) / E3 (55 años~). Nivel educativo. N1 (Educación básica ~10 años) / N2 (Educación secundaria ~16-18 años) / N3 (Educación superior ~21-22 años).

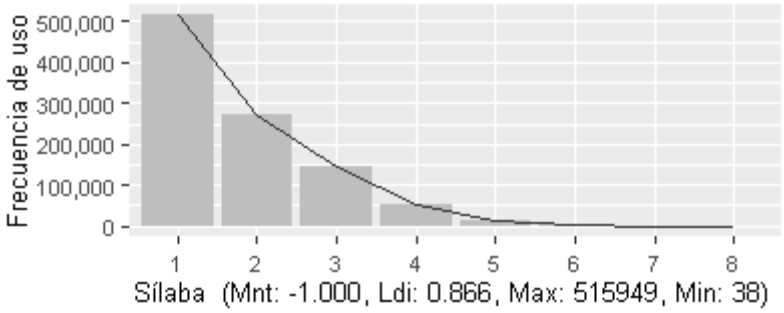

Fig. 1. X: Número de sílabas (S): Y: Frecuencia de uso

de lemas en los diccionarios, las preposiciones como *a, de, en, con* y *por* y las conjunciones como *y, e, que* se cuentan una vez cada una, pero varios diccionarios de frecuencia muestran que la frecuencia de uso de estas palabras es extremadamente alta. El análisis de nuestro corpus ofrece el número de veces que se usa cada lema, por lo que usamos esta cifra para calcular la frecuencia de uso de los lemas correspondientes a la longitud de la palabra (número de sílabas y fonemas). En primer lugar, observamos el número de sílabas a través de la Figura 1.

Al observar este gráfico, podemos ver que la frecuencia de uso disminuye a medida que aumenta el número de sílabas (S). En una primera aproximación, este hecho parece indicar la Ley de acortamiento, que establece que a medida que aumenta la longitud de una palabra, disminuye la frecuencia de su uso (§2).

En la figura anterior, los vértices de las barras se conectan para formar un gráfico lineal con el objetivo de mostrar que la frecuencia de uso disminuye monótonamente (no aumenta ni disminuye continuamente). Cuantificamos el grado de declive monótono y definimos el "índice de monotonía" (Mnt) (§6). El rango del índice de monotonía es [-1, 1], donde -1 indica *monótonamente decreciente* y 1 indica *monótonamente creciente*. *Ldi*, que se encuentra tras Mnt, es el índice de la distribución en forma de 'L'. Su rango es [0, 1], donde 1 es una distribución perfecta de la forma 'L'.

Para descubrir por qué el número de sílabas de una palabra y su frecuencia de uso tienen esta distribución en forma de 'L', observaremos el contenido de las palabras con cada número de sílabas[5].

(1) Palabras monosílabas de alta frecuencia (en orden descendente de frecuencia de uso): *el*.art 118 997, *de*.prep 76 116, *que*.conj 36 827, *y*.conj 32 648, *a*.prep 30 109, *en*.prep 28 910, *un*.art 22 875, *ser*.vb 21 015, *se*.pro 17 597, *no*.sn 13 365, *con*.prep 11 419, *su*.pos 10 826, *por*.prep 10 656, *lo*.pro 10 348, *más*.ind 5307, *me*.pro 5206, *ir*.vb 5123, *o*.conj 4305, él.pro 3541, *si*.conj 2857...

(2) Palabras bisílabas de alta frecuencia (en orden descendente de frecuencia de uso): *para*.prep 7918, *haber*.vb 7382, *este*.dem 6284, *estar*.vb 5950, *tener*.vb 5901, *como*.conj 5796, *hacer*.vb 4741, *todo*.ind 4643, *poder*.vb 4511, *ese*.dem 4472, *pero*.conj 4297, *decir*.vb 4197, *otro*.ind 3249, *mucho*.ind 2257, *porque*.conj 2163, *año*.sus 2158, *cuando*.conj 1996, *solo*.adv 1995, *también*.adv 1955, *saber*.vb 1954...

(3) Palabras trisílabas de alta frecuencia (en orden descendente de frecuencia de uso): *alguno*.ind 1622, *primero*.adj 1537, *ahora*.adv 1111, *parecer*.vb 1011, *entonces*.adv 953, *encontrar*.vb 936, *persona*.sus 842, *conocer*.vb 802, *momento*.sus 760, *último*.adj 724, *durante*.prep 697, *además*.adv 679, *trabajo*.sus 667, *problema*.sus 611, *esperar*.vb 551, *ejemplo*.sus 543, *amigo*.sus 538, *nacional*.adj 527, *público*.adj 523, *nosotros*.pro 521...

5 Categoría gramatical: adj: adjetivo, adv: adverbio, art: artículo, cif: cifra, conj: conjunción, dem: demostrativo, ind: indefinido, extr: extranjerismo, interj: interjección, interrog: interrogación, num: numeral, pos: posesivo, prep: preposición, pro: pronombre, prop: nombre propio, rel: relativo, sn: sí-no, sus: sustantivo, vb: verbo.

Al observar la lista anterior, podemos ver que la mayoría de las palabras de una sílaba más frecuentes son palabras funcionales como artículos, pronombres, preposiciones y conjunciones[6]. Esas palabras en particular tienen una frecuencia extremadamente alta. Por lo tanto, es posible que la frecuencia de las palabras funcionales aumente la frecuencia de todas las palabras de una sílaba. Sin embargo, si hay una gran cantidad de palabras de contenido de baja frecuencia, esa gran cantidad de palabras de contenido también puede aumentar la frecuencia de las palabras de una sola sílaba en su conjunto. Así pues, debemos observar la distribución total de frecuencia de las palabras funcionales y de contenido según la longitud de la palabra. Lo haremos a través de la Figura 2.

El mismo gráfico muestra que la inmensa mayoría de monosílabos son palabras funcionales. Lo contrario es cierto para 2-3 sílabas, con más palabras de contenido, y para 4 o más fonemas, donde solo hay palabras de contenido. Si observamos la lista anterior, es cierto que las palabras de una sílaba son todas palabras funcionales, y no se encuentran palabras de contenido entre las palabras de alta frecuencia, y en las palabras de dos sílabas, hay palabras funcionales (*para, como, haber, pero*, etc.) y palabras de contenido (*vida, hombre, también, siempre*), y la mayoría de las palabras trisilábicas son palabras de contenido. Por lo tanto, la distribución de frecuencia sigue el siguiente orden: de palabras funcionales (pequeña cantidad de sílabas) → palabras de contenido (mayor cantidad de sílabas). Pensamos que esta es la razón por la cual la distribución de frecuencia disminuye monótonamente a medida que aumenta el número de sílabas.

Además, la característica distribución en forma de 'L' no muestra solo que es un descenso monótono, sino también que es un descenso pronunciado. Se cree que la razón por la cual la frecuencia de uso disminuye rápidamente con el número de sílabas de una palabra puede estribar en la extrema diferencia de frecuencia entre las palabras funcionales y las palabras de contenido.

La razón de esta distribución en forma de 'L' de la frecuencia de uso según el número

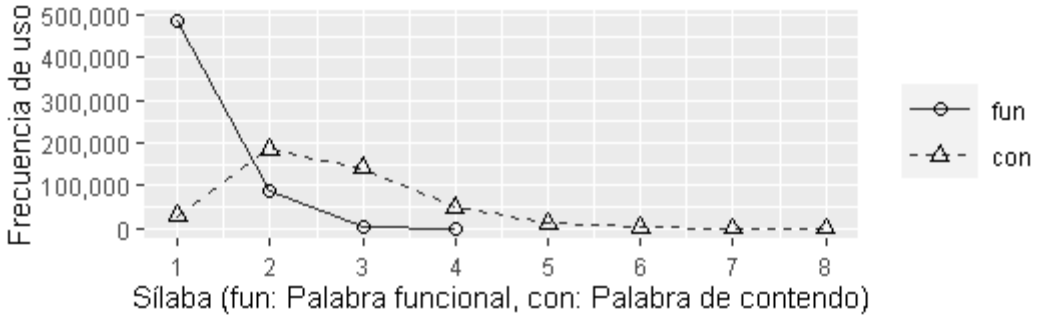

Fig. 2. Distribución de frecuencia por longitud de palabra (sílaba) de palabras funcionales y palabras de contenido

6 Palabras funcionales encontradas en el corpus: *a*.prep, *algo*.pro, *alguien*.pro, *alguno*.adj, *alguno*.pro, *ambos*.adj, *ambos*.pro, *ante*.prep, *aquel*.dem, *aunque*.conj, *bajo*.prep, *bastante*.pro, *cada*.adj, *como*.conj, *cómo*.interrog, *con*.prep, *conmigo*.pro, *consigo*.pro, *contigo*.pro, *contra*.prep, *cual*.rel, *cuál*.interrog, *cualquier*.adj.pro, *cuando*.conj, *cuándo*.interrog, *cuanto*.rel, *cuánto*.interrog, *cuyo*.rel, *de*.prep, *demás*.adj.pro, *demasiado*.pro, *desde*.prep, *donde*.conj.rel, *dónde*.interrog, *durante*.prep, *el*.art, *él*.pro, *en*.prep, *entre*.prep, *ese*.dem, *estar*.vb, *este*.dem, *haber*.vb, *hacia*.prep, *hasta*.prep, *lo*.pro, *mas*.conj, *más*.adj, *más*.adv, *me*.pro, *mediante*.prep, *menos*.adj.adv, *mi*.pos, *mí*.pro, *mientras*.conj, *mismo*.adj.pro, *mucho*.adj.adv.pro, *nada*.adv.pro.sus, *nadie*.pro, *ni*.adv.conj, *ninguno*.adj.pro, *no*.sn, *nos*.pro, *nosotros*.pro, *nuestro*.pos, *o*.conj, *ora*.conj, *os*.pro, *otro*.adj.pro, *pa*.prep, *para*.prep, *pero*.conj, *por*.prep, *por qué*.interrog, *porque*.conj, *pues*.conj, *que*.conj.rel, *qué*.interrog, *quien*.rel, *quién*.interrog, *salvo*.prep, *se*.pro, *según*.prep, *ser*.vb, *si*.conj, *sí*.sn, *sin*.prep, *sino*.conj, *siquiera*.adv.conj, *sobre*.prep, *su*.pos, *tal*.adj.pro, *también*.adv, *tampoco*.adv, *tanto*.adj.adv.pro, *te*.pro, *ti*.pro, *todo*.adj.pro, *tras*.prep, *tu*.pos, *tú*.pro, *un*.art, *uno*.pro, *usted*.pro, *vos*.pro, *vosotros*.pro, *vuestro*.pos, *y*.conj, *ya*.conj, *yo*.pro.

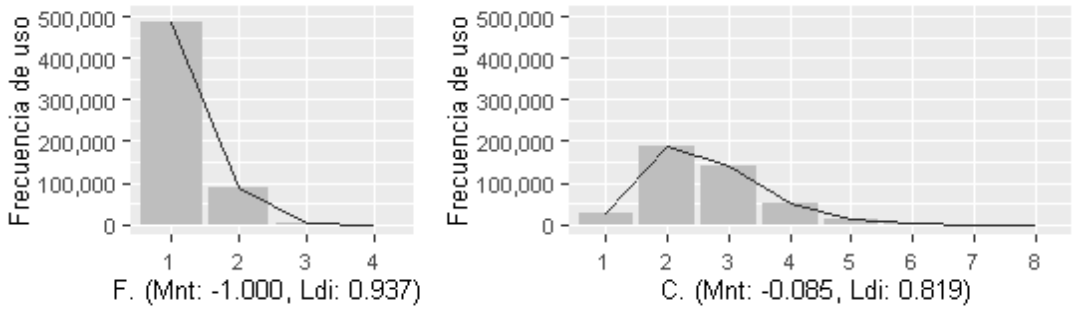

Fig. 3. X: Número de sílabas. Y: Frecuencia de uso.
F: Palabras funcionales, C: Palabras de contenido

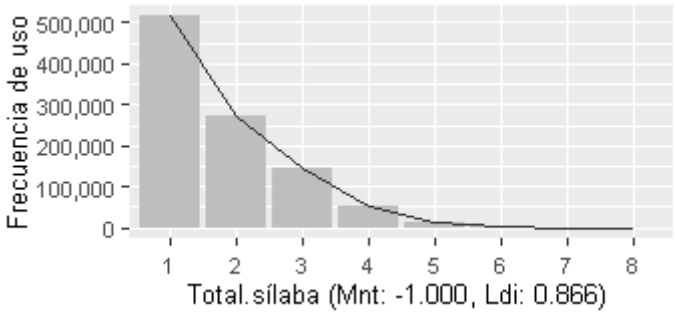

Fig. 4. X: Número de sílabas. Y: Frecuencia de uso. Total

de sílabas de una palabra parece ser la gran diferencia de frecuencia entre las palabras funcionales y las palabras de contenido. Si esto es así, para observar esta diferencia con mayor nitidez, dividimos en dos partes el análisis y observamos la distribución de frecuencia de cada una de manera separada a través de la Figura 3 y del total a través de la Figura 4.

Los tres gráficos anteriores muestran la distribución de frecuencia de F: palabras funcionales, C: palabras de contenido y T: la distribución total. La distribución de frecuencia de las palabras funcionales de 1 a 3 sílabas muestra una tendencia descendente completamente monótona (mnt: -1). Sin embargo, la distribución de frecuencia de las palabras de contenido no es una distribución en forma de 'L', sino que se aproxima a una distribución normal. Por su parte, la distribución total (Fig. 4) vuelve a descender monótonamente. Se puede confirmar aquí que la frecuencia de Total (todas las palabras de una sílaba) refleja la frecuencia de F (las palabras funcionales).

4.2. Análisis a partir del cómputo de fonemas

Muchos estudios que tratan la longitud de las palabras en inglés miden la longitud de las palabras usando sílabas, pero la longitud de las palabras se puede calcular con mayor precisión usando el número de fonemas en vez del número de sílabas, como ya señalamos al comienzo del estudio. A continuación, calculamos la longitud de las palabras en las figuras 5 y 6, usando el número de fonemas.

De esta forma, un hecho que no se puede apreciar al calcular la longitud de la palabra si se usa el número de sílabas queda claro cuando se calcula usando el número de fonemas. Es decir, no se observa una disminución en la frecuencia del primer término a partir de las palabras de un fonema (Total), sino un aumento en la frecuencia de palabras de un fonema a palabras de dos fonemas. Destaca la frecuencia de uso de palabras de dos fonemas en comparación con las palabras con otros números de fonemas cercanos (1, 3,

Fig. 5. Frecuencia de uso por longitud de palabra (fonema). Palabras funcionales (fun) y palabras de contenido (con)

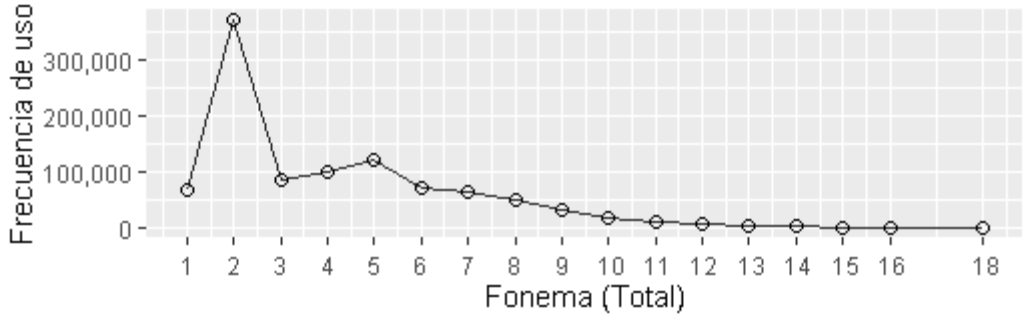

Fig. 6. Frecuencia de uso por longitud de palabra (fonema). Total

4, 5, ... fonemas). A esta distribución, donde destaca alguna parte específica, la llamaremos distribución en forma de 'I'. Esta es muy diferente de la distribución normal y de la de 'L'.

Analizamos ahora la razón por la cual la relación entre la longitud de la palabra medida por el número de fonemas y la frecuencia de uso tiene una distribución en forma de 'I'. El motivo puede apreciarse al observar el número de fonemas que se destaca en 2. La siguiente es una lista que muestra la frecuencia de las palabras con el fonema número 2 (en orden descendente de frecuencia de uso).

(4) Palabras de un fonema de alta frecuencia (en orden descendente de frecuencia de uso): *y*.conj 32 648, *a*.prep 30 109, *o*.conj 4305, *eh*.interj 563, *ah*.interj 197, *oh*.interj 44, *m*.interj 27, *hm*.interj 25, *e*.sus 13.

(5) Palabras de dos fonemas de alta frecuencia (en orden descendente de frecuencia de uso): *el*.art 118 997, *de*.prep 76 116, *que*.conj 36 827, *en*.prep 28 910, *un*.art 22 875, *se*.pro 17 597, *no*.sn 13 365, *su*.pos 10 826, *lo*.pro 10 348, *me*.pro 5206, *ir*.vb 5123, él.pro 3541, *si*.conj 2857, *mi*.pos 2806, *ya*.adv 2631, *yo*.pro 2631, *qué*.interrog 2067, *te*.pro 2017, *sí*.sn 2001, *tu*.pos 1243.

De esta manera, muchas de las palabras difonémicas que aparecen con frecuencia alta son palabras funcionales como artículos (art), preposiciones (prep), conjunciones (conj) y pronombres (pro). Por otro lado, las palabras de contenido, como sustantivos (sus), adjetivos (adj) y verbos (vb), no se encuentran en la lista anterior. Estas palabras funcionales altamente frecuentes en las palabras difonémicas aumentan la frecuencia de todas las palabras difonémicas, lo que hace que la distribución de frecuencia general sea en forma de 'I'. Por lo tanto, si la longitud de la palabra se mide en términos de fonemas en lugar de sílabas, la llamada Ley de acortamiento (según la

cual cuanto más larga es la palabra, menos frecuentemente se usa) no se sostiene. Si la longitud de la palabra se mide en sílabas, la distinción entre un fonema y dos fonemas se eliminará y se combinarán en una sílaba y, como resultado, la distribución general de frecuencia parecerá caer de manera monótona, por lo que la Ley de acortamiento no podría aceptarse. Eso significa que la Ley del acortamiento se refuta midiendo la longitud de las palabras mediante fonemas, que son más precisos que las sílabas.

La Figura 7 muestra el promedio (*m*) y la desviación estándar (*sd*) del número de fonemas de las palabras funcionales (Fun) y palabras de contenido (Con) en todo el corpus (§3). (Fun: m=2,39, sd=1,14; Con: m=6,32, sd=2,29):

Por esta figura, sabemos que el número promedio de fonemas de palabras funcionales es pequeño y el de palabras de contenido es más del doble (2,6 veces). La desviación estándar también es mayor para las palabras de contenido. Esto indica que la longitud de las palabras de contenido que tienen significado varía mucho, mientras que con frecuencia se utilizan palabras funcionales con longitudes cortas, concentrándose en la longitud promedio de las palabras.

A continuación, en las figuras 8 y 9, abordaremos el análisis en el interior de las palabras funcionales y de contenido. Calcularemos el número de fonemas de cada parte principal del discurso (palabras funcionales: artículos, preposiciones, conjunciones, demostrativos, posesivos; palabras de contenido: sustantivos, verbos, adjetivos, adverbios).

De este modo, entre las palabras funcionales (Fig. 8), se utilizan con especial frecuencia artículos y preposiciones en las que el número de fonemas es mayoritariamente 2. Al observar la lista anterior (4), podemos comprobar que los artículos y preposiciones con 2 fonemas son *el*, *de*, *en* y *un*. Por lo tanto, la Ley de acortamiento utiliza la frecuencia de uso como un factor para acortar la longitud de las palabras, pero, en realidad, las palabras funcionales cortas (especialmente artículos y preposiciones *el*, *de*, *en*, *que*) tienen

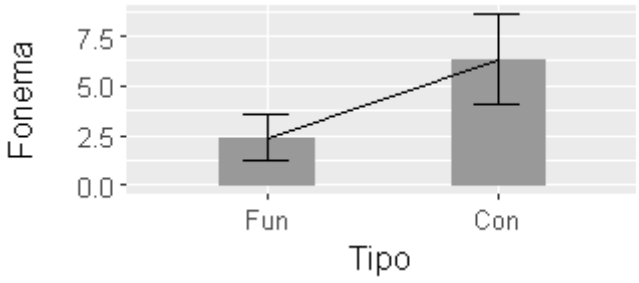

Fig. 7. Promedio y desviación estándar del número de fonemas de palabras funcionales (Fun) y palabras de contenido (Con)

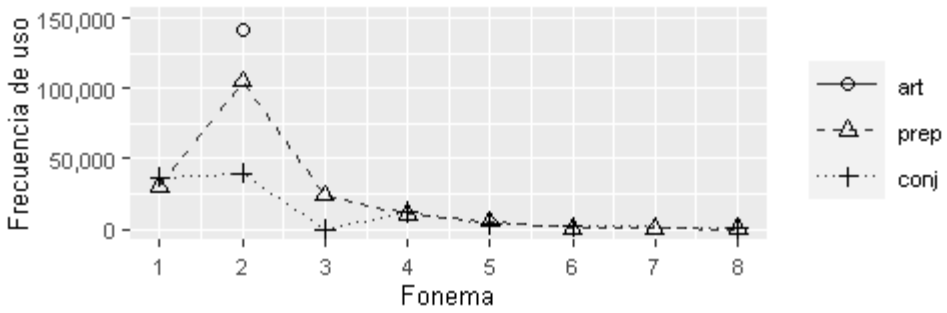

Fig. 8. Palabras funcionales: artículos (art), preposiciones (prep), conjunciones (conj)

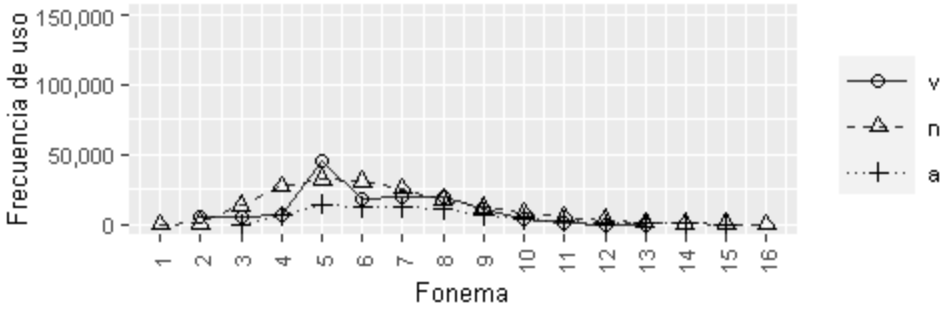

Fig. 9. Palabras de contenido: verbo (vb), sustantivo (sus), adjetivo (adj)

una gran influencia. Por ello, en lugar de observar la relación entre la alta frecuencia (factor) y el acortamiento de la forma de la palabra (efecto, resultado), deberíamos fijarnos en la relación entre las palabras funcionales cortas (causa) y la alta frecuencia (efecto). Hasta aquí, la interpretación de la Ley de acortamiento desde la perspectiva de la distinción gramatical entre palabras funcionales y palabras de contenido.

4.3. Derivación y composición

A continuación, interpretaremos la Ley de acortamiento a partir de la distinción léxica entre palabras simples y palabras complejas. Una palabra simple es una palabra cuya forma verbal consta de un solo morfema (por ejemplo, *noche*, *grande*, *tarde*, *cantar*, *entonces*). Por otro lado, una palabra compleja (derivada y compuesta) se refiere a una palabra en la que se produce el siguiente proceso de formación de palabras:

- Palabra derivada: *re* + *considerar* → *reconsiderar, admirar - ble* → *admirable.*
- Palabra compuesta: *alto* + *voz* → *altavoz, sacar* + *corcho* → *sacacorchos.*

La Figura 10 muestra el promedio y la desviación estándar del número de fonemas de palabras simples y palabras complejas en palabras de contenido (Simple: m=2.99, sd=1.56; Complejo: m=7.77, sd=1.99):

De esta manera, la frecuencia de uso de palabras simples y palabras complejas por la longitud de las palabras se aproxima a una distribución normal centrada en el valor medio. Naturalmente, la distribución de las palabras complejas tiene un mayor número de fonemas que la de las palabras simples, por lo que la línea se desvía hacia la derecha. Además, la curva de frecuencia de palabras simples se eleva por encima de la de palabras complejas.

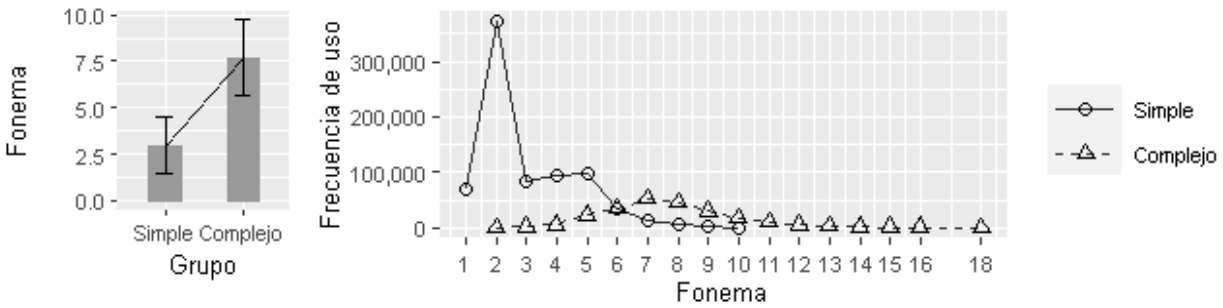

Fig. 10a. Media y desviación estándar del número de fonemas

Fig. 10b. Número de fonemas y frecuencia de uso
(Simple: palabras simples, Complejo: palabras complejas)

Creemos que la razón por la cual las palabras complejas se usan con menos frecuencia que las palabras simples es que las palabras derivadas o compuestas, que son una sola palabra unida con un prefijo o sufijo o la combinación de dos palabras, tienen significados especializados y quedan marcadas. Por ejemplo, según Juilland y Chang Rodríguez (1964), *corto* (frecuencia de uso: 46) > *acortar* (9), *grande* (795) > *engrandecer* (5), *posible* (144) > *imposible* (42); *bien* (577) y *estar* (2651) > *bienestar* (9).

El hecho de que los derivados y compuestos marcados se utilicen con menos frecuencia que las palabras simples es coherente con la universalidad del lenguaje (Greenberg, 1966 [2005], p. 56). Además, mientras que el rango de números de fonemas de palabras simples es estrecho [1, 10], el rango de números de fonemas de palabras complejas es relativamente más amplio [2, 18], aunque con un cierto solapamiento entre los dos casos, concretamente en la región de [2, 10]. Esto muestra que existe una rica variación en la longitud del vocabulario debido a la derivación y composición.

5. VARIACIÓN

Completamos el análisis con la atención puesta en fenómenos de variación relacionados con el estilo de escritura y, en definitiva, con las propiedades estilométricas de los textos.

En las investigaciones estilísticas centradas en obras literarias, la longitud de las palabras ha sido retomada como una de las características propias de los escritores (Yasumoto 1960, pp. 218-223; Yasumoto 1977, p. 417; Hatano 1988, p. 95; Frías Delgado 2009). En la investigación estilística lingüística, la longitud de las palabras a veces se mide en análisis comparativos de publicaciones en diversos campos, no solo de obras literarias, sino también prensa y manuales de texto, entre otros (Yoshioka 1996, pp. 200-201). En la investigación aplicada se utiliza como una de las variables para cuantificar la legibilidad de los textos (Crawford 1985; Alliende 1987; Gómez Guinovart 1999, Ferrer *et al.* 2009; Yasumoto 2009, pp. 255-256; Ríos Hernández 2017).

Abordaremos este análisis del estilo de escritura centrándonos en las palabras de contenido (en oposición a las palabras funcionales) y las palabras complejas (en oposición a las palabras simples).

5.1. Palabras de contenido

A continuación, usaremos tres tipos de materiales para comparar la longitud de las palabras de contenido (sustantivos, adjetivos, adverbios, verbos, etc.) para cada grupo que indican diferencias en grados escolares, publicaciones, sexo, edad y niveles educativos. Por otro lado, con respecto a las palabras funcionales no se han detectado diferencias importantes dentro de ningún grupo.

En general, tal y como se muestra en la Figura 11, se puede observar que en los años inferiores de Justicia (1995) (A1, A2) se utilizan con frecuencia palabras más cortas que la moda (cumbre de la curva de frecuencia) y, a medida que suben los grados (A3), se utilizan con mayor frecuencia palabras más largas que la moda.

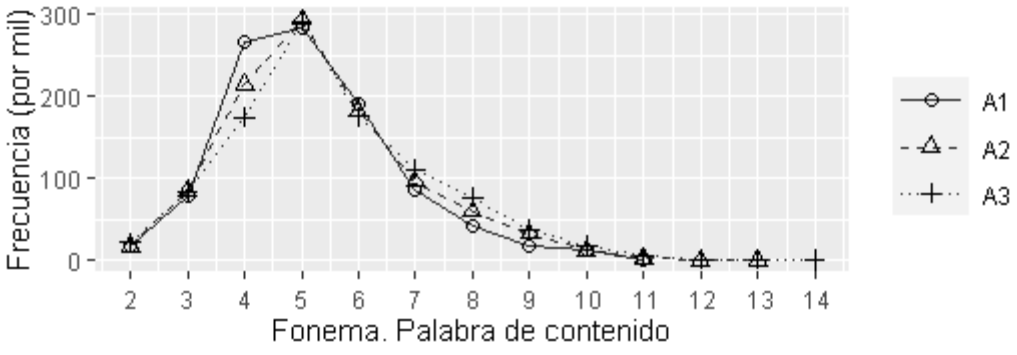

Fig. 11. Diferencias de calificaciones (A.1: Años inferiores; A.2: Años medios; A.3: Años superiores)

La longitud de las palabras del contenido varía mucho según el ámbito de uso, como advertimos en la Figura 12. La longitud de las palabras utilizadas en obras de teatro y novelas es corta, especialmente en las obras de teatro. Las palabras utilizadas en ensayos, prensa (periódicos y revistas) y textos de ciencia y tecnología, recogidos todos ellos en Juilland y Chang-Rodríguez (1964), son, sin embargo, bastante largas.

En la Figura 13, donde advertimos la variación en función del sexo en Martínez y Ueda (2021), se advierte que las palabras utilizadas por los hombres son ligeramente más largas. Las mujeres usan palabras con una longitud un poco por debajo de la moda, y los hombres usan palabras con una longitud ligeramente por encima de la moda.

Apenas hay diferencia si atendemos al parámetro Edad, como se aprecia en la Figura 14. Esta muestra que el vocabulario utilizado por las personas mayores tiene palabras de longitud escasamente mayor.

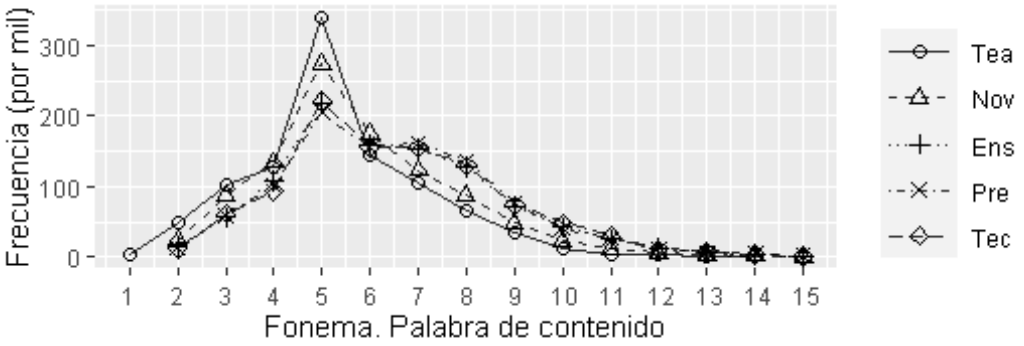

Fig. 12. Publicaciones (Tea: Teatro; Nov: Novela; Ens: Ensayo; Pre: Prensa; Tec: Ciencia y Tecnología)

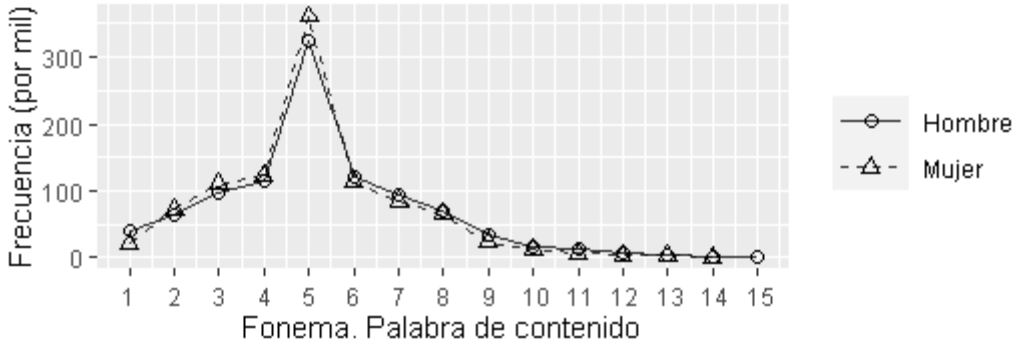

Fig. 13. Diferencias de sexo: Hombre y mujer

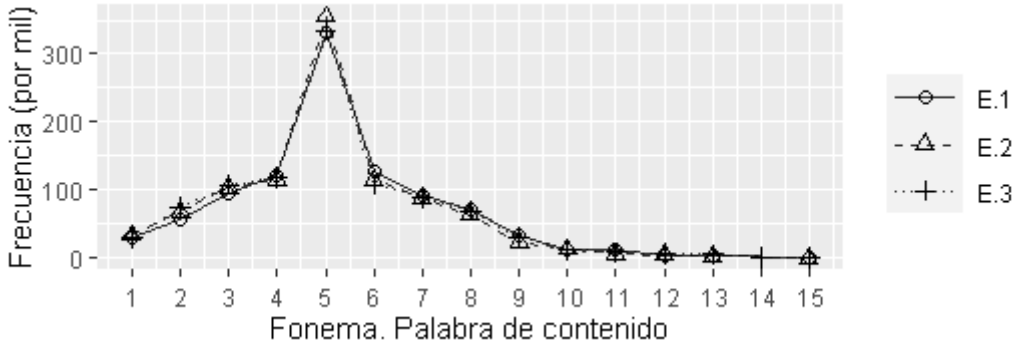

Fig. 14. Diferencia de edad (E-1: Joven; E-2: Mediana edad; E-3: Mayor)

Por último, la Figura 15 nos muestra que hay variación si tomamos en consideración la variable Nivel educativo. En este sentido, el vocabulario utilizado por personas con un alto nivel educativo es un poco más largo.

En general, casi no se advierten diferencias en la distribución de frecuencia de las palabras funcionales. Las palabras funcionales son el núcleo del sistema lingüístico y se utilizan casi al mismo ritmo y frecuencia en cualquier estilo de escritura, mientras que las palabras de contenido se seleccionan libremente dependiendo de variables lingüísticas y temáticas, lo que da lugar a variaciones.

En cuanto a las palabras de contenido, la longitud del vocabulario utilizado en los niveles superiores (grados superiores) es mayor dependiendo del grado de los estudiantes de primaria y del tipo de publicación. Sin embargo, en las grabaciones de entrevistas en español coloquial (Figuras 13, 14, 15), pertenecientes al corpus de Santander, no se percibe mucha variación en las palabras de contenido. Se esperaba que el lenguaje escrito fuera más estable y tuviera menos variaciones, pero los análisis han mostrado lo contrario. Pensamos que el lenguaje hablado se vuelve más uniforme porque está en sincronía con el interlocutor. Por otro lado, dado que el lenguaje escrito es una actividad lingüística individual, puede surgir un estilo de escritura único particular.

5.2. Palabras complejas

A continuación, utilizaremos tres tipos de datos para analizar la longitud de las palabras del vocabulario de cada grupo con respecto a las variaciones que se observan en función del grado escolar, la publicación, el sexo, la edad y el nivel educativo. Nos limitaremos a las palabras complejas (palabras derivadas y palabras compuestas), porque no se han observado grandes diferencias en la frecuencia en lo que respecta a las palabras simples.

En la Figura 16 apreciamos que, en áreas donde el número de fonemas a la izquierda de la moda es relativamente bajo, el orden en

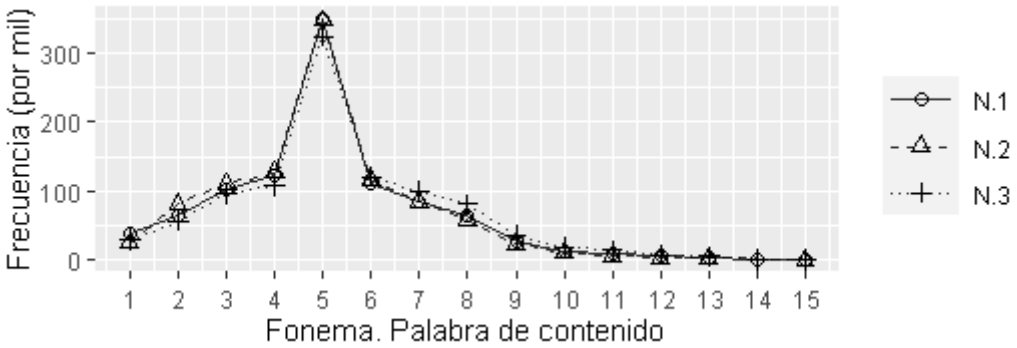

Fig. 15. Diferencia en el nivel educativo (N-1: Bajo; N-2: Medio; N-3: Alto)

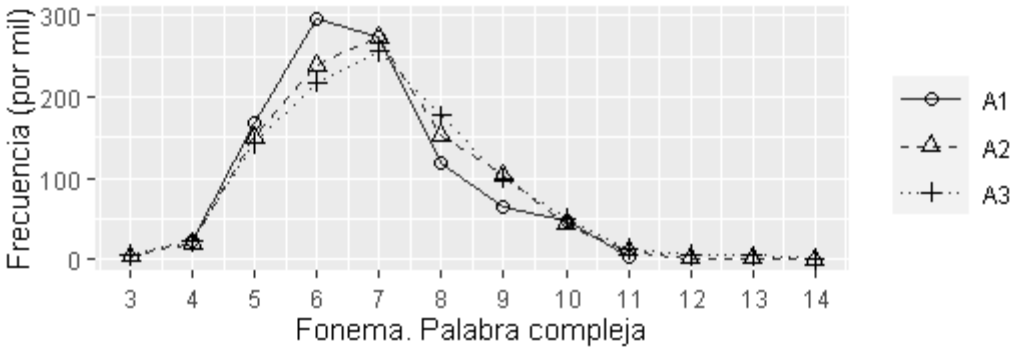

Fig. 16. Diferencias de calificaciones (A.1: Años inferiores; A.2: Años medios; A.3: Años superiores)

el grado educativo es A1 > A2 > A3 y, en áreas donde el número de fonemas a la derecha del ápice es relativamente alto, los números de frecuencia se invierten. Esto indica que el número de fonemas aumenta en los grados superiores, pero, como se trata de palabras complejas, no puede atribuirse simplemente al aumento en el número de fonemas. Naturalmente, se considera que la causa de este fenómeno de inversión es el desarrollo de la formación de palabras, análisis que viene a confirmar la hipótesis de investigación del presente estudio.

En la Figura 17, el lado izquierdo de la distribución (con bajo número de fonemas), muestra que en el ámbito de las publicaciones se utilizan las palabras relativamente breves en el teatro y en el lado derecho (con gran número de fonemas), se utilizan palabras relativamente largas en ensayos y prensa (periódicos y revistas), así como en textos de ciencia y tecnología. Las novelas se encuentran en el punto medio entre los dos.

En la Figura 18, y ya en lo que respecta al sexo, el lado izquierdo de la distribución muestra que la frecuencia del vocabulario breve utilizado por las mujeres es alta y, en el lado derecho de la distribución, la frecuencia del vocabulario largo utilizado por los hombres es alta también. Los posibles factores incluyen la formación de palabras, tal y como hemos señalado líneas arriba.

La Figura 19 nos advierte que es posible entender las características de la diferencia de edad. Las palabras complejas en la mediana edad tienden a ser cortas.

El análisis de la variación correspondiente al nivel educativo (N1, N2, N3) se muestra en la Figura 20. En el lado izquierdo de la distribución, donde el número de fonemas es pequeño, la frecuencia del vocabulario utilizado por personas con un alto nivel educativo es relativamente baja y, a la inversa, en el lado derecho de la distribución, la frecuencia del vocabulario utilizado por personas con un nivel educativo elevado es ligeramente alta.

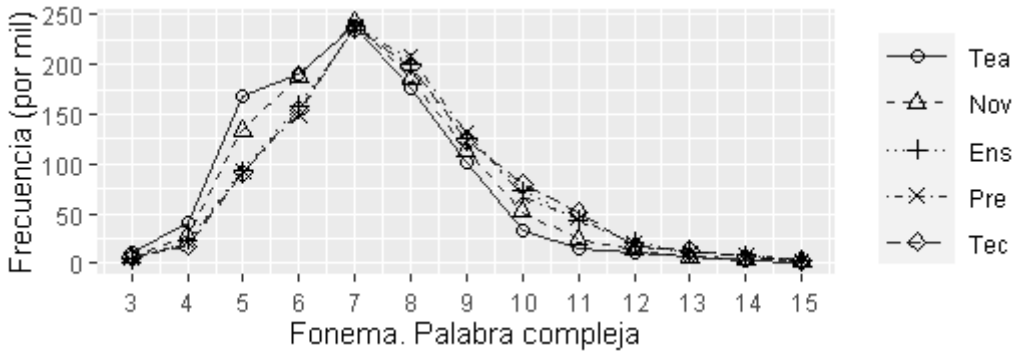

Fig. 17. Publicaciones (Tea: Teatro; Nov: Novela; Ens: Ensayo; Pre: Prensa; Tec: Ciencia y Tecnología)

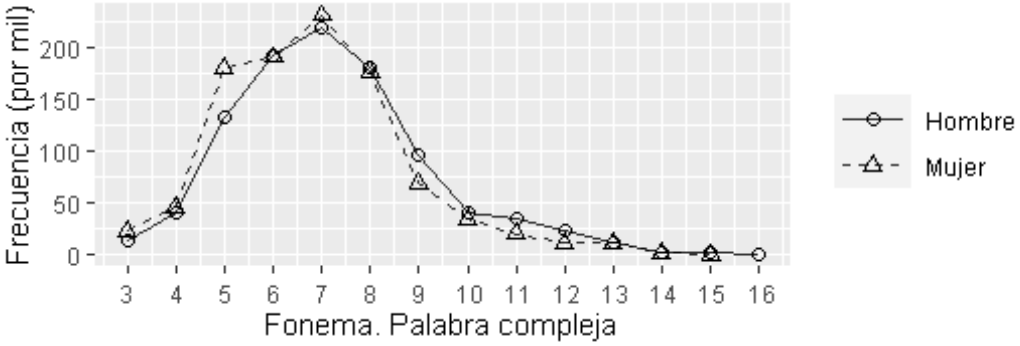

Fig. 18. Diferencias de sexo: Hombre y Mujer

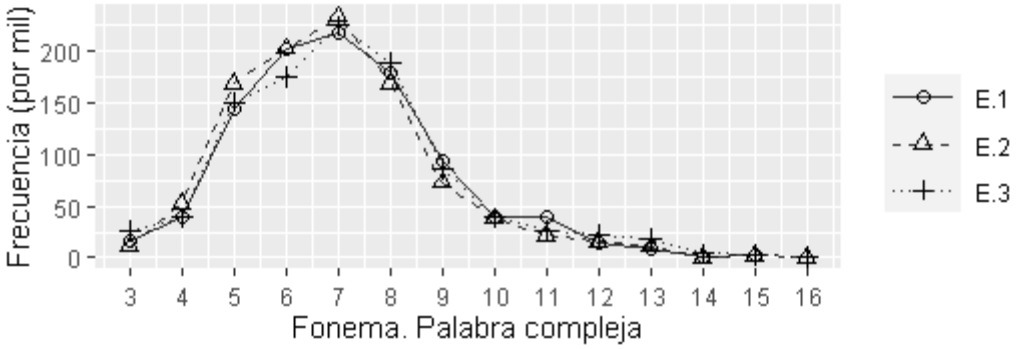

Fig. 19. Diferencia de edad (E-1: Joven; E-2: Mediana edad; E-3: Mayor)

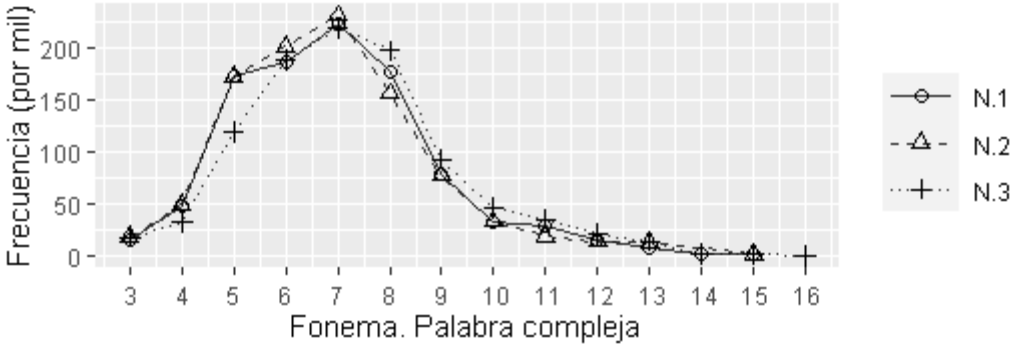

Fig. 20. Diferencia en el nivel educativo (N-1: Nivel bajo; N-2: Nivel medio; N-3: Nivel alto)

Consideramos que esta diferencia se debe al factor en la formación de palabras.

A modo de síntesis para este análisis de la variación, las diferencias destacadas de frecuencia se encuentran en las publicaciones, especialmente entre obras de teatro, ensayos, por un lado, y periódicos, revistas y textos de ciencia y tecnología, por otro. En otros datos, no se encontraron diferencias significativas en palabras simples, y la mayoría de las diferencias se detectaron en palabras complejas, por lo que parece que la formación de palabras es la razón de la relación inversa entre la longitud y la frecuencia de las palabras.

6. CONCLUSIONES

En este estudio, hemos medido, en primer lugar, el número de palabras diferentes (*type*), en sílabas y fonemas, en los lemas de los diccionarios de frecuencia y, de la misma forma, en los datos de nuestro corpus de encuestas grabadas. Observamos que la distribución general se aproxima a una distribución normal y discutimos las razones de esto. A continuación, verificamos si la Ley del acortamiento generalmente aceptada (cuanto más larga es la palabra, menos frecuentemente se usa y cuanto más frecuentemente se usa, más corta se hace la palabra) resulta válida para el número de palabras indicadas. Hemos comprobado que, si la longitud de la palabra se mide en sílabas, la frecuencia de uso muestra una distribución en forma de 'L', lo cual confirmaría la Ley del acortamiento. Sin embargo, esto se debe a que las palabras funcionales con longitudes cortas (artículos, pronombres, preposiciones, conjunciones, etc.) se utilizan con abrumadora frecuencia, mientras las palabras de contenido (sustantivos, adjetivos, adverbios, verbos, etc.), con frecuencia relativamente baja y longitud léxica larga, aparecen mezcladas en los datos del análisis. En este sentido, podemos estar ante el hecho de que la distribución de frecuencias no cae de manera monótona debido a un aumento en el número de sílabas.

Al observar el estado real del vocabulario para cada número de sílabas, encontramos que las palabras funcionales con una frecuencia sumamente alta tienen longitudes de palabras cortas y, por lo tanto, se ubican en el lado izquierdo de la distribución, seguidas de palabras de contenido de frecuencia relativamente baja con longitudes de palabras relativamente largas. Por ello, se encuentra en el lado derecho de la distribución de frecuencias y la curva de distribución parece ser monótonamente descendente. Por lo tanto, la curva monótonamente descendente de la distribución no es causada necesariamente por la Ley del acortamiento.

Todo ello significa que, al analizar los datos separados en palabras funcionales y de contenido usando el número de sílabas, la frecuencia de uso general muestra una distribución en forma de L'; sin embargo, la frecuencia de uso de las palabras de contenido no presenta una distribución en forma de 'L', sino una distribución cercana a la normal. Esto se puede confirmar mostrando la distribución, y analizando de manera similar el número de fonemas y separándolos en palabras funcionales y las de contenido. El resultado general muestra una distribución en forma de 'I' en lugar de una distribución en forma de 'L', mientras las palabras de contenido muestran una distribución normal. Cuando la longitud de la palabra se mide utilizando el número de sílabas, el primer término de la distribución de frecuencia, las palabras de una sílaba, estas se convierten en un grupo que incluye palabras de un fonema y de dos fonemas. El uso de unidades de medida aproximadas, como las sílabas, da como resultado observaciones aproximadas que no proporcionan una imagen precisa de la situación.

Cuando se han observado todos los datos, se aprecian tan solo unos pocos ejemplos donde la Ley del acortamiento se cumple: *bicicleta* > *bici, colegio* >*cole, película* > *peli, zoológico* > *zoo*. Si la misma ley fuera lo suficientemente general como para determinar la distribución de frecuencia de las formas de las palabras, encontraríamos muchos ejemplos en el español moderno. Históricamente, comparando las gramáticas históricas de español e inglés, notamos que las terminaciones en español se han eliminado y acortado con menos frecuencia que en inglés[7]. Por el contrario, hay muchos ejemplos donde las palabras de uso frecuente se han alargado históricamente (lat. *edere* > *cum+edere* > esp. *comer*; lat. *auris* > *auris* + *cula* > *auricula* > esp. *oreja*, lat. *afflare* > esp. *encontrar*, etc.). Una de las razones de este alargamiento de las formas léxicas es que las formas cortas, por no poseer la forma sustancial, son inconvenientes para la comunicación, por lo que se refuerzan añadiendo elementos aparentemente innecesarios.

Además, cambios históricos como *trans* > *tras* > *tra, sub* > *so, illu(m)* > *elo* > *el*, que, a primera vista, parecen dictar la Ley de acortamiento apoyada por el principio del menor esfuerzo, no fueron causados por el mismo principio. Estas palabras funcionales no estaban acentuadas como sí lo están las palabras de contenido. Simplemente, combinadas con palabras acentuadas adyacentes (palabras de contenido), se integraron con ellas, debilitándose fonológicamente en forma breve. Esto también se aplica a palabras funcionales acentuadas como *ipse* > *ese, primero* > *primer*. Por lo tanto, en general, la Ley de acortamiento (debido al uso frecuente) no puede ser aceptada para el español de manera incondicionada.

La razón por la cual las palabras funcionales son generalmente cortas es que las palabras funcionales no son elementos que llevan *significado*, como las palabras de contenido, sino más bien elementos que indican *relación gramatical*. Para que algo tenga un significado, se requiere una forma considerable que corresponda a ese significado, por lo que el número de fonemas aumenta naturalmente. Además, inevitablemente aumentará el número de fonemas de palabras derivadas

7 Veánse, entre otros estudios, Menéndez Pidal (1968) y Brunner (1960 [1962]). Según estos autores, la lengua española mantiene las sílabas átonas finales, excepto los casos de apócope, especialmente de la vocal /e/, mientras que en inglés se registran numerosos casos de acortamiento general en la posición átona. Como una consecuencia de carácter histórico, en la actualidad las palabras bisílabas españolas ocupan el primer lugar en la frecuencia de uso (Navarro Tomás 1966, pp. 54-55), mientras que en inglés, los monosílabos son predominantes, ocupando casi la mitad del texto (Zipf 1936, pp. 22-23).

con afijos para especificar significados y de palabras compuestas que añaden significados.

Por otro lado, puesto que se pone énfasis en las palabras de contenido para indicar claramente la unidad de significado, la forma de la palabra se mantiene con relativa firmeza. En cambio, la esencia de una palabra funcional es indicar la relación gramatical, por lo que basta con tener ese indicador. En muchos casos, ni siquiera el acento es necesario. La razón por la cual muchas palabras funcionales no están acentuadas es que las palabras funcionales no aparecen por sí mismas, sino que se adjuntan antes o después de una palabra de contenido acentuada para indicarle a la palabra de contenido (o a una secuencia de palabras de contenido) su relación gramatical. Como resultado, las palabras funcionales inevitablemente se convierten en formas cortas. Las palabras funcionales relativamente largas como *mediante, durante, unos, según* se usan con menos frecuencia que las palabras funcionales monosilábicas como *el, la, a, de, en, con, por, que*. Además, sin acento, se produce un debilitamiento fonológico, que a menudo resulta en un acortamiento (lat. *illos* > *los, illud* > *lo, illam* > *ela* > *la, et* > *y, aut* > *o*, etc.).

En otro orden de cosas, si comparamos palabras con el mismo contenido, es cierto que, cuanto más larga es la palabra, menos frecuentemente se usa. Esto se debe a cambios estructurales en la formación de palabras debido a la derivación, con prefijos y sufijos, y a la composición, con elementos compositivos. Cuando las palabras se alargan a través de la formación de palabras, los significados de los afijos y elementos compuestos se ponderan y especializan. Es natural que las palabras nuevas con significados especiales sean menos frecuentes que las palabras aisladas con significados generales. La afirmación, por tanto, según la cual cuanto más larga es una palabra, menos frecuentemente se usa, es una observación no empírica, que ignora el proceso mediante el cual se constituyen las formas lingüísticas. Es cierto que no se pueden negar principios explicativos como la ley del mínimo esfuerzo, la economía y el coste, pero estos son efectos resultantes del cambio y la variación del lenguaje, y no son sus causas o factores esenciales. Los factores esenciales de los fenómenos lingüísticos se encuentran en la estructura y el sistema del lenguaje.

En trabajos sobre lingüística general, la Ley de acortamiento se ha discutido utilizando la longitud física de las formas de las palabras (número de letras, sílabas, fonemas). Sin embargo, desde una perspectiva lingüística, la longitud de la forma de una palabra debe observarse lingüísticamente, no físicamente. La longitud lingüística de una palabra debe medirse mediante prefijos y sufijos, y combinación de palabras, que son unidades de morfología derivativa y compositiva.

Desde la perspectiva de la universalidad lingüística, es cierto que existe un contraste entre palabras simples y palabras complejas como no marcadas y marcadas (Greenberg 1966 [2005], p. 56), pero, si tenemos en cuenta la frecuencia de uso, hay un contraste gradual continuo, no dicotómico discreto, entre palabras simples y palabras derivadas. En lugar de una oposición binaria entre no marcado y marcado, la marcación se evalúa progresivamente en el rango [0, 1]. En otras palabras, los afijos utilizados con frecuencia casi no están marcados, mientras que los afijos utilizados con poca frecuencia están muy marcados, lo que da lugar a un valor estilístico. Lo que hace posible este tipo de análisis es el estudio de frecuencia de los afijos. El análisis estilométrico práctico debe tener en cuenta no solo los contrastes teóricos de la lengua, sino también los contrastes basados en el uso práctico del habla en forma de frecuencias de uso.

7. PROGRAMAS EN R[8]

```
(1) Monotonía
Monotony=function(A){
  a=d=0
  for(i in 2:length(A)){
   if(A[i]>A[i-1]) a=a+A[i]-A[i-1] else d=d+A[i-1]-A[i]
  }
 (a-d)/(a+d)
} #Monotony [-1, 1]
```

8 Para el detalle de las operaciones estadísticas, véase Ueda y Moreno Sandoval (2017).

```
(2) Distribución de forma 'L'
Ldi = function(A){
  n=length(A)
  mx=max(A)
  (mx*n - sum(A)) / (mx*(n - 1))
}# L-form distribution
```

8. REFERENCIAS BIBLIOGRÁFICAS

Alliende, F. (1987). Perfil 4, cuatro procedimientos rápidos para determinar la legibilidad de un texto. *Lectura y Vida. Revista Latinoamericana de Lectura*, año 8, (4). http://www.lecturayvida.fahce.unlp.edu.ar/numeros/a8n4/08_04_Alliende.pdf

Brunner, K. (1960 [1962]). *Die englische Sprache*. I und II. Max Niemeyer (Trad. de Matsunami, Tamotsu, Kinshiro Oshitari, Shigeru Ono, Kooichi Zin [1962]. *Eigo hattatsushi*. Taishukan).

Bybee, J. (2007). *Frequency of Use and the Organization of Language*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301571.001.0001

Blasco Pascual, F. J. y Ruiz Urbón, C. (2022). *Análisis de textos desde la estilometría*. Ediciones Universidad de Salamanca.

Crawford, A. N. (1985). Fórmula y gráfico para determinar la comprensibilidad de textos del nivel primario en castellano. *Lectura y Vida. Revista Latinoamericana de Lectura*, año 6, (4). http://www.lecturayvida.fahce.unlp.edu.ar/numeros/a6n4/06_04_Crawford.pdf

Davies, M. (2006). *A Frequency Dictionary of Spanish. Core Vocabulary for Learners*. Routledge. https://doi.org/10.4324/9780203415009

Divjak, D. (2019). *Frequency in Language. Memory, Attention and Learning*. Cambridge University Press. https://doi.org/10.1017/9781316084410

Ferrer García, C., Pascual Gaspar, E. y Laínez Gadea, J. A. (2009). Legibilidad y comprensibilidad de la información individual y consolidada en las empresas cotizadas españolas. *XV Congreso AECA*, 2009. https://www.aeca1.org/pub/on_line/comunicaciones_xvcongresoaeca/cd/27a.pdf

Frías Delgado, A. (2009). Distribución de frecuencias de la longitud de las palabras en español: aspectos diacrónicos y de estilometría. En P. Cantos Gómez y A. Sánchez Pérez (eds.). *A Survey on Corpus-based Research. Panorama de investigaciones basadas en corpus* (pp. 756-770). Asociación Española de Lingüística del Corpus. https://www.um.es/lacell/aelinco/contenido/pdf/51.pdf

García Hoz, V. (1953). *Vocabulario usual, vocabulario común y vocabulario fundamental*. Consejo Superior de Investigaciones Científicas.

Gómez Guinovart, J. (1999). *La escritura asistida por ordenador. Problemas de sintaxis y de estilo*. Servicio de Publicaciones de la Universidad de Vigo.

Greenberg, J. H. (1966, 2005). *Language Universals*. Walter de Gruyter GmbH & Co. https://doi.org/10.1515/9783110899771

Hatano, K. (1988). *Introducción a la psicología de lengua y escritura* (en japonés). Shogakukan.

Herdan, G. (1956). *Language as Choice and Change*. Noordhoff N. V.

Juilland, A. y Chang-Rodríguez, E. (1964). *Frequency Dictionary of Spanish Words*. Mouton. https://doi.org/10.1515/9783112415467

Justicia Justicia, F. (1995). *El desarrollo del vocabulario. Diccionario de frecuencias*. Universidad de Granada.

Kabashima, T. (1968). *Anatomía de expresión* (en japonés). Sanseido.

Kin, A. (2018). Léxico. En T. Ogino (ed.). *Introducción a la lingüística japonesa actual* (en japonés). Meijishoin.

Martinet, A. (1970). *Elementos de lingüística general* (Trad. de J. Calonge Ruiz, 2.ª ed.). Gredos.

Martínez Martínez, I. y Ueda, H. (2021). *Inventario léxico de PRESEEA-Santander*. https://zenodo.org/records/10620777

Martínez Martínez, I. y Ueda, H. (2023). *Inventario morfológico de PRESEEA-Santander*. https://zenodo.org/records/10620852

Menéndez Pidal, R. (1968). *Manual de gramática histórica española* (13.ª ed.). Espasa-Calpe.

Moreno, Antonio, De la Madrid, G., Alcántara, M., González, A., Guirao, J. M. y De la Torre, R. (2005). The Spanish corpus. Enin E. Cresti y M. Moneglia (eds.). *CORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. John Benjamins. https://doi.org/10.1075/scl.15.06mor

Miller, G. (1951 [1979]). *Language and Communication* (Trad. de E. Goigorsky y S. Delpy [1979]. *Lenguaje y comunicación*. Amorortu editores). https://doi.org/10.1037/11135-000

Navarro Tomás, T. (1966). *Estudios de fonología española.* Las Americas Publishing Company.

Ríos Hernández, I. N. (2017). *Un acercamiento a la legibilidad de textos relacionados con el campo de la salud*. CIESPA. https://doi.org/10.16921/chasqui.v0i135.2892

Saporta, S. (1963). Phoneme Distribution and Language Universals. En J. H. Greenberg (ed.). *Universals of Language* (pp. 61-72). The MIT Press.

Tanaka, K. (2021). *Language and fractal* (en japonés). Tokyo *Daigaku* Shuppankai.

Takefuta, Y. (1981). *Konpyuuta no mita gendai eigo. Bokyaburari no kagaku* (Inglés moderno visto por el ordenador. Ciencia del vocabulario, en japonés).

Ueda, H. (1987). *Frecuencia y dispersión del vocabulario español.* https://h-ueda.sakura.ne.jp/kenkyu/goi/frec-disp/frec-disp-0.pdf

Ueda, H. (2021). Parte final y acentuación de palabras españolas. Análisis de diccionarios, corpus grandes y datos sociolingüísticos, geográficos e históricos. *Estudios de geolingüística,* (1), 51-105 (en japonés).

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

Urrutibéheity, H. N. (1972). The statistical properties of the Spanish lexicon. *Cahiers de lexicologie,* (20), 79-95.

Whatmough, J. (1956 [1960]). *Language* (Trad. de H. Toshio y K. Akira [1960]. Iwanamishoten).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. https://doi.org/10.1007/978-3-319-24277-4_9

Wierzbicka, A. (1967 [2011]). *I jezyku dla wszystkich*. Warszawa. Ogawa Masatoshi, Ishii tetsushiro y Abe Yuuko (trad.) *Anna sensei no gengogaku nyuumon* (Introducción a la lingüística de la profesora Anna (en japonés). Tokyo gaikokugo daigaku shuppankai.

Yasumoto, B. (1960). *Nuevos campos de la sicología de la escritura* (en japonés). Sogensha.

Yasumoto, B. (1977). Estudios estilísticos actuales. En *Estilo. Lengua japonesa* (en japonés) (pp. 395-423). Iwanami Shoten.

Yasumoto, B. (2009). Estilística cuantitativa. Sicología de escritura. En *Enciclopedia de lingüística japonesa cuantitativa* (en japonés). Asakurashoten.

Yoshioka, K. (1996 [1982]). Perspectiva de estudios estilísticos cuantitativos (Trad. de K. Anthony [1982]. *The Computation of Style* (pp. 196-237), en japonés). Nanundo.

Zipf, G. K. (1936). *The Psychobiology of Language. An Introduction to Dynamic Philology.* Routledge.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Mansfiels Addison-Wesley Press.