Anuario de Psicología/The UB Journal of Psychology
2013, vol. 43, nº 1, 67-82

La ley de Zipf y la detección de la evolución verbal en la enfermedad de Alzheimer*

Antoni Hernández-Fernández[1-2]
Faustino Diéguez-Vide[1]
[1] *Universitat de Barcelona*
[2] *Universitat Politècnica de Catalunya*

Antecedentes: *en la sociedad actual, es innegable el aumento de sujetos con enfermedad de Alzheimer (EA) y el incremento de enfermos en estadios diferentes de la patología. Aunque es fundamental obtener un diagnóstico precoz, también lo es poder detectar automáticamente la evolución de la enfermedad. La ley de Zipf es una herramienta que permite, por medio de la frecuencia de uso de las palabras, describir lingüísticamente esa evolución.* Método: *se han utilizado un conjunto de corpora de 20 pacientes con EA (10 GDS4 y 10 GDS5) y 10 controles obtenidos a partir de tres pruebas de producción oral y se han analizado estadísticamente.* Resultados: *se han observado desviaciones del exponente de Zipf en las palabras de frecuencia media para pacientes GDS5, pero no para enfermos GDS4.* Conclusiones: *la desviación del exponente de Zipf en los pacientes GDS5 respecto al grupo control muestra que es posible predecir la evolución de un estadio a otro en la EA y permite deducir cuándo existe una alteración en la sintaxis a partir de la simple producción oral del enfermo. En otras palabras, las variaciones en la ley de Zipf puede predecir la evolución sintáctica de estos pacientes. Mediante futuros sistemas de detección automática se pretende conseguir describir la evolución de ciertas enfermedades con alteraciones verbales.*

Palabras clave: *ley de Zipf, enfermedad de Alzheimer, detección de la evolución neurodegenerativa.*

* *Agradecimientos:* Gracias a Ramon Ferrer por sus comentarios y a todos los sujetos que han permitido generar los corpora. Gracias también a las sugerencias y aclaraciones propuestas por dos revisores anónimos.
Correspondencia: Faustino Diéguez-Vide. Departament de Lingüística General. Universitat de Barcelona. Gran Via de les Corts Catalanes 585, 08007 Barcelona. Correo electrónico: fdieguez@ub.edu.

Zipf's law and the detection of the verbal evolution in Alzheimer's disease

Background: *In our society, it is undeniable the increase of prevalence of Alzheimer's Disease (AD) and of patients in different disease stages. Although it is essential to obtain an early diagnosis, it is also desirable to automatically detect the progression of the disease. Zipf's law is a tool that allows, through the analysis of words frequency, to describe the linguistic evolution of patients with AD.* Methods: *A set of corpora of 20 patients with AD (10 GDS4 and 10 GDS5) and 10 controls, derived from three tests of oral production, have been used and studied statistically. Results: Deviations from the Zipf's exponent in the words of mid-frequency for GDS5 patients have been observed, but not for GDS4.* Discussion: *Deviations on Zipf's exponent in GDS5 versus control group show that it is possible to predict the evolution from one disease stage to another in the AD and determine when syntax is altered, exploring the simple oral production of the patient. In other words, variations in Zipf's law can predict the syntactic evolution of these patients. Through future automatic detection systems we aim to describe the evolution of certain diseases with verbal alterations.*

Keywords: *Zipf's law, Alzheimer's disease, automatic detection of neurodegenerative evolution.*

Introducción

La enfermedad de Alzheimer (EA) es una entidad caracterizada por un deterioro cognitivo de inicio insidioso y progresivo y que en los países desarrollados es la forma más frecuente de demencia neurodegenerativa. Aunque existen casos de inicio precoz (< 60 años), suele aparecer en personas mayores y aumenta su incidencia con la edad.

En 2011, en EE.UU. (Alzheimer's Association, Thies y Bleiler, 2011), la EA fue la sexta causa de mortalidad (la quinta en mayores de 65 años) y, a diferencia de otras enfermedades, es la única en que la tasa de mortalidad aumentó: un 66% en el período 2000-2008. También se estima un aumento de nuevos enfermos: en el año 2050, hasta un 62% más en EE.UU. (Alzheimer's Association *et al.*, 2011) y hasta 16,2 millones en Europa (Wancata, Musalek, Alexandrowicz y Krautgartner, 2003).

Ante esta situación, es fundamental obtener herramientas que permitan un diagnóstico precoz de la enfermedad, tanto desde una perspectiva biológica (Berthier y Dávila, 2010; Valls-Pedret, Molinuevo y Rami, 2010) como neuropsicológica (Cuetos-Vega, Menéndez-González y Calatayud-Noguera, 2007). Pero también lo es el poder detectar y, a ser posible, por medios automáticos, la posible evolución de esta enfermedad, tanto respecto a otros cuadros clínicos similares –como el Deterioro Cognitivo Leve (Valls-Pedret *et al.,* 2010; Cuetos-Vega *et al.,* 2007; Mulet *et al.,* 2005) o algunas alteraciones cognitivas mnésicas (Fleisher *et al.,*

2007)– como dentro de la propia EA. Además, es necesario que los instrumentos de diagnóstico sean lo más sencillos posible y lo más rápidos posible en su administración.

La forma más sencilla y rápida de realizar un diagnóstico pasa por la propia producción oral del paciente, pero las escalas que valoran el deterioro cognitivo (como la GDS; Reisberg, Ferris, De León y Crook, 1982) apenas incluyen componentes verbales. La mayoría de pruebas verbales que se administran a enfermos con EA son pruebas de fluidez verbal, tanto semántica como fonológica. En el primer caso, se demanda la denominación de nombres de animales (Carnero y Lendínez, 1999; Carnero *et al.,* 2000) o frutas (Cuetos -Vega *et al.,* 2007), o bien cosas que se pueden encontrar dentro de algún lugar, como una casa (Fernández *et al.,* 2002) o un supermercado (Garcés, Santos, Pérez y Pascual, 2004). Existen, incluso, baterías relacionadas con esta denominación (Peraita, González, Sánchez y Galeote, 2000). En el segundo caso, se demanda que los sujetos digan palabras que comiencen por una letra concreta. Aunque son pruebas rápidas de realizar (normalmente es un minuto), se trata de pruebas que evalúan aspectos concretos y parciales de la lengua del enfermo. Y, de hecho, algunas investigaciones trabajan aun con aspectos lingüísticos más concretos, como la denominación de nombres propios (Semenza, Mondini, Borgo, Pasini y Sgaramella, 2003; Cuetos-Vega *et al.,* 2007)

No obstante, más allá de la fluidez verbal y de la denominación (junto con la detección de categorías; Díaz-Mardomingo, Peraita-Adrados y Garriga-Trillo, 2000), no existen pruebas neuropsicológicas que valoren aspectos diagnósticos con la propia producción oral del paciente. Y esto teniendo presente que la alteración verbal de los sujetos con EA en la producción oral afecta a todos los niveles lingüísticos. Los componentes más resistentes serían la articulación-percepción (se reduce en niveles avanzados de la enfermedad) y la lectura (Patterson, Graham y Hodges, 1994), mientras que el más afectado sería el semántico. En el resto de componentes –sintáctico y discursivo– los resultados son algo más heterogéneos, sobre todo por la disfunción progresiva de los mismos con el avance de la enfermedad.

Más en concreto, el habla de estos enfermos se ha caracterizado como "habla vacía" porque contiene una elevada proporción de expresiones y palabras con bajo contenido semántico (Aronoff, Gonnerman, Almor, Kempler y Andersen, 2006; Almor, Kempler, MacDonald, Andersen y Tyler, 1999; Kempler, 1995; Hier, Hagenlocker y Schindler, 1985). Estas palabras son, a menudo, palabras de alta frecuencia y se corresponden con las denominadas palabras funcionales (preposiciones, artículos, conjunciones, auxiliares, etc.) o con palabras que se utilizan como comodines informativos –palabras ómnibus– (eso, cosa, aquello, etc.). Son palabras que incrementan su aparición en los corpora ante la dificultad de encontrar la palabra adecuada (Almor *et al.,* 1999; Kempler, 1995). Por el contrario, los enfermos presentan una alteración en palabras de baja frecuencia (Patterson *et al.,* 1994; Shuttle-

worth y Huber, 1988), sobre todo en nombres. Una de las consecuencias es, al menos para el inglés, el uso excesivo de pronombres (Almor *et al.,* 1999), aunque no siempre se utilizan de forma correcta.

El objetivo del presente estudio consiste en la descripción de la evolución de la producción oral en pacientes con EA en dos fases clínicas (GDS4 y GDS5), en relación con dos grupos controles, y la propuesta de un método de detección de la evolución de alteraciones –específicamente léxicas y sintácticas– a través del simple análisis del léxico y de la distribución de frecuencias de palabras. Para ello, se hará en la siguiente sección una sucinta revisión a los métodos cuantitativos tradicionales de detección y estudio de la producción lingüística en la EA; posteriormente se presentará la ley de Zipf y sus implicaciones para el lenguaje; por último, se analizará y discutirá la distribución de frecuencias de palabras en los pacientes con EA con el fin de observar si existen o no diferencias entre las dos fases clínicas.

Detección de la evolución de la EA a partir de datos cuantitativos

Una primera aproximación a la detección verbal cuantitativa se ha centrado en comparar la escritura de Iris Murdoch (escritora británica con EA, diagnosticada histológicamente post-mortem en 1999) con la producción de otros escritores (Garrad, Maloney, Hodges y Patterson, 2005; Pakhomov, Chacon, Wicklund y Gundel, 2011). No obstante, se trata más de descripciones que de propuestas diagnósticas.

Las principales magnitudes estadísticas definidas para intentar procedimientos de detección automática de la EA (Thomas, Keselj, Cercone, Rockwood y Asp, 2005; Bucks, Singh, Cuerden y Wilcock, 2000), que normalmente se aplican a las mil primeras palabras producidas en el habla espontánea de los pacientes, se dividen fundamentalmente en tres bloques:

1. La proporción de types-tokens (TTR) y otros estadísticos –como el índice de Brunét (W)– se relacionan con la riqueza léxica en la producción del sujeto. La tabla 1 (ver página siguiente) resume la definición y la manera de calcular ambas magnitudes. Generalmente, la TTR es menor en los pacientes con EA que en los controles, mientras que W debería ser mayor en pacientes que en controles, dada su definición matemática.
2. El promedio de ocurrencias de nombres, adjetivos, verbos y pronombres daría cuenta de la proporción de los distintos tipos de palabras. Al respecto Bucks *et al.* (2000) determinaron que los enfermos de Alzheimer poseían en general una mayor proporción de pronombres, adjetivos y verbos que los controles, junto con una menor presencia de nombres.

TABLA 1. RESUMEN DE ALGUNAS MAGNITUDES QUE DETERMINAN LA RIQUEZA DE LA PRODUCCIÓN LÉXICA.

Magnitud	*Símbolo*	*Ecuación*
Proporción *Type-Token*	TTR	$TTR = \frac{V}{N}$
Índice de Brunét	W	$W = N^{V^{-0.165}}$

Nota: A partir del análisis para la detección automática de la EA de Bucks *et al.* (2000) y adaptado de Thomas *et al.* (2005). V: Vocabulario, número de palabras diferentes (*types*). N: Número total de palabras (*tokens*).

3. Por último, la magnitud CSU (*Clause-like Semantic Unit*) mide la cohesión semántica de las frases y caracteriza la fluidez discursiva. La CSU se puede definir como una cadena de palabras que están semánticamente conectadas formando una unidad de significado. Se contabiliza el número de CSUs que aparecen en un corpus de mil palabras.

Otras aproximaciones apuestan por estudios de los N-gramas más frecuentes en la producción (Keselj, Peng, Cercone y Thomas, 2003), es decir, los conjuntos de palabras (generalmente sintagmas) que más se repiten. Se pueden utilizar como marcadores discursivos las palabras más frecuentes, puesto que se ha determinado que las palabras funcionales (determinantes, preposiciones, conjunciones, etc.) ayudan en la predicción de las deficiencias lingüísticas de los enfermos (Thomas *et al.,* 2005), aunque no se especifique cómo.

La ley de Zipf

Las frecuencias de palabras en el lenguaje siguen el patrón estadístico de la ley de Zipf (Zipf, 1949/1972). Si P*(f)* es la probabilidad de aparición en un corpus de palabras de frecuencia *f*, entonces decimos que el corpus sigue la ley de Zipf si:

$$P(f) \propto f^{\beta} \qquad (1)$$

siendo β el exponente de la ley de Zipf, con β>0. La ecuación anterior mostraría una línea recta cuando dibujamos la probabilidad P(*f*) en una escala logarítmica doble (figura 1). Aunque diferentes funciones han sido propuestas para modelizar la distribución de frecuencias de palabras (Li, Miramontes y Cocho, 2010; Tuldava, 1996; Chitashvili y Baayen, 1993), la ecuación (1) parece describir de forma bastante aproximada la distribución de frecuencias en la producción adulta.

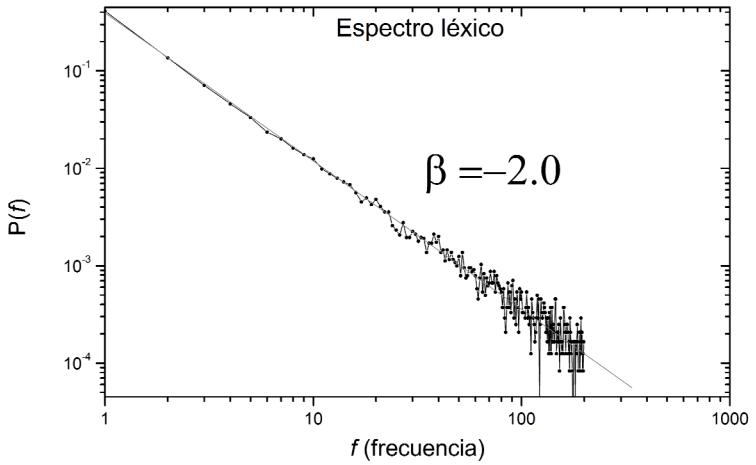

Figura 1. Representación de las palabras de frecuencia f, frente a su probabilidad P(f), en escala logarítmica, para un corpus de 8.000 palabras. Se recupera la ley de Zipf (ecuación 1) con exponente β≈2, típica en la producción adulta (Hernández-Fernández, 2005).

Típicamente tenemos β≈2 para las frecuencias de palabras de muestras de un autor (Ferrer, 2005a; Montemurro, 2001; Montemurro y Zanette, 2002; Zipf, 1942), aunque se han encontrado desviaciones importantes a la ley de Zipf en el habla: por ejemplo, en niños pequeños, β<2 (Ferrer, 2005a; Hernández-Fernández, 2005); en esquizofrenia se presentan exponentes β>2 en algunos pacientes cuando su discurso es muy variado y caótico; exponentes 1<β<2 en pacientes con discurso obsesivo en el que abundan las repeticiones (Ferrer, 2005a; Piotrowska y Piotrowska, 2004; Piotrowski, Pashkovskii y Piotrowski, 1994). Todos estos datos invalidarían las explicaciones de esta ley como un fenómeno debido meramente a azares estadísticos (Ferrer y Elvevag, 2010; Ferrer, 2005b) y lo presentarían como indiscutible universal comunicativo (Ferrer y Solé, 2003; Ferrer, 2005a).

Por otra parte, si representamos gráficamente las probabilidades acumuladas de P(*f*) *versus* la frecuencia (figura 2), entonces el exponente que se determina típicamente para la producción adulta es β'=1 (Ferrer, Solé y Köhler, 2004), siendo:

$$\beta=\beta' + 1 \qquad (2)$$

con β' el exponente del espectro de probabilidades acumuladas de la ley potencial de Zipf, y β el clásico exponente de Zipf del espectro de las palabras (sin acumular). Las frecuencias de palabras del lenguaje siguen esta regularidad (Ferrer y Solé, 2003; Ferrer, 2005b).

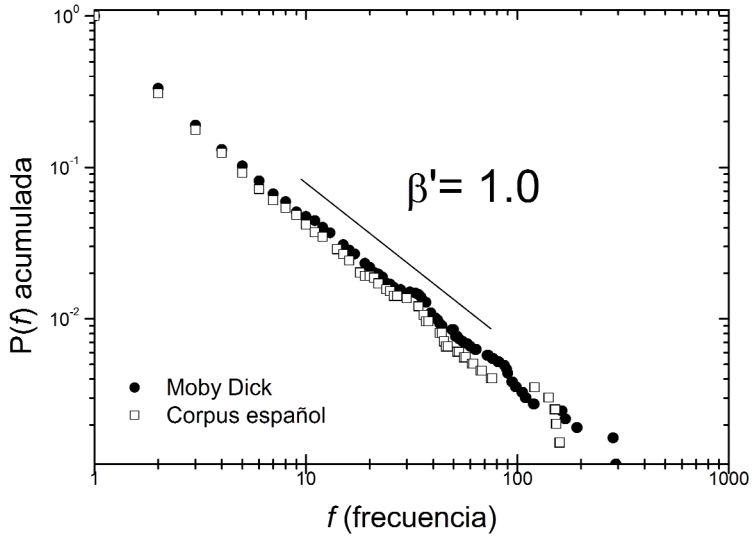

Figura 2. Recuperación de la ley de Zipf en un corpus escrito en español (Hernández-Fernández 2005) comparado con un corpus inglés de similar tamaño (10.000 tokens de la obra Moby Dick). Se representa en escala logarítmica la frecuencia de las palabras frente a la probabilidad acumulada de aparición de una palabra de frecuencia f, y obtenemos β'≈1 (Hernández-Fernández, 2005).

La ley de Zipf se fundamenta en lo que Zipf (1949/1972) denominó el *Principio del mínimo esfuerzo*: si un repertorio comunicativo es demasiado unificado o repetitivo, entonces solo son posibles unos pocos mensajes para expresar todo un surtido de informaciones y, entonces, la complejidad comunicativa será baja. La ley de Zipf maximiza esta eficiencia comunicativa y minimiza el coste de comunicación (Ferrer y Solé, 2003), con un rango de exponentes al parecer limitado (Ferrer, 2005b; 2006) y donde las palabras más conectadas de la red generalmente desempeñan papeles sintácticos relevantes (Ferrer *et al.,* 2004). Esta ley también implica la conectividad entre palabras (Ferrer, Bollobás y Riordan, 2005), un requisito para la sintaxis.

En relación con la conectividad, la gramática y el léxico son aspectos emergentes del lenguaje ligados a la conectividad de los elementos lingüísticos (Bates y Goodman, 1999). Así, las palabras de alta frecuencia suelen ser palabras de bajo contenido semántico y enorme importancia sintáctica por estar muy conectadas en la red lingüística (Ferrer, 2006), mientras que las de baja frecuencia suelen tener menor relevancia sintáctica y en cambio un mayor significado (Ferrer *et al.,* 2005).

Numéricamente, en general la desestructuración sintáctica puede conducir a exponentes de Zipf $\beta>2$, o lo que es lo mismo $\beta'>1$, y en corpus con un exceso de nombres a exponentes incluso $\beta>3$ (Ferrer *et al.,* 2005; Ferrer, 2005a para una revisión de la diversidad de exponentes de la ley de Zipf).

Metodología

A partir de lo expuesto, es plausible suponer que existe una correspondencia entre la densidad sináptica y β. La hipótesis que se plantea aquí es que existirán diferencias en el exponente de Zipf en dos fases clínicas de la EA: GDS4 y GDS5. El análisis léxico y la distribución de frecuencias debería permitir observar exponentes de Zipf $\beta'=1$ en pacientes con GDS4 y en sujetos controles (con algunas desviaciones en palabras de baja frecuencia en los primeros) y $\beta'>1$ en pacientes con un GDS5 (Ferrer, 2006).

Corpus

Los corpora estudiados corresponden a la producción en español de 20 pacientes diagnosticados de Alzheimer, 10 de ellos GDS4 y otros 10 GDS5 (Reisberg *et al.,* 1982). Una parte de la muestra se obtuvo de una investigación doctoral sobre coherencia textual y enfermedad de Alzheimer (Brandao, 2005). Los pacientes se emparejaron con 10 controles emparejados en relación con el sexo, la lengua materna, la edad y los años de escolarización. No obstante, participaron más mujeres (n=19) que hombres (n=11), y su distribución fue desigual en todos los grupos (GDS4: cinco hombres y cinco mujeres, GDS5: dos hombres y ocho mujeres, control: cuatro hombres y seis mujeres). La lengua materna permitió una diferenciación igual (15 catalán, 15 español) y, prácticamente también, el resto de variables: edad (GDS4=78,66, *DE*=4,08; GDS5=81,83, *DE*=2,04; control=77,5, *DE*=3,2) y escolaridad (GDS4=6,5, *DE*=1,22; GDS5=4,33, *DE*=3,01; control=4,33, *DE*=2,33). Aparte de los test neuropsicológicos administrados y de la prueba específica para valorar la fluencia verbal, también se obtuvieron las puntuaciones con el *Mini-Mental State Examination* (Folstein, Folstein y McHugh, 1975): GDS4=23,33, *DE*=3,72; GDS5=16,83, *DE*=0,75; control=28,83, *DE*=1,16. Dadas las características de las pruebas administradas, no se han tenido en cuenta otras variables como la dominancia manual o el nivel socioeconómico.

A todos los sujetos se les administraron tres contextos tradicionales de producción oral:

1. Evocación narrativa sin pistas informativas (evocación espontánea): relato de un recuerdo o acontecimiento de su vida pasada (memoria episódica). Esta prueba se administraba sin ayuda verbal.

2. Producción con pistas verbales: relato de algún suceso o hecho del sujeto, haciéndole preguntas para que fuese avanzando y diera detalles.

3. Producción con pistas visuales: relato del cuento de Caperucita Roja con ayuda de dibujos y mínimas ayudas verbales.

De los 10 pacientes GDS4, dos fueron excluidos del estudio, dos no realizaron la segunda tarea y uno no realizó la tercera tarea. De los 10 pacientes GDS5, cinco no realizaron la segunda tarea y dos no realizaron la tercera.

Procedimiento

Todas las producciones (pacientes y controles) se transcribieron a partir de la grabación de las mismas. Para garantizar un audio correcto, realizaron la transcripción dos personas. Se transcribieron todas las palabras emitidas (incluyendo repeticiones o pseudopalabras), salvo algunas expresiones iniciales que el sujeto realizaba para planificar la emisión (pausas llenas o titubeos). Para facilitar la cuenta de la muestra se realizó una transcripción ortográfica

El corpus obtenido en los pacientes con GDS4 fue de 4.225 palabras, mientras que para el GDS5 fue de 1.528. Para poder realizar un análisis comparativo se seleccionaron corpora similares de los sujetos controles: así, se seleccionó un corpus A con 4.913 palabras y un corpus B con 1.481. Los corpora de los sujetos controles no se igualaron para garantizar que las frases comenzadas tuvieran un sentido final. La comparación se establece, por tanto, entre los pacientes con GDS4 y sus respectivos controles, y los GDS5 y los suyos, de forma que se descartó reducir el corpus de los GDS4 y sus controles para igualarlo a los GDS5.

A partir de estos datos se determinó automáticamente el número de palabras totales, W y la TTR, efectuándose *a posteriori* un análisis manual para evitar errores de conteo de palabras y en las estadísticas de frecuencias. Se midieron estos dos indicadores cuantitativos tradicionales (TTR y W), capaces, ya de *per se*, de discriminar a pacientes de controles, sin tener que recurrir a otros algo más complejos (como la CSU).

Una vez seleccionadas las cuatro muestras, se representó gráficamente para cada uno de los cuatro corpus (GDS4, GDS5, control A, control B) en escala log-log la probabilidad acumulada de determinar palabras de frecuencia f (P(f)) versus la frecuencia (f) y se hizo un primer ajuste con todos los datos mediante regresión lineal de manera que la pendiente de la recta nos diese el exponente β'. Tras este primer ajuste se realizó un segundo ajuste en el que se eliminaron los puntos correspondientes a las palabras de mayor y menor frecuencia, tratando de evitar la distorsión que dichos puntos pueden causar en el ajuste lineal con todos los puntos, como suele ser habitual (Newman, 2005; Goldstein, Morris y Yen, 2004).

Resultados

La tabla 2 recoge los principales resultados obtenidos. Los pacientes con EA presentan una proporción *types/tokens* claramente menor que los controles, así como un índice de Brunet mayor, como era de esperar (Bucks *et al.,* 2000).Estos índices tradicionales nos sirven para comparar la validez del exponente de Zipf como medio de detección de alteraciones lingüísticas. En todos los casos se sigue la ley de Zipf tanto en el primer ajuste como en el segundo, con coeficientes de correlación relativamente altos (>,988).

TABLA 2. ESTADÍSTICA OBTENIDA EN EL ANÁLISIS DE ZIPF DE LOS CORPORA DE LOS ENFERMOS DE ALZHEIMER GDS4, GDS5 Y CONTROL.

	Pacientes con EA			
	GDS4	*GDS5*	*Control A*	*Control B*
Tamaño corpus (*tokens*)	4225	1528	4913	1481
Palabras diferentes (types)	996	427	1927	671
Types/tokens (*TTR*)	**,236**	**,279**	,392	,453
Índice de Brunet (W)	**14,48**	**14,86**	11,47	12,11
Primer ajuste (con todos los puntos)				
Exponente de Zipf (β')	1,15±0,03	1,29±0,02	1,17±0,02	1,31±0,04
Coeficiente de correlación (ρ)	,988	,996	,995	,989
Puntos ajuste (*N*)	42	26	41	25
Segundo ajuste (eliminando primeros y últimos puntos)				
Exponente de Zipf (β')	1,07±0,05	**1,40±0,05**	1,10±0,02	1,14±0,04
Coeficiente de correlación (ρ)	,989	,991	,998	,993
Puntos ajuste (*N*)	28	18	29	17

Nota: p<,001.

Los exponentes encontrados en el primer ajuste para los GDS4 y GDS5 no difieren significativamente del de los controles correspondientes: así, tenemos para los GDS4 β'=1,15±0,03, mientras que los controles muestran β'=1,17±0,02; y en los GDS5 se obtuvo β'=1,29±0,02, mientras que su control es β'=1,31±0,04.

En el ajuste de la zona central de la distribución de frecuencias, en general se acercaron más los exponentes al teórico esperado, aunque con desviaciones al alza debidas al pequeño tamaño de los corpora, y mientras no se obtuvieron diferencias en los GDS4 (β'=1,07±0,03, frente a β'=1,10±0,01), sí hubo una variación significativa del exponente de los GDS5, en los que se β'=1,40±0,05, siendo el control β'=1,14±0,03. Por otra parte, el hecho de tener exponentes superiores a uno es habitual en muestras pequeñas (Lu, Zhang y Zhou, 2010): lo relevante es la desviación respecto a controles del mismo tamaño.

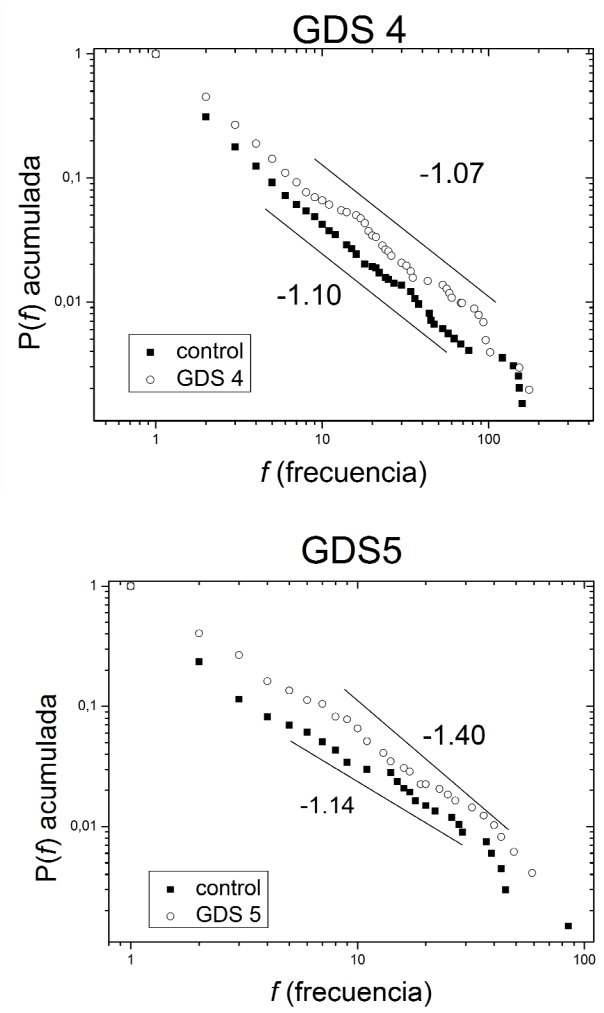

Figura 3. Representación de la frecuencia de la distribución de palabras, f, frente a su probabilidad acumulada P(f), para los pacientes diagnosticados GDS5 y GDS4 con sus respectivos controles.

Por otra parte, una revisión a las palabras más frecuentes (tabla 3) permite constatar que los enfermos con EA (GDS4 y GDS5) muestran la presencia de pronombres personales de primera persona (*yo, me, mí, nos*) como palabras más frecuentes, lo que no se aprecia en los sujetos controles. También en los controles se hallan contracciones (*al, del*) que, si bien en un principio se pensó "recontar" junto a las preposiciones y artículos correspondientes (*al=a+el, del=de+el*), se han dejado en la tabla 3 por no aparecer en los controles como palabras frecuentes (para los GDS4 tenemos 25 *al* y ocho *del*, y en los GDS5 se tienen siete *del* y cuatro *al*).

TABLA 3. PALABRAS MÁS FRECUENTES EN LOS CORPORA Y FRECUENCIA.

GDS4		**GDS5**		**Control A**		**Control B**	
palabra	*f*	*palabra*	*f*	*palabra*	*f*	*palabra*	*f*
y	257	la	59	y	172	de	86
que	180	y	59	de	159	que	45
la	153	que	49	la	153	el	43
de	102	no	43	a	151	la	39
se	96	de	40	en	141	y	37
a	95	el	36	que	121	por	29
el	94	se	32	se	76	en	28
no	87	**me**	27	el	68	a	26
en	69	a	26	un	62	se	22
me	60	sí	23	al	57	al	20
yo	58	esta	20	con	53	con	18
lo	57	aquí	17	por	47	Caperucita	17
los	53	en	17	esta	45	del	17
una	43	**mí**	17	no	44	un	16
era	35	**nos**	16	del	44	no	15

Experimentalmente se ha constatado que el exponente de Zipf en corpora pequeños siempre suele estar por encima del valor típico (en este caso $\beta'>1,0$) y ser más aproximado al teórico $\beta'=1,0$ en la zona central de ajuste (la tomada en el segundo análisis) que en la periferia o en el ajuste total de datos.

Discusión

El presente estudio es un primer avance de lo que puede suponer el análisis de Zipf en el diagnóstico de la patología del lenguaje, diagnóstico además realizado a partir de la producción oral, lo que supone una exploración con más ventajas que otras propuestas actuales.

El resultado más relevante de nuestra exploración, que hasta donde sabemos es la primera aproximación del análisis de Zipf a enfermos de Alzheimer, es el hecho de considerar el exponente de Zipf como un elemento más dentro de la investigación de la producción en patología del lenguaje, completando otros estudios de búsqueda de técnicas objetivas para la evaluación de trastornos lingüísticos en la demencia (Bucks *et al.,* 2000). A tenor de los resultados obtenidos, si bien parámetros tradicionalmente robustos como el TTR o el índice de Brunet discriminan de forma directa a los pacientes con EA de los controles, es curioso que sean las divergencias en el exponente de Zipf para la zona central de la distribución las que discriminen al grupo GDS4 del GDS5. En contra de lo que se podría esperar, no se han encontrado desviaciones en el exponente de Zipf de los datos para los enfermos GDS4 respecto de sus controles: esto podría indicar una cierta preservación de la sintaxis. En los pacientes diagnosticados GDS5, si bien el primer ajuste total de datos no mostró ninguna desviación significativa, sí se obtuvo un exponente β'=1,40 en la zona central de la distribución de frecuencias, que coincide con el límite superior establecido por Ferrer (2006) para la esquizofrenia con un discurso obsesivo.

Futuras investigaciones deben dilucidar si la presencia de palabras de alusión al propio sujeto (pronombres personales, reflexivos o determinantes posesivos) como étimos de alta frecuencia, así como la no aparición de contracciones dentro de esas palabras de alta frecuencia, permiten crear algún otro tipo de coeficiente estadístico que complemente los existentes para la detección automática de la EA, al menos para el español. La cuestión de las contracciones quizá haya podido pasar inadvertida en la literatura por la preponderancia de estudios en inglés.

A tenor de los datos y al no hallar ninguna desviación en el exponente de Zipf para enfermos diagnosticados con GDS4, y sí en los GDS5, el exponente de Zipf tal vez podría actuar como detector de un cierto agravamiento de la enfermedad, aunque faltaría verificarlo con corpora más completos y en estudios longitudinales de pacientes individuales. La brevedad y fragmentación de los corpora de los pacientes con EA es una de las dificultades para realizar estudios estadísticos concluyentes y puede forzar al agrupamiento de datos de diversos enfermos. No obstante, el análisis de Zipf realizado aquí parecería reforzar la preservación de la sintaxis hasta estadios relativamente avanzados de la enfermedad, como ya apuntaron Kempler, Curtiss y Jackson (1987).

También se confirman las predicciones de Ferrer (2006) y se refuerza la correspondencia entre densidad sináptica y β, confirmando la predicción de encon-

trar exponentes de Zipf β'>1. Además, es un resultado en sí mismo que, como toda desviación de la ley de Zipf, no es fácilmente explicable para los detractores de la ley de Zipf (Ferrer, 2005b).

Entendemos, para concluir, que lo que se ha realizado es una primera aproximación, hasta donde sabemos, de la aplicación de la ley de Zipf en el análisis de una patología verbal. Esta primera aproximación indica claramente que es necesaria una futura aplicación, al menos en dos sentidos consecutivos. Primero, se deberían obtener corpora más extensos para validar estos resultados y, sobre todo, para obtener un intervalo seguro de diferentes herramientas y medidas de magnitudes. Además, sería también necesario contrastar los resultados aquí obtenidos con estudios longitudinales y con pacientes en el intervalo GDS2-GDS6.

Segundo, y en el terreno más práctico, pero conectado con el anterior, es necesario también poner al alcance de las personas que trabajan en el diagnóstico de pacientes con EA este procedimiento de detección. Para conseguir este propósito se plantea, una vez conseguidos los datos, la creación de un programa informático que permita valorar la evolución verbal de estos pacientes. El único tratamiento manual por parte de la persona que realice el diagnóstico sería, entonces, la transcripción del habla.

REFERENCIAS

Almor, A., Kempler, D., MacDonald, M.C., Andersen, E.S. y Tyler, L.K. (1999). Why do Alzheimer patients have dificulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's Disease. *Brain and Language*, *67*(3), 202–227.

Alzheimer's Association, Thies, W. y Bleiler, L. (2011). 2011 Alzheimer's disease facts and figures. *Alzheimers Dementia, 7*(2), 208-244.

Aronoff, J.M., Gonnerman, L.M., Almor, A., Kempler, D. y Andersen E.S. (2006). Information content versus relational knowledge: Semantic deficits in patients with Alzheimer's disease. *Neuropsychologia*, *44*(1), 21-35.

Bates, E. y Goodman, J. (1999). On the emergence of grammar from the lexicon. In B. MacWhinney (Ed.), *The emergence of language* (pp. 29-79). Mahwah, NJ: Lawrence Erlbaum.

Berthier, M.L. y Dávila, G. (2010). Anticipando el futuro: diagnóstico de la enfermedad de Alzheimer en las fases predemencia y prodrómica. *Revista de Neurología, 51,* 449-450.

Brandao, L. (2005). *Produçao do discurso de portadores da Doença de Alzheimer em tres tarefas narrativas* (Tesis doctoral no publicada). Universidade Federal do Rio Grande do Sul, Brasil, Porto Alegre.

Bucks, R., Singh, S., Cuerden, J.M. y Wilcock, G. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analyzing lexical performance. *Aphasiology*, *14*(1), 71-91.

Carnero, C. y Lendínez, A. (1999). Utilidad del test de fluencia verbal semántica en el diagnóstico de la demencia. *Revista de Neurología, 29*, 709-714.

Carnero, C., Maestre, J., Marta, J., Mola, S., Olivares, J. y Sempere, A.P. (2000). Validación de un modelo de predicción de fluidez verbal semántica. *Revista de Neurología, 30*, 1012-1015.

Chitashvili, R.J. y Baayen, R.H. (1993). Word frequency distributions. In G. Altmann y L. Hrebcek (Eds.), *Quantitative Text Analysis* (pp. 54-135). Trier: Wissenschaftlicher Verlag Trier.

Cuetos-Vega, F., Menéndez-González, M. y Calatayud-Noguera, T. (2007). Descripción de un nuevo test para la detección precoz de la enfermedad de Alzheimer. *Revista de Neurología, 44*(8), 469-474.

Díaz-Mardomingo, C., Peraita-Adrados, H. y Garriga-Trillo, A.J. (2000). Problemas metodológicos al analizar datos de producción de ejemplares y de atributos en un estudio sobre deterioro semántico en enfermos de Alzheimer. *Psicothema, 12*(2), 192-195.

Fernández, T., Rios, C., Santos, S., Casadevall, T., Tejero, C., López-García, A., ... Pascual, L.F. (2002). 'Cosas en una casa': una tarea alternativa a 'animales' en la exploración de la fluidez verbal semántica: estudio de validación. *Revista de Neurología, 35,* 520-523.

Ferrer, R. (2005a). The variation of Zipf's law in human language. *European Physical Journal B, 44*, 249-257.

Ferrer, R. (2005b). Decoding least effort and scaling in signal frequency distributions. *Physica A: Statistical Mechanics and its Applications, 345*, 275-284.

Ferrer, R. (2006). When language breaks into pieces. A conflict between communication through isolated signals and language. *BioSystems 84*, 242-253.

Ferrer, R., Bollobás, R. y Riordan, O. (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceeding of the Royal Society of London, Series B, 272*, 561-565.

Ferrer, R. y Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE, 5*(3), e9411.

Ferrer, R. y Solé, R.V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences, 100*, 788-791.

Ferrer, R., Solé, R.V. y Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E, 69*, 051915.

Fleisher, A.S., Sowell, B.B., Taylor, C., Gamst, A.C., Petersen, M.D. y Thal, L.J. (2007). Clinical predictors of progression to Alzheimer disease in amnestic mild cognitive impairment. *Neurology, 68*(19), 1588-1595.

Folstein, M.F., Folstein, S.E. y McHugh, P.R. (1975). "Mini-Mental State": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189-198.

Garcés, M., Santos, S., Pérez, C. y Pascual, L.F. (2004). Test del supermercado: datos normativos preliminares en nuestro medio. *Revista de Neurología, 39*, 415-418.

Garrad, P., Maloney, L.M., Hodges, J.R. y Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain, 128*(2), 250-260.

Goldstein, M.L., Morris, S.A. y Yen G.G. (2004). Problems with fitting to the power-law distribution. *European Physical Journal B, 41*, 255–258.

Hernández-Fernández, A. (2005). *La ley de Zipf en el método comparativo* (Tesina no publicada). Universidad de Barcelona.

Hier, D.B., Hagenlocker, K. y Shindler, A.G. (1985). Language disintegration in dementia: Effects of etiology and severity. *Brain and Language, 25*(1), 117–133.

Kempler, D. (1995). Language changes in dementia of the Alzheimer type. In L. Lubinski (Ed.), *Dementia and communication: Research and clinical implications* (pp. 98-114). San Diego: Singular Publishing Group.

Kempler, K.D., Curtiss, S. y Jackson, C. (1987). Syntactic preservation in Alzheimer's disease. *Journal of Speech and Hearing Research, 30*(3), 343-350.

Keselj, V., Peng, F., Cercone, N. y Thomas, C. (2003). N-gram based author profiles for authorship contribution. *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, Dalhousie University, Halifax, Nova Scotia, Canada, agosto 2003, 255-264.

Li, W., Miramontes, P. y Cocho, G. (2010). Fitting ranked linguistic data with two-parameter functions, *Entropy*, *12*, 1743-1764.

Lu, L., Zhang, Z.K. y Zhou, T. (2010). Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE 5*(12), e14139. doi:10.1371/journal.pone.0014139

Montemurro, M.A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A*, 300, 567-578.

Montemurro, M.A. y Zanette, D. (2002). Frequency-rank distribution in large samples: Phenomenology and models. *Glottometrics*, *4*, 87-98.

Mulet, B., Sánchez-Casas, R.M., Arrufat, M.T., Figuera, L., Labad, A. y Rosich, M. (2005). Deterioro Cognitivo Ligero anterior a la enfermedad de Alzheimer: tipologías y evolución. *Psicothema, 17*(2), 250-256.

Newman, M.E.J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics,46*, 323-351. Disponible en: arXiv:cond-mat/0412004v3

Pakhomov, S., Chacon, D., Wicklund, M. y Gundel, J. (2011). Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of Iris Murdoch's writing. *Behavioral Research Methods, 43*(1), 136-144.

Patterson, K.E., Graham, N. y Hodges, J.R. (1994). Reading in dementia of the Alzheimer type: A preserved ability? *Neuropsychology, 8*(3), 395-412.

Peraita, H., González, M.J., Sánchez, M.L. y Galeote, M.A. (2000). Batería de evaluación del deterioro de la memoria semántica en Alzheimer. *Psicothema, 12*(2), 192-200.

Piotrowska, W., y Piotrowska, X. (2004). Pathological text and its statistical parameters. *Journal of Quantitative Linguistics*, *11*(2), 133-140.

Piotrowski, R.G., Pashkovskii, V.E. y Piotrowski, V.R. (1994). Pshychiatric linguistics and automatic text processing. *Automatic Documentation and Mathematical Linguistics, 28*(5), 28-35.

Reisberg, B., Ferris, S.H., De León, M.D. y Crook, T. (1982).The global deterioration scale for assessment of primary degenerative dementia. *American Journal of psychiatry, 139*, 1136-1139.

Semenza, C., Mondini, S., Borgo, F., Pasini, M. y Sgaramella, M.T. (2003). Proper names in patients with early Alzheimer's disease. *Neurocase, 9,* 63-69.

Shuttleworth, E.C. y Huber, S.J. (1988). The naming disorder of dementia of Alzheimer type. *Brain and Language, 34*(2), 222–234.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K. y Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *Proceedings of IEEE ICMA 2005,* Niagara Falls, Ontario, Canada, julio 2005.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, *3*(1), 38-50.

Valls-Pedret, C., Molinuevo, J.L. y Rami, L. (2010). Diagnóstico precoz de la enfermedad de Alzheimer: fase prodrómica y preclínica. *Revista de Neurología, 51*, 471-480.

Wancata, J., Musalek, M., Alexandrowicz, R. y Krautgartner, M. (2003). Number of dementia sufferers in Europe between the years 2000 and 2050. *European Psychiatry, 18*(6), 306-313.

Zipf, G.K. (1942). Children's speech. *Science*, *96*, 344-345.

Zipf., G.K. (1949). *Human behaviour and the principle of least effort. An introduction to human ecology*. Cambridge, MA: Addison-Wesley; reimpreso en Zipf, G.K. (1972). *Human behaviour and the principle of least effort: An introduction to human ecology* (1ª ed., pp. 19-55). New York: Hafner.

anuario de psicología / The UB Journal of Psychology

Artículos

P.V.P.: 14€ (I.V.A. incluido)

ISSN 1988-5253
9 771988 525007

ISSN 00665126
9 770066 512007
36105