Las leyes de la lingüística en los sistemas de comunicación

Antoni Hernández-Fernández

Las leyes de la lingüística en los sistemas de comunicación

Tesis doctoral presentada por Antoni Hernández-Fernández para la obtención del título de doctor por la Universidad de Barcelona.

Codirectores: Ramon Ferrer-i-Cancho y Faustino Diéguez-Vide.
Tutora: Maria Antònia Martí Antonín

Programa de doctorado: Ciencia Cognitiva y Lenguaje (CCiL).

Departamento: Lingüística General.

Agradecimientos

A todos los que habéis hecho posible este trabajo con vuestra ayuda, tanto en lo intelectual como en lo biológico y en lo emocional.

Índice

INTRODUCCIÓN

Más allá del lenguaje

En esta introducción resumimos los propósitos y las principales aportaciones de los artículos que estructuran este trabajo. En los capítulos no se ha pretendido repetir lo que, con más detalle, especialmente en lo referente a la metodología y la estadística, se detalla en cada artículo. Dada la diversidad y transversalidad de nuestro enfoque se han realizado sucintas presentaciones generales que los especialistas tal vez encuentren triviales pero que pretenden ser de ayuda para el no experto. Como se verá, en definitiva, se ha tratado de extender la lingüística cuantitativa más allá del lenguaje. Ha sido un trabajo con un elevado componente colaborativo, sinergético, habitual en la ciencia moderna donde el total, creemos, es más que la suma de las partes.

Si me preguntan qué hay tras estos años de trabajo, en una frase, les diría que muchas ganas de comprender la naturaleza. La física y la lingüística me han conducido y ayudado en la exploración, que prosiguió formativamente con los diversos cursos de doctorado en ciencia cognitiva y lenguaje, y que culmina ahora con estas páginas que, no obstante, no suponen un final sino parte del camino. Porque el recorrido, el proceso de aprendizaje y descubrimiento, suele ser más gratificante que la meta cuando nos apasionamos por lo que hacemos.

Si concreto un poco más, otro propósito fundamental ha sido el ánimo de aportar un granito de arena para lograr que la lingüística crezca y se desarrolle como ciencia de pleno derecho, al defender y promulgar la extensión de los rasgos que, según veremos en el primer capítulo, hacen –o harán– de la lingüística una ciencia. La perspectiva científica en el estudio del lenguaje y la comunicación la hallé por fortuna en mis dos directores de tesis, Ramón Ferrer-i-Cancho y Faustino Diéguez-Vide, que me ayudaron a descubrir el potencial de la lingüística cuantitativa y la neurolingüística, ámbitos en los que encontré el humanismo en sentido estricto, en los que al fin se diluía y superaba la clásica división entre ciencias y letras, y que configuran la temática de este trabajo,

Decidir presentar la tesis como compendio de publicaciones deja fuera de la misma bastante material no publicado, en general relacionado con la ley de Zipf en la adquisición del lenguaje y la comunicación animal, material que, en parte, fue presentado para la obtención de la suficiencia investigadora (Diploma de Estudios Avanzados) y que supuso un esfuerzo formativo y también un punto de partida previo no reflejado aquí. No obstante, se incluye material no publicado, especialmente en los capítulos primero y quinto, y en este último un artículo bajo revisión.

Las publicaciones son la punta del iceberg de una obra colectiva, en las que se refleja mínimamente las muchas horas de trabajo invertido, personalmente o por otros coautores, en el tratamiento y recogida de datos y en el desarrollo de algoritmos y programas informáticos. Por otra parte, también hemos estado a merced de la calidad de los datos que otros investigadores (a los que estamos muy agradecidos) han recopilado. La ciencia moderna es una labor colectiva. Se han añadido, además de las publicaciones, algunos materiales originales en las introducciones a los artículos realizadas en cada capítulo, de carácter más divulgativo.

En esta tesis se ha hecho algo bastante sencillo: contar elementos y medir su tamaño. Navaja de Ockham. En ocasiones nos hubiese gustado, por ejemplo, medir la "longitud" de las palabras de otra manera (temporalmente, no en número de letras), y

para contarlas sabemos que algunos criterios seguidos pueden considerarse arbitrarios. Es el caso de la escritura: al trabajar con corpus escritos el espacio en blanco es, por convención, el delimitador de palabras, aunque seguro que no es el ideal conceptual. La tecnología de la escritura sigue influyendo demasiado en el estudio de las lenguas. Además de palabras, hemos contado y medido cromosomas (tomando las bases nitrogenadas como unidades de medida) e infoquímicos (tomando el número de átomos y la masa atómica como unidad de medida), en un camino de sencillez en la aproximación que, sin embargo, no ha supuesto una trivialidad estadística o matemática posterior: desde el desarrollo de la física del caos y el estudio de la dinámica de sistemas se ha comprobado que las aproximaciones o asunciones sencillas pueden dar como resultado pautas complejas.

0.1. Presentación de los trabajos

Repasemos de forma sucinta qué contiene cada uno de los capítulos y artículos del presente trabajo. En la tabla 0.1 se resume la temática y objetivos iniciales de cada uno. Se deja para el capítulo final la discusión y un compendio de los resultados y conclusiones más relevantes.

En el primer capítulo presentamos un breve resumen de la filosofía que subyace tras nuestro enfoque. No es fundamental ni esencial, ni tampoco se solicitaba en las normas de presentación, pero responde también tanto al carácter interdisciplinar de este doctorado como al momento de transición que vive la lingüística actual. Es un posicionamiento filosófico personal que ubica a la lingüística entre la biología y la teoría de la información, y que posee el respaldo de buena parte de la comunidad científica. El lector que lo desee puede saltárselo.

En el segundo capítulo hacemos una breve revisión de la lingüística cuantitativa y algunas de sus leyes: la ley de Zipf, la ley de Menzerath-Altmann y la ley de brevedad. Se contribuye aquí con un capítulo de libro y tres artículos:

- En el capítulo de libro (Ferrer-i-Cancho y Hernández-Fernández, 2008) se presenta la ley de Zipf y se revisa la relación entre el exponente de la distribución de frecuencias y el exponente de la relación potencial entre la frecuencia y su rango. Se demuestra que ambos exponentes coinciden únicamente cuando su valor es el número de oro (Ferrer-i-Cancho y Hernández-Fernández, 2008).

- Tras revisar las desviaciones de la ley de Zipf en el lenguaje, se presenta el primer estudio de estas desviaciones realizado con corpus de enfermos de Alzheimer, con el ánimo de detectar la evolución verbal de la enfermedad, lo que en el futuro podría ayudar a mejorar la detección precoz de la patología (Hernández-Fernández y Diéguez-Vide, 2013).
- Al explicar la ley de Menzerath-Altmann se creyó oportuno presentar una revisión general sobre su no trivialidad estadística, y de la capacidad de los test que se aplican, ya que es un tema de actualidad (Ferrer-i-Cancho *et al.*, 2014), aunque el artículo está pendiente de publicación.
- Finalmente, tras exponer la ley de brevedad, se añade un artículo en el que se corroboró su presencia en los corpus de siete lenguas y se exploró, con diversos resultados, en el repertorio de delfines y en las emisiones de primates no humanos y cuervos (Ferrer-i-Cancho y Hernández-Fernández, 2013).

El tercer capítulo presenta algunos principios generales que rigen la comunicación, entre ellos el principio de compresión, que tiene como consecuencia la ley de brevedad trabajada en el capítulo anterior. El principio de compresión, originario de la teoría de la información, se propone para la ciencia cognitiva y la comunicación en el único artículo de este capítulo (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013), que es quizá la contribución más importante de esta tesis, si consideramos el impacto y relevancia de su publicación en *Cognitive Science*.

El cuarto capítulo se centra en la investigación sobre la presencia de la ley de Menzerath-Altmann en el genoma, en un ejemplo de aplicación de la biolingüística. Tras una introducción general a la genética, y específica sobre el nivel cromosómico, se recogen cuatro artículos y un *erratum*, que suponen la extensión de las leyes y métodos de la lingüística cuantitativa al estudio de la genética:

- En el primero de ellos (Hernández-Fernández *et al.*, 2011) se siguió la línea de trabajo de Ferrer-i-Cancho y Forns (2009) y se escudriñó la presencia de la ley de Menzerath-Altmann en el cariotipo de plantas angiospermas y gimnospermas, mamíferos, reptiles, pájaros, hongos, insectos, anfibios y diversos tipos de peces. Se aportó además por primera vez un diagrama de fases bidimensional en el que se representa el número de cromosomas haploide respecto el tamaño total del genoma,

una representación en la que se distinguen y separan claramente algunas agrupaciones de puntos de los diversos tipos de especies estudiados.

- Posteriormente, tras la crítica de Solé (2010) al artículo de Ferrer-i-Cancho y Forns (2009), se revisó la metodología y la estadística de nuestro trabajo y se respondió en Ferrer-i-Cancho y colaboradores (2012), artículo en el que se rebaten, matemáticamente y conceptualmente, los argumentos de Solé (2010). Es sin duda una muestra de la relevancia del debate constructivo en la ciencia moderna.
- Finalmente, en Baixeries y colaboradores (2012) se analiza el modelo de juguete presentado por Solé (2010), y se aprovecha para hacer una revisión crítica de los modelos de fragmentación aleatoria (Sankoff y Ferreti, 1996), utilizados en algunas aproximaciones a la genómica. El cuarto artículo de este capítulo es el *erratum* de Baixeries y colaboradores (2012) en el que se precisaron los teoremas y corolarios del artículo, afinando en el uso del concepto de "independencia" e "independencia media" estadística (Ferrer-i-Cancho, Baixeries, *et al.*, 2013), aunque sin consecuencias para las tesis y argumentos generales de Baixeries y colaboradores (2012).

En el quinto capítulo se explora el fenómeno de la comunicación química, más allá del ADN. Tras definir los conceptos de infoquímico, feromona y aleloquímico, en el único artículo de este capítulo (Hernández-Fernández y Ferrer-i-Cancho, 2014, bajo revisión) se analizó cuantitativamente la base de datos Pherobase (El-Sayed, 2012), encontrando dos regímenes de la ley de Zipf en la distribución de infoquímicos según el grado o número de especies que utiliza cada sustancia, lo que demostraría que también hay un repertorio químico nuclear y otro periférico en la comunicación química, en analogía a lo que sucede en el lenguaje (Ferrer-i-Cancho y Solé, 2001).

Para concluir, se plantean algunas consideraciones generales y se discuten algunas líneas abiertas de investigación: es inevitable que, al trabajar con datos, siempre ansiemos implementar los que disponemos y se abran las puertas de trabajos futuros en los que mejorar cada análisis. A mejores datos, mejores resultados. Porque, como decíamos al principio, presentar esta tesis no es un punto final sino parte de nuestro camino, del sendero que recorremos con ganas de comprender mejor la Naturaleza.

Artículos de esta tesis y **temática** de cada uno	Capítulo	Objetivos generales
1. Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2014). The infochemical core. *Journal of Quantitative Linguistics*, enviado el 30 de julio de 2013, pendiente de publicación. **Lingüística cuantitativa, comunicación animal, comunicación química.**	5.3	Análisis cuantitativo de la base de datos *Pherobase*, del grado o distribución de uso de infoquímicos. Selección del mejor modelo para la relación entre grado y rango. Exploración de la relación entre el lenguaje humano y la comunicación química.
2. Hernández-Fernández, A. y Diéguez-Vide, F. (2013). La ley de Zipf y la detección de la evolución verbal en la enfermedad de Alzheimer, *Anuario de Psicología/The UB Journal of Psychology*, 43 (1), 67-82. **Lingüística cuantitativa, neurolingüística, patología del lenguaje.**	2.4.	Estudio de la evolución verbal y de la distribución de frecuencias de palabras en enfermos de Alzheimer. Breve revisión de las desviaciones de la ley de Zipf en la neurolingüística. Contribuir al desarrollo de métodos cuantitativos de diagnóstico.
3. Hernández-Fernández, A., Baixeries, J., Forns, N. y Ferrer-i-Cancho, R. (2011). Size of the whole versus number of parts in genomes. *Entropy*, 13 (8), 1465–1480, doi:10.3390/e13081465. **Lingüística cuantitativa, genómica.**	4.4.	Valoración de la relación existente entre el tamaño de los cromosomas y el genoma de diversos grupos de especies y estudio de la presencia de la ley de Menzerath-Altmann. Refutar argumentos contra la relevancia de la ley de Menzerath en el genoma.
4. Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2013). The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, 4 (1), 45-55. **Lingüística cuantitativa, comunicación animal, lingüística comparativa.**	2.8.	Estudio comparativo de la ley de brevedad de Zipf en el lenguaje y en algunos sistemas de comunicación animal. Justificación de las excepciones a la ley de brevedad en el caso de dos primates.
5. Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J. y Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*. DOI: 10.1111/cogs.12061. **Lingüística teórica, ciencia cognitiva, comunicación animal.**	3.3.	Presentación, justificación y argumentación sobre la operatividad del principio de compresión en los sistemas de comunicación y en el comportamiento animal. Mostrar la relación del principio de compresión con las leyes de la comunicación.
6. Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G. y Baixeries, J. (2013). The challenges of statistical patterns of language: the case of Menzerath's law in genomes, *Complexity*, 18, 11–17. **Lingüística cuantitativa, genómica, biolingüística.**	4.6.	Argumentación y demostración matemática de la relevancia de la presencia de la ley de Menzerath-Altmann en el nivel cromosómico del genoma y refutación de los argumentos en contra. Comparación del ADN con el lenguaje.
7. Ferrer-i-Cancho, R., Baixeries, J., Hernández-Fernández, A., Debowski, L. y Macutek, J. (2014). *When is Menzerath-Altmann law mathematically trivial? A new approach*. Pendiente de publicación. Disponible en: http://arxiv.org/abs/1210.6599 **Lingüística cuantitativa, lingüística matemática, genómica.**	2.6.	Revisión y mejora de los test estadísticos de correlación entre el número de pares de bases y el número de cromosomas, que permiten valorar la presencia de la ley lingüística de Menzerath-Altmann en el genoma.
8. Ferrer-i-Cancho, R., Baixeries, J. y Hernández-Fernández, A. (2013). *Erratum* to "Random models of Menzerath-Altmann law in genomes" (*BioSystems* 107 (3), 167–173). *BioSystems* 111 (3), 216-217. **Lingüística matemàtica, genómica.**	4.8.	Mejora y puntualizaciones matemáticas y estadísticas de los argumentos dados en Baixeries *et al.* (2012), en la revisión de los modelos de fragmentación aleatoria en el genoma. Comparación entre los sistemas lingüísticos y los genéticos.
9. Baixeries, J., Hernández-Fernández, A., Forns, N., y Ferrer-i-Cancho, R. (2013). The parameters of Menzerath-Altmann law in genomes," *Journal of Quantitative Linguistics*, 20, 94–104. **Lingüística cuantitativa, genómica, biolingüística.**	4.6.	Cálculo de los parámetros de la ley de Menzerath-Altmann en el nivel cromosómico del genoma de diversos grupos de especies, suponiendo su validez. Comparación cualitativa y cuantitativa entre modelos y grupos de especies.
10. Baixeries, J., Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2012). Random models of Menzerath-Altmann law in genomes. *BioSystems* 107 (3), 167–173. **Lingüística cuantitativa, lingüística matemática, genómica, biolingüística.**	4.8.	Revisión conceptual y cuantitativa de los modelos de fragmentación aleatoria en el genoma y su consistencia con la presencia de la ley de Menzerath-Altmann. Refutación de argumentos contra la ley de Menzerath-Altmann.
11. Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2008). Power laws and the golden number. En: G. Altmann, I. Zadorozhna y Y. Matskulyak (eds.), *Problems of general, germanic and slavic linguistics* (pp. 518-523). Chernivtsi: Books – XXI. **Lingüística cuantitativa, lingüística matemática.**	2.2.	Presentación general de la ley de Zipf y revisión de la relación matemática existente entre el exponente de Zipf de la distribución de frecuencias de palabras y el exponente de la distribución de rangos.

Tabla 0.1.: Ubicación de los trabajos que conforman esta tesis, indicando su temática y objetivos generales.

CAPÍTULO 1

Una filosofía subyacente

Toda actividad intelectual posee una filosofía subyacente. Cabe situarse antes de abordar toda tesis para no perturbar la experimentación y el desarrollo teórico o, como mínimo, conocer cómo se influye en el sistema de estudio. Como es habitual en biología, y dado que se considera la lingüística una ciencia, se defiende el realismo gnoseológico, bajo el cual el mundo existe por sí mismo y puede ser investigado, y el materialismo naturalista: luego hay que fundamentar con evidencias empíricas toda hipótesis.

Las ciencias no tratan de explicar y casi no intentan interpretar: se consagran a hacer modelos de la realidad (Von Newmann, 1955). Por modelo se entiende una construcción matemática que describe los fenómenos observados en la naturaleza y cuya justificación es única y se asienta tanto sobre la experiencia como sobre elementos formales. Y citando a Von Newmann y Morgenstern (1944):

> What is important is the gradual development of a theory, based on a careful analysis of the facts. (...) The theory finally obtained must be mathematically rigorous and conceptually general. Its first applications are necessarily to elementary problems where the result has never been in doubt and no theory is actually required. At this early stage the application serves to corroborate the theory. The next stage develops when the theory is applied to somewhat more complicated situations in which it may already lead to a certain extent beyond the obvious and familiar. Here theory and application corroborate each other mutually. Beyond lies the field of real success: genuine prediction by theory. It is well known that all mathematized sciences have gone through these successive stages of evolution.

Se ha intentado, por tanto, que los datos vayan de la mano de los modelos teóricos propuestos como explicativos. Sin ánimo de ser exhaustivos, introduciremos en este apartado algunos conceptos básicos sobre la filosofía subyacente de los modelos que más han influido en el marco de nuestro trabajo y que están tras los artículos que vertebran esta tesis.

Mario Bunge (2010) acusa a muchos científicos de no considerar en sus estudios los postulados filosóficos que defienden: no cometeremos el mismo error de partida. Como mínimo hay que ser consciente de dónde estamos y a dónde desearíamos llegar, para lo cual es imprescindible posicionarse desde el principio, también filosóficamente. A tal efecto, adoptaremos el marco de la epistemología bungeana y lo aplicaremos a la lingüística.

Sostiene el propio Bunge (1978) que toda actividad intelectual, sea auténtica o sea falsa, posee una filosofía subyacente, y la ciencia no es una excepción. En concreto, posee una gnoseología (teoría del conocimiento) y una ontología (teoría sobre el ser y devenir). Aquí defenderemos el realismo gnoseológico, según el cual el mundo existe por sí mismo, independientemente de la mente del observador, y el materialismo naturalista, propio de la biología evolutiva, que junto al realismo nos conduce a defender la centralidad de la evidencia empírica en todo constructo teórico que se pretenda fundamentar o desarrollar. El naturalismo implica que los elementos reales, nuestros objetos de estudio, poseen energía o materia, y se pueden ajustar a leyes

(causales o probabilísticas), además de considerar los procesos mentales como procesos cerebrales (Bunge, 2001), con un claro correlato neuronal (Rumelhart y McClelland, 1986).

En definitiva, necesitamos pruebas, experimentales y reproducibles, para que nuestras hipótesis y modelos estén fundamentados: conjeturar no es lo mismo que descubrir, y aunque en muchas ocasiones son las conjeturas las que nos conducen al hallazgo y, parafraseando a Jorge Waggensberg (2007), el gozo intelectual está en ambas facetas de la ciencia. El placer del descubrimiento está más en demostrar con rigor que en enunciar verdades probables.

En el fondo, hay tres ideas que quizá fuerzan a que nuestros fundamentos filosóficos sean el realismo y el naturalismo:

1. Para empezar, la idea de que la lingüística es una ciencia (o el deseo de que llegue a serlo, como se verá), una ciencia que se podría fundamentar en la física y en la matemática. Centrándose en la "biolingüística", Fitch (2009) afirma en esta línea:

> Biolinguistics is not yet a science — it is more a loosely-defined collection of questions and approaches — but it certainly has the potential to become a science.

2. Seguidamente la concepción –que se podría tildar de reduccionista, aunque argumentaremos que no lo es– de la navaja de Ockham, según la cual ante dos explicaciones de un mismo fenómeno deberíamos apostar por la más sencilla: la idea es no complicar sin motivo el punto de partida, la forma de aproximarnos al problema de estudio, aunque para ello se complique el posterior análisis de datos. La navaja de Ockham es aplicable siempre que la teoría sea explicativa, es decir el reduccionismo emerge cuando la simplicidad implica perder potencialidad y capacidad en la comprensión fenomenológica.

3. Por último, entendemos que la teoría matemática de la comunicación, tal y como Claude Shannon (1948) la constituyó en su día, es otro de los pilares conceptuales de la ciencia del lenguaje, fundamental en la formalización de la lingüística cuantitativa.

Así pues, vayamos por partes con cada uno de estos tres puntos.

1.1. La ciencia del lenguaje

Se sostiene que la lingüística es una ciencia. Deberíamos responder entonces a la pregunta gnoseológica *¿Qué es la ciencia?*, cuestión ya formulada por Platón y profusamente debatida a lo largo de la historia de la filosofía (Bunge, 2010) con esfuerzos notables como el del círculo de Viena que presentó a la filosofía como la encargada de discernir qué es conocimiento científico y que no (Kraft, 1986) y con múltiples aproximaciones epistemológicas (para una revisión, Chalmers, 2000).

No es nuestro objetivo aquí extendernos en lo que daría para otra tesis, pero definiremos nuestra postura de forma sucinta. En muchas ocasiones se reduce la definición de ciencia a un único rasgo: la aplicación del método científico. Y es de nuevo Bunge el que nos alerta de que no podemos caer en semejante reduccionismo: si la lingüística es una ciencia, sus investigaciones deben entonces poseer como mínimo una décupla (Bunge, 2010) de elementos generales. Aplicados a la lingüística estos son:

1. **Comunidad de investigadores**: es un hecho que la lingüística posee en todo el mundo una comunidad científica que se ha formado en los fundamentos de la lingüística, se comunica y posee una tradición de investigación, aunque todavía buena parte de dicha comunidad se mantenga dividida, en cuanto a considerar a la lingüística una ciencia. Para empezar muchos lingüistas no actúan como científicos, en lo referente a algunos de los puntos siguientes.
2. Hay una **sociedad** que alberga a la comunidad científica y fomenta (o tolera) la actividad de la lingüística.
3. El **dominio** de investigación de la lingüística consta de entidades reales. Aquí, en la definición de dominio, surgen algunas discrepancias con parte de la comunidad científica, a saber:
    a) La lingüística se puede definir como la ciencia del lenguaje, del que tenemos evidencias empíricas suficientes (las lenguas), como para no tratarlo meramente como una idea o entidad subjetiva. O, de forma más extensa, la lingüística se podría definir como la ciencia de los sistemas de comunicación, e incluiría entonces, además del lenguaje, a los sistemas de comunicación de otros seres vivos (comunicación animal, vegetal y de otros reinos) y máquinas. Así, aunque la lingüística nace históricamente como el estudio de las lenguas, amplia su campo de estudio a medida que el ser humano se va liberando de su antropocentrismo y va conociendo la

complejidad de la comunicación de otros seres vivos, y va siendo capaz, especialmente en el siglo XX, de desarrollar tecnologías como la informática o la electrónica, a través de la cual diseña y construye sensores que le permiten escudriñar la realidad más allá de los límites de sus sentidos. La información y sus procesos, pasan pues a formar parte de una lingüística mucho más amplia, centrada en la comunicación tanto de máquinas como de seres vivos (Gleick, 2011). La lingüística ya no solo emplea los computadores para el estudio de las lenguas (lingüística computacional), sino que es capaz de emplear trasversalmente sus fundamentos en otras disciplinas y dominios en los que aparece la información, como es el caso de la genética, la computación o la etología (Bel-Enguix, Dahl y Jiménez-López, 2011).

b) El lenguaje puede ser entendido como el dominio de la lingüística, bien como una realidad cognitiva exclusivamente humana que, pese a la subjetividad –por definición– de las mentes que poseen el lenguaje, tiene una realidad palpable (las lenguas); o bien como una realidad más extensa, de la que las lenguas no serían sino una evidencia empírica más (en humanos), de un mundo en el que todos los seres vivos se comunican, con rasgos más o menos compartidos, según las especies y las modalidades de comunicación (Riba, 1990, para una revisión zoosemiótica), y en el que el lenguaje humano destaca, en muchas ocasiones, no tanto de forma cualitativa como cuantitativa.

4. La **perspectiva general** de la lingüística posee una gnoseología realista, ya que podemos investigar los hechos reales del lenguaje, pese a las diferencias que pueda haber para algunos en su dominio. El trasfondo filosófico de la lingüística se fundamenta en una ontología de objetos que poseen una dinámica, pero sujeta a leyes. En este punto, en defender la existencia de leyes lingüísticas y que éstas sean o no universales, o derivadas de principios universales, también existe cierta controversia entre parte de la comunidad científica.
5. El **trasfondo formal** de la lingüística debe estar constituido por la matemática y la lógica, y sus teorías aceptadas y actualizadas, y no teorías formales obsoletas. El método científico, con su formulación y contrastación de hipótesis, está tras el trasfondo formal. Este punto también divide a la comunidad científica, pues todavía existe un subconjunto de lingüistas que se aferran a modelos tradicionales

descriptivos de las lenguas pero que no utilizan en ningún caso ni la lógica ni la matemática como trasfondo. La lingüística matemática es la rama de la lingüística que se ha especializado en este trasfondo formal, con la lingüística cuantitativa como subdisciplina especializada en el análisis estadístico de datos. Según Köhler (2005) las propiedades de los elementos lingüísticos y sus relaciones se infieren de leyes universales que se pueden formular de manera análoga a otras ciencias naturales (cuyos datos provienen también a menudo de procesos estocásticos), ciencias que han abandonado su paradigma causal y determinista para pasar a una perspectiva estadística y probabilística, sin perder por ello un ápice de rigor y permitiendo la formulación de leyes formales.

6. El **trasfondo específico** de la lingüística lo conforman los datos, sobre los que se formulan hipótesis y teorías confirmadas, aplicando metodologías contrastadas. No obstante, las teorías se corrigen paulatinamente a medida que surgen nuevos y mejores datos. En la actualidad, la lingüística cuantitativa aglutina ambos trasfondos, formal y específico. Los datos configuran un dominio restringido de estudio (las lenguas o la comunicación animal, por ejemplo), y permiten realizar hipótesis sobre el dominio general (el lenguaje o los sistemas de comunicación).
7. La **problemática** de la lingüística se compone de problemas referentes a cualquiera de los puntos anteriores, problemas cognitivos referentes a la naturaleza y sus leyes más generales, aplicadas. Así el paradigma de la información y sus leyes se erige en uno de los problemas básicos de la naturaleza que debe ser resuelto. Aunque la lingüística nació del estudio de las lenguas humanas, y posteriormente se abundó durante siglos en la concepción antropocéntrica del lenguaje, las teorías de la información se extendieron para mostrarnos que la información forma parte inherente de la naturaleza, como lo pueda ser la energía, la masa o la gravedad y el resto de fuerzas que gobiernan la dinámica del Cosmos. Pensemos que también otras ciencias, como la física o la química, tardaron cientos de años en definirse y consolidarse como tales, generalmente partiendo de concepciones animistas y antropocéntricas: ¡todavía en la Edad Media a los médicos se les conocía como físicos! Es reciente el postulado que afirma que la física estudia una realidad que existe y es independiente del ser humano.
8. El **fondo de conocimiento** de la lingüística lo conforman teorías, hipótesis y datos actualizados y comprobables, compatibles con el trasfondo específico y que se pueden organizar de forma que constituyen ya un acervo intelectual revisable y

sobre el que ir desarrollando la ciencia del lenguaje. Al respecto, cabe notar que todavía parte de la comunidad científica insiste en construir modelos y teorías no revisables, que describen fenómenos puntuales y que sus autores principales deben periódicamente reformular (a veces prácticamente por completo), sin haber logrado en años establecer un fondo de conocimiento organizado y fundamental, por ejemplo capaz de estructurar un curso de lingüística ordenado que permita al estudiante iniciarse de forma sólida y progresiva, como puede por ejemplo hacer cualquier estudiante de física que empiece a aprender cinemática. El fondo de conocimiento todavía es limitado en la lingüística, debido tanto a la juventud de la lingüística (en comparación con otras ciencias) como al hecho de que parte de la comunidad científica prosiga formulando constructos teóricos no revisables. Andersen (2001) ya argumentó que en las ciencias desarrolladas los libros de texto que los principiantes deben abordar contienen el conocimiento común compartido (respecto del cual no hay versiones distintas o contrapuestas), sin entrar en disquisiciones históricas ni en divergencias, propias de las ciencias inmaduras en el sentido kuhneano, como puede ser el caso de la lingüística (Otero, 2004).

9. En consecuencia, para la creación de un fondo de conocimiento revisable, los **objetivos** de los miembros de la comunidad científica incluyen la sistematización del dominio de estudio y de la problemática, el descubrimiento y la utilización de leyes, y el refinamiento metodológico para abordar todo problema. En el caso de la lingüística, sistematizar el dominio de estudio implicaría partir de los datos y su estudio pormenorizado como soporte de las hipótesis.
10. La **metódica** de la lingüística debe constar de procedimientos justificables y escrutables, que puedan ser analizados y controlados, y sujetos a la crítica, como es el caso del método científico. Del fondo de conocimiento surgen problemas que se deben abordar metódicamente, definiendo hipótesis verificables y probándolas o comprobándolas con los datos adquiridos en la investigación, y de forma reproducible. Acudiendo de nuevo a Fitch (2009):

> Furthermore, progress will be aided by comparing and contrasting multiple hypotheses, not simply rejecting implausible null hypotheses in favor of single pet hypotheses. Ultimately, as for physics, what biolinguistics needs most are creative empirical tests of hypotheses.

Bunge (2010) completa esta décupla con otros postulados, como son la necesidad de la existencia de campos afines, básicos y contiguos, que en el caso de la

lingüística podrían ser tanto la neurología como la acústica, y su inclusión –parcial o total– en otros campos de investigación científica fáctica, como podría ser el caso de la psicología, la cibernética o la biología. Así, la obvia potencia explicativa de la física acústica fundamenta la fonética y la fonología, mientras que la teoría de la información se erige como paradigma científico explicativo de cualquier sistema de comunicación (Cover y Thomas, 2006).

No obstante, es imprescindible (Bunge, 2001) que como resultado de las investigaciones efectuadas haya una evolución, una ampliación de los problemas que se explican y se comprenden, en relación con todos los elementos generales vistos. La lingüística está en ese camino de avance, no en un sentido positivista puro sino en un sentido pragmático y de comprensión cada vez mejor de sus problemas.

Por tanto, tal y como se ha expuesto, la existencia de algunas controversias entre parte de la comunidad científica, especialmente en cuanto al dominio, la perspectiva general, el trasfondo formal y la generación de un fondo de conocimiento consensuado, hace que todavía pueda considerarse a la lingüística como una **ciencia en desarrollo**, análogamente a la psicología (Bunge, 2010), aunque no totalmente dentro de ella si se entiende un dominio extendido de la lingüística, como se ha argumentado. No creemos que la lingüística merezca el calificativo de acientífica, como argumenta Bunge para la teología o la crítica literaria, ni de protociencia (como la economía y la politología) en el sentido bungeano.

Sin embargo, algunos investigadores tratan aún a la lingüística como una protociencia, mientras otros defienden que no sea una disciplina científica, en la línea del anarquismo epistemológico de Feyerabend (1975), al no tener necesidad de abordar su problemática según los postulados planteados por el método científico. Otero (2004) cita a Thom (1993, pp.142-143) para dar cuenta del confuso panorama de la lingüística:

> La situación de la lingüística en Francia, sobre todo desde el punto de vista sociológico, es un auténtico desastre. Hay una miríada de pequeñas iglesias: los chomskianos, los funcionalistas, y así sucesivamente. Parroquias que tienen sus revistas, en las que se publican artículos cuyo único propósito es demoler las tesis de las sectas rivales... Siempre me ha sorprendido este carácter feudal de la mayor parte de las ciencias humanas. Una situación que no es ni de lejos comparable a la de las matemáticas: en este campo, afortunadamente no existe feudalismo alguno; hay rivalidades, evidentemente, entre las distintas especialidades, pero no tienen el carácter de lucha organizada que es característico de las ciencias humanas.

Evitar la "lucha organizada" a la que alude Thom (1993), y el lamentable espectáculo que da en muchas ocasiones (véase Fitch (2009) para algunos ejemplos), así como la fragmentación en lo que debería constituir el fondo de conocimiento de la lingüística, será un requisito esencial para lograr que la lingüística se constituya como ciencia. El conflicto es debido en parte a que mientras un segmento de la comunidad está trabajando en su constitución como ciencia, otro mantiene posiciones pseudocientíficas ancladas en el pasado.

1.2. La navaja de Guillermo de Ockham

Ante dos explicaciones válidas para un mismo fenómeno, la navaja de Ockham nos invita a quedarnos con la más simple de ambas: *entia non sunt multiplicanda praeter necessitatem*. Es una máxima que se ha seguido en las investigaciones de esta tesis, con consecuencias tanto experimentales como epistemológicas.

Si bien esta afirmación formaría parte de la metodología y está ampliamente difundida en las ciencias, la navaja de Ockham nos invita a reflexionar sobre el concepto de simplicidad en la lingüística, tanto en su trasfondo formal matemático, como empírico y filosófico. Se ha defendido en la lingüística que partiendo de pocos principios se puede dar cuenta de la complejidad del lenguaje (Chomsky, 1986). Pero no debe confundirse la sencillez conceptual de partida con la complejidad y potencia explicativa de los constructos y modelos teóricos: así, por ejemplo, podemos acercarnos a un sistema de comunicación de una forma simple, contando palabras o elementos comunicativos y, sin embargo, la modelización teórica subyacente a las distribuciones de frecuencias obtenidas puede ser muy compleja y dar para libros enteros (Baayen, 2001; Popescu, 2007).

Sencillez aproximativa es una cosa y otra sencillez en la modelización o el análisis. Podemos puntualizarlo en diversos niveles:

a) **Nivel matemático**: Se tiende a pensar en la simplicidad matemática –algebraica– de una ecuación que ajusta estadísticamente un conjunto de pares de datos *(x,y)*, y por ejemplo se asume que una función lineal (tipo $y=ax+b$, con a y b números reales, constantes en muchos casos) es algebraicamente más sencilla que un polinomio de grado 8 ($y= ax^8+bx^7+cx^6+\ldots$), por obviamente contener menos parámetros, o que una función exponencial ($y=a\cdot e^x+b$), por considerarse la función lineal más sencilla que la exponencial.

Debemos plantearnos un criterio empírico, como parte de M, revisable, en el que intervengan otros elementos, además del número de parámetros del sistema, como es lo bien que la función ajusta los datos y por tanto, por ejemplo, minimiza las desviaciones de la nube de puntos *(x,y)* respecto a la curva del ajuste. Tanto el número de parámetros como la bondad del ajuste deberían compensarse para ofrecernos un criterio formal, o una hipótesis de trabajo básica con la que trabajar con datos.[1]

No obstante, la consolidación del estudio de los sistemas complejos y las teorías del caos, llevó en el siglo XX a revisar el determinismo en la ciencia y condujo a los científicos a descubrir que las funciones que ajustan en un intervalo el comportamiento de un conjunto de datos, pueden no hacerlo en intervalos de estudio más amplios (Gleick, 1989).

Sin ir más lejos, alguno de los artículos de esta tesis surgió como respuesta a una supuesta aplicación de la navaja de Ockham: Solé (2010) proponía un modelo matemáticamente más sencillo que el planteado en Ferrer-i-Cancho y Forns (2009) y supuestamente igual de general que éste, aunque como se demostró (Baixeries *et al.*, 2012; Ferrer-i-Cancho, Forns *et al.*, 2013) su modelo se apoyaba en cierta evidencia cualitativa basada en una reducida muestra de grupos taxonómicos (mamíferos y un subconjunto de plantas), dejando aparte el resto de datos sin analizar. Es un claro ejemplo de aplicación errónea de la navaja de Ockham: cuando se ve limitada la potencia explicativa o la capacidad de generalización por utilizar un modelo excesivamente simple en un conjunto reducido de datos.

Aunque, por otra parte, huelga decir que el debate es constante en ciencia, especialmente en biología: ¿aumento la complejidad de un modelo para generalizar o no, y me quedo con un modelo sencillo pero que explique menos fenómenos? Luego la sencillez de un ajuste algebraico puede ser reduccionista o simplemente válida en un contorno concreto, y no hay que confundir el minimalismo de los elementos del sistema lingüístico –o de la pregunta que nos formulamos– con la simplicidad matemática o la complejidad en su tratamiento posterior (Chomsky, 1995).

b) **Nivel experimental**: además de los aspectos matemáticos de los modelos, se ha evitado la complejidad innecesaria de los diseños experimentales. Hay ocasiones en las que uno no puede escapar de complicar las experiencias (como sucede con los modelos

[1] Veremos en los capítulos siguientes los detalles de los criterios seguidos para la optimización de ajustes.

matemáticos), para evitar el reduccionismo en los fenómenos a abarcar. Mezclar grupos de datos –aunque se ha hecho en algún caso– permite analizar conjuntamente, por ejemplo, el habla de sujetos con características similares, aunque en general no es una práctica idónea, ya que la estadística del conjunto puede producir sesgos y hacernos llegar a conclusiones erróneas, lo que es habitual en biología cuando se mezclan grupos de datos (Magurran, 2004).

La estadística en cambio no puede fundamentarse en un único caso, aunque hay casos en los que la exhaustividad de estudios pioneros lo justifica (Roy *et al*., 2006), y así tenemos de nuevo una encrucijada empírica que afecta sobremanera a la psicología y a la lingüística, que se enfrentan al dilema de presentar estudios de un único caso o de poblaciones más extensas, como cuando se analizan corpus de un autor o diversos, o diferentes obras de un mismo autor.

La buena formulación de protocolos experimentales, con datos contrastados con sus correspondientes grupos control, y la mejora metodológica es la única vía para efectuar experiencias válidas –aunque mejorables, sin duda– y aceptables[2]. Los estudios de poblaciones, en los que se mira más el bosque que el árbol, han sido mayoritarios en nuestro caso, entendiendo que se podían inferir pautas, e incluso leyes, que serían difíciles de obtener en análisis de casos aislados. Debemos ser conscientes, empero, de las limitaciones epistemológicas de los estudios generales, y blindarnos estadísticamente al respecto.

En definitiva, si bien la navaja de Ockham nos lleva a buscar la sencillez metodológica y en los diseños experimentales, no siempre la sencillez conceptual se corresponde con la obviedad matemática o con el protocolo en el que menos interviene el experimentador. Todo dependerá de qué queramos comprender o estudiar en el bosque del lenguaje.

c) **Nivel filosófico**: Se ha defendido de partida el realismo y el naturalismo. La lingüística debería también aproximarse con la mayor simplicidad posible a la realidad que analiza. Se ha debatido mucho sobre la adecuación de la navaja de Ockham desde la filosofía de la ciencia (Feyerabend, 1975) o desde algunos enfoques de la lingüística (Boeckx, 2006).

[2] Como ejemplo puede revisarse la web del *U.S. National Institutes of Health*, con una exposición de los protocolos en la investigación clínica: http://clinicaltrials.gov/

Intervenimos en el objeto de estudio, aumentado la complejidad de la realidad observada, en el momento en el que aplicamos constructos teóricos sobre ella, y son estos constructos los que después estudiamos, y sobre los que formulamos modelos matemático-físicos y contrastamos, o falsamos en el sentido popperiano, hipótesis. A veces definimos conceptos con un sentido que se nos antoja claro, como es el caso de la definición de palabras, morfemas o fonemas, y su definición nos parece básica e imprescindible, incluso obvia, para la lingüística. Después descubrimos que las definiciones más usuales tienen problemas, y si nos encallamos en este punto, si concluimos que no sabemos ni qué es una palabra, es fácil caer en la tentación del escepticismo radical.

Deberíamos ser conscientes de que, si nos centramos por ejemplo en la oralidad de las lenguas, lo ideal sería ser capaces de analizar y escudriñar directamente las características físicas de las ondas y deducir sobre sus parámetros físicos (frecuencia, intensidad, energía, etc.) las leyes y principios básicos de la lingüística. Han sido los estudiosos de la teoría de la información, parte esencial de la lingüística, tal y como debería entenderse, los que más han indagado en esta dirección, junto con físicos e ingenieros. Se obtienen de esa forma correlatos y definiciones explicativas, directamente fundamentados en las características físicas de los estímulos que medimos y procesamos, como sucede en la fonética acústica. La lingüística debería evitar enmarañarse en definiciones innecesarias que aumentan la complejidad del objeto de estudio, a la vez que la alejan cada vez más de la realidad que se pretendía representar.

En nuestro caso, aunque se ha intentado ser bastante estricto en la aplicación de la navaja de Ockham en el nivel experimental, y en especial en los ajustes matemáticos efectuados y en la elección de modelos, no hemos podido evitar utilizar algunas unidades teóricas básicas – aceptadas y establecidas en la lingüística, aunque todavía problemáticas– como son las palabras, las letras (caso del análisis de corpus escritos) o los fonemas, debido a que los corpus a los que se ha tenido acceso las emplean habitualmente. Y es que aunque estén comúnmente aceptadas por la comunidad científica unidades como la palabra o el fonema poseen sus propios problemas de definición y no son los elementos más sencillos que podrían teóricamente estudiarse, si se pretende, como es el caso, aplicar la navaja de Ockham en la aproximación al objeto de estudio. Hay una tendencia, por ejemplo, a hablar de *sonidos* en lugar de *fonemas* en la neurociencia actual (Cuetos, 2012), lo que aproxima las unidades de estudio de la lingüística a la física subyacente.

De nuevo puede que una palabra, abundando en el ejemplo, nos parezca algo muy sencillo cognitivamente. De hecho, se nos ha estimulado con palabras desde que nacimos (e incluso antes, véase Sovilj (2011) para una revisión), ¿no será esa la causa de la aparente sencillez del concepto? Ha costado mucho que las máquinas aprendan a segmentar palabras, y nosotros mismos tenemos dificultades cuando aprendemos otras lenguas. Estamos gobernados por nuestros sentidos y nuestro sistema cognitivo. Poseemos una cognición entrenada y dotada genéticamente para el lenguaje y la comunicación. Y, salvo patologías, lo hacemos muy bien cualitativamente. Sin embargo, no podemos limitar la lingüística a la percepción cualitativa, categorial, que opera en nuestra cognición.

También tenemos percepciones sobre el tiempo, la velocidad o la aceleración, pero la cinemática hace siglos dio el salto cuantitativo: debimos desarrollar instrumentos que permitiesen medidas numéricas. Sin ellas no se puede efectuar un desarrollo científico formal. Y como no poseemos una cognición ni unos sentidos capaces de detectar numéricamente amplitudes acústicas, frecuencias o tiempos, el futuro de la lingüística debería intentar realizar el salto cuantitativo definitivo, el que la conduzca a medir directamente, mediante la tecnología a nuestro alcance, variables y parámetros numéricos.

Las pocas experiencias de grandes tomas de datos cuantitativas han demostrado su tremenda potencialidad, como es el caso de *The Human Speechome Project*, (Roy *et al.*, 2006) en el que se obtuvieron miles de horas de grabación de la evolución de la adquisición de la lengua inglesa de un niño desde su nacimiento[3]. Este tipo de experiencias van más allá de definiciones en las que la intervención humana hace en apariencia más sencillos los objetos de estudio, porque muchas veces en realidad lo que logramos es hacernos más fácil –técnicamente– nuestro trabajo de análisis, cuando no disponemos de la tecnología (ni el presupuesto) de proyectos cómo *Speechome* (Roy *et al.*, 2006).

La navaja de Ockham debería hacernos comprender que ante dos aproximaciones a la realidad deberíamos quedarnos con la que menos la distorsiona, la que minimiza nuestra intervención, aunque sea la más cara. Es el ánimo de toda ciencia de base empírica el pulir hasta la saciedad sus experimentos reduciendo los errores y la

[3] Puede el lector visitar la página de las publicaciones de Deb Roy, cuyo hijo es el protagonista de la mayor toma de datos de adquisición del lenguaje de la historia: http://dkroy.media.mit.edu/publications/

intervención del observador: también debe serlo para el lingüista. Heisenberg (1956, pp.129-130), uno de los que más ahondó sobre la influencia del observador en la ciencia, afirmó tras formular su célebre principio de incertidumbre:

> Los procesos biológicos deben ser tratados con una experimentación científica mucho más cautelosa que los procesos de la física y la química. Tal como ha advertido Bohr, puede ocurrir que no sea posible dar una descripción del organismo vivo que pueda considerarse completa desde el punto de vista del físico ya que ello requeriría experimentos que interfieren demasiado en las funciones biológicas.

No obstante, no nos ceguemos con el principio de incertidumbre de Heisenberg: la lingüística, por ahora, no investiga en el nivel cuántico, y aquí no hay paradoja del observador, más allá de la influencia de los procesos cerebrales sobre nuestra percepción o del carácter inescrutable –sin intervención demasiado invasiva– de algunos fenómenos biológicos (Heisenberg, 1956).

El problema de la subjetividad de la lingüística proviene precisamente de la subjetividad de algunas definiciones y unidades de estudio aceptadas y con las que se trabaja alegremente. Lograr la matematización de las unidades de estudio es uno de los retos en el proceso de hacer de la lingüística una ciencia. En nuestra defensa cabe argumentar que no se ha ido más allá de los elementos más sencillos, como las palabras[4], cuando buena parte de la comunidad de lingüistas, alejados de los postulados bungeanos, sigue empantanada en definiciones complejas que no aportan demasiada luz, que no están suficientemente justificadas y que no acaban configurando un fondo de conocimiento como tal.

1.3. La teoría matemática de la comunicación

La que se ha denominado posteriormente como teoría de la información (Gleick, 2011) fue bautizada literalmente como teoría matemática de la comunicación por Shannon (1948). El conocido esquema que explicó Shannon, si bien había sido propuesto por otros autores del círculo Praga conceptualmente, fue dotado de un modelo matemático general, de un trasfondo formal, aplicable a todos los sistemas de comunicación (figura 1.1.). Nacía además la era del bit (*binary digit*), cuando Shannon recordaba el nombre que J.W. Tukey había dado para la mínima unidad lógica de

[4] El concepto de 'palabra' tiene sus problemas de definición, como se vio en Hernández-Fernández(2006).

información para un sistema de base 2. El modelo es de una concepción simple, tanto que se explica en la enseñanza secundaria, aunque sin embargo su desarrollo e implicaciones todavía colean en la ciencia actual. Un ejemplo más de que de aproximaciones conceptuales sencillas a la realidad pueden inferirse teoremas matemáticamente complejos.

Shannon (1948) empezó analizando un canal discreto de comunicación (citó explícitamente el ejemplo de la telegrafía y posteriormente las lenguas en su modalidad escrita), un sistema en el que se parte de un conjunto finito de elementos $\{s_1, s_2, ..., s_n\}$ que pueden ser transmitidos secuencialmente de un punto a otro y donde cada elemento s_i posee una duración temporal t_i. Shannon (1948) ya planteaba la duración del elemento comunicativo como básica en el caso discreto: no todas las secuencias de s_i son igualmente transmisibles, y de hecho sólo algunas de ellas están permitidas. Cada s_i se emite con una cierta probabilidad p_i: ya Shannon comprendió que la fuente discreta de información generaba el mensaje, símbolo a símbolo, mediante un proceso estocástico gobernado por un conjunto de probabilidades. Fue capaz, sin la potencia de los ordenadores actuales, de simular el inglés escrito mediante aproximaciones iterativas probabilísticas.

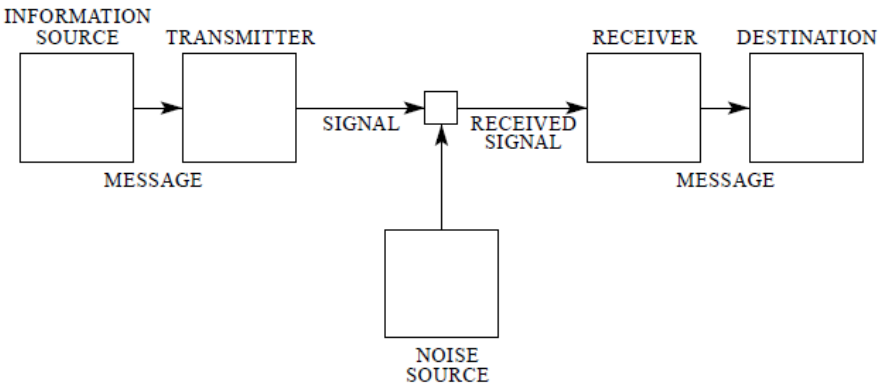


Figura 1.1.: Diagrama esquemático de un sistema general de comunicación (Shannon, 1948).

Pero, ¿qué sucede con los sistemas de comunicación que no son discretos? Según la teoría de Shannon (1948), la radio o la televisión son ejemplos de sistemas de comunicación continuos, mientras que el habla es un sistema mixto en el que se transmiten a la vez variables discretas y continuas. Ambos casos son más complejos matemáticamente que los discretos, al entrar más variables en juego, aunque abordables.

La ciencia cognitiva actual (Hernández-Fernández, 2006) apunta a que nuestro cerebro es capaz de establecer un límite entre las secuencias de sonidos de la cadena hablada, analizando en tiempo real las probabilidades de coocurrencia de elementos, y después las probabilidades transicionales entre elementos adyacentes (proceso de Markoff, ya citado por Shannon (1948)), tanto para la detección de fonemas como para la detección de palabras (Saffran *et al.*, 1996).

No debe confundirse un enfoque probabilístico del lenguaje con un enfoque azaroso. La física estadística ha demostrado de sobras que se pueden inferir leyes de fenómenos que no son deterministas. Sean discretos o continuos, cada sistema de comunicación posee una velocidad máxima de transmisión de información y un número máximo de elementos por segundo que pueden enviarse, que ya Shannon y Weaver (1949) relacionaron con la capacidad del canal y la entropía.

Además, Shannon y Weaver (1949) añadieron que las lenguas contienen más del cincuenta por ciento de elementos redundantes, sonidos o letras superfluos para transmitir un mensaje (Gleick, 1989). ¿Tiene sentido desde la lingüística analizar frases y formular sobre ellas teorías y modelos cuando la mitad de su contenido es redundante? ¿No será redundante el modelo teórico así generado y por tanto simplificable? No obstante, el propio Shannon reconocía la necesidad de la redundancia para una transmisión informativa con éxito, luego ¿cuánta redundancia es necesaria?

Los estudiosos de los sistemas complejos percibieron la inutilidad de estudiar las partes haciendo salvedad del todo: para ellos el caos implicaba la desaparición del programa reduccionista de la ciencia (Gleick, 1989). En la lingüística, "estudiar el todo" es harto complejo pues, para empezar, hay que definir bien el "todo" y después, gracias a un conjunto limitado de datos experimentales, contrastar los modelos teóricos propuestos. La estadística emerge como una vía única e inevitable, una tarea onerosa a la que se dedican muchas horas de trabajo (Baayen, 2007).

Es necesaria una vuelta a los orígenes de la teoría de la *comunicación* de Shannon (1948) y Shannon y Weaver (1949), más allá de la *información*. Estamos con Fitch (2009), en la idea de que el rompecabezas del lenguaje seguramente se resolverá de una vez por todas a través de la teoría de la comunicación de Shannon.

CAPÍTULO 2

Leyes de la lingüística cuantitativa

La lingüística cuantitativa investiga el lenguaje mediante las matemáticas. Como sucede con la física y otras ciencias básicas, la lingüística requiere de la matematización de la realidad para la formulación de leyes. Los datos deben pasar por el cedazo de la estadística, la contrastación empírica y su ulterior replicación para que las regularidades y patrones puedan ascender el peldaño que les otorga el rango de ley. Por otra parte, las leyes de la lingüística cuantitativa se han encontrado en otros fenómenos ajenos al lenguaje pero relacionados con la teoría de la información.

El marco de la lingüística cuantitativa aúna la concepción realista y naturalista de la lingüística con la matemática y la generación de un trasfondo formal. En cuanto se cuantifican y se pueden medir las unidades de estudio se puede construir una teoría y pasar entonces a su contrastación empírica. En definitiva, la lingüística será cuantitativa o no será (ciencia), aunque su formalización todavía no escape de la controversia. En palabras de Kornai (2008):

> It is hard to find any aspect of linguistics that is entirely uncontroversial, and to the mathematician less steeped in the broad tradition of the humanities it may appear that linguistic controversies are often settled on purely rhetorical grounds.

La lingüística cuantitativa analiza las estructuras de las lenguas, sus propiedades e interrelaciones con otros sistemas de comunicación (lingüística comparativa), con ánimo de descubrir las leyes subyacentes del lenguaje, mediante el uso de técnicas cuantitativas como la estadística (Köhler, 2005). Se asume que la perspectiva de la lingüística cuantitativa es la más próxima a la de las ciencias clásicas empíricas, como apuntan Liu y Huang (2012):

> The language laws it discovers contribute to more accurate description and explanation of relevant language phenomena and are vitally important and necessary for the establishment of a type of linguistic theory in the modern scientific sense. As an empirical discipline based on authentic language data, the mode of thinking and research methodology practiced in quantitative linguistics are generally in line with those in other empirical disciplines. The most representative accomplishments of quantitative linguistics are the various language laws concerning the structure and evolution of human languages, which constitute the basis of relevant theories.

La transdisciplinariedad de la lingüística cuantitativa llevó en su día a la formulación de la lingüística sinergética (Köhler, 2005b), que en el paradigma de la sinergética de Haken (1983) permite la integración en una teoría general de las leyes que se infieren del comportamiento cuantitativo de las lenguas (Liu y Huang, 2012). La lingüística sinergética fue definida por Köhler (1986) con el ánimo de aunar e integrar las diferentes leyes observadas en la lingüística en un modelo que no sea meramente descriptivo de los fenómenos lingüísticos sino que además sea explicativo. Para lograrlo, Köhler (2005) apunta a la necesidad de considerar el lenguaje como un sistema

autoregulado y autoorganizado con una interdependencia dinámica entre la función y la forma de los elementos lingüísticos.

La modelización en lingüística sinergética impone la necesidad de partir de determinados axiomas que todo sistema semiótico debe poseer (Köhler, 1990) y que son necesarios para realizar investigaciones sólidas[5]:

1. Para empezar, hace falta un código que dé cuenta de la oposición de fuerzas que se establece entre la tendencia a la estabilidad del sistema lingüístico y a su adaptación a diversos contextos comunicativos, entendiendo que nos centramos en que el lenguaje tiene como objetivo la comunicación, en el sentido de la especialización de Hockett (1963). Por tanto, el código entre otras características debe:
   a) Poder crear elementos con significado.
   b) Permitir el almacenamiento de información (memoria).
   c) Ser eficiente en el proceso de codificación/decodificación.
   d) Minimizar la energía de los procesos involucrados, en la medida de lo posible, pues puede haber otros procesos involucrados (por ejemplo la reproducción).
   e) Transmitir información de forma segura, preservando la integridad del emisor.
   f) Garantizar la comunicación.
2. Determinar las unidades, variables y nivel de estudio de la investigación. Generalmente estos elementos se forman con los elementos del código. Así, por ejemplo, las unidades de estudio pueden ser la palabra, el fonema, el texto, etc.
3. Efectuar hipótesis sobre dependencias entre variables y analizar sus consecuencias y efectos teóricos, así como buscar equivalentes funcionales, experiencias y ejemplos concretos.
4. Formular modelos matemáticos que den cuenta de las hipótesis y datos experimentales y que sean contrastables estadísticamente mediante test rigurosos. Cuando las hipótesis se corroboran reiteradamente entonces obtienen el estatus de ley.

[5] Puede consultarse un resumen en http://www.glottopedia.org/index.php/Synergetic_linguistics

Como es de esperar bajo la perspectiva bungeana de una ciencia en desarrollo (Bunge, 2010), la lingüística todavía necesita implementar el cuarto punto. Como se verá, varias de las contribuciones de esta tesis han ido en la línea de que algunas regularidades sean consideradas leyes, es el caso de la ley de brevedad (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013) o de la ley de Menzerath-Altmann en el genoma (Hernández-Fernández *et al*., 2011), así como analizar las excepciones a leyes universales como la ley de Zipf (Corominas-Murtra y Solé, 2010), en el caso de la demencia de tipo Alzheimer (Hernández-Fernández y Diéguez-Vide, 2013), o a la propia ley de brevedad en la comunicación animal (Ferrer-i-Cancho y Hernández-Fernández, 2013).

Luego tras la especificación del código, de las variables y nivel de estudio (1 y 2) y tras la formulación hipotética del problema a estudiar (3), los métodos cuantitativos resultan imprescindibles para poder completar la aproximación sinergética (3 y 4) en todas las ramas de la lingüística (Johnson, 2008). Herdan (1964) ya expuso la relevancia de los métodos cuantitativos, de la combinatoria y la estadística en las diferentes ramas de la lingüística y se anticipó, en una dura crítica al innatismo más duro del primer Chomsky, a la dicotomía que se estableció en la lingüística de la segunda mitad del siglo XX, entre los que siguieron el camino bungeano, de desarrollo de la ciencia de la lingüística, y los que no (Herdan, 1964, p.11):

> …Chomsky completely forgets the practical aspect of establishing statistical regularities. Wheter they are "very" or "less surprising", they enable us to make predictions, and thus to know what to expect under normal circumstances.

Sin duda, hacer de la lingüística una ciencia implica contar con un conjunto de leyes que expliquen los fenómenos del lenguaje y la comunicación, que permitan realizar predicciones. Se ha discutido mucho sobre cuáles de las leyes deberían aceptarse ya, es decir, hay evidencia empírica suficiente, y sobre cuáles aún son hipótesis que contrastar mejor experimentalmente. El debate es habitual en una ciencia en desarrollo.

Así por ejemplo, cuando la Universidad de Trier implementó en 2005 una wiki en la que recoger las leyes de la lingüística cuantitativa, admitía en su página principal[6]:

[6] Cita de la web: http://lql.uni-trier.de/index.php/Main_Page

The quest for laws of language and text during the last decades has resulted in a wealth of law hypotheses and accepted laws. It has become difficult to achieve a systematic overview of the relevant studies and findings.

A continuación presentaremos brevemente los fundamentos de las leyes que se han estudiado en esta tesis, de las que la bibliografía es ahora muy extensa[7] y que además se han ido encontrando más allá del lenguaje humano. Como apuntábamos en el capítulo anterior probablemente estamos ante leyes generales de los sistemas de comunicación y que, por tanto, aparecen en las lenguas como en otros sistemas en los que se codifica información.

Puede admitirse la controversia respecto a las causas que originan cada una de ellas. Pero las leyes se dan, aunque no tengamos todavía una explicación: se puede admitir el problema de cuál es el mejor modelo para ajustar las distribuciones de frecuencias de palabras (Li *et al*., 2010; Hernández-Fernández y Ferrer-i-Cancho, 2013; Baayen, 2001; Herdan, 1964), pero la ley de Zipf, la ley de brevedad y la ley de Menzerath-Altmann son ya leyes de pleno derecho en la lingüística y aparecen más allá de ella, como se verá en los artículos de esta tesis. Se observan en la naturaleza.

Así por ejemplo, es bien conocido el clásico caso en el que se recupera la ley de Zipf en el fenómeno del *random typing* (Miller, 1957). Miller (1957), y Mandelbrot (1952, 1953, 1953b) con anterioridad, propuso la irrelevancia de la ley de Zipf al recuperar la misma mediante la generación de textos, con la construcción de *palabras* mediante el añadido aleatorio de caracteres y espacios en blanco. No obstante, que un proceso azaroso o una simulación recupere una ley no implica que ésta deje de tener sentido: podríamos inventar al azar un universo en el que se cumpliera alguna ley de la física y, sin embargo, eso no haría que las leyes de la física del universo conocido dejasen de tener sentido, si tenemos evidencia empírica de ellas. Volviendo al *random typing* ya fue Howes (1968) el que apuntó que los supuestos del *experimento* de Miller eran contradictorios directamente con algunas propiedades básicas del lenguaje.

[7] Por ejemplo la página antigua de referencias sobre la Ley de Zipf de Wentian Li (la última, más actualizada no se puede enlazar en la actualidad:
http://ccl.pku.edu.cn/doubtfire/NLP/Statistical_Approach/Zip_law/references%20on%20zipf's%20law.htm
O la página web de Ramon Ferrer-i-Cancho en la que ya se lista la presencia de las leyes del lenguaje más allá de las lenguas humanas:
http://www.lsi.upc.edu/~rferrericancho/laws_of_language_outside_human_language.html

En concreto, Howes (1968, p.271) notó como el modelo de Miller (1957) implicaba que las palabras de una misma longitud eran equiprobables, lo que no sucede en las lenguas humanas, y respecto a la ley de brevedad, definida por Zipf (1949) para las *frecuencias medias* de palabras, aclaró:

> Zipf demonstrated that in natural languages the average frequency of words of length *i* is greater than the average frequency of words of length *(i + 1)*. Miller has converted this empirical finding into the assumption that all words of length i have probabilities greater than all words of length *(i + 1)*, which is obviously untenable.

Sin embargo, Howes (1968) dejó la puerta abierta a que el modelo anterior de Mandelbrot (1952, 1953, 1953b) de maximización de información fuese una objeción seria a la relevancia de la ley de Zipf. La controversia se alimentó durante la segunda mitad del siglo XX, y sigue viva (Piantadosi *et al*. 2011; Ferrer-i-Cancho y Del Prado Martín, 2011) aunque siempre bajo el paradigma cuantitativo y admitiéndose tanto la ley de Zipf como la de brevedad, como evidencias experimentales a explicar, y se ha visto que los modelos de Mandelbrot, centrados en la eficiencia y la redundancia de la comunicación en el sentido de Shannon (1948), presentan problemas especialmente por dar cuenta de forma poco realista de los textos de las lenguas reales (Manin, 2009). La redundancia informacional de Shannon (1948) y su vinculación con la ley de Zipf sigue siendo otro punto caliente de la investigación actual, en especial en la adquisición del lenguaje (Bannard y Lieven, 2008; Baixeries *et al.*, 2013).

No obstante, los datos están ahí para ser escudriñados, analizados y contrastados. Como se verá, no se ha querido idealizar ninguna de estas leyes. Todo lo contrario, debe ampliarse el abanico de posibilidades para, test estadístico mediante, comprobar qué ley es la que se ajusta mejor a los datos experimentales.

2.1. Frecuencias y ley de Zipf

Si bien es cierto que hubo antecedentes similares (como Pareto en 1897, Estoup en 1916 o Condon en 1928[8]), fue George Kingsley Zipf (1902-1950), el que formalizó en la lingüística la llamada en su honor, ley de Zipf. En 1932 Zipf escribe su *Selective*

[8] Ver referencias de antecedentes en:
http://ccl.pku.edu.cn/doubtfire/NLP/Statistical_Approach/Zip_law/references%20on%20zipf's%20law.htm
http://www.positivedeviance.org/pdf/publicationgeneralpd/Herschbach_Zipf_Law.pdf

Studies and the Principle of Relative Frequency in Language (recogido en Zipf, 1935), obra en la que efectúa una observación estadística basada en las frecuencias de palabras en escritos diversos, hallando una relación empírica entre la frecuencia (*f*) de ocurrencia normalizada de una palabra (la frecuencia de una palabra dividida por el número total de palabras) y su rango (*r*), u orden (en el listado de palabras de mayor a menor frecuencia), de forma que:

$$f(r) = \frac{C}{r^a} \tag{1}$$

Con *C* una constante de normalización y *a* el exponente que en la ley de Zipf (1935) es aproximadamente la unidad, $a \approx 1$. Así pues, dado un conjunto de datos, si no consideramos la constante de normalización, la ley de Zipf es una ley con un parámetro, el exponente, a ajustar matemáticamente.

La propuesta estadística de Zipf (1935) se puede completar –a la vez que se desmitifica un poco– con otras posibilidades (tabla 2.1), cuyo desarrollo matemático y peculiaridades estadísticas en el ajuste resumen muy bien Li *et al.*(2010), y a las que pueden añadirse más opciones, como por ejemplo el modelo de dos regímenes propuesto por Ferrer-i-Cancho y Solé (2001). De hecho, una de las contribuciones de esta tesis es aportar al conjunto de funciones analizadas habitualmente (tabla 2.1.) un modelo de dos regímenes de Zipf, pero aplicando la aproximación de Baayen (2001) para detectar el *breakpoint* o punto de transición entre zonas (Hernández-Fernández y Ferrer-i-Cancho, 2013).

En todos los casos debe considerarse el número de parámetros de la función, ya que mejorar cualquier ajuste estadístico de datos aumentando el número de parámetros no tiene mérito: ¡si suponemos una función polinómica de grado *n* siempre ajustaremos perfectamente $n+1$ puntos! Por eso es necesario establecer un criterio con el que discernir si un ajuste, con una función concreta, es comparativamente mejor que otro, considerando no sólo lo bien que la función ajusta los datos sino además el número de parámetros de la función. A tal efecto Akaike (1974) propuso el llamado en su honor Akaike Information Criterion (AIC), una medida que debe minimizarse para considerar un modelo mejor que otro.

El AIC está fundamentado en la verosimilitud (*likelihood*) *L* de un modelo (Venables y Ripley, 1999) que se define, en el marco de la regresión en el que se

pretende minimizar errores entre los datos reales y la curva teórica de cada modelo (Li *et al.*, 2010), como

$$L = C - \frac{n}{2}\log\left(\frac{SSE}{n}\right) - \frac{n}{2} \qquad 2)$$

Con n el máximo rango, $SSE \equiv \sum_{r=1}^{n}(f(r)-F)^2$ la suma de errores cuadráticos que minimiza $y = F$, y C una constante que depende del modelo (tabla 2.1).

MODELO	PARÁMETROS (*K*)	FUNCIÓN (*r*=rango, *f*=frecuencia, *n*=tamaño repertorio)	TRANSFORMACIÓN LOGARÍTMICA	Referencias bibliográficas
Zipf	1	$f = \frac{C}{r^a}$	$\log f = C_0 + C_1 \log r$	*Zipf(1935)* *Zipf(1949)*
Beta	2	$f = \frac{C(n+1-r)^b}{r^a}$	$\log f = C_0 + C_1 \log r + C_2 \log(n+1-r)$	*Li et al..(2010)*
Yule	2	$f = \frac{C \cdot b^r}{r^a}$	$\log f = C_0 + C_1 \log r + C_r \cdot r$	*Yule (1925)*
Menzerath-Altmann	2	$f = C \cdot r^b \cdot e^{-a/r}$	$\log f = C_0 + C_1 \log r + C_3/r$	*Menzerath(1954)*
Zipf-Mandelbrot	2	$f = \frac{C}{(1+Ar)^a}$	$\log f = C_0 + C_m \log(1+Ar)$	*Mandelbrot(1952)* *Mandelbrot(1966)*
Doble Zipf	3	1er régimen: $f = \frac{C}{r^a}$; 2º régimen: $f = \frac{C'}{r^{a'}}$	*primer régimen:* $\log f = C_{01} + C_{11} \log r$ *segundo régimen:* $\log f = C_{02} + C_{12} \log r$	*Ferrer-i-Cancho y Solé (2001)* *Hernández-Fernández y Ferrer-i-Cancho(2013)*

Tabla 2.1: Diversas funciones con las que ajustar un conjunto de datos de frecuencia (f), ordenados por rango (r), en un repertorio comunicativo de n elementos. El resto de constantes de cada ajuste (C, a, a', b, C_0, C_1…) son parámetros que se obtienen del ajuste estadístico, lineal o logarítmico, según el caso. Adaptada de Hernández-Fernández y Ferrer-i-Cancho (2013), disponible en este trabajo, donde se pueden consultar detalles sobre la obtención de coeficientes.

Como exponen Li *et al.*(2010) toda constante se puede eliminar al comparar dos modelos, luego el AIC definido por Akaike (1974) será entonces:

$$AIC = -2L + 2K = n\log\left(\frac{SSE}{n}\right) + 2K \qquad 3)$$

Siendo K el número de parámetros libres del modelo (ver tabla 2.1), lo que permite definir la diferencia entre AICs de dos modelos, Δ, como la resta entre el AIC del modelo planteado respecto el mejor (el que minimiza el valor de la ecuación 3).

Luego Δ=0, trivialmente, para el mejor modelo (Hernández-Fernández y Ferrer-i-Cancho, 2013).

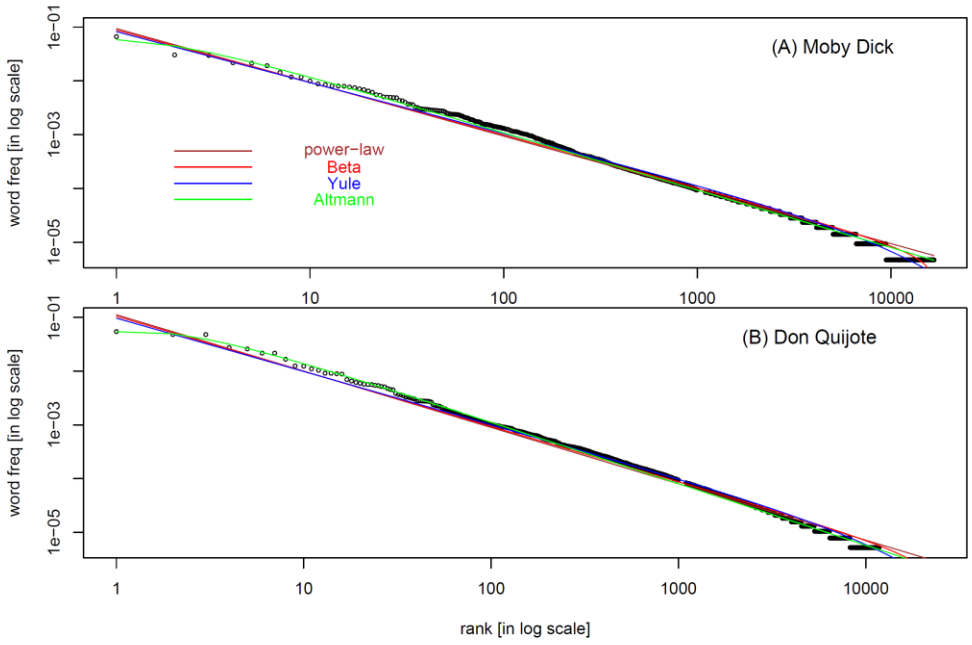


Figura 2.1.: Comparación entre la ley de Zipf (*power law*), la función Beta, la de Yule y la de Menzerath-Altmann (imagen extraída de Li *et al.*, 2010), aplicadas al ajuste de la distribución de frecuencias de palabras de Moby Dick (inglés) y El Quijote (español). En corpus extensos la diferencia entre ajustes es relativamente pequeña (Li *et al.*, 2010).

Las palabras más frecuentes suelen ser palabras funcionales y junto con las palabras menos frecuentes, de rangos elevados, ambos extremos a menudo se desvían del comportamiento de la ley de Zipf (Montemurro, 2001). Mandelbrot (1952) modificó la ley de Zipf, introduciendo un parámetro más, para ajustar mejor los extremos de la distribución de frecuencias, como revisa Montemurro (2001), de forma que obtuvo la llamada ley de Zipf-Mandelbrot, también de dos parámetros (*K*=2):

$$f(r) = \frac{C}{(1+Ar)^a} \qquad 4)$$

Sin embargo esta función no es reducible, como otras, a un ajuste logarítmico sencillo (de manera que tengamos el logaritmo de las frecuencias *log f* en función del logaritmo de los rangos *log r*) que permita la regresión entre frecuencias y rangos (tabla 1), lo que reduce sin duda su presencia en la literatura actual en lingüística.

Además, si consideramos la frecuencia *f* de un elemento comunicativo (una palabra, por ejemplo), como una variable discreta, entonces ésta sigue una distribución zeta (Ferrer-i-Cancho y Hernández-Fernández, 2008) cuando la distribución de probabilidad, el llamado espectro de frecuencias (Tuldava, 1996), es

$$P(f) = C \cdot x^{-b} \qquad 5)$$

Con $C = \frac{1}{\sum_{x=1}^{\infty} x^{-b}} \equiv \frac{1}{\zeta(b)}$, y *ζ(b)* la función zeta de Riemann y *b* el exponente de la distribución de probabilidades. El exponente de Zipf (*a* en la ecuación *1*) se relaciona con el de su distribución de probabilidades a través de la expresión (Ferrer-i-Cancho y Hernández-Fernández, 2008)

$$b = \frac{1}{a} + 1 \qquad 6)$$

Lo que se demuestra con detalle en el primer artículo de esta tesis (Ferrer-i-Cancho y Hernández-Fernández, 2008), citado con posterioridad por otros autores (Font-Clos *et al.*, 2013, la más reciente). En Ferrer-i-Cancho y Hernández-Fernández (2008) apuntamos una trivial pero intrigante conclusión: cuando se igualan los exponentes $a=b$ la ecuación 6 conduce a la ecuación cuadrática $b^2+b+1=0$, que tiene como solución positiva el famoso número de oro $b = \varphi = \frac{1+\sqrt{5}}{2} \approx 1.618...$ Aunque el resultado podría ser fruto de una casualidad estadística y quizá sea una aportación mínima de esta tesis, no es menos cierto que hemos sido los primeros, que sepamos, en notarla.

La aparición de la sección áurea o número de oro, *φ*, se vincula a la serie de Fibonacci y a procesos de optimización en la naturaleza, y en concreto se ha explorado en la ley de Zipf de distribución de rangos de ciudades según su tamaño (Gabaix, 1999), que ya arranca en el pionero estudio *Das Gesetz der Bevölkerungskonzentration* ("La ley de concentración de la población") de Félix Auerbach (1913), precedente de la ley de Zipf, o en la distribución zipfiana de las páginas web de Internet (Dominich[9] y Horváth, 2008).

[9] Trabajo póstumo del profesor Sándor Dominich (1954-2008), RIP.

El número de oro, o sección áurea, aparece en otros estudios de lingüística cuantitativa, como así recogen Popescu *et al.* (2009). Estos autores definen el punto h, inspirándose en Hirsch (2005), como el punto fijo de la distribución de frecuencias de palabras, es decir, el punto en el que el rango coincide con la frecuencia, $r = f(r)$. El punto *h* posee diversas aplicaciones en el análisis textual, y se puede definir en otros términos (Popescu *et al.*, 2009) como:

$$h = \begin{cases} r, & si \quad \exists f(r) | r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & si \quad \nexists f(r) | r = f(r) \end{cases} \qquad 7)$$

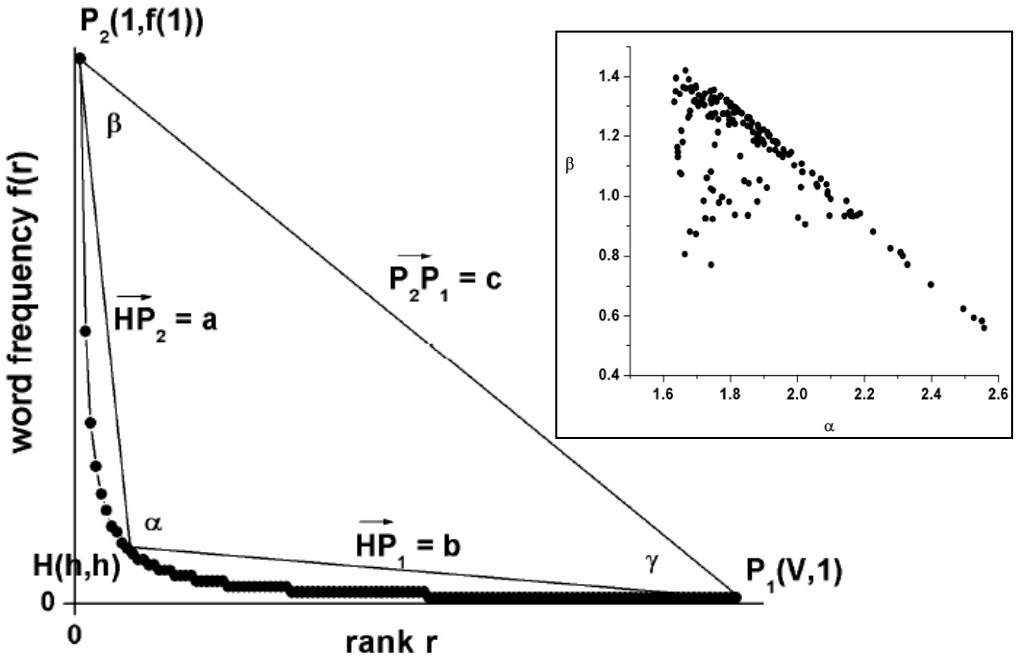


Figura 2.2: Detalle (figuras extraídas de Popescu *et al.*, 2009) de los ángulos *α, β,* y *γ* de la distribución de frecuencias de palabras, y de la situación del punto fijo *H*, en el que *r=f(r)*. Popescu *et al.* (2009) dan como límite inferior para el ángulo α (en radianes) el número de oro, lo que se aprecia con claridad en el gráfico superior.

Si tomamos el punto *H(h,h)* y el primer y el último punto de la distribución de frecuencias, gráficamente (figura 2.2) se definen tres ángulos asociados a estos puntos *(α, β, γ)*, de forma que Popescu *et al.* (2009) mostraron –con datos de más de 20 lenguas– que el ángulo *α* asociado al punto *H* posee como límite inferior, en radianes, el número de oro.

Este es únicamente otro ejemplo más de la intrigante aparición del número de oro en lingüística cuantitativa. ¿Es el número de oro algún tipo de umbral para los exponentes de Zipf? ¿Una consecuencia de procesos de optimización? La idea subyacente es que si el número de oro emerge en la naturaleza como resultado de procesos de optimización quizá también lo haga en el lenguaje o en los sistemas de comunicación por el mismo motivo. Hay que seguir explorando para averiguarlo.

2.2. Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2008). *Power laws and the golden number*.

Se incluye aquí el capítulo de libro:

- Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2008). Power laws and the golden number. En: G. Altmann, I. Zadorozhna y Y. Matskulyak (eds.), *Problems of general, germanic and slavic linguistics* (pp. 518-523). Chernivtsi: Books – XXI.

Ferrer i Cancho, R. & Hernández Fernández, A. (2008). Power laws and the golden number. In: "Problems of general, germanic and slavic linguistics", Altmann, G., Zadorozhna, I. & Matskulyak, Y. (eds.), Chernivtsi: Books - XXI. pp. 518-523.

Power laws and the golden number

Ramon Ferrer i Cancho [a,*]

[a] Departament de Física Fonamental, Universitat de Barcelona. Martí i Franquès 1, 08028 Barcelona, Spain.

Antoni Hernández Fernández [b]

[b] Departament de Lingüística General. Universitat de Barcelona. Gran Via de les Corts Catalanes, 585. 08007 Barcelona, Spain.

Abstract

The distribution of many real discrete random variables (e.g., the frequency of words, the population of cities) can be approximated by a zeta distribution, that is known popularly as Zipf's law, or power law in physics. Here we revisit the relationship between power law distribution of a magnitude and the corresponding power relationship between the magnitude of a certain element and its rank. We show that the exponents of the two power laws coincide when its value is the famous golden number, $\varphi = (1+\sqrt{5})/2$.

Key words: Zipf's law, power laws, zeta distribution, Golden number and Fibonacci series.

1 Introduction

A discrete random variable x (such that $x \geq 1$) follows a zeta distribution (Wimmer and Altmann, 1999) if

$$P(x) = b(\beta)x^{-\beta}, \quad (1)$$

where $b(\beta)$ is a normalization function and β is the so-called exponent.

* Corresponding author. Phone: +34 934137870. Fax: +34 934137787.
Email address: rferrericancho@lsi.upc.edu (Ramon Ferrer i Cancho).

We have that $b(\beta) = \zeta(\beta)^{-1}$, where

$$\zeta(\beta) = \sum_{x=1}^{\infty} x^{-\beta} \tag{2}$$

is the Riemann zeta function. Therefore, the exponent β is the only parameter of the distribution. Since $\zeta(\beta)$ converges only for $\beta>1$, Eq. 1 is only well-defined when $\beta>1$. Hereafter we assume $\beta>1$. Examples of real discrete random variables following this distribution are word frequencies and the population size of cities (Gisiger, 2001; Newman, 2005). The zeta distribution receives equivalent names: power law distribution in physics or Zipf's law in general (Newman, 2005). See Newman (2005); Ferrer i Cancho and Servedio(2005); Simkin and Roychowdhury (2006); Mitzenmacher (2003) for a review of the most popular explanations of the zeta distribution.

If the discrete random variable x is a frequency of occurrence (e.g., the frequency of a word), then its probability distribution is called a frequency spectrum (Tuldava, 1996). Historically, besides considering the probability of a discrete magnitude x, researchers have also been concerned about the relationship between x or the normalized value of x versus rank (Zipf, 1972; Wimmer et al., 1999). Imagine that we have a set of elements (e.g., words) that are characterized by a certain magnitude x (e.g., their frequency of occurrence in a text). Imagine that we sort these elements decreasingly by their selected magnitude (e.g., decreasingly by their frequency). The element with the largest value of x is assigned rank i=1, the element with the second largest value of x is assigned rank i=2 and so on. We define $x(i)$ as the value of x of the element of rank i (e.g., $x(i)$ is the frequency of the i-th most frequent word of a text). It is known that if x is zeta distributed then $x(i)$ obeys (Chitashvili and Baayen, 1993; Adamic, 2000; Pietronero et al., 2001; Adamic and Huberman, 2002)

$$x(i) \sim i^{-\alpha}. \tag{3}$$

The organization of the remainder of this article is as follows. Firstly, we will review how $b(\beta)$ can be calculated (Section 2). This will provide us with some basic strategies for the next step. Secondly, we will review the relationship between the exponents α and β (Section

3). Thirdly, we will show that both exponents equate when their value is the golden number (Section 4). Finally, we will discuss the finding (Section 5).

2 Calculation of $b(\beta)$

For certain natural values of β, exact values of $b(\beta)$ can be obtained. It is known that (Spiegel and Liu, 1999)

$$\zeta(2k) = \frac{2^{k-1}\pi^{2k}B_k}{(2k)!}, \tag{4}$$

where k=1,2,3,... and B_k is the k-th Bernoulli number. Thus, trivially

$$b(2k) = \frac{(2k)!}{2^{k-1}\pi^{2k}B_k}, \tag{5}$$

with k=1,2,3,.... Recalling B_1=1/6, B_2=1/30 and B_3=1/42 and Eq. 2, we obtain for example $b(2) = 6/p^2$, $b(4) = 90/p^4$ and $b(6) = 945/p^6$.

As for other values of β, we can find tight bounds using integrals. Knowing that any (integrable) monotonically decreasing function $f(k)$ satisfies (Cormen et al., 1990)

$$\int_m^{n+1} f(x)dx \le \sum_{k=m}^{n} f(k) \le \int_{m-1}^{n} f(x)dx, \tag{6}$$

we obtain

$$b(\beta) \le \beta\text{-}1 \tag{7}$$

for β>1. Notice that we cannot obtain a lower bound for $b(\beta)$ using Eq. 6 because x≥1. As for a lower bound for $b(\beta)$, it is easy to see that if $f(x)$ is an integrable monotonically decreasing function we have that

$$\sum_{k=m}^{n} f(k) \le f(m) + \int_{m}^{n} f(x)dx \tag{8}$$

and thus

$$b(\beta) \ge 1 - \frac{1}{\beta}. \tag{9}$$

In sum,

$$1 - \frac{1}{\beta} \le b(\beta) \le \beta - 1. \tag{10}$$

3 The equivalence between α and β

The relationship between the exponents β and α is well-known (Chitashvili and Baayen, 1993; Mandelbrot, 1997; Adamic, 2000; Pietronero et al., 2001; Adamic and Huberman, 2002)

$$\beta = \frac{1}{\alpha} + 1 \tag{11}$$

for β>1.

Now we provide a simple but not too simplified derivation of the previous equation (something in-between the sophisticated maths of (Chitashvili and Baayen, 1993; Mandelbrot, 1997) and the informal approaches of (Adamic, 2000; Pietronero et al., 2001; Adamic and Huberman, 2002)). We focus on the estimation of $x(i)$ from a sample of T occurrences of elements knowing that these elements are distributed according to $P(x)$ (Eq. 1) with β>1. $N(x)=TP(x)$ is the expected number of different elements whose magnitude is x in a sample of T occurrences distributed according to $P(x)$ (Eq. 1). The expected smallest rank of an element of magnitude x is

$$i_{\min}(x) = 1 + \sum_{x'=x+1}^{\infty} N(x') \tag{12}$$

and the expected largest rank is

$$i_{\max}(x) = \sum_{x'=x}^{\infty} N(x'). \tag{13}$$

Imagine that we assign an arbitrary rank within the interval [i_{min}(x), $i_{max}(x)$], to each of the $N(x)$ elements of magnitude x. Then, the mean rank of elements of frequency x is

$$\bar{i}(x) = \frac{1}{N(x')} \sum_{j=1}^{N(x)} \left(j + \sum_{x'=x+1}^{\infty} N(x') \right). \tag{14}$$

Knowing that (Spiegel and Liu, 1999)

$$\sum_{j=1}^{k} j = \frac{k(k+1)}{2}, \tag{15}$$

we obtain

$$\bar{i}(x) = \frac{N(x)+1}{2} + \sum_{x'=x+1}^{\infty} N(x') \tag{16}$$

after some algebra.

Applying the integral bounds of summations given in Eq. 6 to Eq. 16 with β>1 we obtain, after some work,

$$\bar{i}(x) \geq \frac{1}{2}\left[c(\beta)\left((x+1)^{1-\beta} + x^{1-\beta}\right)+1\right] \tag{17}$$

and

$$\bar{i}(x) \leq \frac{1}{2}\left[c(\beta)\left(x^{1-\beta} + (x-1)^{1-\beta}\right)+1\right], \tag{18}$$

where

$$c(\beta) = \frac{Tb(\beta)}{\beta - 1} \tag{19}$$

Knowing that $x^{1-\beta} + (x-1)^{1-\beta} \leq 2(x-1)^{1-\beta}$ and $(x+1)^{1-\beta} + x^{1-\beta} \geq 2(x+1)^{1-\beta}$ we can rewrite Eq. 18 as

$$\bar{i}(x) \geq c(\beta)(x+1)^{1-\beta} + \frac{1}{2} \tag{20}$$

$$\bar{i}(x) \leq c(\beta)(x-1)^{1-\beta} + \frac{1}{2}. \tag{21}$$

Writing x as a function of $\bar{i}$ in the two previous Eqs. we obtain

$$x(\bar{i}) = \left[\frac{1}{c(\beta)}\left(\bar{i} - \frac{1}{2}\right)\right]^{-\frac{1}{\beta-1}} \pm 1. \tag{22}$$

Put differently, in the previous Eq. we have derived the relationship between the magnitude x and the mean rank of elements whose magnitude is zeta distributed. Interestingly, our error in the value of such magnitude is of only ± 1, which can be neglected for large $\bar{i}$.
Applying the method above, we also obtain

$$\left[\frac{1}{c(\beta)}(i_{\min} - 1)\right]^{-\frac{1}{\beta-1}} - 1 \leq x(i_{\min}) \leq \left[\frac{1}{c(\beta)}(i_{\min} - 1)\right]^{-\frac{1}{\beta-1}} \tag{23}$$

and

$$\left(\frac{1}{c(\beta)} i_{\max}\right)^{-\frac{1}{\beta-1}} \leq x(i_{\max}) \leq \left(\frac{1}{c(\beta)} i_{\max}\right)^{-\frac{1}{\beta-1}} + 1. \tag{24}$$

4 The golden number

Equating the l.h.s. of Eq. 24 and the r.h.s. of Eq. 3 we obtain

$$\alpha = \frac{1}{\beta - 1} \tag{25}$$

for β>1. Similarly, comparing Eq. 22 with Eq. 3, we obtain that Eq. 25 also holds approximately.

Knowing Eq. 25, β equates α when β=1/(β-1), which leads to the quadratic equation

$$\beta^2 - \beta - 1 = 0\,. \tag{26}$$

The previous Eq. has two solutions, i.e.

$$\beta = \frac{1 \pm \sqrt{5}}{2} \tag{27}$$

of which the negative must be discarded for two reasons. First, Eq. 25 is obtained assuming β>1. Second, $x(i)$ is, by definition, monotonically decreasing. The positive solution is $\varphi = \left(1 + \sqrt{5}\right)/2$, the famous golden number (Ghyka, 1977; Walser, 2001).

5 Discussion

We have seen that the exponent φ is the value where the exponents of the probability distribution of a discrete magnitude and that of the value of the magnitude versus its rank coincide. This is one among many contexts in which the golden number appears (Ghyka, 1977; Walser, 2001). Probably, one of the most famous places where this number appears is the Fibonacci series, that is defined by the recurrence relation

$$F_{n+1} = F_{n-1} + F_n \tag{28}$$

for $n \geq 2$ with $F_0=0$ and $F_1=1$. The begining of the series is 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181, 6765, 10946, 17711, 28657, 46368, 75025, 121393, 196418, 317811,... It is well-known that the golden number is the limit ratio between two consecutive numbers of the series, i.e.

$$\lim_{n\to\infty} \frac{F_{n+1}}{F_n} = \varphi \tag{29}$$

At this time, we believe that it cannot yet be said if our discovery of φ is just a mathematical curiosity a the beginning of a series of discoveries that may change the way in which scientists study power laws in mathematical (Mitzenmacher, 2003; Mandelbrot, 1997) and natural sciences (Gisiger, 2001; Stanley, 1999; Newman, 2005). We hope that our finding stimulates further research.

Acknowledgements

RFC's work was funded by the projects FIS2006-13321-C02 and BFM2003-08258-C02-02 of the Spanish Ministry of Education and Science under a Juan de la Cierva contract from the Spanish Ministry of Education and Science.

References

Adamic, L., (2000). Zipf, power-law, Pareto - a ranking tutorial. Available from: http://www.hpl.hp.com/research/idl/papers/ranking/.

Adamic, L. A., Huberman, B. A., (2002). Zipf's law and the internet. Glottometrics 3, 143–150.

Chitashvili, R. J., Baayen, R. H., (1993). Word frequency distributions. In: Altmann, G., Hřebíček, L. (Eds.), Quantitative Text Analysis. Wissenschaftlicher Verlag Trier,

Trier, pp. 54–135.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., (1990). Introduction to algorithms. The MIT Press, Cambridge, MA.

Ferrer i Cancho, R., Servedio, V., (2005). Can simple models explain Zipf's law for all exponents? Glottometrics 11, 1–8.

Ghyka, M., (1977). The geometry of art and life. Dover, New York, 1st appeared as Ghyka, C. M. (1927). Esthétique des proportions dans la nature et dans les arts. Paris: Gallimard.

Gisiger, T., (2001). Scale invariance in biology: coincidence of footprint of a universal mechanism. Biol. Rev. 76, 161–209.

Mandelbrot, B., (1997). Fractals and scaling in finance: discontinuity, concentration, risk. Springer, New York, Ch. Rank-size plots, Zipf's law, and scaling.

Mitzenmacher, M., (2003). A brief history of generative models for power law and lognormal distributions. Internet Mathematics 1, 226–251.

Newman, M. E. J., (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323–351.

Pietronero, L., Tosatti, E., Tosatti, V., Vespignani, A., (2001). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. Physica A 293, 297–304.

Simkin, M. V., Roychowdhury, V. P., (2006). Re-inventing Willis. physics/0601192.

Spiegel, M., Liu, J., (1999). Mathematical handbook of formulas and tables. McGraw-Hill, New York, 2nd edition.

Stanley, E., (1999). Scaling, universality and renormalization: three pillars of modern critical phenomena. Reviews of Modern Physics 71 (2), S358–S366.

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

Walser, F., (2001). The golden section. The Mathematical Association of America, Washington.

Wimmer, G., Altmann, G., 1(999). Thesaurus of univariate discrete probability distributions. Stamm, Germany.

Wimmer, G., Sidlik, P., Altmann, G., (1999). A new model of rank-frequency distribution. Journal of quantitative linguistics 6, 188–193.

Zipf, G. K., 1972. Human behaviour and the principle of least effort. An introduction to human ecology. Hafner reprint, New York, 1st edition: Cambridge, MA: Addison-Wesley, 1949.

2.3. Relevancia de la ley de Zipf en la comunicación

Se ha presentado en el apartado anterior, con intención aproximativa y para evitar dogmatismos, la ley de Zipf como una función matemática más, entre otras posibles, de las capaces de ajustar las frecuencias de palabras con respecto al rango (Li *et al.* 2010). ¿Hay pues alguna función mejor que la ley de Zipf que se erija como 'ley' fundamental de la lingüística? ¿Es relevante la ley de Zipf? ¿Qué implicaciones teóricas tiene la ley de Zipf? Empezaremos por las dos últimas preguntas, dejando la primera para el final del apartado.

Desde la lingüística cuantitativa nadie discute la relevancia de la ley de Zipf en ámbitos muy diversos: se han dado argumentos tanto cuantitativos como cualitativos como puede revisarse en el volumen especial de *Glottometrics* (VV.AA., 2002), en honor a G.K. Zipf, por ejemplo. Pero no se trata ahora de mostrar idolatría o ser presa de cierto dogmatismo romántico, también presente a veces en la ciencia y en algunas ramas de la lingüística, respecto la ley de Zipf.

La revisión de Powers (1998) destacaba la relevancia de la ley de Zipf en diversos contextos de la lingüística computacional:

a) **Aprendizaje estadístico en corpus**. La ley de Zipf nos dice cuánto texto, como mínimo, necesitamos analizar en un corpus dado para ser significativo el estudio y qué precisión debe tener nuestra estadística. Puede ser una herramienta fundamental en el diseño de sistemas lingüísticos en máquinas.

b) **Semántica y recuperación de información**. La ley de Zipf proporciona un modelo de base con el que determinar la relevancia de un término aparecido en un corpus y puede proporcionar información importante lexicométrica.

c) **Evaluación de segmentadores (*parsers*).** De nuevo la ley de Zipf proporciona un modelo de referencia en el que probar la eficacia de los sistemas computacionales de segmentación.

d) **Psicolingüística computacional.** La ley de Zipf establece la distribución de frecuencias a la que se debería exponer a todo aquel que aprendiese una lengua y, por ende, permite la evaluación de modelos de aprendizaje lingüístico.

La perspectiva pragmática de Powers (1998) también lleva al autor a conectar la obra de Zipf (1949) con la teoría de la información de Shannon y Weaver (1949), en especial en lo referente a la minimización del esfuerzo del emisor y el receptor, y a

analizar posteriormente las desviaciones de la ley de Zipf debidas al tamaño del corpus y al tipo de muestra, para concluir, siguiendo las tesis de Mandelbrot (1953):

> Zipf's theory requires effort to be constant independent of frequency, however Information Theory and Psychological experiments both indicate that this ought not to be the case, and that it in fact decreases in a way consistent with an optimal strategy for an unbounded lexicon.

La ley de Zipf (1949) se fundamenta en lo que Zipf denominó el 'Principio del mínimo esfuerzo' en el cual se propone que el habla humana y el lenguaje se estructura óptimamente bajo la presión de dos fuerzas opuestas: unificación y diversificación. Es decir, si un repertorio comunicativo es demasiado unificado o repetitivo, entonces los mensajes posibles son solo unos pocos para expresar todo un surtido de informaciones, luego habrá una baja complejidad comunicativa. En un caso extremo tendríamos una única señal para expresar todo mensaje.

La teoría de la comunicación de Shannon (1948) ofrece herramientas cuantitativas para aproximarse al problema. Si el repertorio es muy diverso, o distribuido al azar, tenemos en el caso extremo una señal diferente para cada mensaje (McCowan *et al.*, 1999; Zipf, 1949). La ley de Zipf emerge del conflicto entre dos principios sin que ninguno de los dos se cumpla totalmente: la minimización de la entropía de las palabras y la maximización de la información mutua entre palabras y significados, pues aunque la información transmitida no se maximiza cuando se llega a un equilibrio entre la unificación y la diversificación (Ferrer-i-Cancho y Solé, 2003; McCowan *et al.*, 1999), se llega a un óptimo comunicativo.

La teoría de la información de Shannon (1948) nos alerta de que es necesaria cierta redundancia para la transmisión de información en un canal con ruido. En este marco, el principio del mínimo esfuerzo implica que energéticamente de entrada lo más económico para un emisor sería emitir una única señal para toda información a transmitir, mientras que lo mejor para el receptor –que minimizaría su esfuerzo– sería que cada mensaje se transmitiera en una señal diferente (Ferrer-i-Cancho y Solé, 2003; Zipf, 1949), facilitando así su decodificación del mensaje.

En este sentido, Ferrer-i-Cancho y Solé (2003) partiendo de un modelo sencillo que conecta señales y objetos de referencia mediante una matriz binaria, mostraron, en el marco de la teoría de la información, que la ley de Zipf es una consecuencia del equilibrio de fuerzas que se establece entre el emisor y el receptor en la comunicación y que es una función que permite la optimización energética del sistema lingüístico

formado por el emisor y el receptor, y se apuntaba a su relevancia para la comunicación simbólica y la sintaxis, afirmación que se apuntaló en Ferrer-i-Cancho y colaboradores (2005).

Aparte de la teoría de la información y de la visión de optimización comunicativa de Ferrer-i-Cancho y Solé (2003), la ley de Zipf se ha derivado de otros supuestos matemáticos para leyes de potencias (Saichev, Malevergne y Sornette, 2010; Newman, 2005, para una revisión), que suelen caracterizarse por la invariancia de escala, también en la ciencia cognitiva (Kello *et al.*, 2010; Petersen *et al.*, 2012), y la universalidad (Corominas-Murtra y Solé, 2010).

Como se vio al principio del capítulo, aunque procesos estocásticos de escritura (como los propuestos por Mandelbrot (Mandelbrot, 1953; Mandelbrot, 1962) y Simon (Simon, 1955; Simon, 1960) o el silencio intermitente (Miller, 1957) sean capaces de recuperar la ley de Zipf, más recientemente se ha demostrado que ello no implica que la ley de Zipf carezca de significado (Manin, 2008, 2009). Puede aparecer un patrón –como la ley de Zipf– en un proceso de aleatorización y ello no implica que en un sistema lingüístico real no sea significativa (Ferrer-i-Cancho, 2005b; Ferrer-i-Cancho, 2005c).

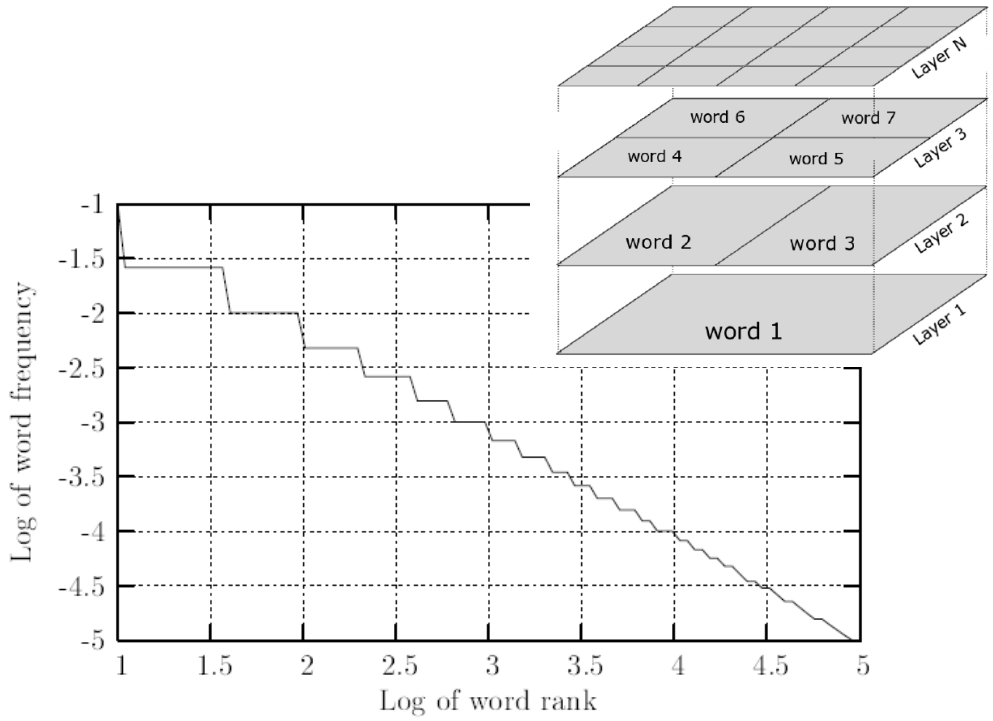


Figura 2.3: Recuperación de la ley de Zipf en un modelo jerárquico de distribución del espacio semántico (imágenes de Manin, 2008).

En este sentido, la organización cognitiva de los elementos semánticos (Manin, 2008) podría explicar la ley de Zipf, sin ser contradictoria con los conceptos de optimización ya propuestos por Zipf (Zipf, 1949) y sus revisiones posteriores (Mandelbrot, 1962; Ferrer-i-Cancho y Solé, 2003; Ferrer-i-Cancho, 2005b; Ferrer-i-Cancho, 2005c; Baixeries *et al.*, 2013)).

Manin (2008) cita el trabajo pionero de Guiraud (1968) para justificar la ley de Zipf como producto de la estructura del significado, que se reflejaría en su correspondiente significando. Manin (2008) interpreta la semántica estructuralista de Guiraud (1968) asignando a la forma de cada palabra un subconjunto de elementos básicos de significado (*semes*) que configuran su significado, así como su espacio semántico y sus relaciones (hiperonimia, homonimia, polisemia, sinonimia…). La segmentación jerárquica del espacio semántico en palabras conduce a una distribución zipfiana (figura 2.3) y según Manin (2008) equilibra fenómenos semánticos como la sinonimia y la polisemia, en una interesante interpretación de la oposición de fuerzas de diversificación/unificación (Ferrer-i-Cancho y Solé, 2003) ya establecida por Zipf (1949).

Por otra parte, un aspecto crucial es la variabilidad del exponente de Zipf: si la ley de Zipf fuese una mera casualidad o curiosidad estadística, ¿cómo se explicarían las regularidades y desviaciones halladas en diversos sistemas comunicativos?

La ley de Zipf presenta desviaciones muy interesantes, tanto por las implicaciones teóricas que suponen como por los casos en los que se dan. Así frente al típico exponente $b = 2$ ($a=1$ si partimos de la ecuación 1) que tenemos para el lenguaje adulto, las desviaciones encontradas en el lenguaje son diversas. Ferrer-i-Cancho (2005) recoge algunas de ellas (acúdase a la fuente para una revisión bibliográfica de cada desviación) detectadas hasta la fecha de su publicación:

- En patología del lenguaje, la producción de los esquizofrénicos ha sido estudiada al detalle encontrando desviaciones de la ley de Zipf tales que $b > 2$ (entre 2.12 y 2.42) para esquizofrénicos en una primera fase de discurso fragmentado y con empleo de léxico caótico, y $1 < b < 2$, para esquizofrénicos en fase más avanzada y con un discurso marcado por la fuerte presencia de términos obsesivos.
- Si se analizan solo nombres, en sujetos adultos sanos, se obtiene un exponente $b = 3.35$, lo que sería normal cuando nos restringimos al

estudio de un único tipo de palabras, puesto que la balanza comunicativa se decanta del lado del receptor (Ferrer-i-Cancho y Solé, 2003).

- La variabilidad de los exponentes de Zipf también se ha documentado en textos legales escritos, con un exponente algo menor al esperado (Ha *et al.*, 2002; Smith y Devine, 1985), mientras que un muestreo deficiente lleva en general a b >>2, (Ferrer-i-Cancho, 2005)

Ferrer-i-Cancho (2005) resumía también las desviaciones de la ley de Zipf halladas hasta aquel momento en la adquisición del lenguaje, y que se estudiaron con mucha más profundidad y exhaustividad en Baixeries *et al.* (2013), donde además de la producción infantil longitudinal se analizó el habla de los progenitores y otros adultos (como los investigadores) que intervinieron en la recogida de datos de las diversas fuentes de la base de datos CHILDES[10].

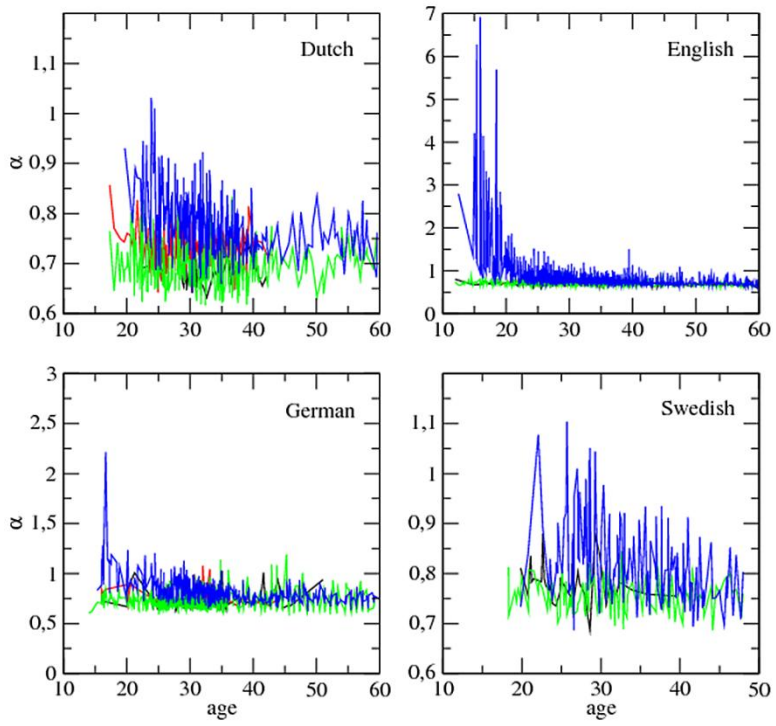


Figura 2.4: Evolución del exponente *a* (α en las figuras) de Zipf en la adquisición del lenguaje, para el inglés, holandés, alemán y sueco, según el estudio de Baixeries *et al.* (2013). En negro se dibuja la evolución infantil, en gris la de las madres, investigadores y padres.

[10] Parte de la base de datos CHILDES fue la base del estudio de Baixeries *et al.* (2013). Accesible en: http://childes.psy.cmu.edu

La variabilidad y evolución de los exponentes de Zipf demostrada por Baixeries *et al.* (2013) nos conduce necesariamente a una visión dinámica del exponente de Zipf (figura 2.4) en la adquisición del lenguaje y, por extensión, en todo fenómeno comunicativo. Pero no solo eso: el exponente de Zipf obtenido es $a<1$, inferior a la unidad, para adultos y para niños (en promedio en los niños $0.71<a<0.87$). Todavía quedan cuestiones abiertas como la influencia de la modalidad (oralidad versus escritura) o del habla dirigida a niños (lenguaje maternal y habla dirigida a niños) en el exponente de Zipf (Hernández-Fernández, 2006).

Una explicación a las desviaciones encontradas en la producción infantil se enmarcaría en las modernas teorías de la mente: en la producción infantil más temprana el niño no tiene en cuenta las necesidades del receptor; el niño expresa sus necesidades del aquí-y-ahora, y comunicativamente tiende a comportarse como un emisor 'egoísta' (Hernández-Fernández, 2006). Otros han considerado que las limitaciones cerebrales de un cerebro aún inmaduro (McCowan *et al.*, 1999) explicarían las desviaciones del exponente de Zipf (en el caso de McCowan *et al.*(1999), en comunicación animal para los exponentes de delfines jóvenes), de forma que la emergencia sintáctica podría ir de la mano de la recuperación del exponente "adulto" de la ley de Zipf (Hernández-Fernández, 2006; Ferrer-i-Cancho *et al.*, 2005). En la producción animal, para el delfín mular (*Tursiops truncatus*) McCowan *et al.* (1999) hallaron un exponente a = 0.95, muy próximo al teórico a= 1 humano, y desviaciones en delfines jóvenes que evolucionan con la edad (a = 0.82 para delfines de menos de 1 mes, a = 1.07 para delfines de entre 2 y 8 meses) y recuperan a = 0.95 para delfines de entre 9 y 12 meses. Introducir aquí la referencia a McCowan *et al.* (1999) es añadir además la relevancia y la potencialidad de la ley de Zipf para analizar corpus en comunicación animal.

Baixeries *et al.* (2013) demostraron también la capacidad de la lingüística cuantitativa para enfrentarse también a corpus pequeños. Las desviaciones de la Ley de Zipf no pueden ser explicadas por los modelos de Simon o el silencio intermitente de Mandelbrot (Ferrer-i-Cancho, 2005). Presenta enorme interés, por tanto, ver qué desviaciones en concreto se dan en la adquisición y la pérdida del lenguaje y en la comunicación animal, y más cuando hay varios modelos teóricos detrás de la ley de una ley de Zipf, ávida de datos empíricos que permitan su consolidación.

Nuestra intención en Hernández-Fernández y Diéguez-Vide (2013) fue realizar una primera aproximación zipfiana al fenómeno de la pérdida del lenguaje. Pese a la potencia estadística de aproximaciones como la de Baixeries *et al.* (2013), el tamaño de

la muestra, sin embargo, sigue siendo un problema empírico, más en estudios de pérdida del lenguaje, caso de la afasia (Van Ewijk, 2011) o de la demencia (Hernández-Fernández y Diéguez-Vide, 2013), que en adquisición del lenguaje (Baixeries *et al.*, 2013; Roy *et al.*, 2006).

En afasiología la producción de los pacientes a menudo no es extensa, lo que puede conducir a muestreos pequeños, aunque ajustables y en los que se recupera la ley de Zipf para pacientes individuales pero con alteraciones en el exponente (Van Ewijk, 2011) o en grupos de pacientes similares (Hernández-Fernández y Diéguez-Vide, 2013). Por ende, entendemos que la concepción dinámica del exponente de Zipf (Baixeries *et al.*, 2013) es necesaria en especialmente en afasiología (Van Ewijk, 2011) y en el estudio de la demencia (Hernández-Fernández y Diéguez-Vide, 2013).

Aunque hay otras escalas en clínica para la clasificación de la demencia, como la FAST de evaluación funcional o la CDR (véase De León y Reisberg, 1999), en Hernández-Fernández y Diéguez-Vide (2013) se utilizó la clásica Escala Global de Deterioro o GDS (Reisberg *et al*., 1982), que se resume en la tabla 2.2. Una de las dificultades actuales en el estudio de la demencia, y en concreto en la demencia tipo Alzheimer, es la detección precoz (Cuetos-Vega *et al.*, 2007). En Hernández-Fernández y Diéguez-Vide (2013) aplicamos la ley de Zipf, por primera vez, como herramienta en el estudio de la evolución de la producción oral en pacientes con Alzheimer en dos fases clínicas (GDS4 y GDS5). Así, a través del simple análisis del léxico y de la distribución de frecuencias de palabras, fue posible discriminar el grupo GDS4 del GDS5, y de sus respectivos controles.

La relevancia de la ley de Zipf para los sistemas de comunicación, tanto a nivel sintáctico (Ferrer-i-Cancho *et al.*, 2005) como semántico (Manin, 2008), justifica teóricamente la realización del estudio práctico de la distribución de frecuencias de palabras y la descripción de la evolución de la producción oral en pacientes con enfermedad de Alzheimer y a proponer el exponente de Zipf como un estadístico más en los estudios de la pérdida del lenguaje (Hernández-Fernández y Diéguez-Vide, 2013). ¿Puede el exponente de Zipf ayudarnos a la detección precoz de la EA? ¿Qué implicaciones teóricas tienen las desviaciones de la ley de Zipf para la ciencia cognitiva? ¿Hay funciones estadísticamente mejores para ajustar los datos empíricos de los que disponemos? Nuestro trabajo es solo un primer paso en una línea en la que debe sin duda profundizarse más.

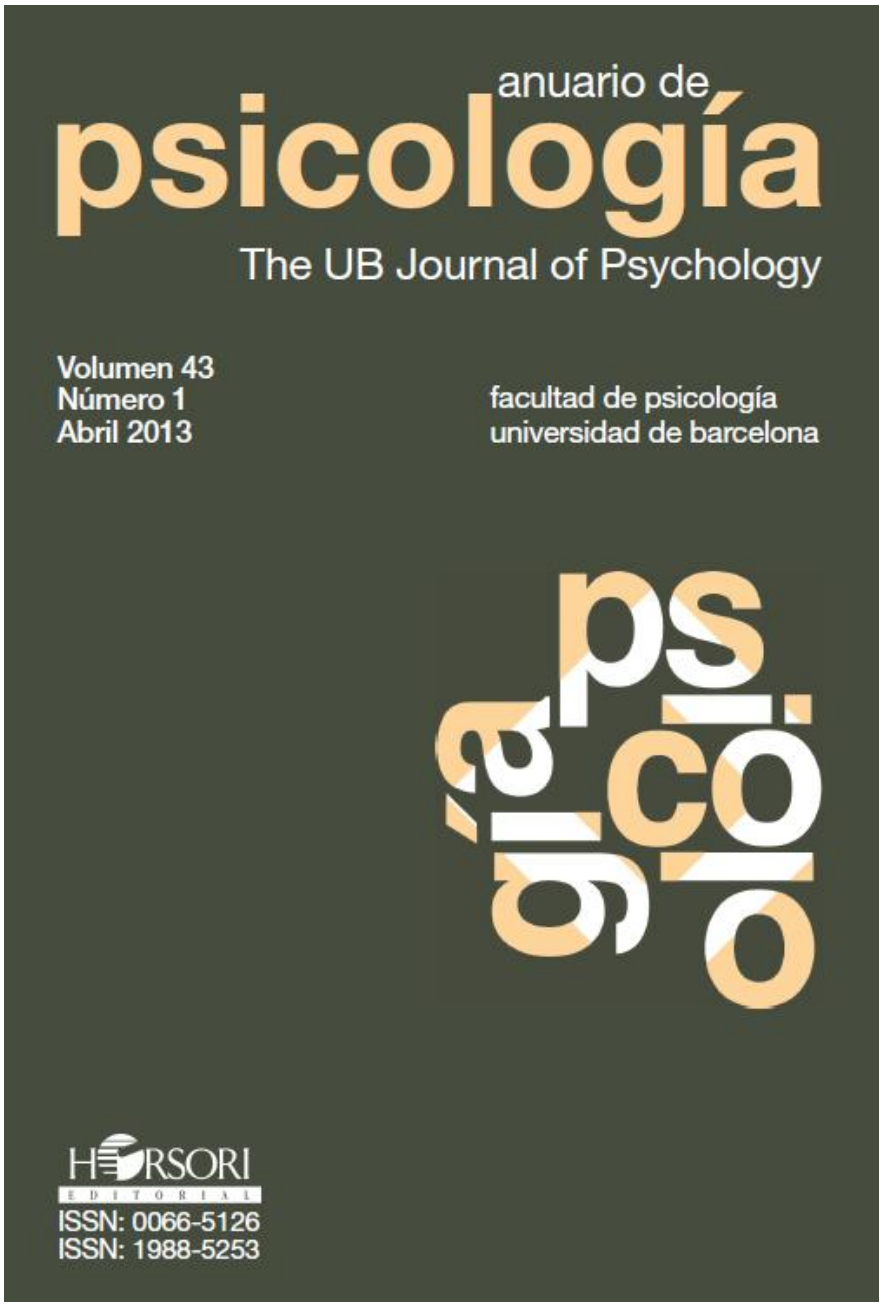
Diagnóstico	Fase	Señales y Síntomas
Falta de demencia	**Fase 1: Ausencia de alteración cognitiva**	En esta fase la persona tiene una función normal, no hay pérdida de memoria. Hay salud mental y no hay demencia.
Falta de demencia	**Fase 2: Disminución cognitiva muy leve**	Olvido normal asociado al envejecimiento (nombres y ubicación de objetos familiares). Los síntomas no son evidentes a los seres queridos ni al médico.
Falta de demencia	**Fase 3: Declive cognitivo leve**	Falta de memoria creciente, capacidad de trabajo mental disminuida. Más dificultad para encontrar las palabras correctas, con síntomas evidentes para los más cercanos.
Etapa temprana	**Fase 4: Declive cognitivo moderado**	Dificultades de concentración y para viajar solo a lugares nuevos, disminución de la memoria de eventos recientes y problemas para ejecutar tareas complejas de precisión. Puede no querer reconocer sus síntomas y recluirse porque las interacciones sociales se hacen más difíciles.
Etapa media	**Fase 5: Declive cognitivo grave**	Deficiencias serias de memoria, necesita ayuda para realizar las actividades de la vida diaria (vestirse, preparar comida, asearse…). Síntomas cognitivos claros durante la evaluación y entrevista clínica. Tendencia a refugiarse en la familia.
Etapa media	**Fase 6: Declive cognitivo severo (demencia media)**	Las personas en esta fase requieren ayuda de tercera persona para hacer las actividades diarias. Graves problemas de memoria (olvido de nombres de familiares, muy poco recuerdo de eventos recientes...) y aritmético-lógicos. Incontinencia, agitación, alteraciones de personalidad, compulsiones y ansiedad.
Etapa avanzada	**Fase 7: Declive cognitivo muy severo (demencia avanzada)**	Se añade a la situación anterior la pérdida de las capacidades comunicativas y de habilidades motoras básicas como caminar. Esa pérdida motora de sistemas básicos del organismo suele conducir a la muerte.

Tabla 2.2: Escala Global del Deterioro o *Global Deterioration Scale* (GDS). Adaptada de Reisberg *et al*. (1982) y De León y Reisberg (1999).

2.4. Hernández-Fernández, A. y Diéguez-Vide, F. (2013). La ley de Zipf y la detección de la evolución verbal en la enfermedad de Alzheimer.

Se incluye en este apartado el artículo:

- Hernández-Fernández, A. y Diéguez-Vide, F. (2013). La ley de Zipf y la detección de la evolución verbal en la enfermedad de Alzheimer, *Anuario de Psicología/The UB Journal of Psychology*, 43 (1), 67-82.



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.

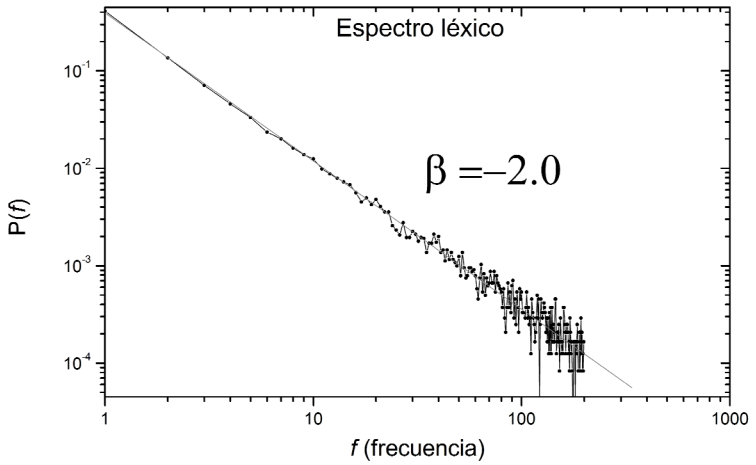


Figura 1. Representación de las palabras de frecuencia f, frente a su probabilidad P(f), en escala logarítmica, para un corpus de 8.000 palabras. Se recupera la ley de Zipf (ecuación 1) con exponente β≈2, típica en la producción adulta (Hernández-Fernández, 2005).

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).

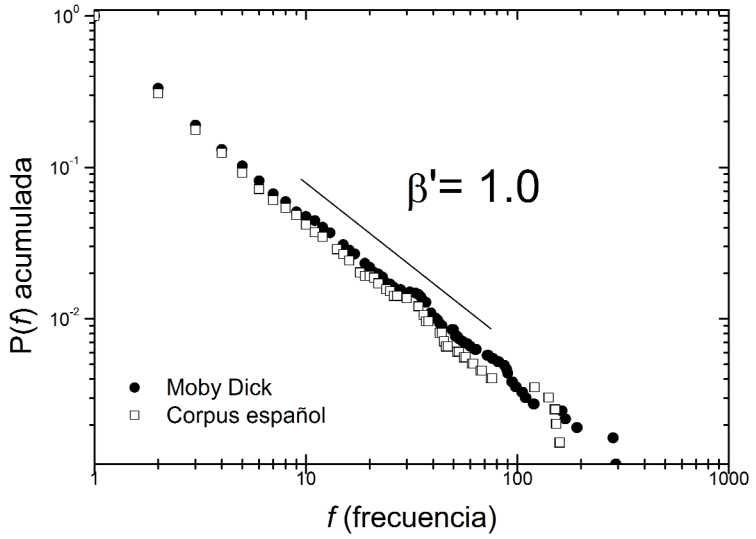


Figura 2. Recuperación de la ley de Zipf en un corpus escrito en español (Hernández-Fernández 2005) comparado con un corpus inglés de similar tamaño (10.000 tokens de la obra Moby Dick). Se representa en escala logarítmica la frecuencia de las palabras frente a la probabilidad acumulada de aparición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

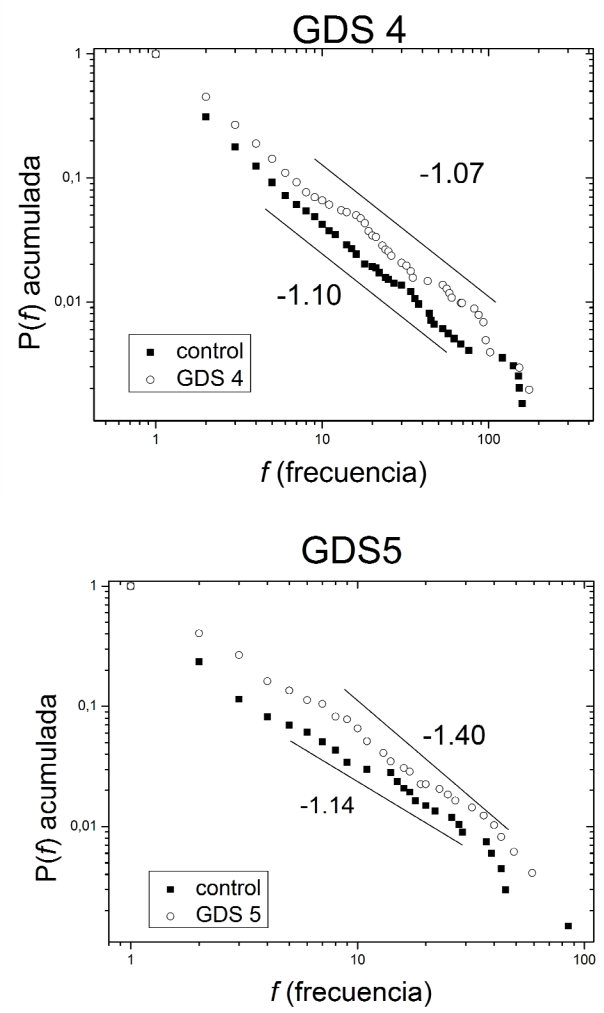


Figura 3. Representación de la frecuencia de la distribución de palabras, f, frente a su probabilidad acumulada P(f), para los pacientes diagnosticados GDS5 y GDS4 con sus respectivos controles.

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.

2.5. La ley de Menzerath-Altmann

Una de las regularidades fundamentales del lenguaje (Polikarpov, 2006) es la tendencia empírica que Paul Menzerath encontró por primera vez en 1954: la longitud media de las sílabas de una palabra, medida en letras (o en fonemas) se correlaciona negativamente con la longitud total de la palabra, medida en número de sílabas (Menzerath, 1954). Este hallazgo se generalizó en la lingüística cuantitativa (Hrebicek, 1995) al encontrarse en múltiples niveles de estudio (morfemas, frases, textos…) y se reformuló posteriormente (Polikarpov, 2006; Li, 2012) especialmente por Gabriel Altmann (Altmann, 1980; Teupenhayn y Altmann, 1984): la conocida como ley de Menzerath-Altmann sostiene que a mayor tamaño de un constructo lingüístico, menores serán sus partes o constituyentes. Así, cuanto más larga es una frase tienden a ser más cortos sus sintagmas (medidos en número de palabras), y en palabras más largas tienden a ser más breves sus sílabas.

Matemáticamente, si X es el número de partes de un constructo lingüístico (como el número de sintagmas de una frase, por ejemplo), e Y el tamaño total de cada constituyente del constructo (el número total de palabras de la frase), entonces entendiendo que ambas variables son discretas podemos definir $Z= Y/X$, para $X\neq0$, como el tamaño medio del constructo (la longitud media de cada sintagma en número de palabras). En este caso, la ley de Menzerath-Altmann afirma que Z tiende a decrecer a medida que X se incrementa, y se puede modelizar como (Altmann, 1980; Cramer, 2005):

$$Z = aX^{b}e^{cX} \qquad 8)$$

Siendo a, b y c los parámetros del modelo. La complejidad del lenguaje y de sus componentes y subsistemas se sustenta por correlaciones cuantitativas que, como la ley de Menzerath-Altmann, han sido verificadas para lenguas muy diversas, incluyendo la ecuación exponencial (8) tal y como Fenk-Oczlon y Fenk (2008) constataron:

a) A más sílabas por palabra, menos fonemas por sílaba (la original ley de Menzerath (1954)).

b) A menos fonemas por sílaba, más sílabas por sintagma.

c) A más sílabas por frase, más sílabas por palabra. Este es un caso de correlación positiva (no se cumple la Ley de Menzerath-Altmann).

d) A más sílabas por palabra, menos palabras por frase.

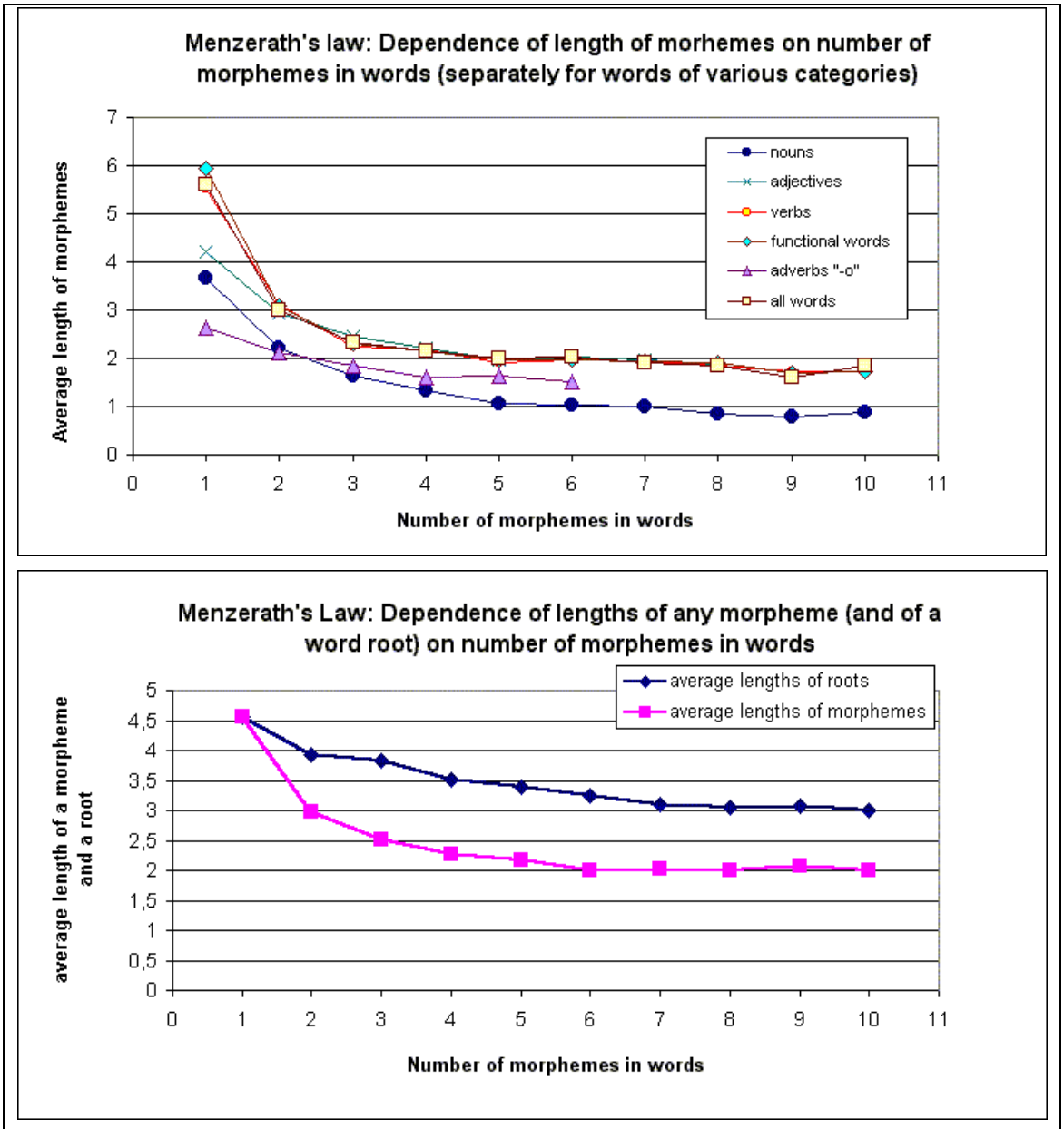


Figura 2.5: Disminución de la longitud media de los morfemas y las raíces léxicas según el número de morfemas por palabra (ley de Menzerath-Altmann), para diversas categorías. Extraída de Polikarpov (2006).

Así por ejemplo Polikarpov (2006) recupera para el ruso la ley de Menzerath-Altmann a nivel morfológico, de forma que la longitud de los morfemas –incluidas raíces léxicas– disminuye cuando las palabras son más largas (figura 2.5).

Fenk-Oczlon y Fenk (2008, pp.58-59) van más allá y relacionan los parámetros cuantitativos anteriores con la complejidad morfológica y semántica de las lenguas, tras

encontrar un aumento exponencial del número de monosílabos, al aumentar el número de sílabas diferentes posibles:

> We think that low morphological complexity especially in word structure favours a higher semantic complexity -e.g., a tendency to homonymous and polysemous expressions encoding a higher number of senses– which in turn requires and favours a rather formulaic speech, i.e., stable fragments of speech that allow a quick identification of the context-relevant meaning. Therefore the context should be stored and memorized together with the homonymous and polysemous words. A high proportion of idioms and a tendency to formulaic speech increase the cognitive costs in the acquisition of the respective language.

Tal y como revisaron Ferrer-i-Cancho y Forns (2009), la ley de Menzerath-Altmann se ha encontrado, además de en el lenguaje (Altmann, 1980; Teupenhayn y Altmann, 1984; Fenk-Oczlon y Fenk, 2008) y en la música (Boroda y Altmann, 1991), también en el genoma (Wilde y Schwibbe, 1989; Ferrer-i-Cancho y Forns, 2009; Li, 2012; véase el capítulo 3 de esta tesis). En el caso del genoma se abrió un debate (Solé, 2010; Ferrer-i-Cancho, Forns *et al*., 2013, en esta tesis) sobre la relevancia de la ley de Menzerath-Altmann y sus parámetros (Baixeries *et al*., 2013, en esta tesis), así como su nivel de aplicación en el genoma (Li, 2012). Como se verá en el capítulo siguiente, si bien la ley no se recupera en todos los genomas estudiados, sí parece un camino interesante a seguir explorando y que dista mucho de ser trivial (Solé, 2010).

De hecho, Fenk-Oczlon y Fenk (2008, p. 59) prosiguen por ejemplo citando a Jespersen para relacionar la frecuencia de monosílabos en las lenguas con la homonimia y la polisemia, y por extensión con la complejidad semántica:

> According to Jespersen (1933) there are about four times more monosyllabic than polysyllabic homonyms: "The shorter the word, the more likely is it to find another word of accidentally the same sound". Homonymy affects of course also "parts of speech" and most of the "grammatical homophones" such as love (verb and noun) or round (noun, adjective, adverb, preposition, and verb) are again monosyllables. Because of the well known association between frequency and polysemy on the one hand and frequency and shortness on the other, polysemy should also be a frequent phenomenon in monosyllabic words. Both homonymy and polysemy may be viewed, as already mentioned, as dimensions of semantic complexity.

Luego relaciones como la ley de Menzerath-Altmann están detrás de la complejidad del lenguaje y de fenómenos como la relación entre la frecuencia de las palabras y el tamaño de las mismas (ley de Zipf de brevedad, véase apartado 2.7.).

Altmann (1980) se aproximó al problema partiendo de una ecuación diferencial (Andres *et al.*, 2012) con la que derivar la ecuación 8:

$$\frac{\dot{Z}}{Z} = \frac{b}{X} + c \qquad 9)$$

Con *b* y *c*, los mismos parámetros que en (8) y $\dot{Z} = \frac{dZ}{dX}$. Como muestran Andres *et al.* (2012), integrando se llega a la forma anterior (8) de la ley de Menzerath-Altmann[11], con *a>0* necesariamente, y se pueden recuperar otras expresiones de la ley, como su versión *truncada*, cuando *c*=0. Andres *et al.* (2012) muestran las diferencias cuantitativas entre cuatro versiones de la ley de Menzerath-Altmann, según los parámetros escogidos, al aplicar el método de mínimos cuadrados en la optimización de parámetros. En su caso, el mejor modelo que resulta es el que contiene todos los parámetros (ecuación 8), aunque no hay ninguna valoración bajo criterios como el AIC que permitan discernir si el resultado es mejor debido a la mayor complejidad del ajuste. Es un claro ejemplo de cómo aumentar la complejidad matemática de un modelo pude ajustar mejor los datos, pero sin un contraste claro entre modelos como apuntan Li y colaboradores (2010) o Ferrer-i-Cancho y colaboradores (2014).

Eroglu (2013), recientemente, ha realizado su propia interpretación, desde la mecánica estadística, a la ley de Menzerath-Altmann, proponiendo un modelo (SMMA) de cuatro parámetros en el que se recupera la distribución de la ley de Menzerath-Altmann, al menos para dos corpus (uno del inglés y otro del turco), pero en el que los parámetros de (8) quedan como parámetros independientes de la estructura del sistema.
La interesante idea de Eroglu (2013), todavía por desarrollar desde la lingüística teórica, es asociar los parámetros de su modelo SMMA con una interpretación física del sistema que se conecte con sus propiedades termodinámicas, se trate o no de un sistema lingüístico. Y concluye (Eroglu, 2013):

> The derivation procedure may suggest that the reason for the MA[12] law behavior's inevitable presence in many natural and artificial organizations might be the discrete and energy-preserving nature of such constructs' constituent configuration.

[11] En su caso toman –*b* en vez de *b*, lo que no es relevante aquí.

[12] MA, Menzerath-Altmann, nota no incluida en la cita de Eroglu (2013).

La conexión de la ley de Menzerath-Altmann con aspectos como la optimización energética es sin duda, un asunto sobre el que seguir investigando y que ya cuenta con evidencia empírica suficiente (Eroglu, 2013b; artículos del capítulo 3 de esta tesis), como para considerarse una trivialidad. La importancia actual de la investigación interdisciplinar en biología, lingüística y computación (Bel-Enguix *et al.*, 2011), así como la bondad de las aproximaciones que exploran la presencia de leyes potenciales en múltiples áreas de la genómica (Luscombe *et al.*, 2002) y el hecho de los recientes avances en el llamado *conectoma* y el estudio de las variaciones del genoma neuronal (McConell *et al.*, 2013), son un añadido que debe motivar a investigaciones en esta línea que, como decíamos, veremos en detalle en el capítulo 3.

Cabe todavía preguntarse qué sorpresas puede deparar en campos como la genética la presencia de la ley de Menzerath-Altmann u otras procedentes de la lingüística cuantitativa, pero lo que está claro es que toda propuesta teórica que parta de una aproximación cuantitativa debe aposentarse sobre una sólida base matemática, que es lo que aporta el artículo que se incluye en el apartado siguiente (Ferrer-i-Cancho *et al.*, 2014), en el que se revisan los fundamentos estadísticos de la ley de Menzerath-Altmann.

2.6. Ramon Ferrer-i-Cancho, Jaume Baixeries, Antoni Hernández-Fernández, Łukasz Debowski y Ján Macutek. (2014). *When is Menzerath-Altmann law mathematically trivial? A new approach.*

Aquí se incluye el artículo pendiente de publicación:

- Ferrer-i-Cancho, R., Baixeries, J., Hernández-Fernández, A., Debowski, L. y Macutek, J. (2014). *When is Menzerath-Altmann law mathematically trivial? A new approach.* Pendiente de publicación.
  Disponible en: http://arxiv.org/abs/1210.6599

When is Menzerath-Altmann law mathematically trivial? A new approach

Ramon Ferrer-i-Cancho[1], Jaume Baixeries[1], Antoni Hernández-Fernández[1,2], Łukasz Dębowski[3] and Ján Mačutek[4]

[1] Complexity & Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain
[2] Departament de Lingüística General, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona (Catalonia), Spain
[3] Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
[4] Department of Applied Mathematics and Statistics, Comenius University, Mlynska dolina, 84248 Bratislava, Slovakia

Abstract

Menzerath's law, the tendency of Z, the mean size of the parts, to decrease as X, the number of parts, increases is found in language, music and genomes. Recently, it has been argued that the presence of the law in genomes is an inevitable consequence of the fact that $Z = Y/X$, which would imply that Z scales with X as $Z \sim 1/X$. That scaling is a very particular case of Menzerath-Altmann law that has been rejected by means of a correlation test between X and Y in genomes, being X the number of chromosomes of a species, Y its genome size in bases and Z the mean chromosome size. Here we review the statistical foundations of that test and consider three non-parametric tests based upon different correlation metrics and one parametric test to evaluate if $Z \sim 1/X$ in genomes. The most powerful test is a new non-parametric based upon the correlation ratio, which is able to reject $Z \sim 1/X$ in ten out of eleven taxonomic groups. Rather than a fact, $Z \sim 1/X$ is a baseline that real genomes do not meet. The view of Menzerath-Altmann law as inevitable is seriously flawed.

Keywords: Menzerath-Altmann law – power-laws – genomes – Monte Carlo methods

Corresponding author: R. Ferrer-i-Cancho, rferrericancho@lsi.upc.edu

1 Introduction

Consider that X and Y are two discrete random variables and that $Z = Y/X$ with $X \neq 0$. For the particular case that Z is a mean, Menzerath's law is the tendency of Z to decrease as X increases (Menzerath, 1954, Li, 2012): X stands for the number of parts of the construct (e.g., the number of clauses of a sentence), Y stands for the size of the whole in parts (e.g., the length in words of the sentence) and Z stands for the mean size of the construct (e.g., the mean length of the clauses in words). The scaling of Z with X that Menzerath's law describes qualitatively is typically modelled by means of Menzerath-Altmann law (Altmann, 1980, Cramer, 2005), i.e.

$$Z = aX^b e^{cX}, \tag{1}$$

where a, b and c are the parameters of the model. Menzerath's law has been found in language (Menzerath, 1954) and genomes (Ferrer-i-Cancho and Forns, 2009, Li, 2012) and also indirectly through a wide range of studies where Menzerath-Altmann law is fitted to human language e.g., (Altmann, 1980, Teupenhayn and Altmann, 1984), and music (Boroda and Altmann, 1991) and genomes (Wilde and Schwibbe, 1989, Li, 2012) which yields parameters a, b and c that support a negative correlation between Z and X at least for sufficiently large X (see Cramer (2005) for a review of parameter values). Note that in our definition of these laws, Menzerath's law is a light and model neutral law. In our view, the only constraint imposed by that law is that Z and X are negatively correlated (the dependency between Z and X might not be functional) while Menzerath-Altmann law is a strong assumption for two reasons: it assumes that dependency between Z and X is functional and takes the form defined by Eq. 1.

Recently, it has been argued that $Z = Y/X$ leads inevitably to a power-law of the form (Solé, 2010)

$$Z = aX^{-1}, \tag{2}$$

that is Menzerath-Altmann law with $b = -1$ and $c = 0$. If the argument was correct, Menzerath-Altmann law would be a trivial scaling law at least from a mathematical perspective. Being X the number of chromosome of a species, Y the genome size in bases of that species and Z its mean chromosome size, agreement with Eq. 2 has been claimed simply by fitting Menzerath-Altmann law with $c = 0$ and obtaining $b \approx 1$ (Solé, 2010). However, more than a decade of statistical research on presumable power laws in biology indicates that looks can be deceiving (May and Stumpf, 2000, Tanaka, Yi, and Doyle, 2005, Stumpf and Ingram, 2005, Khanin and Wit, 2006). Sometimes, the divergence of the degree distribution of biological networks from a power-law is obvious upon visual inspection if a convenient representation of the data is employed (Tanaka et al., 2005). In general,

the hypothesis of a power-law for the degree distribution of biological networks has been rejected for different kinds of biological networks even when an exponentially truncated power-law similar to Eq. 1 was considered (Khanin and Wit, 2006). Modern model selection methods indicate that simple power-law models do not provide an adequate description of the degree distribution of protein interaction and metabolic networks (Stumpf and Ingram, 2005, Stumpf, Ingram, Nouvel, and Wiuf, 2005). The same has happened to other hypotetical power-laws after careful inspection (May and Stumpf, 2000, Tjørve, 2003). The dependency between the number of different species as a function of area is described more accurately by functions that are not power-laws (May and Stumpf, 2000). Interestingly, the power-law model (as well as the exponential model) is seen as lacking biological depth for describing the species-area relationship with regard to other possible functions (Tjørve, 2003). In a similar vein, Eq. 2 has been argued to imply assumptions that jeopardize chromosome well-formedness in the context of the relationship between chromosome number and mean chromosome length (Baixeries, Hernández-Fernández, and Ferrer-i-Cancho 2012, Ferrer-i-Cancho, Forns, Hernández-Fernández, Bel-Enguix, and Baixeries, 2013b).This suggests that rather than indicators of complexity, pure power laws (or purepower-laws with a certain exponent, e.g., -1 in Eq. 2) might play the role of base-lines in certain circumstances.The main goal of this article is presenting a powerful and statistically rigor-ous methodological framework to test if a real sample follows the particular case of power-law defined by Eq. 2. Notice that the equation is a particular case of power-law for two reasons: the exponent of the power-law is -1 and, more importantly, the response variable Z is $Z = Y/X$, being X the predictor.

Whether Eq. 2 holds in genomes for the relationship between chromosome number and mean chromosome size, has been debated (Solé, 2010, Ferrer-i-Cancho et al., 2013b, Hernández-Fernández, Baixeries, Forns, and Ferrer-i-Cancho, 2011, Baixeries et al.,2012, Ferrer-i-Cancho, Baixeries, and Hernández-Fernández, 2013a, Baixeries, Hernández-Fernández, Forns, 2013). This is part of a long-running debate on the depth and importance of statistical laws of language in science (e.g., Miller and Chomsky (1963), Li (1992),Suzuki, Tyack, and Buck (2005), McCowan, Doyle, Jenkins, and Hanser (2005), Solé(2010), Ferrer-i-Cancho et al. (2013b)).

These laws are seen by many as inevitable (Miller, 1968, Solé, 2010), useless (Suzuki et al., 2005) or lacking mechanistic so-phistication (Li, 1992, Stumpf and Porter, 2012). Here we aim to contribute to this general debate from the perspective of Menzerath-Altmann law with new theoreti-cal insights and new experiments on genomes. In particular, we will provide some theoretical foundations for testing if Eq. 2 holds. It will be shown that rejecting Eq. 2 if X and Y are correlated (Baixeries et al., 2012, Hernández-Fernández et al., 2011) is correct but conservative. Furthermore, a new test that rejects Eq. 2 in all taxonomic groups considered so far except one will be presented. The view of

Menzerath-Altmann law as inevitable (Solé, 2010) is seriously flawed.

2 Statistical foundations

2.1 The meaning of $Z = a/X$

According to standard modelling, claiming that Z scales with X following Eq. 2 can be recast as (Ritz and Streibig, 2008, pp. 1),

$$E[Z|X = x] = a/x, \tag{3}$$

for any x, being $E[Z|X = x]$ the conditional expectation of Z given x (a concrete value of X). Testing if Eq. 2 holds reduces to testing if Y is mean independent of X (Poirier, 1995, pp. 67), namely $E[Y|X = x] = E[Y]$ for any x (Ferrer-i-Cancho et al., 2013a). Formally, this is supported by the following theorem (Ferrer-i-Cancho et al., 2013a)

Theorem 2.1 *Consider a constant a and two random natural variables, X and Y, and a third random number Z, such that $X > 0$ and $Z = Y/X$. Then, $E[Z|X = x] = a/x$ if and only if Y is mean independent of X, i.e. $E[Y|X = x] = E[Y]$ for any x.*

Therefore, one condition for a trivial Menzerath-Altmann law is that Y is mean independent of X. Mean independence is well-known in econometrics (Cameron, and Trivedi, 2009, Wooldridge, 2010). Another more obvious mathematically trivial version of
the law occurs when Z is mean independent of X, i.e.

$$E[Z|X = x] = E[Z], \tag{4}$$

which is equivalent to constant $E[Z|X = x]$ (Ferrer-i-Cancho et al., 2013a). The analysis of the correlation between Z and X in genomes discarded this mean constancy of Z for nine out of eleven taxonomic groups (Ferrer-i-Cancho and Forns, 2009) and the result was confirmed with an updated dataset (Baixeries et al., 2012). Therefore, Menzerath-Altmann law as a model of $E[Z|X = x]$ has two trivial versions:

- $b = c = 0$: Z is mean independent of X.
- $b = -1$ and $c = 0$: Y is mean independent of X.

Interestingly, b lies between 0 and -1 when $c = 0$ is assumed: e.g., $b = -0.27 \pm 0.11$ in language, being Z is the mean clause length in sentences and X is the number of sentences (Teupenhayn and Altmann, 1984), and $b = -0.44 \pm 0.09$ in music, being Z is the mean F-motif length in tones and X is the number of F-motifs

(Boroda and Altmann, 1991). In both cases, we report $b = \mu \pm \sigma$, where b is the exponent of a sample while μ and σ are, the mean and the standard deviation of b in an ensemble of samples, respectively.

2.2 Three definitions of lack of association between X and Y

For the remainder of sections, it is important to bear in mind the definition of three statistical relations between X and Y (Poirier, 1995, pp. 67-68):

- *X and Y are independent:* $p(Y = y|X = x) = p(Y = y)$ for any x and y.
- *Y is mean independent of X:* $E[Y|X = x] = E[Y]$ for any x.
- *X and Y are uncorrelated:* $COV(X,Y) = 0$ where $COV(X,Y) = E[XY] - E[X]E[Y]$ is the covariance between X and Y. Notice that uncorrelation, i.e. $\rho(X,Y) = 0$, being $\rho(X,Y)$ the Pearson correlation coefficient, is equivalent to zero covariance as (DeGroot and Schervish, 2012)

$$\rho(X,Y) = \frac{COV(X,Y)}{\sigma(X)\sigma(Y)}, \tag{5}$$

  with $\sigma(X)$ and $\sigma(Y)$ as the standard deviation of X and Y, respectively.

As X and Y are uncorrelated if and only if $\rho(X,Y) = 0$ (or $COV(X,Y) = 0$), Y is mean independent of X if and only if $\eta(Y,X) = 0$, where $\eta(Y,X)$ is a less-known association metric: the correlation ratio (Crathorne, 1922, Kruskal, 1958). $\eta(Y,X)$ derives from the variance of $E[Y|X = x]$, which is by definition,

$$Var[E[Y|X = x]] = E[(E[Y|X = x] - E[E[Y|X = x]])^2]. \tag{6}$$

By the law of total probability for expectations (DeGroot and Schervish, 2012, pp. 258), $E[E[Y|X = x]] = E[Y]$ and thus

$$Var[E[Y|X = x]] = E[(E[Y|X = x] - E[Y])^2]. \tag{7}$$

From this variance, the correlation ratio of Y on X is defined as (Kruskal, 1958)

$$\eta(Y,X) = \left(\frac{Var[E[Y|X = x]]}{Var[Y]}\right)^{1/2} = \frac{\sigma[E[Y|X = x]]}{\sigma[Y]}, \tag{8}$$

where $\sigma(...)$ indicates the standard deviation. Notice that $0 \leq \eta(Y,X) \leq 1$ whereas $-1 \leq \rho(X,Y) \leq 1$ (Kruskal, 1958, pp. 816-817). As $\rho(X,Y)$ is a normalized $COV(X,Y)$, $\eta(Y,X)$ is a normalized $Var[E[Y|X = x]]$. Interestingly, the correlation ratio satisfies the following properties (Kruskal, 1958, pp. 816-817):

- $\eta(Y,X) = 1$ if and only if Y is a perfect function of X.
- $|\rho(X,Y)| \leq \eta(Y,X)$ with equality if and only if Y is a linear function of X.
- $0 \leq \eta(Y,X) \leq 1$ (whereas $-1 \leq \rho(X,Y) \leq 1$). As $\rho(X,Y)$ is a normalized $COV(X,Y)$, $\eta(Y,X)$ is a normalized $Var[E[Y|X=x]]$.

It is well-known that (Kolmogorov, 1956, Poirier, 1995):

$$\begin{array}{c} X \text{ and } Y \text{ are independent} \\ \Downarrow \\ Y \text{ is mean independent of } X \quad (\eta(Y,X) = 0) \\ \Downarrow \\ X \text{ and } Y \text{ are uncorrelated} \quad (\rho(X,Y) = 0) \end{array}$$

Proofs of the top to bottom implications have been provided by, e.g., (Kolmogorov, 1956, pp. 60) or (Poirier, 1995, pp. 67). Mean independence implies uncorrelation but the reverse (uncorrelation implies mean independence) is not necessarily true. To see it consider that

$$p(X=x,Y=y) = \begin{cases} 1/2 & \text{if } x=0 \text{ and } y=-1 \\ 1/4 & \text{if } x=-1 \text{ and } y=1 \\ 1/4 & \text{if } x=1 \text{ and } y=1 \end{cases} \quad (9)$$

Thus $E(X) = E(Y) = 0$ and $E(XY) = (1/4)(-1) + (1/2)0 + (1/4)1 = 0$. Therefore $COV(X,Y) = 0$ but Y is not mean independent of X because $E(Y|X=-1) = 1 \neq E(Y|X=0) = -1$. Similarly, independence implies mean independence but the reverse (mean independence implies independence) is not necessarily true (see (Ferrer-i-Cancho et al., 2013a) for a counterexample).

In the next section it will be shown that the correlation ratio is indeed more powerful than a correlation coefficient for testing whether Eq. 2 holds.

3 How to test that $Z = a/X$

3.1 Non-parametric tests

Here we consider three sample correlation statistics: the Pearson correlation $\rho(X,Y)$, the Spearman correlation $\rho_S(X,Y)$ and the correlation ratio $\eta(X,Y)$, and evaluate if they are significantly different from zero using a permutation test (a particular case of randomization test (Sokal and Rohlf, 1995, pp. 803-819)). In a sample of size n, X and Y can be seen as vectors, i.e. $X = \{x_1,...,x_i,...,x_n\}$ and $Y = \{y_1,...,y_i,...,y_n\}$, being (x_i,y_i) the information associated to the i-th element of the sample. The

Table 1: Analysis of the correlation between Y (genome size in bases) and X (chromosome number). N is the number of species (the sample size), $\rho(X,Y)$ is the sample Pearson correlation coefficient, $\rho_S(X,Y)$ is the sample Spearman correlation coefficient and $\eta(Y,X)$ is the sample correlation ratio of Y on X. The p-values were estimated using a Monte Carlo permutation test. The correlation statistics and the corresponding p-values were rounded to leave only two or only one significant digit, respectively.

Taxonomic group	N	$\rho(X,Y)$	p-value	$\rho_S(X,Y)$	p-value	$\eta(Y,X)$	p-value
Fungi	56	0.41	0.002	0.28	0.04	0.74	0.03
Angiosperms	4706	-0.0024	0.9	-0.039	0.008	0.27	0.001
Gymnosperms	170	0.1	0.2	0.32	$3 \cdot 10^{-5}$	0.74	$< 10^{-6}$
Insects	269	0.09	0.1	0.22	0.0003	0.46	0.02
Reptiles	170	0.31	$4 \cdot 10^{-5}$	0.24	0.001	0.48	0.003
Birds	99	-0.029	0.8	-0.033	0.7	0.78	0.001
Mammals	371	0.3	$< 10^{-6}$	0.3	$< 10^{-6}$	0.68	$< 10^{-6}$
Cartilaginous fishes	52	0.014	0.9	-0.13	0.4	0.76	0.1
Jawless fishes	13	-0.76	0.003	-0.74	0.005	0.98	0.05
Ray-finned fishes	647	0.47	$< 10^{-6}$	0.49	$< 10^{-6}$	0.69	$< 10^{-6}$
Amphibians	315	0.13	0.02	0.45	$< 10^{-6}$	0.58	$< 10^{-6}$

p-value of $\rho(X,Y)$ is the probability that X', a random permutation of X, yields $|\rho(X',Y)| \geq |\rho(X,Y)|$. The test is two-sided as both positive and negative values of the sample $\rho(X,Y)$ that are significantly different from 0 are indicative of non-zero correlation (both $\rho(X,Y) = 1$ and $\rho(X,Y) = -1$ are indicative of maximum correlation). The p-value of $\rho_S(X,Y)$ is calculated as that of $\rho(X,Y)$. The p-value of $\eta(Y,X)$ is the probability that X' (a random permutation of X) yields $\eta(Y,X') \geq \eta(Y,X)$. The test is one-sided as only positive values of the sample $\eta(X,Y)$ that are significantly large are indicative of non-zero correlation ($\eta(X,Y)$ is indicative of maximum correlation only when $\eta(X,Y) = 1$). The p-values were estimated with a Monte Carlo method generating $R = 10^6$ uniformly random permutations. In the present article, we adopt a significance level of 0.05. For the analyses of this section, we used the same dataset of Refs. (Hernández-Fernández et al., 2011, Baixeries et al., 2012, 2013, Ferrer-i-Cancho et al., 2013b). As the fact that Y is mean independent of X implies uncorrelation, i.e. $COV(X,Y) = 0$ (or $\rho(X,Y) = 0$), Eq. 2 can be tested by means of the following procedure: if $COV(X,Y)$ (or $\rho(X,Y) = 0$) is significantly different from 0 then reject Eq. 2, otherwise accept it. That procedure can be used to reject Eq. 2 in genomes, being Z the mean chromosome length in bases of a species and X being the number of chromosomes of that species (Hernández-Fernández et al., 2011).

Table 1 summarizes the results of the analysis of the Pearson correlation between X (chromosome number) and Y (genome size in bases) in genomes. A significant correlation is found in six out of eleven taxonomic groups. The test is conservative as it rejects Eq. 2 indirectly by means of a necessary condition for this equation to hold: $COV(X,Y) = 0$. Therefore, the five groups where the Eq. 2 could not be rejected might be false negatives. Pearson correlation is a measure of linearity between variables and has difficulties for capturing non-linear dependencies. A possible improvement is using a more powerful correlation metric such as $\rho_S(X,Y)$, the Spearman rank correlation coefficient, which is a measure of monotonic (linear or non-linear) dependency (Zhou, Tuncali, and Silverman, 2003). The Spearman rank correlation test revealed that the majority of taxonomic groups (nine out of eleven) exhibit a significant correlation between X and Y that is incompatible with Eq. 2 (Table 1). The exceptions are birds and cartilaginous fishes. These findings confirm qualitatively the previous results with similar methods (Hernández-Fernández et al., 2011). Interestingly, there is a more powerful way of testing Eq. 2: testing directly if Y is mean independent of X from its definition, i.e. $E[Y|X = x] = E[Y]$ for any x, which is equivalent to $\eta(X,Y) = 0$ (Kruskal, 1958).

Table 1 also summarizes the results of the analysis of the correlation ratio of Y (genome size in bases) on X (chromosome number) in genomes using a permutation test. $\eta(Y,X)$ was not significantly large in one taxonomic group: cartilaginous fishes. Mean independence and, equivalently, Eq. 2, cannot be rejected for only one

Table 2: Summary of the analysis of the correlation between X (genome size in bases) and Y (chromosome number) in genomes. Three statistics are considered: the sample Pearson correlation coefficient ($\rho(X,Y)$), the sample Spearman correlation coefficient ($\rho_S(X,Y)$) and the sample correlation ratio ($\eta(Y,X)$). 'Yes' indicates that the corresponding correlation test indicates a significant correlation at a significance level of 0.05.

Taxonomic group	$\rho(X,Y)$	$\rho_S(X,Y)$	$\eta(Y,X)$
Fungi	Yes	Yes	Yes
Angiosperms		Yes	Yes
Gymnosperms		Yes	Yes
Insects		Yes	Yes
Reptiles	Yes	Yes	Yes
Birds			Yes
Mammals	Yes	Yes	Yes
Cartilaginous fishes			
Jawless fishes	Yes	Yes	Yes
Ray-finned fishes	Yes	Yes	Yes
Amphibians	Yes	Yes	Yes

out of eleven groups.

Table 2 summarizes the results of the Pearson, Spearman and correlation ration test. The number of taxonomic groups for which a test rejects Eq. 2 is 6, 9, and 10, respectively. The qualitative summary for $\rho(X,Y)$ and $\eta(Y,X)$ comes, respectively, from Table 1. The power of the correlation ratio test can be easily explained by the fact that the theoretical correlation ratio can only be zero when mean independence fails. Table 2 suggests that Spearman rank correlation has an intermediate power between Pearson correlation and correlation ratio.

3.2 A parametric test

The hypothesis of $Z = a/X$ has been accepted with the only support that the fit of $Z = aX^b$ yields $b \approx -1$ (Solé, 2010). This procedure is very prone to type II error (accepting a false null hypothesis) as it needs that $Z = aX^b$ holds first (Baixeries et al., 2013, Ferrer-i-Cancho et al., 2013b). Our analysis shows that for $Z = a/X$ to hold, it is not only necessary that $b \approx -1$ is retrieved but also $a \approx \mu[Y]$, where $\mu[Y]$ is the mean of Y, an estimate of $E[Y]$ (recall Theorem 2.1; see also

(Ferrer-i-Cancho et al., 2013a)). Even if $a \approx \mu[Y]$ and $b \approx -1$, type II errors are not excluded and minimizing them needs evidence that $Z = aX^b$ is well-supported by data.

4 Discussion

We have argued that claiming that Z scales with X following a very particular form of Menzerath-Altmann law, i.e. Eq. 2, is indeed equivalent to claiming that $E[Z|X] = E[Y]/X$, which is indeed equivalent to claiming that Y is mean independent of X. We have also presented a new correlation ratio test revealing that Eq. 2 could only hold in cartilaginous fishes. Therefore, the trivial scaling defined by Eq. 2 is the exception, not the rule.

The random breakage model where X and Y are independent and uniformly distributed (Solé, 2010) fails to fit the majority of taxonomic groups because independence is a particular case of mean independence (Ferrer-i-Cancho et al., 2013a) and mean independence fails in at least ten out of eleven taxonomic groups (Table 2). Furthermore, it has been argued that independence between X and Y is problematic as it can lead to organisms with empty chromosomes or empty chromosome parts (Baixeries et al., 2012). Interestingly, Y does not need to be the size of genome in bases. It could be the size in units between the base and the chromosome.

The problem of empty components also concerns mean independence. The condition for not expecting empty chromosomes for a given x (a concrete value of X) is

$$\begin{aligned} E[Z|X=x] &\geq 1 \\ \frac{1}{x}E[Y|X=x] &\geq 1. \end{aligned} \tag{10}$$

For that x, the condition in Eq. 10 becomes $E[Y] \geq x$ when Y is mean independent of X as $E[Y|X=x] = E[Y]$ in that case. Thus, under mean independence, empty chromosomes are expected in an organism of x chromosomes if $E[Y] < x$. Notice that expecting that Eq. 10 holds on average for any x, leads to

$$\begin{aligned} E[E[Y|X=x]] &\geq E[X] \\ E[Y] &\geq E[X] \end{aligned} \tag{11}$$

thanks to the law of total probability for expectations (DeGroot and Schervish, 2012, pp. 258). The restrictions defined by Eqs. 10 and 11 are perhaps very simple but Baixeries et al. (2012) considered more elaborated constraints for the viability of an organism based upon the parts making an ideal chromosome: a centromere, two

telomeres and a couple of intermediate regions. Those viability constraints lead to a deviation from Menzerath-Altman law with $b = -1$ and $c = 0$ (see Fig. 3 of Baixeries et al. (2012)), which Theorem 2.1 allows one to interpret unequivocally as a departure from mean independence. Therefore, the viability and well-formedness of chromosomes is not compatible with mean independence either. The negative correlation between Z and X, known as Menzerath's law, defies a trivial explanation in genomes (Ferrer-i-Cancho and Forns, 2009, Wilde and Schwibbe, 1989). Claiming that the scaling defined by Eq. 2 is inevitable (Solé, 2010) is equivalent to claiming that Y must be mean independent of X in any circumstance, a very strong requirement for real language, music and genomes. Eq. 2 should be regarded as a baseline instead of a fact for the relationship between chromosome number and mean chromosome length. Rather than signs of complexity, certain power laws (the value of the exponent can be crucial) might indicate the control that must be considered before any claim of "sufficient biological complexity" can be made.

Acknowledgements

We are grateful to P. Delicado, R. Gavaldà and E. Pons for their valuable mathematical insights. We owe the counterexample showing that uncorrelation does not imply mean independence to P. Delicado. We are also grateful to G. Bel-Enguix and N. Forns for helpful discussions. This work was supported by the grant *Iniciació i reincorporació a la recerca* from the Universitat Politècnica de Catalunya, the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation and the grant 2/0038/12 from the VEGA funding agency (JM).

References

Altmann, G. (1980): "Prolegomena to Menzerath's law," *Glottometrika 2*, 2, 1–10.

Baixeries, J., A. Hernández-Fernández, and R. Ferrer-i-Cancho (2012): "Random models of Menzerath-Altmann law in genomes," *Biosystems*, 107, 167âĂŘ173.

Baixeries, J., A. Hernández-Fernández, N. Forns, and R. Ferrer-i-Cancho (2013): "The parameters of Menzerath-Altmann law in genomes," *Journal of Quantitative Linguistics*, 20, 94–104.

Boroda, M. G. and G. Altmann (1991): "Menzerath's law in musical texts," *Musikometrica*, 3, 1–13.

Cameron, A., , and P. K. Trivedi (2009): *Microeconometrics: Methods and Applications*, Cambridge: Cambridge University Press.

Cramer, I. M. (2005): "The parameters of the Altmann-Menzerath law," *Journal of Quantitative Linguistics*, 12, 41–52.

Crathorne, A. R. (1922): "Calculation of the correlation ratio," *Journal of the American Statistical Association*, 18, 394–396.

DeGroot, M. H. and M. J. Schervish (2012): *Probability and statistics*, Boston: Wiley, 4th edition.

Ferrer-i-Cancho, R., J. Baixeries, and A. Hernández-Fernández (2013a): "Erratum to "Random models of Menzerath-Altmann law in genomes" (BioSystems 107 (3), 167âĂŞ173)," *Biosystems*, 111, 216–217.

Ferrer-i-Cancho, R. and N. Forns (2009): "The self-organization of genomes," *Complexity*, 15, 34–36.

Ferrer-i-Cancho, R., N. Forns, A. Hernández-Fernández, G. Bel-Enguix, and J. Baixeries (2013b): "The challenges of statistical patterns of language: the case of Menzerath's law in genomes," *Complexity*, 18, 11–17.

Hernández-Fernández, A., J. Baixeries, N. Forns, and R. Ferrer-i-Cancho (2011): "Size of the whole versus number of parts in genomes," *Entropy*, 13, 1465–1480.

Khanin, R. and E. Wit (2006): "How scale-free are biological networks," *Journal of Computational Biology*, 13, 810–818.

Kolmogorov, A. N. (1956): *Foundations of the theory of probability*, New York: Chelsea Publising Company, 2nd edition.

Kruskal, W. H. (1958): "Ordinal measures of association," *Journal of the American Statistical Association*, 53, 814–861.

Li, W. (1992): "Random texts exhibit Zipf's-law-like word frequency distribution," *IEEE T. Inform. Theory*, 38, 1842–1845.

Li, W. (2012): "Menzerath's law at the gene-exon level in the human genome," *Complexity*, 17, 49–53.

May, R. M. and M. P. H. Stumpf (2000): "Species-area relations in tropical forests," *Science*, 290, 2084–2086.

McCowan, B., L. R. Doyle, J. M. Jenkins, and S. F. Hanser (2005): "The appropriate use of Zipf's law in animal communication studies," *Anim. Behav.*, 69, F1–F7.

Menzerath, P. (1954): *Die Architektonik des deutschen Wortschatzes*, Bonn: Dümmler.

Miller, G. A. (1968): "Introduction," in *The Psycho-Biology of Language: an Introduction to Dynamic Psychology (by G. K. Zipf)*, Cambridge, MA, USA: MIT Press, v–x.

Miller, G. A. and N. Chomsky (1963): "Finitary models of language users," in R. D. Luce, R. Bush, and E. Galanter, eds., *Handbook of Mathematical Psychology*, volume 2, New York: Wiley, 419–491.

Poirier, D. J. (1995): *Intermediate Statistics and Econometrics: A Comparative Approach*, Cambridge: MIT Press.

Ritz, C. and J. C. Streibig (2008): *Nonlinear regression with R*, New York: Springer.

Sokal, R. R. and F. J. Rohlf (1995): *Biometry. The principles and practice of statistics in biological research*, New York: W. H. Freeman and Co., 3rd edition.

Solé, R. V. (2010): "Genome size, self-organization and DNA's dark matter," *Complexity*, 16, 20–23.

Stumpf, M., P. Ingram, I. Nouvel, and C. Wiuf (2005): "Statistical model selection methods applied to biological network data," *Trans. Comp. Syst. Biol.*, 3, 65âĂŞ77.

Stumpf, M. P. H. and P. J. Ingram (2005): "Probability models for degree distributions of protein interaction networks," *Europhysics Letters*, 71, 152.

Stumpf, M. P. H. and M. A. Porter (2012): "Critical truths about power laws," *Science*, 335, 665–666.

Suzuki, R., P. L. Tyack, and J. Buck (2005): "The use of Zipf's law in animal communication analysis," *Anim. Behav.*, 69, 9–17.

Tanaka, R., T.-M. Yi, and J. Doyle (2005): "Some protein interaction data do not exhibit power law statistics," *{FEBS} Letters*, 579, 5140 – 5144.

Teupenhayn, R. and G. Altmann (1984): "Clause length and Menzerath's law." *Glottometrika*, 6, 127–138.

Tjørve, E. (2003): "Shapes and functions of species-area curves: a review of possible models," *Journal of Biogeography*, 30, 823–832.

Wilde, J. and H. Schwibbe (1989): "Organizationsformen von Erbinformation Im Hinblick auf die Menzerathsche Regel," in G. Altmann and M. H. Schwibbe, eds., *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, Hildesheim: Olms, 92–107.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.

Zhou, K., K. Tuncali, and S. G. Silverman (2003): "Correlation and simpler linear regression," *Radiology*, 227, 617–628.

2.7. La ley de brevedad

La ley de brevedad de Zipf, o simplemente ley de brevedad, es la regularidad estadística que da cuenta de que en las lenguas las palabras más frecuentes tiendan a ser más cortas (Zipf, 1949; Strauss *et al.*, 2007; Jayaram y Vydia, 2009). En la tabla 2 se recoge por ejemplo la comprobación estadística de la ley que se realizó en Ferrer-i-Cancho y Hernández-Fernández (2013), partiendo de corpus de la base de datos *Childes* (MacWhinney, 2000). La ley de brevedad puede generalizarse y extenderse a todo sistema de comunicación, y definirse entonces como la tendencia a que los elementos más frecuentes del sistema comunicativo sean más breves o cortos (Ferrer-i-Cancho y Hernández-Fernández, 2013, en esta tesis, apartado siguiente), extendiéndose por tanto a la comunicación animal (Ferrer-i-Cancho y Lusseau, 2009; Semple *et al.*, 2010).

	Tokens	*Types*	r_s	p
Ruso	54104	7908	-0.204	$< 10^{-9}$
Español	980797	27479	-0.229	$< 10^{-9}$
Croata	298038	15381	-0.269	$< 10^{-9}$
Griego	52158	4203	-0.191	$< 10^{-9}$
Sueco	511191	19164	-0.161	$< 10^{-9}$
Inglés americano	2674937	24101	-0.227	$< 10^{-9}$
Indonesio	2417587	30461	-0.227	$< 10^{-9}$

Tabla 2.3: Correlación entre la frecuencia y la longitud de las palabras en algunas lenguas (Ferrer-i-Cancho y Hernández-Fernández, 2013, apartado siguiente de esta tesis). El tamaño de la muestra se indica en número total de palabras (*tokens*) y número de palabras diferentes (*types*). Se añade además el factor de correlación de Spearman, r_s, entre la frecuencia de cada palabra y su longitud (en número de letras) y el p-valor correspondiente. El valor negativo de r_s demuestra que la ley de brevedad se sigue para estas lenguas, es decir, a menor frecuencia las palabras tienden a ser más cortas.

Cuando la ley de brevedad no fue encontrada en el sistema de comunicación de dos pequeños primates (Bezerra *et al.*, 2011), decidimos reanalizar aquellos datos y explorar otras especies (Ferrer-i-Cancho y Hernández-Fernández, 2013) y lenguas (tabla 2.3) como controles. Para empezar, se revisaron las unidades de estudio: lo habitual era trabajar con duraciones medias de las vocalizaciones de los primates (Bezerra *et al.*, 2011; Semple *et al.*, 2010), de manera que si f es la frecuencia de un tipo de vocalización y D es la duración total de las vocalizaciones de un tipo, entonces las correlaciones se realizaban entre f y $\langle d \rangle = D/f$. El hecho de que Bezerra y colaboradores (2011) no encontrasen la ley de brevedad al analizar en global todo el repertorio de titíes ni de uakaris, repertorios en los que se incluían por cierto las llamadas a larga distancia, no fue óbice para no encontrar la ley de brevedad en un subconjunto del repertorio y, es más: en el subconjunto de vocalizaciones en donde no

se halló la ley de brevedad se encontraban todas las llamadas de larga distancia (Ferrer-i-Cancho y Hernández-Fernández, 2013). En las llamadas de larga distancia se espera menor presión para que se cumpla la ley de brevedad, ya que aumenta la energía necesaria para que la llamada sea comunicativamente efectiva.

El lenguaje y las vocalizaciones de primates no humanos tradicionalmente se han distinguido cualitativamente respecto a la semántica y en la manera en la que los individuos adquieren y emplean las señales acústicas (Fitch, 2000). No obstante, en ambos casos rigen las leyes de la acústica y, por tanto, no todo está en la brevedad o no de una señal (Ferrer y Cancho *et al.*, 2013): si atendemos a la energía de las ondas mecánicas, además de la duración hay que considerar la intensidad de la onda y su frecuencia (figura 2.6). Es lo que hacemos cuando gritamos para que nos oiga alguien que está lejos y pretendemos tener éxito en la comunicación. Transcribiríamos: "¡Ramoooooon!". Aumentamos la intensidad de la señal (y por tanto los decibelios), a la vez que alargamos la vocalización –donde fisiológicamente se puede, en las vocales, generalmente– para disminuir el efecto de la atenuación de la señal en el aire.

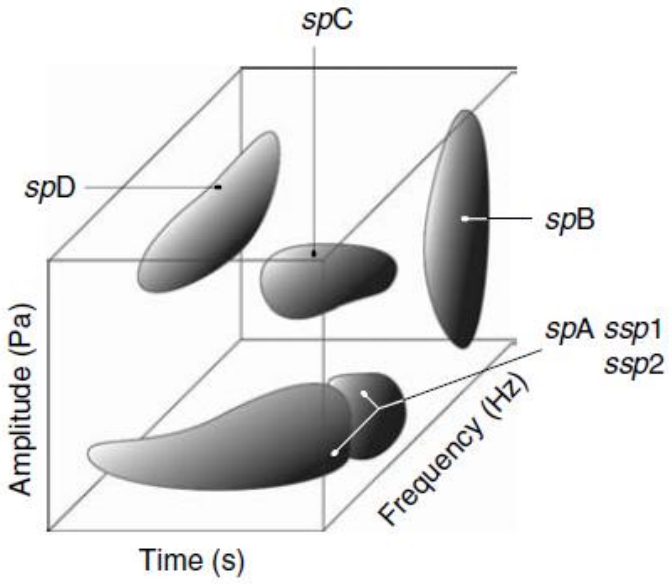


Figura 2.6: Ejemplo de espacio acústico tridimensional (tiempo, frecuencia y amplitud), en el que se distribuyen diferentes especies (spA, spB, spC, spD) y subespecies de spA (ssp1, ssp2). Extraída de Sueur (2006, p.210).

Es lógico que lo que pasa cuando queremos comunicarnos a larga distancia sea justo lo contrario que lo que sucede con las palabras de alta frecuencia, que emitimos habitualmente a corta distancia, y que suelen ser más breves acústicamente, no solo en

cuanto al número de letras de la cadena escrita (Tomaschek *et al*., 2013). Las palabras de alta frecuencia son más cortas pero también poseen vocales más cortas, al menos para el inglés (Aylett y Turk, 2006).

De hecho, todas las especies con comunicación acústica o vibracional pueden analizarse en un espacio de fases en el que se incluyen las tres dimensiones mecánicas (amplitud, tiempo y frecuencia), de manera que cada especie suele ocupar un volumen diferenciado (figura 2.6) típicamente dentro de este espacio, afectada por múltiples factores intraespecíficos e interespecíficos (para una revisión Sueur, 2006, p.210-211). Por tanto, la brevedad temporal es únicamente una de las posibilidades de economía energética en un problema mucho más complejo (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013).

En este espacio de fases acústico, los humanos, como otros primates, son capaces de extraer estadísticas secuenciales de la señal acústica (aprendizaje estadístico) de modo que pueden emplear esta información para decidir qué secuencias pueden formar potenciales palabras: se computan las probabilidades de coocurrencia de elementos y después las probabilidades transicionales entre elementos adyacentes, tanto para la detección de fonemas como para la detección de palabras (Saffran *et al*., 1996; Hauser *et al*., 2001). Las coincidencias en las estructuras acústicas de humanos y primates no humanos han quedado más que demostradas (Zimmermann *et al.*, 1995), así como los paralelismos fisiológicos y neuronales en la percepción auditiva (Zoloth y Green, 1979), y la lateralización y asimetría hemisférica (Cantalupo y Hopkins, 2001).

Tanto en humanos como en otras especies que se comunican acústicamente no conviene gastar energía innecesariamente, y la brevedad es buena tanto para el emisor como para el receptor, siempre que la comunicación tenga éxito (Ferrer-i-Cancho y Solé, 2003), tras considerar los múltiples factores contextuales (presencia de depredadores, apareamiento,…) que pueden rodear la comunicación y que influyen tanto a emisor como receptor (Sueur, 2006; Seyfarth y Cheney, 2003).

Bajo el paraguas de las leyes evolutivas y de la física queda más que justificada la búsqueda de universales comunicativos, fundamentados empíricamente en las vocalizaciones y otro tipo de modalidades comunicativas (gestuales, por ejemplo) de humanos y primates no humanos, como es el caso de la ley de brevedad, cuyas excepciones probablemente encontramos en las llamadas a larga distancia, cuando el éxito en la comunicación se antepone al gasto energético siendo necesaria la redundancia (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013; Ay *et al.*, 2007), o

como sucede también en entornos en los que el ruido fuerza a aumentar la energía de las emisiones (efecto Lombard), tanto en humanos como en otras especies (Zöllinger y Brumm, 2011; Brumm y Zöllinger, 2011). Las evidencias de relaciones entre la duración, las frecuencias de emisión, y los contextos de emisión son claras (figura 2.7), aunque todavía deben estudiarse con más profundidad (Sueur, 2006).

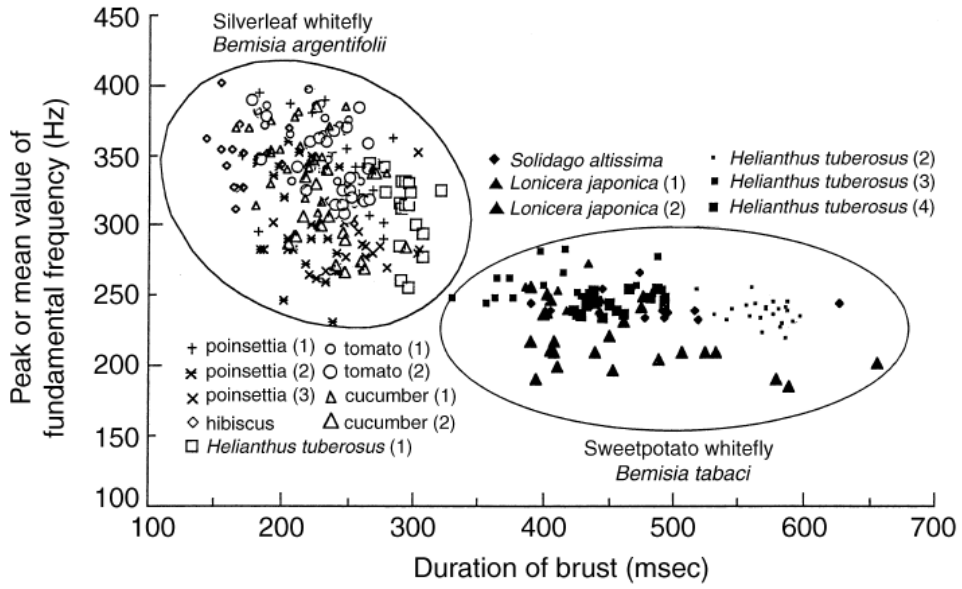


Figura 2.7: Ejemplo de diagrama de dispersión de la duración respecto la frecuencia fundamental de las emisiones de Bemisia Tabaci y Bemisia Argentifolii en diversos contextos (plantas en las que habitan) y poblaciones. Extraída de Sueur (2006).

En el apartado siguiente se incluye el artículo (Ferrer-i-Cancho y Hernández-Fernández, 2013) en el que se contrastan y argumentan éstos y otros motivos por los que la ley de brevedad no fue recuperada en las vocalizaciones de dos pequeños primates por Bezerra *et al.* (2011), y que se explicaron posteriormente bajo el paradigma del principio de compresión (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013; Cover y Thomas, 2006). Como se verá, los argumentos contra la ley de brevedad de Bezerra *et al.* (2011) se pueden refutar en parte por incluir llamadas de larga distancia: la calidad de los ajustes estadísticos es fundamental para obtener conclusiones correctas, más allá de la navaja de Ockham (Ferrer-i-Cancho y Hernández-Fernández, 2013), que aplicamos en la sencillez aproximativa al problema (contabilizar tipos de vocalizaciones y sus duraciones) pero no en el análisis estadístico posterior y en el tratamiento de los datos. La ley de brevedad, como se verá más adelante, no es fruto del azar ni una trivialidad empírica: se fundamenta en el principio de compresión (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013).

2.8. Ramon Ferrer-i-Cancho y Antoni Hernández-Fernández (2013). *The Failure of the Law of Brevity in Two New World Primates. Statistical Caveats.*

Aquí se incluye el artículo:

- Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2013). The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, 4 (1), 45-55.

Akademie Verlag

GLOTTOTHEORY

International Journal
of Theoretical Linguistics

Volume 4 • 2013 • Number 1

ISSN 1337-7892

The Failure of the Law of Brevity in Two New World Primates. Statistical Caveats.[1]

Ramon Ferrer-i-Cancho, Antoni Hernández-Fernández

Abstract: *Parallels of Zipf's law of brevity, the tendency of more frequent words to be shorter, have been found in bottlenose dolphins and Formosan macaques. Although these findings suggest that behavioral repertoires are shaped by a general principle of compression, common marmosets and golden-backed uakaris do not exhibit the law. However, we argue that the law may be impossible or difficult to detect statistically in a given species if the repertoire is too small, a problem that could be affecting golden backed uakaris, and show that the law is present in a subset of the repertoire of common marmosets. We suggest that the visibility of the law will depend on the subset of the repertoire under consideration or the repertoire size.*

1 Introduction

World languages exhibit many statistical patterns that qualify as candidates for linguistic universals: they hold in any language where they have been tested. A popular example is Zipf's law for word frequencies, namely that the probability of the r-th most frequent word in a text is approximately $p(r) \sim r^{-a}$, where a is the exponent of the law (Zipf 1949). Languages show many other statistical regularities such as the law of abbreviation, i.e. the tendency of more frequent words to be shorter (Zipf 1949, pp. 63; Strauss et al. 2007), a law of meaning distribution indicating that more frequent words tend to have more meanings (Zipf 1949, pp. 28–31), or Menzerath-Altmann law stating that the longer a construct (e.g., a sentence), the shorter its components or constituents (e.g., the shorter the clauses) (Altmann 1980; Teupenhayn & Altmann 1984; Boroda & Altmann 1991). Statistical patterns of language defy the claim that linguistic universals are a myth (Evans & Levinson, 2009).

Over the last decade, many parallels between human language and the behavior of other species have been established by means of statistical 'laws' of language. Zipf's law and a parallel of the law of meaning distribution have been found in dolphin whistles (McCowan et al. 1999; Ferrer-i-Cancho & McCowan 2009). Parallels of the law of abbreviation have been reported in chick-a-dee calls (Hailman et al. 1987), vocalizations of Formosan macaques (Semple et al. 2010) and surface behavioral patterns of dolphins (Fer-

1 This work was supported by the grant BASMATI (TIN2011-27479-C04-03) from the Spanish Ministry of Science and Innovation. We thank D. Lusseau for the opportunity to reanalyze the dolphin data (Ferrer-i-Cancho & Lusseau 2009) and S. L. Vehrencamp for making us aware of the research in ravens by Conner (1985). We are grateful to R. Dale, D. Lusseau, L. Doyle and B. Elvevåg for helpful comments.

rer-i-Cancho & Lusseau 2009). Beyond the realm of animal behavior, qualitative agreement with Menzerath-Altmann law has been found in genomes at different levels of organization (Ferrer-i-Cancho & Forns 2009; Hernández-Fernández et al. 2011; Li 2012). The focus of the present article is shedding light on recent research on the law of brevity or abbreviation in two New World primates (Bezerra et al. 2011).

Zipf's law of brevity, i.e. the tendency of more frequent words to be shorter (Zipf 1936; Zipf 1949; Straus et al. 2007), can be generalized as the tendency of more frequent elements to be shorter or smaller, following an abstraction akin to the one that led to the current formulation of Menzerath-Altmann law in terms of constructs and constituents (Altmann 1980). Compression (Salomon 2007; Dawkins 1976), the information theoretic principle of assigning shorter codes to more frequent elements, is argued to underlie the generalized law of brevity across species (Zipf 1949; Ferrer-i-Cancho & Lusseau 2009), and is independent from modality or whether the behavior is communicative or not. The law is not a perfect rule (Zipf 1949; Strauss et al., 2007; Ferrer-i-Cancho & Lusseau 2009; Semple et al. 2010) and its presence does not imply that compression is the only principle involved (Bezerra et al. 2011).

Recently, it has been argued that the failure to find the law in the vocalizations of two New World primates: common marmosets and golden-backed uakaris, means that the law is not widely applicable in animal communication (Bezerra et al. 2011). In contrast, we propose here that the law of brevity has not surfaced in these two primates due to two not necessarily exclusive reasons: insufficient sampling and mixing of call types with distinct compression pressures.

The organization of the remainder of the article is as follows. Section 2 presents a summary of the presentence of the law of brevity in humans and other species unifying methodologies. Section 3 unravels some limits of the methods used to detect the presence of the law in a given species. Section 4 reviews important statistical issues for research on the law of brevity across species.

2 The Law of Brevity in Various Species

Here, we aim to provide a methodologically homogenous summary of the statistical evidence about the law of brevity in various species that will be the basis of a meta-analysis in the next section.

2.1 Materials and Methods

2.1.1 Materials

The frequency and duration data for common marmosets and golden-backed uakaris that is used in this article comes from the electronic supplementary material of Bezerra et al. (2011), which, in both cases, covers a subset of the whole repertoire of those species (see Table 1). The frequency and duration data for common ravens comes from the main text of Conner (1985), who did not study the law of brevity in his article. Unfortunately, the full

data from the pioneering research on the law of brevity in chick-a-dee calls (Hailman et al. 1987; Hailman et al. 1985; Ficken et al. 1978) is not available for the kind of reanalysis intended here.

In humans, the law of brevity has been typically studied in written language (e.g., Zipf 1949; Strauss et al. 2007). The use of written language could give a prior advantage to human language in terms of the degree of adherence to the law of brevity. In order to study the law of brevity in oral human language and to allow for a fairer comparison with animal data, all available data (15 April, 2011) in the Childes Database (MacWhinney 2000) for the following languages was used: Russian, Croatian, Greek, Swedish and Indonesian. A selection of corpus of American English (Bloom, Bates, Brown, HSLLD and MacWhinney) and Spanish (Aguirre, BecaCESNo, DiezItza, Irene, OreaPine, Ornat and SerraSole) had to be used to keep the total number of words close to the other languages examined. All the corpora are freely available at http://childes.psy.cmu.edu.

For the analysis of each language, all the speakers were considered (children and adults, regardless of their role). This condition was used to approximate the mixing of individuals of the majority of animal studies reviewed here. For simplicity, the units of the analysis were word forms.

Table 1: Summary of the results of the exploration of the law of brevity in various species. r_s and p were rounded to leave, respectively, two and one significant digits.

Species	n	Law of brevity	r_s	p	Reference
Golden-backed uakaris	7 (9)	No	-0.36	0.4	Bezerra et al. (2011)
Common marmosets	12 (17)	No	0.056	0.9	Bezerra et al. (2011)
Common ravens	18	No	-0.060	0.8	This article.
Dolphins	31	Yes	-0.51	0.003	Ferrer-i-Cancho & Lusseau (2009)
Formosan macaques	35	Yes	-0.43	0.01	Semple et al. (2010)
Humans	>> 35	Yes	(-0.26, -0.17) approx.	<< 0.001	Table 2 of this article. See also: Zipf (1935), Zipf (1949), Strauss et al. (2007).

Table 2: The correlation between frequency and length in oral human language.

	Tokens	Types	r_s	p
Russian	54,104	7,908	-0.204	$< 10^{-9}$
Spanish	980,797	27,479	-0.229	$< 10^{-9}$
Croatian	298,038	15,381	-0.269	$< 10^{-9}$
Greek	52,158	4,203	-0.191	$< 10^{-9}$
Swedish	511,191	19,164	-0.161	$< 10^{-9}$
US English	2,674,937	24,101	-0.227	$< 10^{-9}$
Indonesian	2,417,587	30,461	-0.227	$< 10^{-9}$

2.1.2 Methods

The results of the correlation between frequency and length for dolphins in Table 1 were obtained by a reanalysis of the data of Ferrer-i-Cancho & Lusseau (2009), where a Pearson correlation test was used, through a Spearman rank correlation test.

Notice that the Spearman rank correlation test is a very powerful tool to study and compare the law of brevity from heterogeneous sources. For instance, Semple et al. (2010) studied the law of brevity as a relationship between mean call type duration and frequency while in our analysis of human data (Table 2) we are studying the relationship between word length in letters and frequency. The results of the Spearman rank correlation test on human data will not change if it is assumed that mean word duration is a strictly monotonically increasing function of word length. In the end, the relevant unit of measurement under a compression or coding efficiency hypothesis is not duration or length but the energetic cost of the word. The results of the Spearman rank correlation tests will not change if that cost is a strictly monotonically increasing function of duration in seconds or length in discrete units.

2.2 Results

Table 1 summarizes the results of the study of the law of brevity in various species using the same notation as in Bezerra et al. (2011): n for the repertoire size considered in the references of the last column, r_s for the Spearman rank correlation between frequency and duration/length and p for the p-value. In parenthesis, the size of the whole repertoire according to Bezerra et al. (2011) is also shown. Table 2 summarizes the results of the study of the law for word frequencies and their length (in letters) in various languages.

3 The Pressure for Brevity Can Be Hidden

A key issue in the quest for exceptionless universals (Evans & Levinson 2009) or universal principles (Köhler 2005; Zipf 1949) is understanding the limits and other subtleties of the statistical methods that are being used.

3.1 Pressure for Brevity without Statistical Significance

Table 1 shows that the law of brevity has not been found in two New World primates (golden-backed uakaris and common marmosets) and common ravens. The law has therefore not been found in 3 out of 6 cases thus far. The small repertoire sizes of these three species may have caused type II statistical errors when trying to reject the null hypothesis that frequency and duration are unrelated. We consider the probability that these three cases coincide with the three cases that have the smallest repertoire, as it is the case according to Table 1, simply by chance. This probability is $q = (3!3!)/6! = 0.05$. Thus the null hypothesis that the coincidence is accidental can be rejected at a significance level of 0.05.

Notice that our meta-analysis for type II errors is conservative: we consider that all the languages where the law of brevity has been reported (e. g., Zipf 1935; Strauss et al. 2007) have collapsed into the category "humans" in Table 1. If we considered that each of the L languages where the law has been found must contribute separately, the p-value would become $q = (3!(2 + L)!)/(5 + L)!$, which cannot exceed the p-value of our initial calculation, i.e. 0.05, for $L \geq 1$. The fact that $q = 6/((5 + L)(4 + L)(3 + L))$ indicates that q drops as L grows. For simplicity, let us consider only the positive reports of the law in Table 2, which gives $L = 7$. Then we would have $q = 6/(12 \cdot 11 \cdot 10) \approx 0.0045$, namely a stronger support for the hypothesis of type II errors that could, nevertheless, be unfairly biased by human languages.

Table 3 shows different values of n^*, the minimum repertoire size that is needed to achieve significance in correlation tests for different significance levels a (see Appendix for the precise mathematical argument) depending on whether the test is one-tailed ($k = 1$) or two-tailed ($k = 2$). Absence of ties was assumed for calculating Table 3. Knowing that the law of brevity is not a perfect rule, we suggest that the repertoire size of golden-backed uakaris, 7 calls (Bezerra et al. 2011), is dangerously close to 5, the value of n^* for a two-tailed correlation test at a significance level of 0.05 (Table 3). There are no ties in the datasets of the two New World primates that we reanalyze from Bezerra et al. (2011).

Table 3: The minimum sample size for significance (n^*) versus the significance level (a).

	n^*	
a	$k = 1$	$k = 2$
0.05	4	5
0.01	5	6
0.001	7	7
0.0001	8	8

3.2 The Law of Brevity Hidden in a Subset of the Repertoire

The fact that the law of brevity has not been found for the whole repertoire does not a priori imply that it cannot be found for a subset of the repertoire. Simpson's paradox indicates that statistically significant correlations may emerge when the sample is partitioned according to a certain criterion (DeGroot 1989).

Notice that Semple et al. (2010) and Bezerra et al. (2011) study the correlation between the frequency of occurrence f of a call type and its mean duration $\langle d \rangle = D/f$, where D is the sum of the durations of the occurrences. Since $\langle d \rangle = D/f$, it is mathematically convenient to consider the dependency between D and f (Hernández-Fernández et al. 2011; Li 2012), which is shown in Fig. 1 for common marmosets, to investigate the possibility of different compression pressures. Error bars indicate standard errors. The standard error of D was inferred from the standard error of $\langle d \rangle = D/f$ (that is available in the supplementary online information of Bezerra et al. (2011)) through $\sigma(D) = f\,\sigma(\langle d \rangle)$. For this species, Fig. 1 A suggests two different clusters of call types: a low D cluster ending at the 6-th call type with the smallest D and a high D cluster beginning at the 7-th call type with the smallest D. The

cluster boundaries can be determined quantitatively. If D_i is defined as the i-th call type with the smallest total duration, the boundary between clusters is defined by the value of i that maximizes the difference in order of magnitude between consecutive total durations, i.e. $\Delta_i = \log(D_i/D_i\text{-}1)$, where log is a natural logarithm. Table 4 shows that D_7 is maximum. Δ_i was rounded to leave only two decimal digits.

A negative correlation between f and $\langle d \rangle$ is found for the low D cluster ($n = 6$, $r_s = -0.886$, $p = 0.019$) but not for the high D cluster ($n = 6$, $r_s = -0.086$, $p = 0.872$). The low D cluster contains the call types "tisik", "very brief whistle", "chatter", "submissive squeal", "egg" and "tse". Unfortunately, the same analysis cannot be extended to golden-backed uakaris because the repertoire considered by Bezerra et al. (2011) is too small. The possibility that the law of brevity has emerged in the low D cluster for a trivial reason will be examined further.

Table 4: The difference in order of magnitude between the i-th and the $(i-1)$-th signal with the smallest D.

i	Δ_i	i-th signal with the smallest D
1	-	Tisik
2	0.25	Very brief whistle
3	0.25	Chatter
4	0.06	Submissive squeal
5	1.03	Egg
6	0.08	Tse
7	1.35	Brief phee call level 3
8	0.18	Brief phee call level 2
9	0.47	Brief phee call level 1
10	1.07	Long phee call
11	0.34	Twitter
12	0.50	Trill

A negative correlation between frequency and duration could be obtained simply when the expectation of D given f is constant, e.g., D and f are independent (Hernández-Fernández et al. 2011; Li 2012), giving $\langle d \rangle = a/f$, where a is a constant, which means a very strong correlation between $\langle d \rangle$ and f. A constant expectation for D given f has been excluded as an explanation of the law of brevity in the vocalizations of Formosan macaques by showing that D and f are indeed correlated (Semple et al. 2012). The problem of constant expectation does not concern the presence of the law in dolphin surface behavioral patterns (Ferrer-i-Cancho & Lusseau 2009) or human word lengths (Table 2) because in those cases the law is studied between fixed length λ (in letters for human words) and frequency and then $D = \lambda f$, i.e. D and f are of course, correlated. Concerning Formosan macaques, a dependency between D and f in the low D cluster could not be supported by a two-sided Spearman rank correlation test ($n = 6$, $r = 0.6$, p-value = 0.2) but could

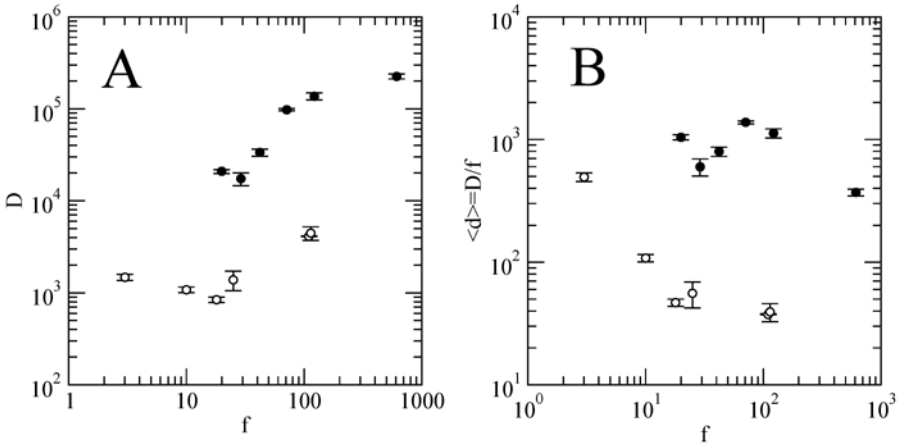


Figure 1: A. *D* versus *f* in common marmosets. B. <*d*> versus *f* in the same dataset. In both subfigures, white and black circles are used, respectively, for the cluster of low *D* call types and that of high *D*.

be sustained by a two-sided Pearson rank correlation test ($n = 6$, $r = 0.97$, p-value = 0.04) at a significance level of 0.05.

The ability of the Pearson correlation statistic to unravel a significant correlation between *D* and *f* in the low *D* cluster may suggest that the dependency between both is trivially linear, i.e. $D = af + b$, where a is the slope and b is the intercept. However, such a linear dependency is unlikely. To see it, notice first that, by the definition of total duration *D*, $b = 0$ as *D* must be zero when $f = 0$. Second, if $b = 0$ then $\langle d \rangle = D/f = a$ but a significant correlation between <*d*> and *f* has already been shown.

4 Discussion

Concerning the failure to find the law of brevity in two New World primates, it was said that "brevity is likely to be vital in memorizing the components of a large vocabulary" as in the case of human languages (Bezerra et al. 2011). However, if a very large repertoire size was crucial for an agreement with the law, one would expect not to find the law of brevity in the small repertoires of macaques and dolphins (Semple et al. 2010; Ferrer-i-Cancho & Lusseau 2009), which are various orders of magnitude smaller than those of human languages (Table 1) or even within a very small subset of the repertoire of common marmosets. Coding efficiency is important even in small repertoires.

The supported hypothesis of false negatives (type II errors) for the law of brevity has various implications for animal behavior research. It should be checked if the vocalization types that were excluded from the analysis of the two New World primates (see Table 1), 5 out of 17 in marmosets and 2 out of 9 in uakaris (Bezerra et al. 2011) would change the results on the law of brevity. Besides, Table 3 shows that the minimum repertoire size needed for significance, n^*, is smaller for one-tailed tests. Indeed, to test the hypothesis test of the law of brevity we only need to show that the correlation between frequency and duration is

significantly small. Thus, we propose that one-tailed tests are adopted for the study of the law of brevity in other species especially for small repertoires.

Our meta-analyses suggest that the law of brevity and the underlying principle of compression may not be detected by conventional statistical approaches in a particular species if the repertoire is not large enough. Although type II errors are normally overcome by increasing the sample size, the invisibility of the law in a certain species may be intrinsic, or hard to avoid, because further sampling can improve the accuracy of the correlation but not the size of the repertoire, which is normally fixed *a priori*. We suggest that the visibility of the law in a given species will vary according to the size of the repertoire or according to our ability to identify the subsets of the repertoire where the pressure for brevity is high enough. At present, there is insufficient evidence for golden-backed uakaris and common marmosets as true exceptions to a statistical universal that is known as the law of abbreviation (Zipf 1949).

In this article we have approached the problem of the law of brevity from a quantitative linguistics point of view. We have reviewed various statistical issues that are important for future research from a biological or ethological perspective. Concerning marmosets, it has not escaped our attention that all the possible long distance calls in the dataset that we reanalyzed ("long phee call" and "twitter" (Bezerra 2011)) belong to the high D cluster, the cluster where the law of brevity is not found. Pressure for brevity is reduced for long-distance communication (Slabbekoorn 2006). Information theory indicates that redundancy and increasing duration, in particular, facilitate transmission (Cover & Thomas 2006, 184) and enhancing transmissibility is more important for long-range than for short range communication (Wiley 2009, 829; Slabbekoorn 2006). The elongation of signals (the opposite of abbreviation) is known to be a solution adopted by common marmosets to fight against background noise (Brumm et al. 2004). However, determining the bioacustical or biological factors hiding the law of brevity in golden-backed uakaris or leading to the two clusters that we have discovered using statistical arguments should be the subject of a further research.

Appendix

We investigate the rank correlation test as a particular case of randomization test (Sokal & Rohlf 1995, 803–819) to shed light on the limitations of rank correlation tests with small repertoires. We start with the calculation of the lower bonds for the p-value p of one-tailed and two-tailed rank correlation tests. Imagine that one wants to calculate $\rho(X, Y)$, the correlation between two vectors $X = x_1,...,x_i,...,x_n$ and $Y = y_1,...,y_i,...,y_n$, and that there are no ties either in X or in Y. X' is used to refer to a permutation of X. In a one-tailed rank correlation test, p is the proportion of permutations of where $\rho(X', Y) \leq r(X, Y)$ (or $\rho(X', Y) \geq \rho(X, Y)$ depending on the tail of interest). We have $p \geq 1/n!$, where n is the repertoire size and 1 is the minimum number of permutations of X yielding a correlation at least as large (or at most as low, depending on the tail of interest) as that between X and Y. The permutation $X' = X$ is at least one of those permutations.

In a two-tailed rank correlation test, we have $p \geq 2/n!$, where n is the repertoire size and 2 is the minimum number of permutations yielding X' such that $|\rho(X', Y)| \geq |\rho(X, Y)|$. The factor 2 comes from the fact that there are at least two of these permutations: $X'=X$ and $X'=X''$, where X'' is the inverse of X, i.e. $x''_i = x_{n-i+1}$. Notice that $\rho(X, Y) = -\rho(X'', Y)$ and thus $|\rho(X, Y)| = |\rho(X'', Y)|$.

For statistical significance, it is required $p \leq a$, where a is the significance level. n^* is defined as the minimum value of n needed to achieve significance. Table 3 shows the different values of n^* as a function of different values of a for one-tailed and two-tailed tests. n^* is the minimum value of n that satisfies $k/n! \leq a$, where k is the number of tails ($k = 1$ for one-tailed and $k = 2$ for two-tailed).

References

Altmann, G. (1980): Prolegomena to Menzerath's law, in: *Glottometrika* 2, 1–10.

Bezerra B. M. (2011): Personal communication.

Bezerra, B. M.; Souto, A. S. Radford, A. N. & Jones, G. (2011): Brevity is not always a virtue in primate communication, in: *Biology Letters* 7, 23–25. (DOI: 10.1098/rsbl.2010.0455).

Boroda, M. G. & Altmann, G. (1991). Menzerath's law in musical texts, in: *Musikometrica* 3, 1–13.

Brumm, H. Voss, K.; Köllmer, I. & Todt, D. (2004): Acoustic communication in noise: regulation of call characteristics in a New World monkey, in: *The Journal of Experimental Biology* 207, 443–448.

Conner, R. N. (1985): Vocalizations of common ravens in Virginia. *Condor* 87, 379–388.

Cover, T. M. & Thomas, J. A. (2006): *Elements of information theory.* Hoboken, NJ, USA: Wiley. 2nd edition.

Dawkins, R. (1976): Hierarchical organization: a candidate principle for ethology, in: *Growing points in ethology* (Bateson, P. P. G. & Hinde, R. A. (eds.)). Cambridge: Cambridge University Press.

DeGroot, M. H. (1989): *Probability and Statistics.* Reading, MA, USA: Addison-Wesley, pp. 215. 2nd edition.

Evans, N. & Levinson, S. C. (2009): The myth of language universals: Language diversity and its importance for cognitive science, in: *Behavioral and Brain Sciences* 32 (5), 429–492.

Ferrer-i-Cancho, R. & Lusseau, D. (2009): Efficient coding in dolphin surface behavioral patterns, in: *Complexity* 14 (5), 23–25. (DOI: 10.1002/cplx.20296).

Ferrer-i-Cancho, R. & McCowan, B. (2009): A law of word meaning in dolphin whistle types, in: *Entropy* 11 (4), 688–701.

Ferrer-i-Cancho, R. & Forns, N. (2009): The self-organization of genomes, in: *Complexity* 15 (5), 34–36.

Ficken, M. S.; Hailman, J. P. & Ficken, R.W. (1978): A model of repetitive behaviour illustrated by chickadee calling, in: *Animal Behaviour* 26 (2), 630–631. (DOI:10.1016/0003-3472(78)90075-1).

Hernández-Fernández, A.; Baixeries, J.; Forns, N. & Ferrer-i-Cancho, R. (2011): Size of the whole versus number of parts in genomes, in: *Entropy* 13 (8), 1465–1480. (DOI: 10.3390/e13081465).

Hailman, J. P.; Ficken, M. S. & Ficken, R. W. (1985): The 'chick-a-dee' calls of Parus atricapillus: a recombinant system of animal communication compared with written English, in: *Semiotica* 56, 191–224.

Hailman, J. (1987): Constraints on the structure of combinatorial 'chick-a-dee' calls, in: *Ethology* 75, 62-80.

Köhler, R. (2005): Synergetic Linguistics, in: *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. (R. Köhler, G. Altmann & R.G. Piotrowski (eds.)). Berlin, New York: Walter de Gruyter, 760–775.

Li, W. (2012): Menzerath's law at the gene-exon level in the human genome, in: *Complexity* 17, (4), 49–53.

MacWhinney, B. (2000): *The CHILDES project: Tools for analyzing talk*. 3rd edition. Volume 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

McCowan, B.; Hanser, S. F. & Doyle, L. R. (1999): Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires, in: *Animal Behaviour* 57, 409–419.

Salomon, D. (2007): *Data compression: the complete reference*. Berlin: Springer Verlag.

Semple, S.; Hsu, M. J. & Agoramoorthy, G. (2010): Efficiency of coding in macaque vocal communication, in: *Biology Letters* 6, 469–471. (DOI:10.1098/rsbl.2009.1062).

Semple, S.; Hsu, M. J.; Agoramoorthy, G. & Ferrer-i-Cancho, R. (2012): The law of brevity in macaque vocal communication is not an artifact of analyzing mean call durations. Submitted. Avaiable at http://arxiv.org/abs/1207.3169

Slabbekoorn, H. (2006): Animal communication: long-distance signaling, in: *Encyclopedia of Language Linguistics*, K. Brown (Ed.). 2nd edition. Oxford: Elsevier, 272-276.

Sokal, R.R. & Rohlf, F.J. (1995): *Biometry: the principles and practice of statistics in biological research*. New York: Freeman.

Strauss, U.; Grzybek, P. & Altmann, G. (2007): Word length and word frequency, in: *Contributions to the science of text and language*, (P. Grzybek (Ed.)). Dordrecht: Springer, 277–294.

Teupenhayn, R. & Altmann, G. (1984): Clause length and Menzerath's law, in: *Glottometrika* 6, 127–138.

Wiley, R. H. (2009): Signal transmission in natural environments, in: *Encyclopedia of Neuroscience*, (Squire, L. R. (ed.)), Volume 8. Oxford: Elsevier, 827-832.

Zipf, G. K. (1935): *The psycho-biology of language: an introduction to dynamic philology*. Cambridge, MA: MIT Press.

Zipf, G. K. (1949): *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Ramon Ferrer-i-Cancho
Complexity and Quantitative Linguistics Lab.
Departament de Llenguatges i Sistemes Informàtics
TALP Research Center.
Universitat Politècnica de Catalunya
Campus Nord, Edifici Omega
Jordi Girona Salgado 1–3
08034 Barcelona (Catalonia)
Spain
E-mail: rferrericancho@lsi.upc.edu

Antoni Hernández-Fernández
Departament de Lingüística General
Universitat de Barcelona
Gran Via de les Corts Catalanes 585
08007 Barcelona (Catalonia)
Spain
E-mail: antonio.hernandez@upc.edu

CAPÍTULO 3

Principios en una teoría de la comunicación

La búsqueda de principios generales de la comunicación y del lenguaje es uno de los objetivos de la lingüística. De la formalización de la teoría de la comunicación de Shannon y Weaver se pueden derivar varios principios de optimización: la minimización de la entropía y la maximización de la información mutua. Añadimos el principio de compresión, una de las aportaciones más relevantes de este trabajo por el que todo sistema comunicativo tiende a la minimización de la longitud de los elementos del sistema. Una consecuencia del principio de compresión es la ley de brevedad.

Múltiples modelos teóricos intentan dar cuenta de las leyes del lenguaje y la comunicación. Lo que solemos hacer es partir de la observación de datos de los que se infieren regularidades, de forma inductiva o deductiva, y posteriormente se formulan leyes o principios teóricos explicativos de los fenómenos observados. Formulados los principios, somos capaces de aventurar nuevas e hipotéticas leyes que contrastar con la experiencia. La contrastación empírica, o la falsabilidad en el sentido popperiano, serán las encargadas posteriormente de validar o refutar los modelos, siempre siguiendo el principio dialéctico de la ciencia: si nuestra hipótesis sostiene *A* y observamos *no A*, deberemos cambiar entonces o nuestra hipótesis o nuestra manera de observar (Wagensberg, 1994).

Sin embargo, las características de los fenómenos empíricos del lenguaje y la comunicación hacen imprescindible una aproximación probabilística que, como apuntan Chater y Manning (2006, p.343):

> Understanding and producing language involves complex patterns of uncertain inference, from processing noisy and partial speech input to lexical identification, syntactic and semantic analysis, to language interpretation in context. Acquiring language involves uncertain inference from linguistic and other data, to infer language structure. These uncertain inferences are naturally framed using probability theory: the calculus of uncertainty.

Por otra parte, la extensión de los patrones estadísticos del lenguaje a la comunicación y al comportamiento animal nos permite reconducir la investigación sobre los universales lingüísticos más allá del mito (Evans y Levinson, 2009), siguiendo los parámetros de la lingüística cuantitativa (Köhler, 2005) fundamentalmente a tres niveles (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013):

1.- Generalizar la tipología de los universales lingüísticos para dar cuenta de la comunicación en otras especies, convirtiéndolos en universales comunicativos.

2.- Considerar la posibilidad de que algunos universales comunicativos puedan ser universales del comportamiento animal.

3.- Trasladar el centro de la investigación del estudio descriptivo de las propiedades universales del lenguaje, o el comportamiento, a la búsqueda de principios explicativos.

Si nos ceñimos a la comunicación acústica, los datos experimentales se corresponden en general con el espacio tridimensional formado por una tríada (tiempo, frecuencia y amplitud) para las diferentes señales. Tal como vimos en la figura 2.6 (página 34) es de esperar que cada especie tenga un nicho comunicativo en el espacio de fases acústico, es

decir, un conjunto de intervalos posibles de frecuencias, amplitudes y duraciones (Sueur, 2006) en los que se comunica.

No obstante, cuando nos adentramos en la comunicación animal sin el amparo que nos otorga conocer de antemano el código –cómo nos sucede en el lenguaje humano–, nos hallamos indefensos ante lo desconocido, delante de un fenómeno complejo en el que descubrimos muchos elementos del espacio acústico tridimensional que es necesario considerar (figura 3.1.) y cuya complejidad hace que todavía se esté avanzando en la automatización de la toma de datos de vocalizaciones (Clemens y Johnson, 2006), en la que ya se han obtenido logros significativos, por ejemplo, en la discriminación e identificación de individuos en llamadas a larga distancia de panteras (Ji *et al.*, 2013).

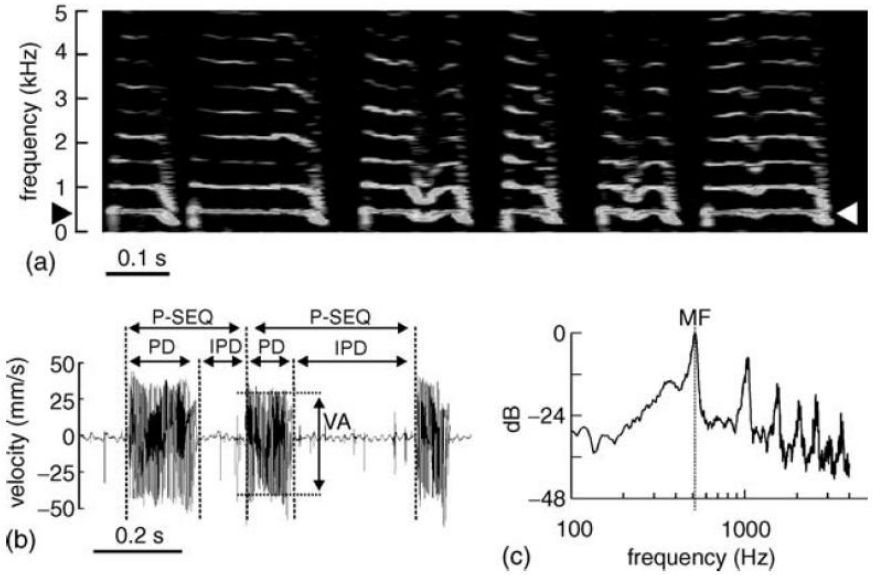


Figura 3.1: Características acústicas de la comunicación de la abeja sin aguijón (*Melipona rufiventris*). Las oscilaciones rítmicas del tórax de la abeja emiten una señal de la que se puede medir la duración de los pulsos (PD), su amplitud (en este caso de la velocidad del pulso, VA) y la frecuencia fundamental (MF, 0 dB), así como el intervalo de pausa entre pulsos (IPD) y su secuencia (P-SEQ). Extraído de Hrncir *et al.* (2006).

Pero, ¿qué determina el espacio acústico de cada especie? Sin duda hay constricciones fisiológicas (Fitch, 2000; Hauser, 1996) y energéticas (Prestwich, 1994; Gillooly y Ophir, 2010), de manera que por ejemplo la frecuencia de emisión se relaciona con el tamaño (Fletcher, 2004), llegando a imposibilitar, por ejemplo, en el medio aéreo la comunicación acústica de los insectos con un sistema como el de los mamíferos (Fletcher, 2004; Bennet-Clark, 1998). De hecho, las leyes alométricas operan

desde el genoma hasta los ecosistemas y podrían tener un principio de optimización común explicativo (West y Brown, 2005). Estas limitaciones para la comunicación acústica con las que se encuentran los insectos, arácnidos y demás seres vivos de pequeño tamaño, se salvan mediante la percusión, la comunicación vibracional o la comunicación química mediante feromonas que se ha estudiado en uno de los artículos de este trabajo (Hernández-Fernández y Ferrer-i-Cancho, 2014).

3.1. El principio de minimización de la entropía y el principio de maximización de la información mutua

El contexto de la teoría de la información proporciona un marco idóneo para adentrarse en la tarea de analizar los principios que gobiernan la comunicación (Ash, 1965; Shannon, 1948; Shannon y Weaver, 1949). Formalizaciones posteriores han avanzado en esa línea, deduciendo por ejemplo la ley de Zipf de la teoría de Kolmogorov, o dando cuenta de la universalidad de la ley de Zipf, quizá como patrón óptimo aunque sin explicar sus desviaciones en el exponente (Corominas y Solé, 2010), como vimos en el capítulo anterior.

De forma sucinta presentaremos aquí un marco para la minimización de la entropía[13] y la maximización de la información mutua. Un posible marco general es definir para un sistema de comunicación un conjunto de n señales $S=\{s_1,\ldots, s_i,\ldots, s_n\}$, que interactúan con un conjunto de m objetos de referencia o estímulos de respuesta $R=\{r_1,\ldots, r_i,\ldots, r_m\}$ (Ferrer-i-Cancho y Solé, 2003; Ferrer-i-Cancho, 2005c; Ferrer-i-Cancho, 2005d) que implica a su vez una red semántica (Ferrer-i-Cancho y Díaz-Guilera, 2007; Baronchelli *et al*., 2013; Ferrer-i-Cancho, 2013). Se define entonces una matriz binaria $A= \{a_{ij}\}$, con $1< i< n$ y $1< j< m$, siendo $a_{ij}= 1$ si la señal i se relaciona con el objeto j, o $a_{ij}= 0$ en caso contrario, y un marco probabilístico en el que $p(s_i)$ es la probabilidad asociada a la señal s_i y $p(r_j)$ la probabilidad del objeto r_j. Si aplicamos la definición de probabilidad condicionada y se define $w_j = \sum_i a_{ji}$ como el número de sinónimos que tenemos para referirnos al objeto j, se puede llegar a (Ferrer-i-Cancho y Solé, 2003)[14]:

[13] En nuestro caso la entropía no es la utilizada en termodinámica sino la definida por Shannon (1948) referida a la cantidad de información promedio que contienen los elementos del sistema de comunicación.
[14] Hay un error tipográfico en el subíndice de *w* en Ferrer-i-Cancho y Solé (2003) p.788.

$$p(r_j, s_i) = a_{ij} \frac{p(r_j)}{w_j} \quad (1)$$

Y entonces, para el caso sin sinonimia, en el que w_j=1, tenemos que (Ferrer-i-Cancho y Solé, 2003; Ferrer-i-Cancho, 2013):

$$p(s_i) = \sum_j a_{ij} p(r_j) \quad (2)$$

La propuesta de Ferrer-i-Cancho y Solé (2003) vincula el esfuerzo del emisor con la diversidad de señales disponibles, relacionada con su entropía $H_n(S)$, y el esfuerzo del receptor $H_m(R|S)$ *condicionado* a la recepción de la señal, de manera que la entropía de primer orden (Cover y Thomas, 2006):

$$H_n(S) = -\sum_{i=1}^{n} p(s_i) \log_n p(s_i) \quad (3)$$

En el caso de que el emisor sólo emitiera una señal para toda comunicación (y referirse a cualquier objeto) se minimizaría la entropía y $H_n(S) = 0$. Esto no es comunicativamente efectivo, pues no se desambigua nada y se violan los principios básicos de cooperación y éxito en la comunicación (Grice, 1989). Minimizar la entropía implica disminuir el tamaño del repertorio, puesto que la entropía de orden cero $H_o \propto \log n$ (Doyle *et al.*, 2008, para una revisión y ejemplo de aplicación). El esfuerzo del receptor, por su parte será (Ferrer-i-Cancho y Solé, 2003):

$$H_m(R|S) = -\sum_{i=1}^{n} p(s_i) \sum_{j=1}^{m} p(r_j|s_i) \log_m p(r_j|s_i) \quad (4)$$

La minimización de la entropía para el receptor debería darse cuando a cada objeto r_j le correspondiese una señal s_i diferente. Como veremos más adelante, ambas minimizaciones de la entropía, para el emisor y el receptor (optimización de las ecuaciones 3 y 4) no pueden darse simultáneamente de forma significativa (sin caer en el caso trivial). En este contexto se puede definir la función de información mutua (Cover y Thomas, 2006) como:

$$I_m(R;S) = H(S) - H(S|R) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(s_i|r_j) \log \frac{p(s_i|r_j)}{p(s_i)\, p(r_j)} \quad (5)$$

El máximo valor de $I_m(R;S)$ implica que hay una –y solo una señal– para cada objeto (Ferrer-i-Cancho y Solé, 2003). El modelo es general y puede aplicarse tanto al lenguaje humano como a la comunicación animal en todo su amplio espectro. La

información mutua da una idea de cómo de conectados están los objetos con sus referencias, o las palabras y sus significados, en el caso del lenguaje. Luego puede haber una tendencia a la maximización de la información mutua que posee como límite superior la entropía de manera que $I_m(R;S) \leq H_n(S)$, lo que se ha visto para el caso de la adquisición del lenguaje (Ferrer-i-Cancho, 2013), y que supone ya un primer conflicto entre principios: la minimización de la entropía tiende a disminuir el repertorio por parte del emisor, pero la maximización de la información mutua tiende a ampliarlo.
Por otra parte, quizá inspirada en los planteamientos de Zipf y Shannon, en su día Eleanor Rosch propuso dos principios de categorización semántica cuando no podían contrastarse todavía experimentalmente (Rosch, 1978, p.28):

> Two general and basic principles are proposed for the formation of categories: The first has to do with the function of category systems and asserts that the task of category systems is to provide maximum information with the least cognitive effort. The second principle has to do with the structure of the information so provided and asserts that the perceived world comes as structured information rather than as arbitrary or unpredictable attributes. Thus maximum information with least cognitive effort is achieved if categories map the perceived world structure as closely as possible.

Rosch (1978) estaba realizando hipótesis en ciencia cognitiva que podían –y debían– haberse fundamentado en su época mediante los principios de la teoría de la información, aunque es interesante comprobar cómo sus planteamientos cognitivos, aunque sin el apoyo de una sólida base matemática, coinciden con los formulados aquí. Los avances en neurociencia, y en especial en las técnicas de neuroimagen y exploración cerebral (Stemmer y Whitaker, 2008), permitieron posteriormente comprobar cómo realmente el cerebro sigue estrategias de optimización en la ubicación de elementos lingüísticos (Pulvermüller, 2002) y contrastar teorías que habían quedado marginadas, como así hizo en su caso Pulvermüller con la teoría de Hebb de la asociación neuronal (para una revisión, Pulvermüller, 1999).

Tanto el principio de minimización de la entropía para el emisor y el receptor, como el de maximización de la información mutua, actúan en la comunicación humana como presiones opuestas que se han modelado como principios integrados de optimización de funciones de energía (como por ejemplo en Ferrer-i-Cancho y Solé, 2003) y que seguramente tendrán su correlato neuronal (Pulvermüller, 1999).

Zipf (1949) ya intuyó la tendencia a unificar y diversificar el repertorio comunicativo, según se atienda a las necesidades del emisor o el receptor (Hernández-Fernández, 2005); se ha demostrado que bajo estos dos principios de teoría de la comunicación, minimización de entropía y maximización de la información mutua, obtenemos como resultado la emergencia de la ley de Zipf en el lenguaje (también si se toma el enfoque de Kolmogorov, véase Corominas y Solé, 2010) y se da cuenta de los efectos de frecuencia en las lenguas (Ferrer-i-Cancho, 2005d; Ferrer-i-Cancho y Solé, 2003). En este contexto nos preguntamos, ¿hay tras la ley de brevedad, observada en el lenguaje y la comunicación animal (Ferrer-i-Cancho y Hernández-Fernández, 2013), algún principio subyacente? Así surgió la propuesta del principio de compresión.

3.2. El principio de compresión

Según la teoría de la información, el principio de compresión hace que los códigos tiendan a minimizar su longitud media, lo que en los sistemas de comunicación se traduce en que hay una tendencia a que las palabras o elementos más frecuentes sean más cortos o breves, siguiendo la ley de Zipf de brevedad (Ferrer-i-Cancho, Hernández-Fernández *et al*., 2013, Ferrer-i-Cancho y Hernández-Fernández, 2013). En Ferrer-i-Cancho, Hernández-Fernández *et al.* (2013) presentamos el principio de compresión para los sistemas de comunicación como la tendencia a la minimización de la longitud de los elementos del sistema (ya sean duraciones, letras o bits). Obviamente no es el único principio que opera en la comunicación (en la sección anterior hemos visto otros dos), pero sí es el responsable de que los elementos del sistema más frecuentes tiendan a disminuir su tamaño y que emerja entonces la ley de brevedad (Ferrer-i-Cancho y Hernández-Fernández, 2013).

En concreto, si tenemos un repertorio S de n elementos (vocalizaciones, palabras…), siendo $p(s_i)$ la probabilidad de emitir el elemento i-ésimo y $l(s_i)$ su longitud (en segundos, letras, fonemas…), entonces la longitud del repertorio S será (Cover y Thomas, 2006; Ferrer-i-Cancho, Baixeries *et al.*, 2013):

$$L(S) = \sum_{i=1}^{n} p(s_i) l(s_i) \tag{6}$$

De manera que, siguiendo la hipótesis que planteamos en Ferrer-i-Cancho, Hernández-Fernández y colaboradores (2013), el principio de compresión consiste en encontrar las longitudes $l(s_i)$ que minimizan $L(S)$, dadas las probabilidades $p(s_i)$; si

establecemos una relación entre la energía que necesitamos para producir cada elemento $e(s_i)$ y la longitud de cada elemento, $l(s_i)$ entonces valorar la eficiencia de un código implica ver si se minimiza la energía necesaria para producirlo[15]. Los correlatos fisiológicos son bien conocidos (Prestwich, 1994).

En todo caso, presuponemos que emitir elementos más largos implica siempre un mayor gasto energético; así por ejemplo, en acústica la energía de una onda sonora está relacionada linealmente con la duración temporal y cuadráticamente con la amplitud de la señal, lo que sugiere que el principio de compresión debería operar tanto en el espacio temporal como en el de intensidades acústicas, buscando así la minimización de la energía, mientras que las frecuencias generalmente están marcadas por las constricciones fisiológicas y del medio de emisión (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013).

Todos estos argumentos nos conducen a la hipótesis del principio de compresión generalizado, que actuaría en las tres dimensiones del espacio de fases comunicativo (figura 3.2) y que va un poco más allá del principio de compresión explorado en Ferrer-i-Cancho, Hernández-Fernández y colaboradores (2013), limitado a la dimensión temporal, al extenderse a todo el espacio espectrográfico (Hernández-Fernández, 2014, en preparación). Definido el volumen espectrográfico ideal, en un espacio de fases tridimensional (frecuencia, intensidad y tiempo) la compresión actúa provocando el achatamiento del volumen en el eje correspondiente. El volumen ideal se ha representado generando la tercera dimensión –en el eje temporal– de la curva perceptiva de Wegel, proyección del volumen en el plano de intensidad y frecuencia, y que probablemente actúa, por el principio de economía, como umbral superior de las posibilidad de emisión, que surge de la superposición de los intervalos perceptivos acústicos (Lewicki, 2002).

Se espera que se dé una presión evolutiva menor en el ahorro energético y el principio de compresión en contextos como el apareamiento, en el que la presión sexual –clave en la evolución por selección natural– conduce a violar el principio en aras de la procreación y la perpetuación genética; ni tampoco en las llamadas a larga distancia

[15] Seguimos aquí el razonamiento inverso al del artículo (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013), en el que partíamos del razonamiento energético para llegar a las longitudes de elementos. La ecuación 6 se traduce pues directamente en la ecuación 1 de Ferrer-i-Cancho, Hernández-Fernández y colaboradores (2013).

donde el éxito comunicativo conduce a aumentar la intensidad y la duración, para evitar la atenuación de la señal como se expuso en Ferrer-i-Cancho y Hernández-Fernández (2013); ni tampoco en situaciones en las que se debe aumentar la duración o intensidad de la señal para paliar efectos ambientales o de ruido (Brumm y Slabbekoorn; 2005; Zöllinger y Brumm; 2011).

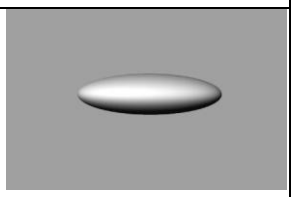
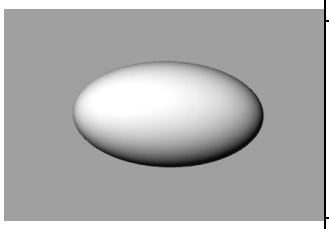
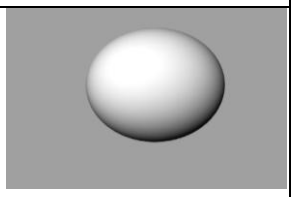
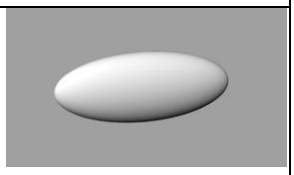
Espacio espectrográfico ideal *(x,y,z) = (f,I,t)*	**Acción de la compresión**	**Consecuencias**	**Efecto en el espacio espectrográfico**
El elipsoide representa una idealización del espacio espectrográfico.	Reducción temporal de la señal	Ley de Brevedad Ahorro energético	
	Aumento de las frecuencias de emisión (Equivale a reducción del periodo) Ajuste al canal	Selección y ajuste de frecuencias de emisión. Límites: fisiología, alometría	
La compresión puede afectar a cada eje por separado o a la vez.	Reducción de la intensidad acústica	Principio de economía Ahorro energético	

Figura 3.2: Características y consecuencias del principio de compresión generalizado para un espacio espectrográfico ideal (elipsoide) en el que se incluyen las tres características de las señales acústicas (frecuencias, intensidades y tiempos). Definido el elipsoide ideal, la compresión actúa provocando el achatamiento del mismo en el eje correspondiente. El elipsoide ideal se ha representado generando la tercera dimensión de la curva perceptiva de Wegel, que probablemente es un umbral superior a la posibilidad de emisión, y surge de la superposición de los intervalos perceptivos (Lewicki, 2002).

De hecho, otros fenómenos acústicos como la minimización de las interferencias hacen que existan ventanas y umbrales para las vocalizaciones (Brown y Sinnot, 2006), que nos llevan a recordar que el principio de compresión es uno más de los que debemos considerar al aproximarnos al estudio de la comunicación, donde impera el éxito en la emisión y recepción (Endler, 1993).

La relación entre la eficiencia de un código y una estimación de la densidad se puede obtener también aplicando el teorema de Shannon para los emisores de código (Shannon, 1948; Lewicki, 2002) según el cual la longitud esperada del código posee como límite inferior la entropía (ecuación 3, $H(p) = -\Sigma p(s) \log p(s)$). No obstante, los códigos reales poseen una longitud superior a los óptimos según la teoría de la

información, luego la entropía es una cota inferior del valor esperado de la longitud del código, *L(s)* es (Cover y Thomas, 2006, apartado 5.3. para una revisión y demostraciones):

$$L(S) \geq H_n(S) \tag{7}$$

El hecho de que se aumente la longitud de los elementos del código en los sistemas reales facilita la percepción, que funciona con cierto margen en el espacio de frecuencias, intensidades y duraciones (Lewicki, 2002). Luego, podemos completar la desigualdad planteada en el apartado anterior, referente a la información mutua y a la entropía, de manera que $I_m(R,S) \leq H_n(S) \leq L(S)$, lo que indica que en teoría de la información la entropía posee como cota inferior al valor esperado de la longitud del código, entendiendo que dimensionalmente se opera en bits (Cover y Thomas, 2006).

Es cierto que ha habido otras aproximaciones en la misma línea desde la filosofía y la ciencia cognitiva, la más cercana de las cuales ha sido el llamado principio de simplicidad desarrollado por Chater (Chater, 1999; Chater y Vitányi, 2002; Chater y Brown, 2008), basado en las ideas cognitivistas de Mach (1959) y Shepard (1957), y con un enfoque filosófico inicial que compartimos (Chater y Vitányi, 2002, p.22, ver las referencias del propio artículo):

> Since Mach, a number of theorists have proposed the sweeping idea that much of cognition concerns compression, or the elimination of redundancy. This 'simplicity principle' has been developed into a mathematically rigorous method for finding patterns in data, has served as the foundation for a broad range of cognitive models, and is consistent with a range of empirical data. We suggest that simplicity is worth pursuing as a potentially important unifying principle across many areas of cognitive science.

Como vimos en el capítulo 1, tanto el grupo de Chater como nosotros enlazamos plenamente con la filosofía de Guillermo de Ockham planteada. De hecho, Chater revisa su relevancia en la historia de la epistemología y en la selección de modelos en ciencia, y en concreto afirma (Chater, 1999, p.273-274):

> The cognitive system must cope with a world that is immensely complex but that is, nonetheless, highly patterned. The patterns are crucial. In a completely random world, prediction, explanation, and understanding would be impossible –there would be no patterns on which prediction could be based, to which explanations could refer, or the comprehension of which could amount to understanding. Even

more fundamentally, without any patterns relating actions to consequences, there would be no basis to choose one action rather than another. (…) The idea that cognition involves a search for simplicity has a long lineage, in the discussion of both normative and descriptive issues. On the normative side, the injunction to favour simple scientific theories can be traced to William of Ockham (1285-1349) and is endorsed by Newton (see Li & Vitányi, 1997, p. 317). Simplicity was also assigned fundamental importance in early positivist epistemology (e.g. Mach, 1883/ 1960), and it remains a standard principle in modern philosophy of science (e.g. Sober, 1975).

En nuestro caso, pese a compartir en buena parte el enfoque de Chater y colaboradores, en particular todo lo referente a la inferencia, la aproximación cuantitativa y de aplicación del método científico, nuestra visión de principios de la comunicación y el comportamiento es externalista, mientras que la visión de Chater y colaboradores es internalista (King, 2000)[16], es decir, Chater pretende conectar las representaciones mentales internas con las evidencias empíricas externas o patrones de comportamiento (como el lenguaje). De hecho, su postura internalista conduce a problemas referentes a la estructura del sistema cognitivo y a un difícil atolladero, al toparse además con el problema de las representaciones mentales, que se teorizan más que analizar bajo una perspectiva neurológica (Pulvermüller, 2002) o matemático-dinámica (Kelso, 2000), pese al uso de formalismos como la complejidad de Kolmogorov (Chater y Vitányi, 2002, p.20):

But how does the simplicity principle stand up to direct empirical testing? This question is difficult to answer, for two reasons:

[16] Citando a Patricia King (2000): "El internalista afirma que el agente cognoscitivo debe tener acceso inmediato a todas las condiciones necesarias y suficientes que determinan la justificación de sus creencias; el externalista, en cambio, sostiene que el agente cognoscitivo no necesita tener acceso inmediato a todas esas condiciones, donde por acceso inmediato a un hecho o estado de cosas casi siempre se entiende, en el contexto de esta discusión, que si este hecho ocurre, necesariamente sabemos que ocurre y viceversa. Uno de los puntos en litigio entre estas dos posiciones consiste en que los externalistas hacen desempeñar un papel epistémicamente justificatorio a la estructura causal del mundo aun cuando el sujeto cognoscitivo no tenga acceso inmediato a ella; en cambio, los internalistas sostienen que el sujeto debe poseer acceso inmediato a todo aquello que le permite justificar alguna de sus creencias, por lo que hechos y sucesos del mundo físico no pueden ser parte de la justificación ya que no son internos a nuestra mente y, por tanto, no son de acceso inmediato. La idea general que subyace en la distinción entre externalismo e internalismo es que el externalista no insiste en que todas las condiciones que determinan la justificación de la creencia de una persona involucran necesariamente estados mentales de esa persona; mientras que el internalista está convencido de que todas las condiciones que determinan la justificación hacen referencia a estados mentales a los cuales, presumiblemente, el sujeto tiene acceso inmediato.".

(1) The representation problem: Although, in the limit, and assuming the brain has universal Turing machine power and Kolmogorov complexity is language invariant, many specific, non-asymptotic empirical predictions from simplicity depend on assumptions about mental representation, which will affect what regularities can be detected. And the mental representation of perceptual and linguistic stimuli is highly contentious in cognitive science.
(2) The search problem: The cognitive system might prefer the simplest interpretation that it can find, but be unable to find a simple pattern of interest. Thus, without creating a full-scale cognitive model, involving assumptions about representation and perhaps also search, precise predictions from the simplicity viewpoint cannot be obtained.

Es cierto que nuestra postura, en la línea bungeana, evita el clásico problema de la percepción y el observador en el momento en el que nos aferramos a los datos, datos que suponemos no manipulados por la mente del observador, y tampoco admitimos principios no comprobables empíricamente. De ahí nuestro posicionamiento fijado en el primer capítulo: suponemos que el mundo exterior existe independientemente de la mente humana, los estados mentales son neurobiología fundamentada en la configuración y dinámica neuronal (Kelso, 2000; Pulvermüller, 1999) y confiamos en nuestros sentidos, parapetados tal vez en los intervalos de confianza que nos proporciona la estadística. No obstante, para la *neurofilosofía* el debate sobre si los estados mentales son reducibles a la neurobiología no se ha cerrado (Smith Churchland, 1985). Esta pequeña disertación sobre los trabajos de Chater y colaboradores ha pretendido diferenciar nuestra aportación de la suya –y de otras similares existentes en la ciencia cognitiva– aunque, cabe reiterar, ambas visiones son compatibles por abordar problemas complementarios y poseer enfoques diferentes.

Se ha demostrado matemáticamente la relación entre la ley de brevedad y el principio de compresión (apartado 2 de Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013): la ley de brevedad (Ferrer-i-Cancho y Hernández-Fernández, 2013) es un epifenómeno del principio de compresión aplicado a la dimensión temporal. Además, hemos entrado en la justificación empírica, pues mediante un test de aleatorización, (Sokal y Rohlf, 1995), se ha visto que el principio de compresión se cumple en diversas lenguas (griego, español, inglés, ruso, croata, sueco e indonesio), y para las vocalizaciones de macacos de Formosa, comunicación superficial de delfines y un subconjunto (el que no incluye llamadas a larga distancia) de vocalizaciones de titíes comunes. Quedan fuera las vocalizaciones a larga distancia de titíes comunes, los uakaris y los cuervos (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013; Ferrer-i-

Cancho y Hernández-Fernández, 2013), aunque en estos casos, posiblemente deberían investigarse más en profundidad, con técnicas automáticas de análisis y discriminación de señales más potentes (Ji *et al.*, 2013) y minimizando la intervención humana, por ejemplo en el análisis manual de espectrogramas (Zollinger *et al.*, 2012). En todo caso, como se ha mentado, otras presiones selectivas pueden operar además del principio de compresión. Es muy probable que la investigación en la percepción acústica y el filtrado auditivo de las señales (Bregman, 1990) nos pueda brindar nuevas evidencias sobre el principio de compresión para proseguir en el futuro con su investigación y aplicación. En palabras de Lewicki (2002):

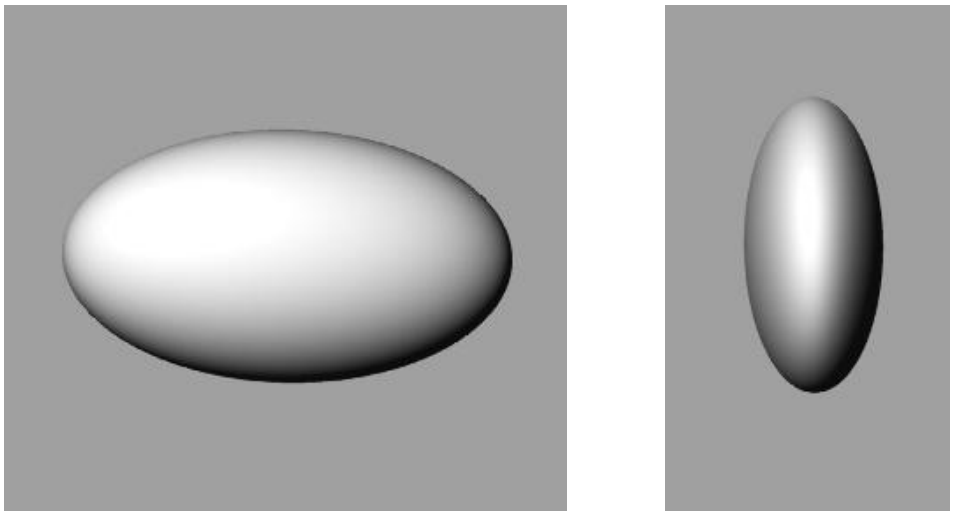
> The notion of an efficient code cannot be separated from the ensemble of signals that are being encoded. To make predictions for sensory codes, it is necessary to make conjectures about what class of stimuli the sensory system has evolved to process. This could range from a broad class of signals in the natural environment to only those crucial for reproduction and survival.

Finalmente, no podemos olvidar que intensidad, duración y frecuencia siguen siendo la tríada acústica y es de esperar que el principio de compresión debiera dar cuenta de todo el espacio de fases de la comunicación, lo que queda pendiente para trabajos futuros, así como profundizar en las posibles implicaciones termodinámicas del principio de compresión (Eroglu, 2013) y su relación con enfoques generales de optimización de la codificación (Wallace y Freeman, 1987; Cover y Thomas, 2006).

3.3. Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. y Semple, S. (2013). *Compression as a Universal Principle of Animal Behavior.*

Este apartado incluye el artículo:

- Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J. y Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*. DOI: 10.1111/cogs.12061.



COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2013) 1–14

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12061

Compression as a Universal Principle of Animal Behavior

Ramon Ferrer-i-Cancho,[a] Antoni Hernández-Fernández,[a,b] David Lusseau,[c,d] Govindasamy Agoramoorthy,[e] Minna J. Hsu,[f] Stuart Semple[g]

[a]*Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya*
[b]*Departament de Lingüística General, Universitat de Barcelona*
[c]*Institute of Biological and Environmental Sciences, University of Aberdeen*
[d]*Marine Alliance Science and Technology for Scotland, University of Aberdeen*
[e]*College of Environmental and Health Sciences, Tajen University*
[f]*Department of Biological Sciences, National Sun Yat-sen University*
[g]*Centre for Research in Evolutionary and Environmental Anthropology, University of Roehampton*

Received 10 April 2012; received in revised form 9 January 2013; accepted 11 February 2013

Abstract

A key aim in biology and psychology is to identify fundamental principles underpinning the behavior of animals, including humans. Analyses of human language and the behavior of a range of non-human animal species have provided evidence for a common pattern underlying diverse behavioral phenomena: Words follow Zipf's law of brevity (the tendency of more frequently used words to be shorter), and conformity to this general pattern has been seen in the behavior of a number of other animals. It has been argued that the presence of this law is a sign of efficient coding in the information theoretic sense. However, no strong direct connection has been demonstrated between the law and compression, the information theoretic principle of minimizing the expected length of a code. Here, we show that minimizing the expected code length implies that the length of a word cannot increase as its frequency increases. Furthermore, we show that the mean code length or duration is significantly small in human language, and also in the behavior of other species in all cases where agreement with the law of brevity has been found. We argue that compression is a general principle of animal behavior that reflects selection for efficiency of coding.

Keywords: Law of brevity; Compression; Animal communication; Language; Animal behavior; Linguistic universals

Correspondence should be sent to Ramon Ferrer-i-Cancho, Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya, Barcelona (Catalonia) 08034 Spain. E-mail: (rferrericancho@lsi.upc.edu)

1. Introduction

Understanding the fundamental principles underpinning behavior is a key goal in biology and psychology (Gintis, 2007; Grafen, 2007). From an evolutionary perspective, similar behavioral patterns seen across different animals (including humans) may have evolved from a common ancestral trait or may reflect convergent evolution; looking for shared quantitative properties of behavior across diverse animal taxa may thus allow identification of the elementary processes constraining or shaping behavioral evolution (e.g., Hailman, Ficken, & Ficken, 1985; McCowan, Doyle, & Hanser, 2002; Sumpter, 2006). Recent evidence of consistent patterns linking human language with vocal communication and other behavior in a range of animal species has pointed to the existence of at least one such general principle (Ferrer-i-Cancho & Lusseau, 2009; Semple, Hsu, & Agoramoorthy, 2010). Here, we detail the nature of this evidence before describing and mathematically exploring the principle in question—compression—that we propose underlies the consistent patterns discovered.

Words follow Zipf's law of brevity, that is, the tendency of more frequently used words to be shorter (Strauss, Grzybek, & Altmann, 2007; Zipf, 1935), which can be generalized as the tendency of more frequent elements to be shorter or smaller (Ferrer-i-Cancho & Hernández-Fernández, 2013). Statistical laws of language have been studied outside human language (see Ferrer-i-Cancho & Hernández-Fernández, 2013 for an overview), and to our knowledge, the first report of conformity to this generalized brevity law outside human language is found in the pioneering quantitative research by Hailman et al. (1985) on chick-a-dee calls. More recently, accordance with the generalized law of brevity has been found in dolphin surface behavioral patterns (Ferrer-i-Cancho & Lusseau, 2009), the vocalizations of Formosan macaques (Semple et al., 2010), and a subset of the vocal repertoire of common marmosets (Ferrer-i-Cancho & Hernández-Fernández, 2013). A lack of conformity to the law has been found in the vocalizations of golden-backed uakaris (Bezerra, Souto, Radford, & Jones, 2011) and ravens (Ferrer-i-Cancho & Hernández-Fernández, 2013).

Statistical patterns of language, and the law of brevity in particular, offer a unique chance to reframe research on language universals (Evans & Levinson, 2009) in three ground-breaking directions, according to the current state of the art. First, in extending typology of linguistic universals beyond human language and considering the possibility of universals of behavior across species. Second, in considering that some language universals may, in fact, not be specific to language or communication but an instance of universals of animal behavior. Third, in changing the stress of the quest from universal properties of language to universal principles of language (or behavior) along the lines of modern quantitative linguistics (Köhler, 1987, 2005; Zipf, 1949). More important than expanding the collection of universal properties or delimiting what is universal or not is (a) a deep understanding of the principles explaining the recurrence of these regularities regardless of the context, and (b) the role of those principles, perhaps hidden, in exceptions to widespread regularities. If these regularities are not inevitable (Ferrer-i-Cancho,

Forns, Hernández-Fernández, Bel-Enguix, & Baixeries, 2013), a parsimonious hypothesis would be a minimal set of principles that are independent from the context and thus universal. An important research goal is defining such a set, if it exists. Furthermore, this third point could contribute to reconciling the wide diversity of languages with the need for unifying approaches, outside the realm of a "language faculty" or innate specializations for language (Evans & Levinson, 2009).

The discovery of conformity to statistical laws of language outside human language raises a very important research question: Are the findings simply a coincidence, or are there general principles responsible for the appearance of the same statistical pattern across species or across many levels of life? As the arguments against the importance or utility of language laws within and outside human language are falling (e.g., Ferrer-i-Cancho & Elvevåg, 2010; Ferrer-i-Cancho & McCowan, 2012; Ferrer-i-Cancho et al., 2013; Hernández-Fernández, Baixeries, Forns, & Ferrer-i-Cancho, 2011), here we propose the principle of compression—the information theoretic principle of minimizing the expected length of a code—as an explanation for the conformity to the law of brevity across species. Hereafter, we consider the term "brevity," the short length of an element, as an effect of "compression."

In his pioneering research, G. K. Zipf argued that the law of brevity was a consequence of a general principle of economy (Zipf, 1949). He used the metaphor of an "artisan who will be obliged to survive by performing jobs as economically as possible with his tools" (p. 57). The artisan is a metaphor for a speaker (or a community of speakers) while tools is a metaphor for words. The artisan should decrease the size or the mass of the tools that he uses more frequently to reduce the amount of work (Zipf, 1949, pp. 60–61). Along similar lines, it is well known that S. Morse and A. Vail optimized the Morse code simply by choosing the length of each character approximately inversely proportionally to its frequency of occurrence in English (Gleick, 2011). Although it has been claimed that conformity to the law of brevity is a sign of efficient coding in both human language (Zipf, 1949) and animal behavior (Ferrer-i-Cancho & Lusseau, 2009; Semple et al., 2010), no strong direct connection with standard information theory, and with coding theory in particular, has been shown.

Imagine that

- n is defined as the number of elements of the repertoire or vocabulary.
- p_i is defined as the probability of producing the i-th most likely element.
- e_i is defined as the energetic cost of that element. e_i could be the number of letters or syllables of the i-th most likely word or the mean duration of the i-th most likely vocalization or behavior of a non-human animal.

Then, the mean energetic cost of a repertoire or vocabulary can be defined as

$$E = \sum_{i=1}^{n} p_i e_i. \tag{1}$$

Indeed, a particular case of the definition of E in Eq. 1 is E_{CL}, the mean code length, which is the cost function that is considered in standard information theory for the

problem of compression, being e_i the length of the code used to encode the i-th element of the set of symbols (Cover & Thomas, 2006, pp. 110). Data compression consists of finding the code lengths that minimize E_{CL} given the probabilities $p_1,\ldots p_i,\ldots,p_n$. E_{CL} provides an objective measure of coding efficiency. In his pioneering research, G. K. Zipf called his own version of Eq. 1 the "minimum equation" (in the sense that it is the function to minimize in order to reduce effort), with p_i being frequency of use of a tool and e_i being work, that is, the product of the "mass" of the tool and the "distance" of the tool to the artisan (Zipf, 1949, p. 59). Therefore, G. K. Zipf's view constitutes a precursor of coding theory. However, he never tested if E_{CL} is significantly small in human languages, using a rigorous statistical approach. Hereafter, the definition of E_{CL} from information theory is relaxed so that e_i can be not only the bits used to code the i-th element but also its size, length, or duration. $e_i > 0$ is assumed here as we focus on "elements" that can be observed in a real system, although some models of behavior produce elements of length zero, and then $e_i = 0$ for such empty elements (see Ferrer-i-Cancho & Gavaldà, 2009 for a review of these models). Pressure for minimizing E_{CL} in animal behavior may arise from a number of different sources.

First, there may be a need to minimize the direct energetic costs of producing a behavior; evidence from a range of vertebrate and invertebrate species indicates, for example, that energy availability may limit the duration of calling behavior (Bennet-Clark, 1998; Fletcher, 1997; Gillooly & Ophir, 2010; Klump, 2005; Thomas, 2002). Furthermore, Waters and Jones (1995) showed theoretically that energy is directly proportional to duration in acoustic communication by integrating the energy flux density, as it is well known in fluid mechanics, according to which the sound energy flux (a) is given by the time integral of the squared sound pressure (Landau & Lifshitz, 1987), and (b) can be roughly approximated by $A \cdot t$, where t is the pulse duration and A is its amplitude, for relatively short stimuli, as Baszczyk (2003) explains. According to general acoustics, the energy of a sound wave is $\xi = P \cdot t$, where P is its sound power and t is its duration, and $P = I \cdot S$, where I is the sound intensity and S is the area of the propagation surface (Kinsler, Frey, Coppens, & Sanders, 2000). For a sound wave of amplitude A, the sound energy becomes (Fahy, 2001; Kinsler et al., 2000):

$$\xi = P \cdot t = I \cdot S \cdot t \propto A^2 \cdot t, \tag{2}$$

namely, the energy of a signal depends linearly on time and the squared amplitude. This suggests a priori that it is not only duration that matters but also amplitude. However, amplitude determines the energy of a sound wave at a given time, and therefore reducing the amplitude reduces the reach of the signal due to the natural degradation of the signal with distance and interference caused by other sounds (Bennet-Clark, 1998). Amplitude fluctuations during the propagation of sound have different effects on receivers, depending on wave frequency, and interfere with their ability to detect directionality (Wiley & Richards, 1978). Minimizing energy by amplitude modulation puts the success of communication at risk. Furthermore, amplitude is highly determined by body size (Bennet-Clark, 1998; Fletcher, 2004) as Gillooly and Ophir (2010) demonstrate because there is a

dependency between sound power and amplitude. Therefore, our energy function E_{CL} is capturing the contribution to sound energy that can be more easily controlled, namely, that of duration.

Second, shorter signals may have advantages independent of energetic production costs. In predation contexts, it may be beneficial for calling prey to decrease conspicuousness to predators (Endler, 1993; Hauser, 1996; Ryan, Tuttle, & Rand, 1982) or for calling predators to decrease conspicuousness to prey (Deecke, Ford, & Slater, 2005). In addition, shorter signals suffer less from problems linked to reverberation, an important phenomenon that degrades signals in environments containing solid structures (Waser & Brown, 1986). For the vocalizations of rainforest primates, for example, there appears to be an upper call duration limit of 200–300 ms, to minimize such interference (Brown & Sinnot, 2006; p. 191). Interestingly in relation to this point, the law of brevity has been found in one subset of the repertoire of common marmosets where all but one call type (the exception being the submissive squeal) are below the 300 ms limit; the other subset, where the law is not found, (a) consists of calls that are above 300 ms in duration, and (b) contains all long-distance communication calls (Ferrer-i-Cancho & Hernández-Fernández, 2013). If sound pressure, P, falls inversely proportional to the distance from the sound source (according to the so-called distance law of sound attenuation), and sound intensity (and then energy, recall $\xi = I \cdot S \cdot t$) falls inversely proportional to the square of the distance (Kinsler et al., 2000; Landau & Lifshitz, 1987), then long-distance calls must generally show an increase either in sound intensity or duration, or both. Therefore, E_{CL} does not measure all the costs of a repertoire, as it does not include intensity, and sometimes it may be advantageous to increase duration rather than reducing it. By focusing on E_{CL} rather than on E, we hope to shed light on the importance of compression in animal behavior.

Here, we aim to provide support for the principle of compression (minimization of E_{CL}) as a general principle of animal behavior. We do not propose that compression is the only principle of animal behavior, or the only principle by which behavior is optimized. The design of language is a multiple constraint engineering problem (Evans & Levinson, 2009; Köhler, 2005) and the same applies to the communication systems of other species (Endler, 1993). The appearance of design in communication systems can exist without a designer or intentional engineering (Cornish, 2010; Kirby, Cornish, & Smith, 2008).

Our notion of principle should not be confused with explanation. Principles are the ingredients of explanations. Using the physical force of gravity as an example helps to illustrate our notion of principle. The universal force of gravity explains why objects fall, but when a rocket flies upward its movement does not constitute an exception to the force. The force is still acting and is involved in explaining, for instance, the amount of fuel that is needed to fly in the opposite direction of the force. As the falling of an object is a manifestation of the force of gravity, we propose that the law of brevity in animal behavior is a manifestation of a principle of compression. Just as the force of gravity is still acting on a rocket moving away from the Earth, so we should not conclude prematurely that compression is not a relevant principle of behavior when the law of brevity is not detected.

Although the Earth and Venus are very different planets, the force of gravity is valid in both (indeed universal) and physicists only care about the difference in its magnitude. Similarly, we do not assume that human language and animal behavior cannot have compression in common, no matter how large the biological, social, cognitive, or other differences. Science is founded on parsimony (Occam's razor), and from an evolutionary perspective hypothesizing that the principle of compression is common to humans and other animal species is a priori simpler than hypothesizing that compression is not shared. We are adopting the perspective of standard statistical hypothesis testing, where the null hypothesis is that there is no difference between two populations, for example, two different species (Sokal & Rohlf, 1995). The research presented here strongly suggests that there is no need to adopt, a priori, the more demanding alternative hypothesis of an intrinsic difference between humans and other species' linguistic and non-linguistic behavior for the particular case of the dependency between the size or length of the units and their frequency. However, according to modern model selection theory, a good model of reality has to comply with a trade-off between its parsimony and the quality of its fit to reality (Burnham & Anderson, 2002). A key and perhaps surprising result that will be presented in this article is that E_{CL} is never found to be significantly high, in spite of apparently clear advantages in certain situations of increasing signal/behavior length (e.g., for long-distance communication). It could be interpreted that even when the direct effect of compression is not observed, compression still has a role, just as a rocket heading to space is still being attracted by the Earth's gravitational field. It is premature to conclude that compression has no role even when there is no evidence that E_{CL} is being minimized; it is important to note that there are serious statistical limits for detecting efficient coding, especially in small repertoires (Ferrer-i-Cancho & Hernández-Fernández, 2013).

The remainder of the article is organized as follows. Section 2 shows that the energetic cost of an element (e.g., the length of a word) cannot increase as frequency increases in a system that minimizes E_{CL}. Thus, the law of brevity is an epiphenomenon of compression. Section 3 shows direct evidence of compression in animal behavior, in particular, by demonstrating that E_{CL} is significantly small in all the cases where conformity to the law of brevity has previously been found. Section 4 discusses these findings.

2. The mathematical relationship between compression and the law of brevity

By definition of p_i, one has

$$p_1 \geq p_2 \geq p_3 \geq \ldots p_n. \tag{3}$$

If Zipf's law of brevity was agreed with fully, one should also have

$$e_1 \leq e_2 \leq e_3 \leq \ldots e_n. \tag{4}$$

Here, a mathematical argument for Zipf's law of brevity as a requirement of minimum cost communication is presented.

Imagine that the law of brevity is not agreed with fully, namely, that there exist some i and j such that $j = i + 1$, $e_i > e_j$, and thus Eq. 4 does not hold perfectly. One could swap the values of e_i and e_j. That would have two important consequences. First, the agreement with the law of brevity (Eq. 4) would be favored. Second, E_{CL} would decrease as is shown next. To see that E_{CL} is reduced by swapping e_i and e_j, we calculate E^s_{CL}, the mean cost after swapping,

$$E^s_{CL} = E_{CL} - p_i e_i - p_j e_j + p_i e_j + p_j e_i. \tag{5}$$

After rearranging Eq. 5, it is obtained that the increment of cost, $\delta = E^s_{CL} - E_{CL}$, is

$$\delta = (p_i - p_j)(e_j - e_i). \tag{6}$$

$\delta < 0$ means that the cost has reduced, that is, compression has increased. To see that $\delta \leq 0$ with equality if and only if $p_i = p_j$, notice that $p_i - p_j \geq 0$ (by definition of p_i and p_j and $i < j$) and $e_j - e_i < 0$ (recall we are considering the case $e_i > e_j$, and thus $e_i = e_j$ is not possible). In sum, a system that minimizes E_{CL} needs to follow Zipf's law of brevity (Eq. 4); otherwise there is a pair of element costs (e_i and e_{i+1}) that can be swapped to reduce E_{CL}. Finally, note that the idea of swapping of costs was introduced to demonstrate the necessity of the law of brevity (Eq. 4) from the minimization of E_{CL}, not to argue that swapping is the evolutionary mechanism through which duration or length is optimized according to frequency.

3. Direct evidence of compression

3.1. Methods

E_{CL} is estimated taking p_i as the relative frequency of the i-th most frequent element of a sample. If E_{CL} is significantly small in animal behavior, that would be a sign of compression to some degree. Whether E_{CL} is significantly small or not can be determined by means of a randomization test (Sokal & Rohlf, 1995, pp. 803–819). To this end, $E_{CL}{}'$, a control E_{CL}, is defined as

$$E_{CL}{}' = \sum_{i=1}^{n} p_i e_{\pi(i)}, \tag{7}$$

where π(i) is a permutation function (i.e., a one-to-one mapping from and to integers 1,2,…,n). The left p-value of the test is defined as the proportion of permutations where $E_{CL}{}' \leq E_{CL}$ and the right p-value is defined as the proportion of permutations where $E_{CL}{}' \geq E_{CL}$. The left p-value is estimated by T_L/T, where T is the total number of random permutations used and T_L is the number of such permutations where $E_{CL}{}' \leq E_{CL}$.

Similarly, the right p-value is estimated by T_R/T, where T_R is the number of random permutations where $E_{CL}' \geq E_{CL}$. $T = 10^5$ uniformly random permutations were used.

3.2. Data

We adopt the term type for referring not only to word types but also to types of vocalization and behavioral patterns. We used a dataset that comprises type frequency and type size/length/duration from the following species: dolphins (*Tursiops truncatus*), humans (*Homo sapiens sapiens*), Formosan macaques (*Macaca cyclopis*), common marmosets (*Callithrix jacchus*), golden-backed uakaris (*Cacajao melanocephalus*), and common ravens (*Corvus corax*). For humans, seven languages are included: Croatian, Greek, Indonesian, US English, Russian, Spanish, and Swedish. The data for dolphins come from Ferrer-i-Cancho and Lusseau (2009), for human languages from Ferrer-i-Cancho and Hernández-Fernández (2013), for Formosan macaques from Semple et al. (2010), for common ravens from Conner (1985), and finally, those for common marmosets and golden-backed uakaris come from Bezerra et al. (2011).

The mean duration of a call type (e in our notation) is defined as $e = D/f$, where f is the number of times that the call has been produced and D is total duration (D is the sum of all the durations of that call). For common marmosets, D defines two partitions of the repertoire: the low D partition, where the law of brevity is found, and the high D partition, where the law is not found (further details on this partitioning are given in Ferrer-i-Cancho & Hernández-Fernández, 2013). For dolphins, the definition of the size of a behavioral pattern in terms of elementary behavioral units—see Ferrer-i-Cancho and Lusseau (2009) for a description of these elementary units—is subject to some degree of arbitrariness. For this reason, two variants of behavioral pattern size are considered: one where the elementary behavioral unit "two" is not used and another where elementary unit "stationary" is not used. A summary of the features of the dataset is provided in Table 1.

3.3. Results

Table 1 summarizes the results of the randomization test in all the species where the law of brevity has been studied so far with the exception of chick-a-dee calls (Hailman et al., 1985), for which data are not available for reanalysis. The conclusions reached by Hailman et al. (1985) on chick-a-dees, namely, that the law of brevity holds in bouts of calls but does not (at least not as clearly) in individual calls must be interpreted carefully. Hailman et al.'s (1985) analysis relies on visual inspection of data, which is highly subjective: Visual conformity with the law of brevity is clear for bouts of calls, but disagreement with the law in individual calls is not obvious. Hailman et al. (1985) did not perform a correlation test to test for conformity to the law of brevity; by contrast, we are determining if the law holds by means of a correlation test between frequency and length/duration. Even raw plots of frequency versus length in human language show substantial dispersion, as shown, for example, in Fig. 1A of Ferrer-i-Cancho and Lusseau (2009).

Table 1
Summary of the features of the dataset and the results of the compression test.

Species	Behavior	Conformity to Law of Brevity	n	E_{CL}	Left p-Value	Right p-Value
Golden-backed uakaris	Vocalizations	No (Bezerra et al., 2011)	7	0.14 s	0.09	0.9
Common marmosets	Vocalizations	No (Bezerra et al., 2011)	12	0.46 s	0.5	0.5
	Vocalizations (low D cluster)	Yes (Ferrer-i-Cancho & Hernández-Fernández, 2013)	6	0.048 s	0.006	1
	Vocalizations (high D cluster)	No (Ferrer-i-Cancho & Hernández-Fernández, 2013)	6	0.59 s	0.1	0.9
Common ravens	Vocalizations	No (Ferrer-i-Cancho & Hernández-Fernández, 2013)	18	0.25 s	0.3	0.6
Dolphins	Surface behavioral patterns	Yes (Ferrer-i-Cancho & Lusseau, 2009)	31	1.3 e.b.u.	0.0002	1
	Id. without "stationary"		31	1.2 e.b.u.	0.001	1
	Id. without "two"		31	1.3 e.b.u.	0.002	1
Formosan macaques	Vocalizations	Yes (Semple et al., 2010)	35	0.18 s	0.008	1
Humans	Greek	Yes (Ferrer-i-Cancho & Hernández-Fernández, 2013)	4,203	3.9 char	$<10^{-5}$	$>1-10^{-5}$
	Russian		7,908	4.6 char	$<10^{-5}$	$>1-10^{-5}$
	Croatian		15,381	3.9 char	$<10^{-5}$	$>1-10^{-5}$
	Swedish		19,164	3.2 char	$<10^{-5}$	$>1-10^{-5}$
	US English		24,101	3.8 char	$<10^{-5}$	$>1-10^{-5}$
	Spanish		27,478	3.9 char	$<10^{-5}$	$>1-10^{-5}$
	Indonesian		30,461	4.2 char	$<10^{-5}$	$>1-10^{-5}$

Notes: Units: s stands for seconds; char stands for characters; e.b.u. stands for elementary behavioral units. E_{CL} was rounded to leave two significant digits. When not bounded, estimated left and right p-values were rounded to leave only one significant digit. Conformity to the law of brevity means that the correlation between frequency and size/duration is negative and significant.

4. Discussion

Table 1 presents a number of interesting results. First, E_{CL} was significantly small in all cases where conformity to the law of brevity (a significant negative correlation between frequency and size/length/duration) had been reported. This provides further support for the intimate relationship between the law of brevity and the information theoretic principle of compression. Second, there is no evidence of redundancy maximization, the opposite of compression, in the species that we have analyzed: E_{CL} was never significantly high. This result was unexpected for two reasons: Pressure for compression is not the only factor shaping repertoires (Bezerra et al., 2011; Endler, 1993; Ferrer-i-Cancho & Hernández-Fernández, 2013), and pressure for compression can cause a signal to be more sensitive to noise in the environment. Coding theory indicates that redundancy must be added in a controlled fashion to combat transmission errors (Cover & Thomas, 2006, pp. 184). In cases where signals can be mistaken for each other, for example, due to noise in the environment, forming words by stringing together subunits (e.g., combining "phonemes" into "words") allows a system to increase its capacity to communicate (Plotkin & Nowak, 2000). Therefore, redundancy maximization is a conceivable alternative to compression (redundancy minimization).

A range of phenomena or situations may select against compression in signals. First, as mentioned above, environmental noise may drive signals in the opposite direction to that of code minimization. Elongation of signals is one of a number of adaptations to noise, and this is known as the Lombard effect in human language and the communication systems of other species (Brumm & Zöllinger, 2011; Zöllinger & Brumm, 2011); another major strategy to combat noise is to increase redundancy, in the temporal or spatial organization of signals (Ay, Flack, & Krakauer, 2007; Richards & Wiley, 1980). Secondly, capacity for compression may be constrained in some species more than others, due to biological features of the species (Gillooly & Ophir, 2010) or their environment (Wiley & Richards, 1978). For example, the ability to use low-frequency signals to reduce attenuation and thus maximize transmission success is not always possible for smaller animals due to allometric constraints that limit the frequency (i.e., pitch) of a signal as a function of an animal's body mass (Fletcher, 2004). The same applies to maximizing transmission success by increasing amplitude (Brumm & Slabbekoorn, 2005): A species with a small body cannot produce a sound with high amplitude (Bennet-Clark, 1998; Fletcher, 2004) because amplitude is limited by body mass (Gillooly & Ophir, 2010). Third, reverberation is one of the fundamental problems posed by the forest environment, and one way to overcome this is to package the overall signal into brief pulses that end prior to the delay time expected for the main first reflection of the pulse (Brown & Sinnot, 2006, p. 191). A continuous signal of duration t could be converted into a package of duration $t + u$, where u is the total duration of the interpulse silences. Finally, in relation to human language, elongation occurs in the context of adults' child-directed speech, which tends to be slower, with syllable lengthening, longer pauses, etc. (Saxton, 2010, p. 81 and references therein).

When considering these processes that may drive an increase in signal size, it is important to note that if elongation affected all the types of elements equally (by a constant proportionality factor), then the law of brevity would remain. This suggests that the law of brevity can only be violated if elongation acts specifically on a subset of the repertoire. Consistent with this idea, common marmoset vocalizations do not conform to the law at the level of the whole repertoire, but do within a subset of the repertoire where none of the long-distance calls—calls highly sensitive to noise in the environment and other signal distorting factors—is found (Ferrer-i-Cancho & Hernández-Fernández, 2013). This finding suggests that analysis of logically selected subsets of the vocal repertoire of ravens and golden-backed uakaris, two species where conformity to the law has not been found, is worthwhile.

If mean length maximization plays a role in the species we have examined, it is not strong enough to surface in a statistical test. It is possible that these species introduce redundancy at other levels: below the type of element level, that is, within the shape or structure of the element, or above the element level, that is, in the way sequences of elements are constructed. It is also possible that individuals can achieve a similar goal to redundancy by varying instead call amplitude (e.g., Brumm & Slabbekoorn, 2005). The trade-off between compression (mean length minimization) and the need for successful signal transmission, where new element type formation (e.g., words) by combining elements (e.g., "phonemes") is a well-known strategy (Plotkin & Nowak, 2000), may explain why a clear manifestation of length maximization is difficult to find at our level of analysis.

The existence of true universals in the sense of exceptionless properties is a matter of current debate in human language (Evans & Levinson, 2009) and animal behavior (Bezerra et al., 2011; Ferrer-i-Cancho & Hernández-Fernández, 2013). We believe it is important to investigate the real scope of statistical patterns of language such as the law of brevity in world languages and animal behavior. The analysis of exceptions to the law of brevity is full of subtle statistical and biological details (Bezerra et al., 2011; Ferrer-i-Cancho & Hernández-Fernández, 2013). However, it is of even greater theoretical importance to investigate which are the universal principles of behavior. They are the arena where unification (universality) and the complex nature of reality, including exceptions to language patterns or peculiar situations where certain language patterns emerge, may reconcile. We are not claiming that the law of brevity is a hallmark of language (as opposed to simpler forms of communication), but an example of constrained or convergent evolution by an abstract principle of compression acting upon communicative and non-communicative behavior and reflecting selection for efficiency of coding. This principle provides an avenue for the optimization of the behavior of a species. The exact nature of the solutions adopted may depend on the chances and needs the species encounters during its evolutionary history (Monod, 1972).

Acknowledgments

We are grateful to S. E. Fisher, S. Kirby, Rick Dale, and three anonymous reviewers for helpful remarks. This work was supported by the grant *Iniciació i reincorporació a la*

recerca from the Universitat Politècnica de Catalunya, and the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation.

References

Ay, N., Flack, J., & Krakauer, D. C. (2007). Robustness and complexity co-constructed in multimodal signalling networks. *Philosophical Transactions of the Royal Society of London B*, *362*(1479), 441–447.

Baszczyk, J. W. (2003). Startle response to short acoustic stimuli in rats. *Acta Neurobiologica Experimentalis*, *63*, 25–30.

Bennet-Clark, H. C. (1998). Size and scale effects as constraints in insect sound communication. *Philosophical Transactions of the Royal Society of London B*, *353*(1367), 407–419.

Bezerra, B. M., Souto, A. S., Radford, A. N., & Jones, G. (2011). Brevity is not always a virtue in primate communication. *Biology Letters*, *7*, 23–25.

Brown, C. H., & Sinnot, J. M. (2006). Cross-species comparisons of vocal perception. In S. Greenberg & W. A. Ainsworth (Eds.), *Listening to speech: An auditory perspective* (pp. 183–201). London: Routledge.

Brumm, H., & Slabbekoorn, H. (2005). Acoustic communication in noise. *Advances in the Study of Behavior*, *35*, 151–209.

Brumm, H., & Zöllinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour*, *148*(11–13), 1173–1198.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer-Verlag.

Conner, R. N. (1985). Vocalizations of common ravens in Virginia. *Condor*, *87*, 379–388.

Cornish, H. (2010). Investigating how cultural transmission leads to the appearance of design without a designer in human communication systems. *Interaction Studies*, *11*(1), 112–137.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: Wiley.

Deecke, V. B., Ford, J. K. B., & Slater, P. J. B. (2005). The vocal behaviour of mammal-eating killer whales (*Orcinus orca*): Communicating with costly calls. *Animal Behaviour*, *69*(2), 395–405.

Endler, J. A. (1993). Some general comments on the evolution and design of animal communication systems. *Philosophical Transactions of the Royal Society of London B*, *340*(1292), 215–225.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–492.

Fahy, F. (2001). *Foundations of engineering acoustics*. Oxford, England: Elsevier.

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

Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G., & Baixeries, J. (2013). The challenges of statistical patterns of language: The case of Menzerath's law in genomes. *Complexity*, *18*(3), 11–17, doi:10.1002/cplx.21429.

Ferrer-i-Cancho, R., & Gavaldà, R. (2009). The frequency spectrum of finite samples from the intermittent silence process. *Journal of the American Society for Information Science and Technology*, *60*(4), 837–843.

Ferrer-i-Cancho, R., & Hernández-Fernández, A. (2013). The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, *4*(1), 44–55.

Ferrer-i-Cancho, R., & Lusseau, D. (2009). Efficient coding in dolphin surface behavioral patterns. *Complexity*, *14*(5), 23–25.

Ferrer-i-Cancho, R., & McCowan, B. (2012). The span of correlations in dolphin whistle sequences. *Journal of Statistical Mechanics*, 2012, P06002.

Fletcher, N. H. (1997). Sound in the animal world. *Acoustics Australia*, *25*(2), 2–69.

Fletcher, N. H. (2004). A simple frequency-scaling rule for animal communication. *Journal of the Acoustical Society of America*, *115*(5), 2334–2338.

Gillooly, J. F., & Ophir, A. G. (2010). The energetic basis of acoustic communication. *Proceedings of the Royal Society B*, *277*, 1325–1331.

Gintis, H. (2007). A framework for the unification of the behavioral sciences. *Behavioural and Brain Sciences*, *30*(1), 1–16.

Gleick, J. (2011). *The information: A history, a theory, a flood*. New York: Pantheon Books, Random House.

Grafen, A. (2007). The formal Darwinism project: A mid-term report. *Journal of Evolutionary Biology*, *20* (4), 1243–1254.

Hailman, J. P., Ficken, M. S., & Ficken, R. W. (1985). The 'chick-a-dee' calls of Parus atricapillus: A recombinant system of animal communication compared with written English. *Semiotica*, *56*(3–4), 191–224.

Hauser, M. D. (1996). *The evolution of communication*. Cambridge, MA: MIT Press.

Hernández-Fernández, A., Baixeries, J., Forns, N., & Ferrer-i-Cancho, R. (2011). Size of the whole versus number of parts in genomes. *Entropy*, *13*(8), 1465–1480.

Kinsler, L. E., Frey, A. R., Coppens, A. B., & Sanders, J. V. (2000). *Fundamentals of acoustics (4th edition)*. Hoboken, NJ: John Wiley & Sons, Inc.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Klump, G. (2005). Evolutionary adaptations for auditory communication. In J. Blauert (Ed.), *Communication acoustics* (pp. 27–46). Berlin: Springer-Verlag.

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics*, *14*, 241–257.

Köhler, R. (2005). Synergetic linguistics. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative linguistics: An international handbook* (pp. 760–775) Berlin, New York: Walter de Gruyter.

Landau, L. D., & Lifshitz, E. M. (1987). *Fluid mechanics* (2nd ed.). Oxford, England: Pergamon Press.

McCowan, B., Doyle, L. R., & Hanser, S. F. (2002). Using information theory to assess the diversity, complexity and development of communicative repertoires. *Journal of Comparative Psychology*, *116*(2), 166–172.

Monod, J. (1972). *Chance and necessity*. London: Knopf.

Plotkin, J. B., & Nowak, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology*, *205*(1), 147–159.

Richards, D. G., & Wiley, R. H. (1980). Reverberations and amplitude fluctuations in the propagation of sound in a forest: implications for animal communication. *American Naturalist*, *115*(3), 381–399.

Ryan, M. J., Tuttle, M. D., & Rand, A. S. (1982). Bat predation and sexual advertisement in a Neotropical frog. *American Naturalist*, *119*(1), 136–139.

Saxton, M. (2010). *Child language. Acquisition and development*. Los Angeles, CA: Sage.

Semple, S., Hsu, M. J., & Agoramoorthy, G. (2010). Efficiency of coding in macaque vocal communication. *Biology Letters*, *6*(4), 469–471.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry: The principles and practice of statistics in biological research*. New York: Freeman.

Strauss, U., Grzybek, P., & Altmann, G. (2007). Word length and word frequency. In P. Grzybek (Ed.), *Contributions to the science of text and language* (pp. 277–294). Dordrecht, The Netherlands: Springer.

Sumpter, D. J. T. (2006). The principles of collective animal behaviour. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *361*(1465), 5–22.

Thomas, R. J. (2002). The costs of singing in nightingales. *Animal Behaviour*, *63*(5), 959–966.

Waser, P. M., & Brown, C. H. (1986). Habitat acoustics and primate communication. *American Journal of Primatology*, *10*(2), 135–154.

Waters, D. A., & Jones, G. (1995). Echolocation call structure and intensity in five species of insectivorous bats. *The Journal of Experimental Biology*, *198*, 475–489.

Wiley, R. H., & Richards, D. G. (1978). Physical constraints on acoustic communication in the atmosphere: Implications for the evolution of animal vocalizations. *Behavioral Ecology and Sociobiology*, *3*(1), 69–94.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Zöllinger, S. A., & Brumm, H. (2011). The Lombard effect. *Current Biology*, *21*(16), R614–615.

CAPÍTULO 4

Un mundo de comunicación

La comunicación se inició con la vida. Todos los seres vivos albergan información y se comunican tanto con su entorno interno o cuerpo, como con su entorno externo. Nuestra comunicación es bioquímica en gran medida, no sólo lingüística. ¿Se pueden establecer paralelismos entre el genoma y los sistemas de comunicación? ¿Podemos explorar el genoma utilizando las herramientas y leyes de la lingüística cuantitativa?

Decir que la comunicación se inició con la vida es poner un punto de partida a todo lo que conocemos, a la existencia misma. Comprender nuestros sistemas de comunicación es mucho más que plantearse cómo funciona el lenguaje. El mismo plural, "sistemas de comunicación" nos responde. No podemos ser reduccionistas y a la vez intentar formular una teoría global sobre los sistemas de comunicación olvidándonos de todo lo que hay más allá de las lenguas humanas o la comunicación acústica. Nuestra comunicación, como la de todos los seres pluricelulares, es mucho más compleja y va mucho más allá del lenguaje. En palabras del filósofo José Antonio Marina (2007):

> Ni los animales ni los humanos nacemos siendo páginas en blanco en las que la experiencia irá escribiendo las individuales biografías. A ustedes y a mí nos parieron con mucho texto ya escrito en nuestros organismos, escritura que, como un palimpsesto, podríamos descubrir a la luz adecuada.

Esa luz de la que nos habla Marina (2007) es la ciencia. Y la lingüística, como ciencia biológica, tiene también mucho que decir. Iniciaremos nuestro viaje por este mundo de comunicación que es la vida con un breve repaso a algunas nociones biológicas que nos conducirán a las preguntas que intentamos resolver en las publicaciones integradas en este capítulo.

4.1. Células, ADN y cromosomas

Para empezar, es bien conocido que estamos conformados por células[17] que albergan en su interior nuestro material genético, en el que se codifica la información que nos permite existir desde nuestra concepción, que nos regula y nos hace madurar como organismos. Es la información según la cual vivimos y morimos. El genoma[18] de una especie incluye la totalidad de su información genética, contenida en las células, que en los seres eucariontes –como nosotros– se compone por el ADN (o ácido

[17] La teoría celular fue propuesta en 1837 para los vegetales, y confirmada para el resto de seres vivos posteriormente, por Matthias Jakob Schleiden y Theodor Schwann. Se postula que todos los organismos están compuestos por células, cuyo ciclo regula la vida del organismo que configuran, y que todas estas células derivan de otras precedentes. Schwann, T. & Schleyden, M.J. (1847). *Microscopical researches into the accordance in the structure and growth of animals and plants*. London: Printed for the Sydenham Society.

[18] Término acuñado por el botánico alemán Hans Winckler en 1920 en su obra "*Verbreitung und Ursache der Parthenogenesis im Pflanzen - und Tierreiche*", como acrónimo de gen y cromosoma. Anteriormente, en 1909, el botánico danés Wilhelm Ludwig Johannsen acuñó el étimo gen para referirse a la unidad física y funcional de la herencia biológica.

desoxirribonucleico) del núcleo celular, y el ADN extranuclear, formado por el ADN mitocondrial y el ADN de los cloroplastos (en las plantas). Las células que tienen núcleo son las eucariotas, mientras que las células sin un núcleo celular diferenciado, son las procariotas[19].

El ADN del núcleo celular se divide a su vez en un conjunto de cromosomas, compuestos por genes, que a su vez están configurados por una secuencia ordenada de nucleótidos. Los nucleótidos son moléculas orgánicas formadas por la unión covalente de un monosacárido de cinco carbonos (pentosa), una base nitrogenada y un grupo fosfato. El número de genes en el ADN mitocondrial humano es de unos 37, frente a los poco más de 25.000 genes de nuestro ADN cromosómico nuclear (Alberts *et al.*, 2008, p.142). Los genes están formados por exones, que constituyen el ARN mensajero en la transcripción y codifican proteínas, y los intrones, que son secuencias de nucleótidos que no se transcriben (figura 4.1.). En el genoma procariota no hay intrones.

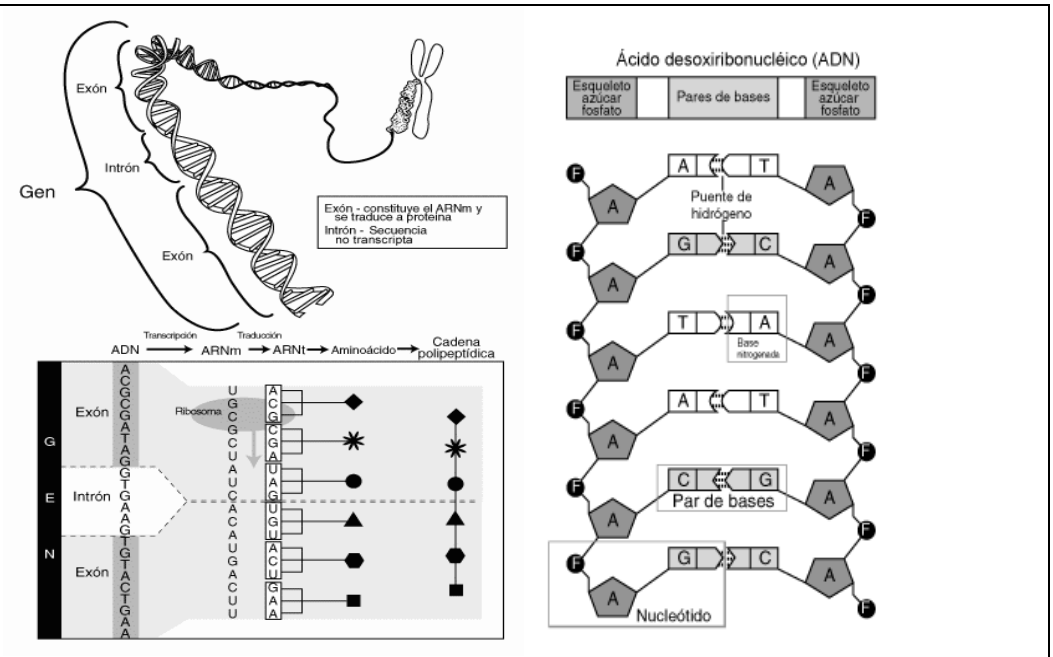


Figura 4.1: Ejemplo de estructura de un gen eucariota. Puede apreciarse la doble hélice de nucleótidos, desenrollada de un cromosoma (izquierda). El esquema corresponde al telómero (final del cromosoma) de un gen eucariota corto (el genoma procariota carece de intrones). Los nucleótidos (derecha) están formados por tres partes: una base nitrogenada (purinas como la adenina o la citosina, o pirimidinas como la timina, citosina y uracilo), una pentosa que da lugar a ácidos nucleicos de dos clases (ribosa en el ARN y desoxirribosa en el ADN) y como mínimo un grupo fosfato que enlaza los nucleótidos (Alberts *et al.*, 2008). Fuente figuras: http://commons.wikimedia.org/wiki/File:Gene.png

[19] Basándose en la secuenciación comparativa del ARN ribosómico existen tres dominios principales de organismos vivos: las eubacterias (*bacteria*), las arqueobacterias (*archaea*) y los eucariontes (*eucarya*). Se consideran procariotas las eubacterias y arqueobacterias. Estos tres reinos tienen un ancestro común universal (Margulis, 1996).

La estructura y características del ADN mitocondrial es muy diferente a la del ADN nuclear (figura 4.2), lo que condujo a la hipótesis de su origen bacteriano: según la teoría endosimbiótica de Margulis (1970) probablemente se dio la endocitosis de una bacteria aeróbica y su posterior transferencia genética hacia el genoma nuclear, aunque todavía hay debate sobre cómo opera el ADN mitocondrial a nivel celular (Galtier *et al.*, 2009). Ya en 1918, empero, el zoólogo francés Paul Portier[20] había llegado a la conclusión de que las mitocondrias de las eucariotas habían sido bacterias en el pasado filogenético (Sapp, 1994). Las mitocondrias son orgánulos que se encuentran en todos los seres eucariotas aeróbicos y contienen las enzimas para las reacciones oxidativas que generan energía para las funciones celulares.

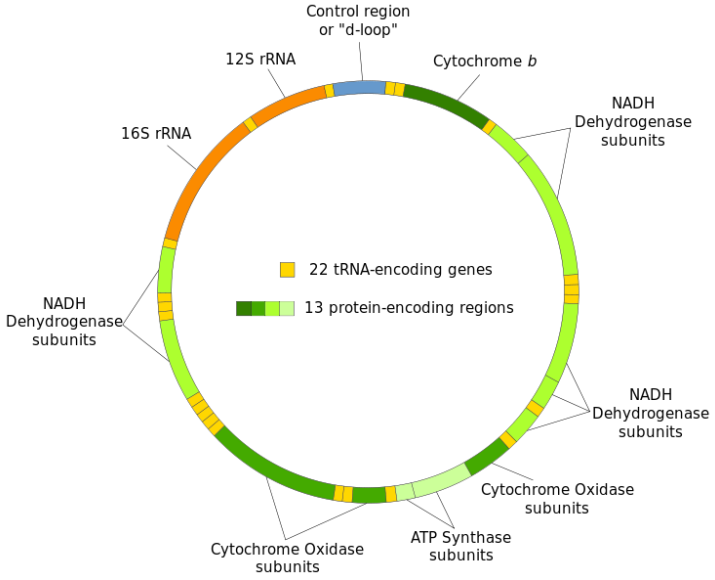


Figura 4.2: ADN mitocondrial humano. Incluye tan solo 37 genes con unos 16600 pares de bases nitrogenadas. Su estructura circular y la ausencia de intrones es análoga al ADN bacteriano, por lo que se sostiene la hipótesis de su origen por endocitosis de una eubacteria aeróbica y la posterior transferencia genética hacia el genoma nuclear (Margulis, 1970), aunque todavía es un debate abierto en biología (Galtier *et al*., 2009). Fuente: http://en.wikipedia.org/wiki/File:Mitochondrial_DNA_en.svg

En realidad es la cromatina (el conjunto de nucleótidos, histonas y proteínas no histónicas que se halla en el núcleo de las células eucariotas) la que constituye finalmente el cromosoma celular[21]. La organización del ADN nuclear en cromosomas

[20] Paul Portier (1866 - 1962) fue más conocido por ser codescubridor de la anafilaxia, junto con el fisiólogo Charles Robert Richet (1850 - 1935). Rojido, G.M. (2001). One hundred years of anaphylaxis. *Alergol Inmunol Clin* 2001; 16: 364-368.

[21] El lector puede revisar cualquier manual de citología para ampliar la breve exposición de este apartado, como es el caso del clásico de Alberts y colaboradores (2008), seguido aquí.

varía según las especies aunque todavía es un misterio su organización. Cada cromosoma posee una longitud característica y tiene una región constreñida, el centrómero, que permite clasificarlos según su posición a lo largo del cromosoma, y unos extremos o telómeros. Las especies poseen un mismo número de cromosomas o número diploide (2n), y se agrupan generalmente en pares. A los miembros de cada par de cromosomas se les denomina cromosomas homólogos.

La estructura interna de los cromosomas es un campo de estudio en el que se ha avanzado mucho desde los estudios pioneros de los años 60 y 70 del siglo XX (Alberts *et al.*, 2008), en los que el interés de las investigaciones se centraba por ejemplo en la estructura cromosómica durante la división celular (véase Marsden y Laemmli, 1979). No obstante, los enfoques que se dan desde la microbiología actual son exhaustivos en la búsqueda de especificidades, aunque, sorprendentemente, no haya estudios que respondan con claridad a una sencilla pregunta: ¿por qué existen los cromosomas? Es decir, ¿qué presiones evolutivas conducen a los organismos eucariotas a dividir su ADN en cromosomas? Las bacterias no lo hacen y normalmente tienen todos sus genes en una sola molécula de ADN, que puede ser circular, como hemos visto en el ADN mitocondrial, probablemente de origen bacteriano (figura 4.2). No obstante, citando a Alberts y colaboradores (2008, p.202):

> Aunque frecuentemente a este conjunto se le denomina el "cromosoma" bacteriano no tiene la misma estructura que los cromosomas de las células eucariotas; conocemos peor la forma en que se empaqueta el DNA de la célula bacteriana, siendo aún menos conocida la forma en que se empaqueta el DNA en las arqueobacterias.

Como exponen Alberts y colaboradores (2008), y otras revisiones (Klug y Cummings, 1999) obviamente la reproducción celular está detrás de la existencia de los cromosomas en eucariotas, pues los cromosomas adoptan diferentes estados a lo largo de la vida de una célula. La cromatina se enrolla formando la típica imagen de los cromosomas durante la mitosis, que no es la habitual de la cadena (figuras 4.1 y 4.3).
Mover el ADN dividido en cromosomas en el fluido de la célula es de entrada aparentemente más sencillo que si el ADN formase una única hebra, simplemente por biofísica pura. Las leyes de la dinámica operan en fenómenos como la segregación cromosómica, y durante la mitosis, las fuerzas generadas en los cromosomas juegan un papel crucial dirigiendo la segregación cromosómica y modulando las señales fundamentales que permiten la división celular; así, estructuras proteicas como el

cinetocoro del centrómero son esenciales para generar la energía que permite a los cromosomas moverse (Rago y Cheeseman, 2013, para una revisión actual).

Pero, si hubiese un óptimo en este sentido, es decir, hubiese una forma ideal para que el material genético se moviese en el fluido del núcleo celular, ¿no la utilizarían todas las especies? Es decir, ¿no tendrían todas las especies el mismo número de cromosomas? No, claramente, si el número de cromosomas y la división de la información que contienen, es relevante para la configuración de la especie. En definitiva, mantenemos la hipótesis de partida de que la información debe también influir en la división del ADN en cromosomas.

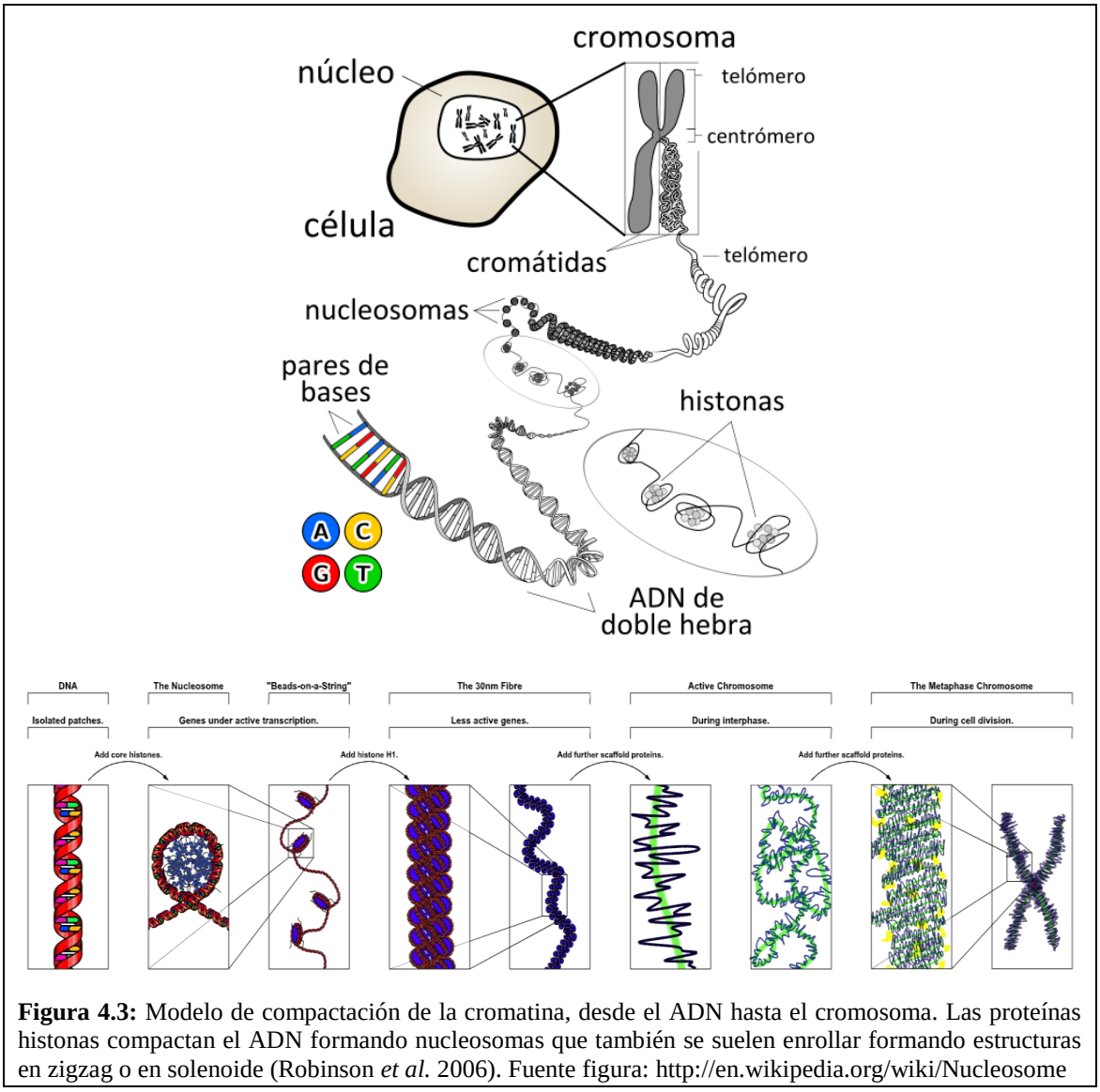


Figura 4.3: Modelo de compactación de la cromatina, desde el ADN hasta el cromosoma. Las proteínas histonas compactan el ADN formando nucleosomas que también se suelen enrollar formando estructuras en zigzag o en solenoide (Robinson *et al.* 2006). Fuente figura: http://en.wikipedia.org/wiki/Nucleosome

4.2. Sobre el tamaño del genoma

La biofísica demuestra que hay un límite superior en el tamaño de los cromosomas para que sea viable el desarrollo de los organismos (Schubert y Oud,

1997). Los organismos eucariotas empaquetan su ADN en cromosomas, que son estructuras dinámicas que configuran su cariotipo. De esta forma, una cadena de nucleótidos que podría ocupar más de un centímetro estirada, se condensa en apenas unos micrómetros, lo que implica una compactación longitudinal de cuatro órdenes de magnitud. Las histonas son las responsables del primer nivel de empaquetamiento del cromosoma: la formación de los nucleosomas que, integrados por ocho histonas y el ADN adyacente, convierten a la molécula de ADN en una hebra de cromatina de un tercio de su longitud original (Alberts *et al.* 2008, pp. 211-212).

Los cromosomas presentan por tanto estructuras sofisticadas con diversos niveles de escala (figura 4.3), según se organizan los millones de pares de bases de ADN en la cromatina (Robinson *et al.,* 2006), y con mecanismos que permiten su movimiento y dinámica (Rago y Cheeseman, 2013). La compresión actúa tridimensionalmente. Se trata de una compresión física, de optimización del espacio, que de entrada no estaría relacionada directamente con el principio de compresión, de tendencia a la minimización de la longitud de los elementos de los sistemas de comunicación, visto en el capítulo anterior (Ferrer-i-Cancho, Hernández-Fernández *et al*., 2013), aunque el paralelismo es evidente y atractivo. En la línea de los argumentos de Chater y Brown (2008), podríamos pensar que los principios de teoría de la información también son válidos para la información química contenida en los cromosomas, y que realmente los genes poseen su propio lenguaje (Searls, 2002).

El grado de optimización de la compactación del ADN depende además del tipo de cromatina: la heterocromatina, se condensa más y el resto, la denominada eucromatina, lo hace menos (Alberts *et al.*, 2008). La heterocromatina está concentrada en centrómeros y telómeros, es decir los extremos (telómeros) y en el punto central (centrómero), de los que hemos hablado previamente, representando el 10% del genoma. La mayor parte del ADN empaquetado en forma de heterocromatina tiene muy pocos genes; los genes eucromáticos que están empaquetados en la heterocromatina resultan inactivados por este tipo de empaquetamiento. Si un gen que normalmente se expresa en la eucromatina es desplazado de forma experimental a una región de la heterocromatina, dejará de expresarse: se dice entonces que el gen está silenciado (Alberts *et al*. 2008, p.220). Dicho de otro modo, la posición de los genes es fundamental para su expresión.

La estructura del ADN en los cromosomas y su dinámica hace ya algún tiempo que condujo a la idea de que operen leyes de escala en la genética (Molina y van

Nimwegen, 2009), tal y como Chater y Brown (2008) sostenían también para los principios de la ciencia cognitiva. A tal efecto, Stanley y colaboradores (1999, p.3) afirmaban:

> In the last decade, scaling analysis (fractal) techniques have been developed for detecting scale-invariant statistical patterns and study physical properties in complex fluids and other random systems. These methods have been successfully applied in a number of disciplines and to a number of problems including stochastic growth processes in physics and chemistry, polymer physics, as well as other problems. Since DNA sequences are long polymer chains, some general scale-invariant properties found in polymer physics may appear in DNA, and alterations of those general properties may serve for characterization of DNA sequences.

La secuenciación genética ha permitido analizar las posibles causas de la evolución del genoma y del cariotipo (Gregory, 2005; Petrov, 2001). El enfoque es ciertamente complejo, aunque están claros bastantes procesos que conducen a mutaciones, inserciones, deleciones, y otros mecanismos fisiológicos (ver figura 4.4, extraída de Gregory, 2005), además de las citadas constricciones biofísicas como el tamaño y longitud de la cadena (Rago y Cheeseman, 2013; Schubert y Oud, 1997).

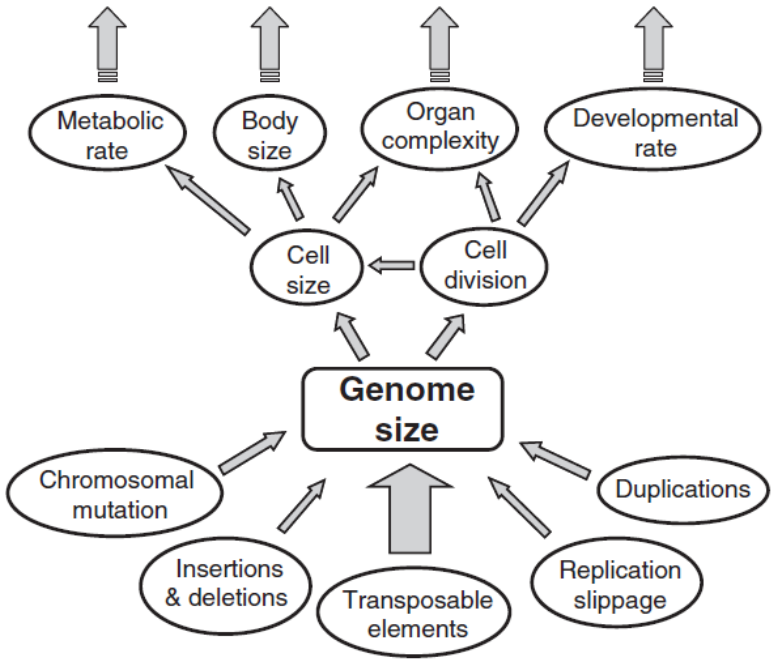


Figura 4.4: Muchos mecanismos en el nivel genómico influyen en el tamaño del genoma y, a su vez, este tamaño está directamente relacionado con el tamaño celular y la biofísica de procesos como la división celular, que posteriormente marcan los procesos fisiológicos y metabólicos del organismo y, finalmente, del ecosistema en el que habita. Extraída de Gregory (2005), p.72.

Un misterio que se añade al problema de la existencia de cromosomas y su tamaño es la cuestión del llamado "ADN basura", es decir, el ADN no codificante. El ADN del genoma de un organismo se divide en dos: el que codifica directamente proteínas y el que no las codifica (ADN basura). En la mayoría de especies, según la visión clásica, al parecer, solo una pequeña parte del ADN codifica proteínas, lo que en el caso de nuestra especie se traduce en que únicamente alrededor del 1,5% del ADN codifica proteínas (Wolfsberg *et al.*, 2001; Alberts *et al.*, 2008). El proyecto Encode[22] (The Encode Project Consortium, 2007) está avanzando en el estudio del ADN no codificante y recientemente argumentó que quizá esta visión no sea del todo acertada, pues el mal llamado ADN basura podría tener funciones bioquímicas muy diversas (Pennisi, 2012; Encode, 2012), citando a Dunham y colaboradores (2012), del proyecto Encode:

> The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.

Pese a ello, el debate está todavía abierto (Doolittle, 2013, para una buena revisión), porque en el fondo se plantean cuestiones lingüístico-filosóficas como qué se entiende por "función" o "evolución" (Doolittle, 2013).

El asunto es complejo y se añade a la llamada *paradoja del valor C*, de la que surgió el concepto de ADN basura, tal vez pecando de cierto antropocentrismo: no puede ser que seres "inferiores" a los humanos, cuyas funcionalidades son más limitadas que las nuestras, posean más genoma que nosotros. El *valor C* se refiere al contenido de ADN nuclear haploide de una especie, que se correlaciona mal con la complejidad o "grado de evolución" de la misma, de manera que los humanos sí tienen miles de genes más que las bacterias, pero muchos menos que muchas plantas o peces (Gregory, 2005). Antropocentrismo aparte, Gregory (2005, p.8-9) nos resume la paradoja, el puzle en su opinión, bajo una perspectiva histórica que merece ser citada:

> The repeated confirmation of a total decoupling of DNA content and organismal complexity (taken as a proxy for gene number) only heightened the confusión surrounding this issue over the two decades that followed. In 1971, C.A. Thomas described this vexing problem as the "C-value paradox," which was (and still is) typically described from one of three different perspectives: (1) Some simple

[22] Las publicaciones y avances de Encode están disponibles en http://www.genome.gov/10005107 y en la web especial de la revista Nature http://www.nature.com/encode/#/threads .

organisms have more DNA than complex ones (e.g., Vendrely, 1955), (2) any given genome seems to contain much more DNA than would be needed to account for the predicted gene number (e.g., MacLean, 1973), or (3) some morphologically similar groups exhibit highly divergent DNA contents (e.g., Gall, 1981). These are three ways of saying the same thing: that DNA content does not correlate with the expected number of genes, which is "paradoxical" in the sensethat C-values were presumed to be constant precisely because DNA is the stuff of genes. As Commoner (1964) put it 40 years ago, "these observations suggest that the DNA/cell does not correspond to the total gene content of the organism, but they fail to explain why, this being the case, the cellular DNA content of a species is, in fact, a fixed inherited characteristic."

Las presiones evolutivas y las fuerzas que conducen al genoma a tener las dimensiones que posee en cada especie son muy diversas (Petrov, 2001; Gregory, 2005). El exceso de ADN no es esencial para el desarrollo o la divergencia evolutiva de los eucariotas (Klug y Cummings, 1999, p.529). Por otra parte, pese a las críticas (Doolittle, 2013), es innegable que ha habido un antes y un después del proyecto Encode (Encode, 2012): la vieja hipótesis del ADN basura debe, cuanto menos, revisarse.

Se abre pues un frente crucial para la lingüística del siglo XXI, insospechado hace apenas medio siglo: entender el genoma, que es comprender la vida y algunos de sus secretos, puede estar en manos de la lingüística genómica. Tras décadas de mejora en las técnicas de secuenciación genética, han aumentado el número de especies y de individuos cuyo genoma ha sido secuenciado (Gregory, 2012; Encode, 2012).

Se han desarrollado diversas bases de datos, así como marcos conceptuales de referencia, que nos permiten ahora aplicar técnicas estadísticas con las que explorar el genoma desde la perspectiva de la lingüística cuantitativa (Bel-Enguix *et al.*, 2011; Köhler, 2005; Searls, 2002; Stanley *et al.*, 1999), para intentar arrojar algo más de luz sobre problemas como el puzle del valor C, el tamaño del genoma o del cariotipo. Esa fue nuestra intención en los artículos incluidos en este capítulo: en nuestro caso, como se verá, básicamente exploramos la presencia de la ley de Menzerath-Altmann en el genoma, así como la relación entre el tamaño del genoma y el número de cromosomas.

4.3. La ley de Menzerath-Altmann en el genoma

En el capítulo 2 se revisó la ley de Menzerath-Altmann, que sostiene que a mayor tamaño de un constructo lingüístico, tienden a ser menores sus partes o constituyentes (Menzerath, 1954; Altmann, 1980; Teupenhayn y Altmann, 1984). Trabajos pioneros como los de Wilde y Schwibbe (1989) o la extensión a más

organismos como en Ferrer-i-Cancho y Forns (2009) supuso una extensión de la ley de Menzerath-Altmann al genoma, y en él concluyeron (Ferrer-i-Cancho y Forns, 2009, p.36), coincidiendo con Chater y Brown (2008) sobre la autoorganización:

> In summary, the genomes with more chromosomes tend to be made of shorter chromosomes in various groups of organisms as words with more syllables tend to be made of shorter syllables in languages. As far as we know, the only previous evidence of Menzerath-Altmann law is only in the genomes of ants. Our findings suggest that selforganization in the sense of organization without global or external control is a general principle of genomes. Organisms within the same group obey the law autonomously, without any group director. Finally, our results suggest that human language and genomes share similar principles of self-organization.

Vinogradov (2001) ya apuntó esta relación en el genoma de plantas monocotiledóneas y dicotiledóneas, aunque sin citar de forma explícita la ley de Menzerath-Altmann. La correlación negativa entre el tamaño del genoma y el número de cromosomas es para Vinogradov (2001) el resultado físico de diferentes mecanismos de recombinación, de forma que por ejemplo en genomas largos aumentan las distancias intergénicas (al haber más ADN no codificante), lo que facilita algunos procesos de recombinación. Curiosamente, Vinogradov (1999) afirmó que este hecho podría extenderse al nivel de los exones: Li (2012) exploró bajo la perspectiva de la lingüística cuantitativa la ley de Menzerath-Altmann en exones, siguiendo la línea de trabajo propuesta en Ferrer-i-Cancho y Forns (2009) y en Hernández-Fernández y colaboradores (2011).

En Hernández-Fernández y colaboradores (2011) se propuso por primera vez, hasta donde sabemos, un diagrama de fases (figura 4.5) en el que representar el número de cromosomas (L_g) de múltiples especies, respecto el tamaño total del genoma en millones de bases nitrogenadas (G). De esta sencilla representación gráfica emergen diversas agrupaciones de puntos, bien diferenciadas en general, en las que se puede trazar un recorrido evolutivo: se aprecia, por ejemplo, una gran dispersión en plantas angiospermas, debida a su comportamiento poliploide[23], mientras que las plantas gimnospermas, más recientes evolutivamente, muestran menos dispersión; los hongos, más antiguos en la evolución, se diferencian claramente, con genomas más breves; los

[23] La poliploidia es el mecanismo por el que se duplica (o multiplica por n) el cariotipo de una especie. Por eso a menudo los estudios cuantitativos se centran en el llamado número haploide de cromosomas, que no considera este efecto.

reptiles están cerca de pájaros y mamíferos; finalmente, los insectos, también filogenéticamente antiguos, muestran también bastante dispersión (Hernández-Fernández *et al.*, 2011).

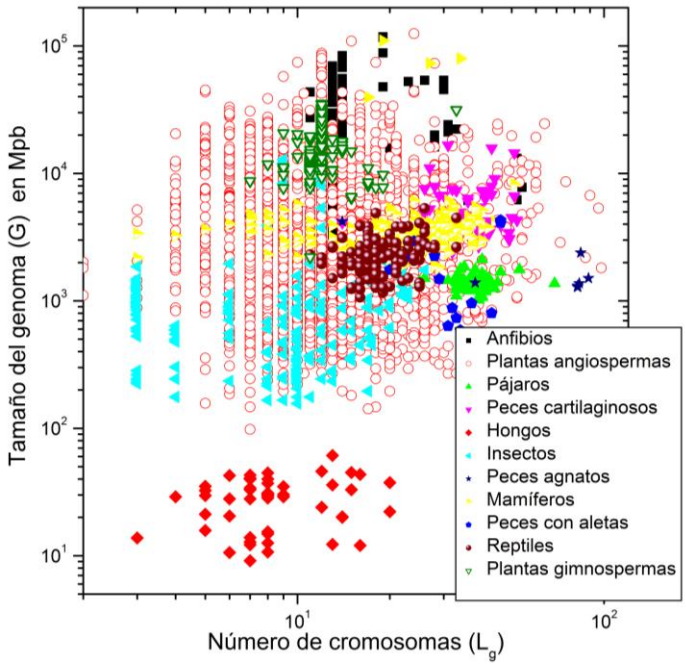


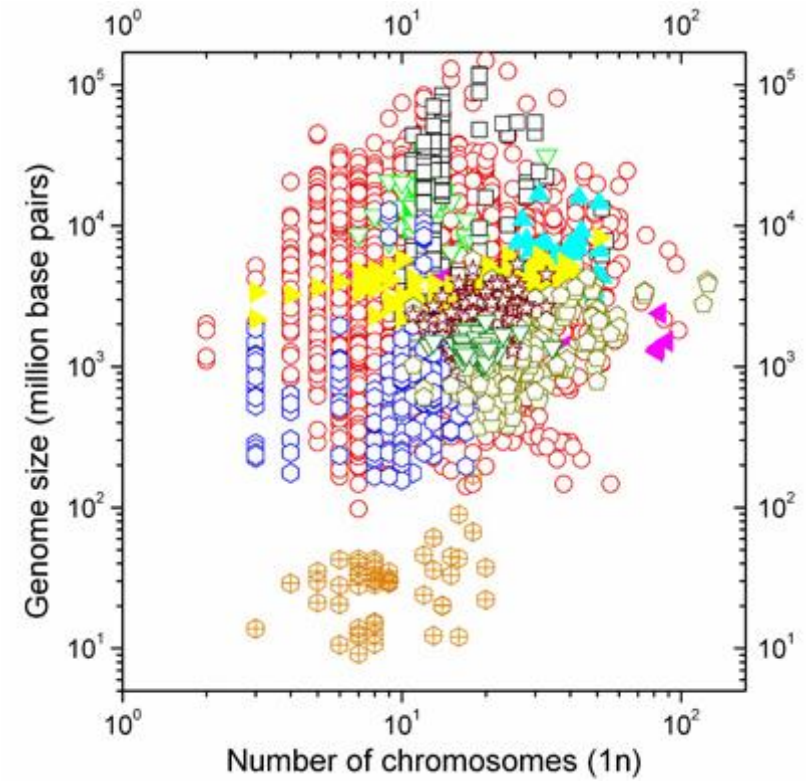
Figura 4.5: Representación del número de cromosomas (*Lg*) respecto el tamaño del genoma (*G*) en millones de pares de bases, de más de 6000 especies secuenciadas (Hernández-Fernández *et al.*, 2011). En Hernández-Fernández y colaboradores (2011) se pueden ver las representaciones para cada grupo por separado.

La evolución del genoma es todavía un misterio, aunque se han apuntado varios correlatos entre el tamaño del genoma y las presiones evolutivas que afectan a algunos rasgos fenotípicos (Petrov, 2001; Gregory, 2005) y el número de cromosomas no se encuentra entre los parámetros considerados. Petrov (2001) en su revisión expone como el tamaño del genoma se relaciona, entre otros rasgos, con la duración de la mitosis y la meiosis, la velocidad del metabolismo basal de mamíferos y pájaros paseriformes, el desarrollo embrionario en algunas salamandras, la complejidad morfológica cerebral de ranas y salamandras o con las emisiones de dióxido de carbono de las plantas (para las citas específicas véase Petrov, 2001, p.24). A la lista de Petrov (2001), pudimos añadir, ampliando el trabajo inicial de Ferrer-i-Cancho y Forns (2009), una correlación significativa (p<0.05) entre el tamaño del genoma (G), en millones de pares de bases, y el número de cromosomas (L_g) para nueve de los once grupos analizados (ver figura 4.5), es decir, todas las especies excepto los pájaros y peces cartilaginosos (Hernández-Fernández *et al.*, 2011).

4.4. Hernández-Fernández, Baixeries, Forns y Ferrer-i-Cancho (2011). *Size of the Whole* versus *Number of Parts in Genomes.*

En este apartado se incluye el artículo:

- Hernández-Fernández, A., Baixeries, J., Forns, N. y Ferrer-i-Cancho, R. (2011). Size of the whole versus number of parts in genomes. *Entropy*, 13 (8), 1465–1480, doi:10.3390/e13081465.



Entropy **2011**, *13*, 1465-1480; doi:10.3390/e13081465

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Size of the Whole *versus* Number of Parts in Genomes

Antoni Hernández-Fernández [1,4], Jaume Baixeries [2], Núria Forns [3] and Ramon Ferrer-i-Cancho [4,*]

[1] Departament de Lingüística General, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona (Catalonia), Spain; E-Mail: antonio.hernandez@upc.edu

[2] LARCA Research Group, Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain; E-Mail: jbaixer@lsi.upc.edu

[3] Departament de Microbiologia, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal 645, 08028 Barcelona (Catalonia), Spain; E-Mail: nuforns@ub.edu

[4] Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain

* Author to whom correspondence should be addressed; E-Mail: rferrericancho@lsi.upc.edu.

Received: 5 July 2011; in revised form: 22 July 2011 / Accepted: 28 July 2011 / Published: 5 August 2011

Abstract: It is known that chromosome number tends to decrease as genome size increases in angiosperm plants. Here the relationship between number of parts (the chromosomes) and size of the whole (the genome) is studied for other groups of organisms from different kingdoms. Two major results are obtained. First, the finding of relationships of the kind "the more parts the smaller the whole" as in angiosperms, but also relationships of the kind "the more parts the larger the whole". Second, these dependencies are not linear in general. The implications of the dependencies between genome size and chromosome number are two-fold. First, they indicate that arguments against the relevance of the finding of negative correlations consistent with Menzerath-Altmann law (a linguistic law that relates the size of the parts with the size of the whole) in genomes are seriously flawed. Second, they unravel the weakness of a recent model of chromosome lengths based upon random breakage that assumes that chromosome number and genome size are independent.

Keywords: Menzerath-Altmann law; genome size; chromosomes

PACS Codes: 87.18.Wd Genomics; 89.75.Da Systems Obeying Scaling Laws; 87.15.A-, Theory, modeling, and computer simulation; 87.16.Sr Chromosomes, histones; 87.14.gk DNA

1. Introduction

Various studies have reported a negative correlation between genome size and number of chromosomes or B chromosomes in angiosperm plants [1,2]. Interestingly, Vinogradov argues that this negative correlation could be explained as a trade-off between different recombination mechanisms [1]. In contrast, it has been argued recently that theoretical models of chromosome length evolution [3,4] "and the current knowledge on the fluid nature of chromosomal rearrangements through time rule **against any special multiscale link between genome-level and chromosome-level patterns.** (boldface is ours)" [5]. Here it will be shown that dependencies between chromosome number and genome size are not a peculiarity of flowering plants, as it may be concluded from the pioneering work of Vinogradov [1], by examining various groups of organisms from different kingdoms: fungi, plants, and animals. As the size of the genomes increases, it will be shown that the number of chromosomes increases in some groups while in others it decreases. Evidence that these dependencies are not simply linear will be provided.

2. Results

N is defined as the number of organisms of a group that is being analyzed. G and L_g are defined, respectively, as length of a genome in million base pairs (Mb) and the size of the genome in chromosomes.

2.1. Correlations between Genome Size and Chromosome Number

Figures 1 and 2 show the relationship between G and L_g for the major groups of organisms analyzed in [6]. It can be seen that certain groups of organisms such as reptiles, birds and fungi, cluster in different regions of the space defined by G and L_g. For certain groups of organisms (e.g., reptiles), a dependency between G and L_g can be seen. However, a rigorous statistical correlation test is necessary. Separate plots of the relationship between G and L_g for each group are provided in Appendix A. Table 1 shows a significant correlation between G and L_g is found in 9 out of 11 groups of organisms at a significance level of 0.05. The only groups where no significant correlation is found are birds and cartilaginous fishes. Therefore, G is not indeed a constant function of L_g for the majority of groups.

2.2. Non-Linearity

Some light on the kind of functional dependency between G and L_g can be shed. If the relationship was purely linear, the point estimation of the slope should not show any dependency with either G or L_g. Table 2 shows that this linearity test (see Methods for further details) rejects the null hypothesis that G is a purely linear function of L_g for all groups (p-value $< 10^{-7}$). Non-linearity is consistent with

the plots in Figures 1 and 2 and in the Appendix A where it can easily be seen that the slope of a linear approximation in double logarithmic scale deviates, in many cases, clearly from one, the expected slope if the relationship was linear. However, our test cannot exclude that linearity is present in some part of the series despite the fact that pure linearity has been rejected for the whole series.

Figure 1. Genome size G (in Mb) *versus* the number of chromosomes L_g (in $1n$) for all the major groups of organisms analyzed in [6] excluding plants, which were plotted separately (Figure 2) due to the high dispersion of angiosperms.

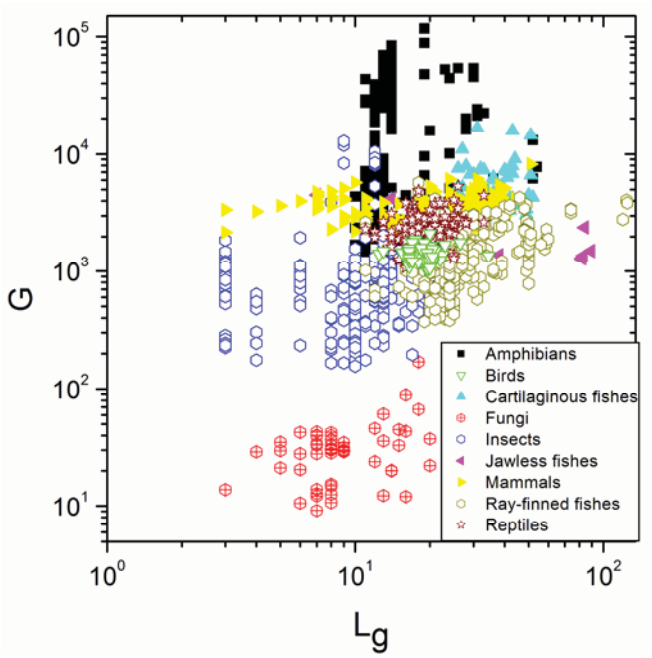


Figure 2. Genome size G (in Mb) *versus* the number of chromosomes L_g (in $1n$) for the major groups of plants analyzed in [6].

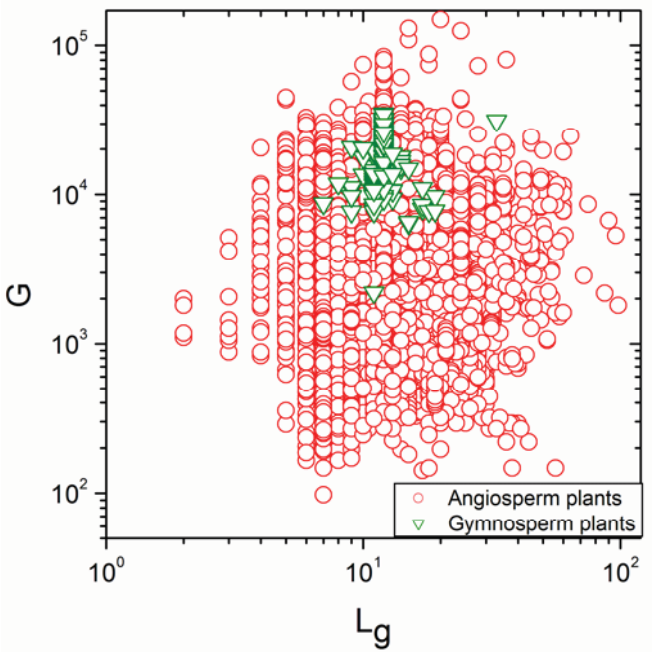


Table 1. Summary of the correlation analysis between genome size G (in Mb) and genome size L_g in number of chromosomes. N, ρ, and p are defined, respectively, as the number of different organisms, the value of Spearman's rank correlation statistic for G *versus* L_g, and the p-value of ρ within a group of organisms. The values of ρ were rounded to leave only three decimals and the p-values were rounded to leave only one significant digit.

Group	***N***	***ρ***	***p***
Fungi	56	0.280	0.04
Angiosperm plants	4706	−0.38	0.008
Gymnosperm plants	170	0.315	3×10^{-5}
Insects	269	0.220	0.0003
Reptiles	170	0.243	0.001
Birds	99	0.008	0.9
Mammals	371	0.297	5×10^{-9}
Cartilaginous fishes	52	−0.129	0.4
Jawless fishes	13	−0.744	0.004
Ray-finned fishes	647	0.487	$<10^{-17}$
Amphibians	315	0.446	9×10^{-17}

Table 2. Summary of the correlation analysis between genome size G (in million base pairs) and $a = (G - c)/L_g$, where L_g is the genome size in number of chromosomes and c is the intercept of a linear approximation of the dependency between G and L_g by a non-parametric linear regression method. N, ρ, and p are defined, respectively, as the number of different organisms, the value of Spearman's rank correlation statistic for G *versus* a, and the p-value of ρ within a group of organisms. The values of ρ were rounded to leave only three decimals and the p-values were rounded to leave only one significant digit.

Group	***N***	***ρ***	***p***
Fungi	56	0.666	2×10^{-8}
Angiosperm plants	4706	0.925	$<10^{-17}$
Gymnosperm plants	170	0.992	$<10^{-17}$
Insects	269	0.802	$<10^{-17}$
Reptiles	170	0.791	$<10^{-17}$
Birds	99	0.771	$<10^{-17}$
Mammals	371	0.278	5×10^{-8}
Cartilaginous fishes	52	0.886	$<10^{-17}$
Jawless fishes	13	0.951	$<10^{-17}$
Ray-finned fishes	647	0.812930	$<10^{-17}$
Amphibians	315	0.983	$<10^{-17}$

3. Discussion

According to Table 1, the dependencies between G and L_g can be classified into three qualitative types:

- "The more parts, the larger the whole"

This is the case of fungi, gymnosperm plants, insects, reptiles, mammals, ray-finned fishes and amphibians.

- "The more parts, the smaller the whole"

This is only the case of angiosperm plants and jawless fishes. A negative correlation between genome size and number of chromosomes in angiosperm plants has previously been reported [1].

- "Other"

Birds and cartilaginous fishes fall into this category, which includes the possibility that the number of parts and the size of the whole are independent. However, independence is not necessarily the only explanation (recall that absence of correlation does not imply independence [7]). We just mention a couple of possibilities. First, the dependency is not monotonic (rank correlation tests of the kind that we have used are more appropriate for strictly monotonically increasing or decreasing functional dependencies). Second, the dataset is not large enough to allow one to unravel the underlying trend for that particular group since only a very small fraction of all the species that actually belong to the groups has been explored (e.g., Table 1.1 of [8]). In sum, absences of correlations are not the rule but the exception in these major groups.

The class "The more parts, the larger the whole" could have simple explanations if G was an increasing linear function of L_g, *i.e.*, $G = aL_g + c$ with $a > 0$. First, imagine that all chromosomes are of about the same size a (and that a does not depend on the number of chromosomes). Then genomes size G would be proportional to L_g, *i.e.*, $G = aL_g$. Second, consider the case of genome duplication. Imagine that a new species is produced by adding k copies copy of the genome of an origin species (with $k = 1$ for genome duplication). The genomes that would be generated by this mechanism would satisfy the relationship $G = aL_g$, where $a = G^0/L_g^0$ would be the ratio between G^0 and Lg^0, respectively, the genome size and the chromosome number of the origin species. Here it has been shown that a linear relationship between G and L_g is not supported for any group. In sum, a purely increasing linear dependency between G and L_g is not supported for any group in our dataset. This has an important biological implication: Simple genome duplication is unlikely to be the only force shaping the class of organisms where "the more parts, the larger the whole".

We have presented a classification into three classes of growth of the whole with regard to its parts at a given taxonomic scale of analysis which does not need to be preserved at lower taxonomic scales. For instance, although angiosperm plants fall into the class "the more parts, the larger the whole", at the level of families, only seven families show this behavior, 22 families show the opposite pattern ("the more parts, the smaller the whole") but an overwhelming number of families, *i.e.*, 194, show no significant part-whole correlation (see the Appendix B for further information on group subdivision). This and other results discussed in the Appendix B mean that these three classes must be interpreted as only valid a priori at their taxonomic scale. The Appendix B also shows that subdividing does not help to unravel a trend in the only two groups where no correlations were found: Birds and cartilaginous fishes.

Our empirical analysis has implications for the debate about the relevance of a connection between human language and genomes through a common pattern: the tendency of the mean size of the parts (syllables or chromosomes) to decrease as the number of parts of the whole (a word or a genome) increases [6]. This pattern is known as Menzerath-Altmann law in quantitative linguistics [9] and is

found not only in language at many levels of description but also in music (see [10] and references therein). According to [5], the finding of this negative correlation between the mean size of the parts and the number of parts in genomes is a trivial consequence of the definition of the size of the parts, L_c as a mean, *i.e.*, $L_c = G/L_g$, which leads to $L_c = a/L_g$ where a is a constant. However, $L_c = a/L_g$ holds if and only if G is a constant function of L_g. In other words, the relationship between the mean size of the parts and the number of parts is trivial if and only if G is constant. In contrast, here it has been shown that G and L_g are significantly correlated in many groups of organisms. The classes "The more parts, the larger the whole" and the classes "The more parts, the smaller the whole" violate the constancy assumption of [5]. Furthermore, it has been shown that, when such significant correlation is not found, the possibility that this is due to the small size of the group sample cannot be denied. Notice that [5] evaluates the goodness of the fit of $L_c = a/L_g$ to actual data with a flawed test, which consists of fitting $L_c = a/L_g^b$ to actual data. If $b = -1$ is obtained this implies that the hypothesis $L_c = a/L_g$ is correct, according to [5]. However, obtaining $b = -1$ from data is a necessary but not a sufficient condition for $L_c = a/L_g$. In contrast, here we have investigated a sufficient condition for $L_c \neq a/L_g$: if G is not a constant function of L_g then $L_c = a/L_g$ cannot be true, at least in some region.

Similarly, our findings unravel the weakness of a random breakage model of chromosome lengths that has been proposed recently [5]. In this model, the information about a certain organism is generated in the following way:

- G is chosen uniformly at random within the interval (G^m, G^M).
- L_g (the number of chromosomes of the organism) is chosen uniformly at random within the interval (L_g^m, L_g^M).
- Chromosome lengths are produced from G and L_g following a random breakage procedure [11,12].

Interestingly, G and L_g are chosen independently in this model. Such independence is totally unrealistic as our analyses and previous research [1] have revealed. Notice that the independence between G and L_g needs (if genomes with chromosomes of length zero are considered as not allowed or totally unrealistic) that the condition $L_g^M \leq G^m + 2$ is satisfied so that all chromosomes can have length greater or equal than one. This condition follows from $L_g \leq L_g^M - 1$, $G^m + 1 \leq G$ and the condition for non-empty chromosomes, *i.e.*, $L_g \leq G$.

Our study is just one among many evidences of the "multiscale link between genome-level and chromosome-level" that the random breakage model above and accompanying arguments deny [5]. Laboratory experiments indicate that "upper and lower tolerance limits for chromosome size are apparently determined by the genome size, chromosome number and karyotype structure of a given species" (see [13] and references therein). Along these lines, a recent statistical study shows that it is possible to predict, for a given species, chromosome sizes by chromosome number, and furthermore, given either genome size or average chromosome length it is possible to predict the size range of all chromosomes of that species [14].

Future work should address the question of the precise mathematical form of the dependency between chromosome number and genome size. By having shown its statistical significance and excluded that it is trivially linear for all groups, the foundations for further research have been established and the actual scope of multiscale links between the genome and the chromosome level has

been clarified. Our selection of groups of organisms was motivated by [5,6] but the same analysis should be extended to other groups of organisms in the future.

4. Methods

4.1. Data

For consistency with [6], the same major groups of organisms (listed on Table 1) were used. The information about each organism was retrieved in June 2011 from the same databases of [6]. The same methods of [6] for filtering incorrect data were applied.

4.2. A Test of Pure Linearity between G and L_g

G is a purely linear function of L_g, if and only if $G = aL_g + c$, where a and c are constants. If G was a purely linear function of L_g, one would have that $a = (G - c)/L_g$ is a constant function of G with c obtained from least squares linear regression. A two-sided Spearman rank correlation test was used to determine if there is a correlation between $(G - c)/L_g$ and G. Notice that here the term 'pure' or 'purely' is not used to mean that the relationship between G and L_g is deterministically linear but to mean indeed that $E[G|L_g]$, the expectation of G given L_g is exactly linear, *i.e.*, $E[G|L_g] = aL_g + c$. The general assumption of regression (and also ours) is that $G = E[G|L_g] + \varepsilon$, where ε is an error that is typically assumed to be normally distributed with mean zero and constant standard deviation [15]. However, a non-parametric linear regression method, Theil's incomplete method [16], was used to estimate a. This method has the following advantages over a simple parametric least squares linear regression [16]:

- It does not assume that all the errors are only in the y-direction.
- It does not assume that either the x- or y-direction errors are normally distributed.
- It is robust in the sense that it is not affected by the presence of outliers.

Acknowledgments

We are grateful to J. Perarnau and X. Messeguer for helpful discussions. This work was supported by the project SESAAME-BAR (TIN2008-06582-C03-01) of the Spanish Ministry of Science and Innovation.

References and Notes

1. Vinogradov, A.E. Mirrored genome size distributions in monocot and dicot plants. *Acta Biotheoretica* **2001**, *49*, 43–51.
2. Trivers, R.; Burt, A; Palestis, B.G. B chromosomes and genome size in flowering plants. *Genome* **2004**, *47*, 1–8.
3. Sankoff, D.; Ferretti, V. Karyotype distributions in a stochastic model of reciprocal translocation. *Genome Res.* **1996**, *6*, 1–9.

4. De, A.; Ferguson, M.; Sindi, S.; Durrett, R. The equilibrium distribution for a generalized Sankoff-Ferretti model accurately predicts chromosome size distribution in a wide variety of species. *J. Appl. Probab.* **2001**, *38*, 324–334.
5. Solé, R.V. Genome size, self-organization and DNA's dark matter. *Complexity* **2010**, *16*, 20–23.
6. Ferrer-i-Cancho, R.; Forns, N. The self-organization of genomes. *Complexity* **2009**, *15*, 34–36.
7. DeGroot, M.H. *Probability and Statistics*, 2nd ed.; Addison-Wesley: Reading, MA, USA, 1989; p. 215.
8. Gregory, T.R. Genome size evolution in animals. In *The Evolution of the Genome*; Gregory, T.R., Ed.; Elsevier: San Diego, CA, USA, 2005; pp. 4–71.
9. Altmann, G. Prolegomena to Menzerath's law. *Glottometrika* **1980**, *2*, 1–10.
10. Boroda, M.G.; Altmann, G. Menzerath's law in musical texts. *Musikometrika* **1991**, *3*, 1–13.
11. Fuquan, K.; Kui, Z.; Yong, Z.; Tianguang, C.; Meinan, N.; Li, S.; Minghui, C.; Yizhong, Z. Analysis of length distribution of short DNA fragments induced by ^{7}Li ions using the random-breakage model. *Chin. Sci. Bull.* **2005**, *50*, 841–844.
12. Becker, T.S.; Lenhard, B. The random *versus* fragile breakage models of chromosome evolution: A matter of resolution. *Mol. Genet. Genomics* **2007**, *278*, 487–491.
13. Schubert, I. Chromosome evolution. *Curr. Opin. Plant Biol.* **2007**, *10*, 109–115.
14. Li, X.; Zhu, C.; Lin, Z.; Wu, Y.; Zhang, D.; Bai, G.; Song, W.; Ma, J.; Muehlbauer, G.J.; Scalon, M.J.; *et al.* Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. *Mol. Biol. Evol.* **2011**, doi:10.1093/nar/gkl828.
15. Ritz, C.; Streibig, J.C. *Nonlinear Regression with R*; Springer: New York, NY, USA, 2008.
16. Miller, J.C; Miller, J.N. *Statistics for Analytical Chemistry*, 3rd ed.; Prentice Hall: London, UK, 1993; pp. 159–161.
17. Pearl, J. *Causality: Models, Reasoning, and Inference*; Cambridge University Press: Cambridge, UK, 2000.

Appendix A

The relationship between genome size and chromosome number is shown in Figure 3 for fungi, Figure 4 for angiosperm plants, Figure 5 for gymnosperm plants, Figure 6 for insects, Figure 7 for reptiles, Figure 8 for birds, Figure 9 for mammals, Figure 10 for cartilaginous fishes, Figure 11 for jawless fishes, Figure 12 for ray-finned fishes and Figure 13 for amphibians.

Figure 3. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for fungi.

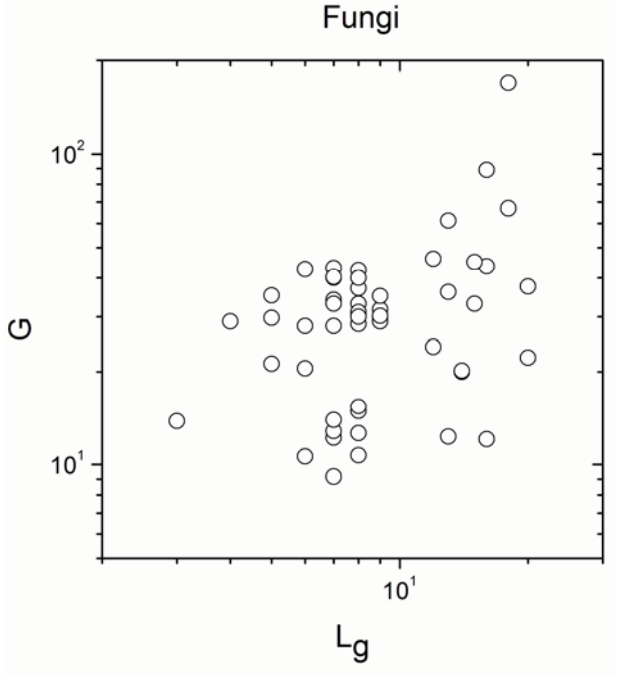


Figure 4. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for angiosperm plants.

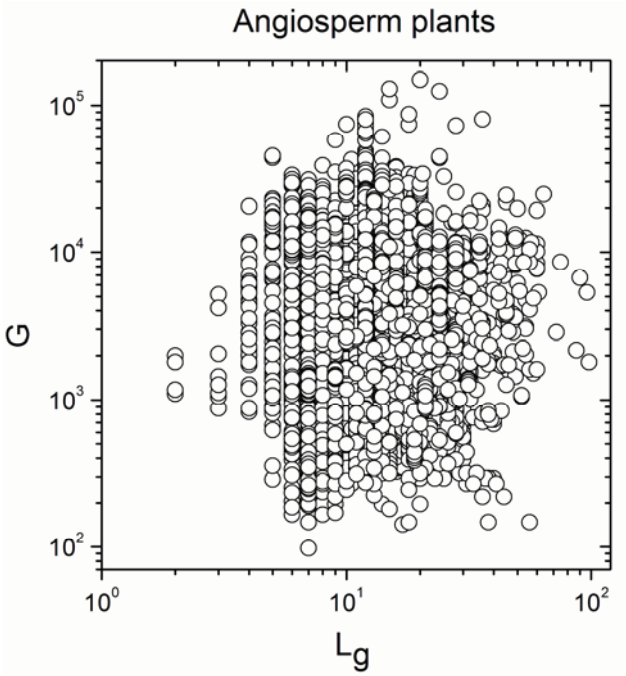


Figure 5. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for gymnosperm plants.

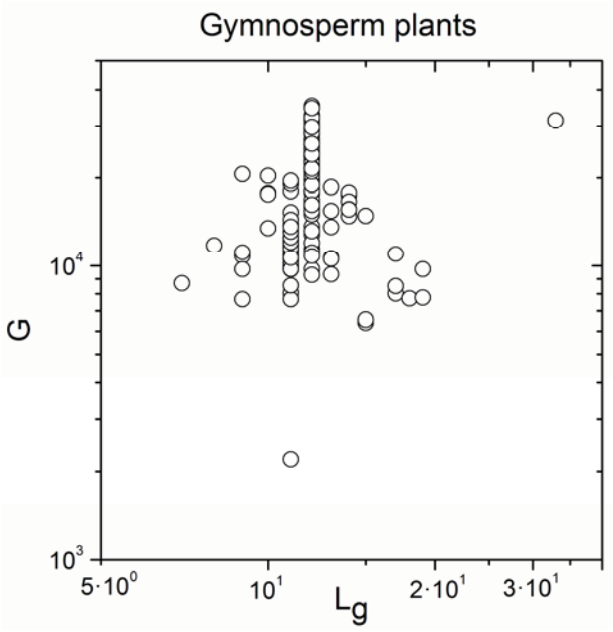


Figure 6. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for insects.

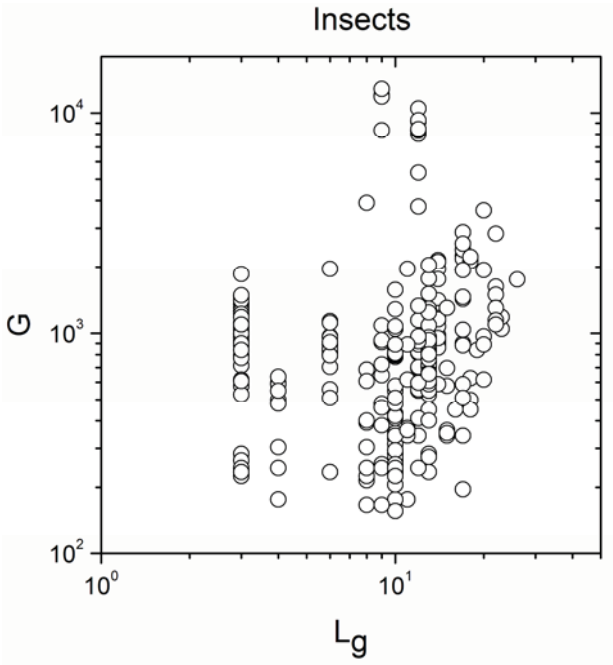


Figure 7. Genome size *G* (in Mb) *versus* the number of chromosomes L_g (in 1*n*) for reptiles.

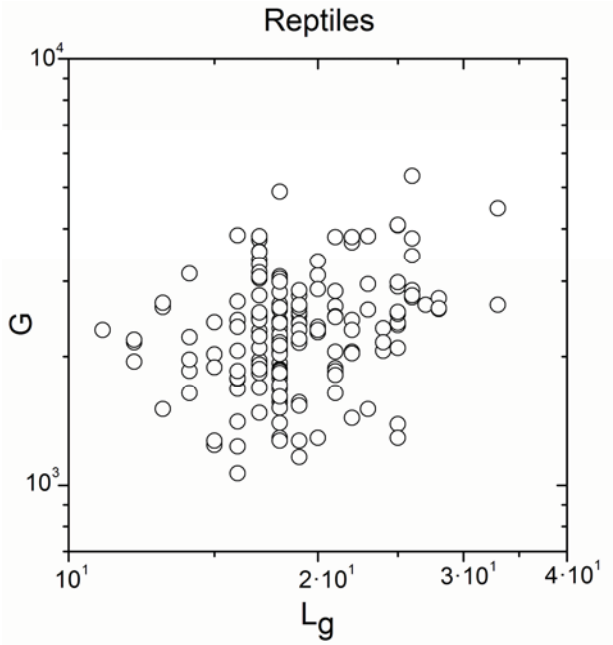


Figure 8. Genome size *G* (in Mb) *versus* the number of chromosomes L_g (in 1*n*) for birds.

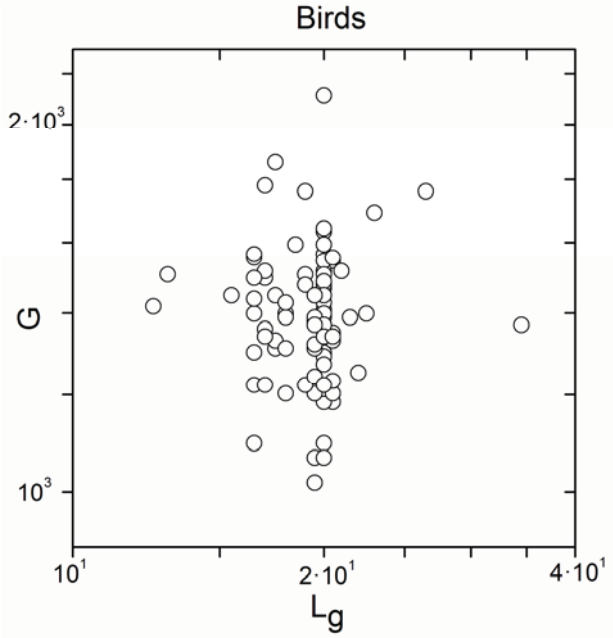


Figure 9. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for mammals.

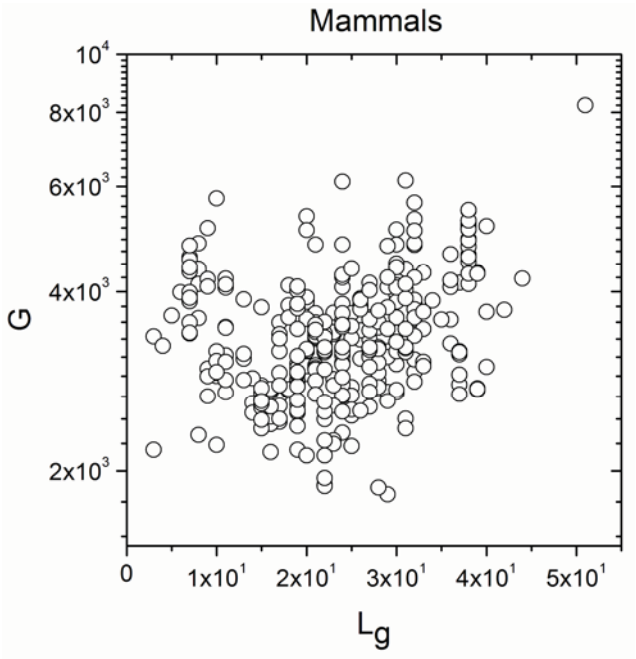


Figure 10. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for cartilaginous fishes.

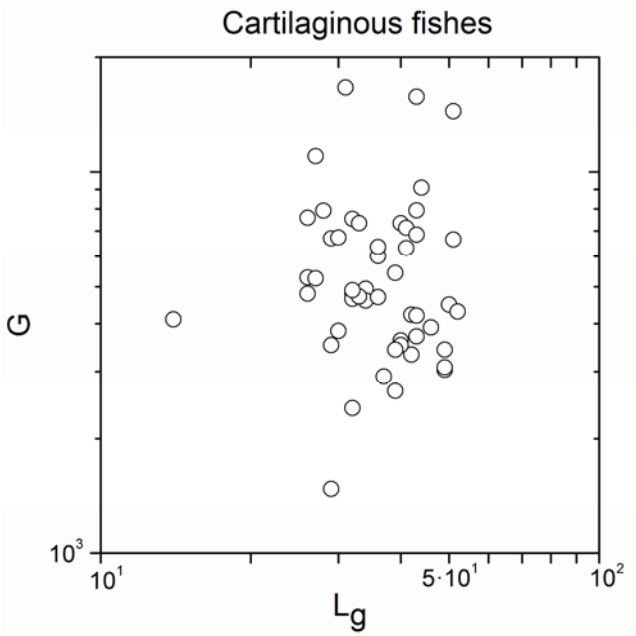


Figure 11. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for jawless fishes.

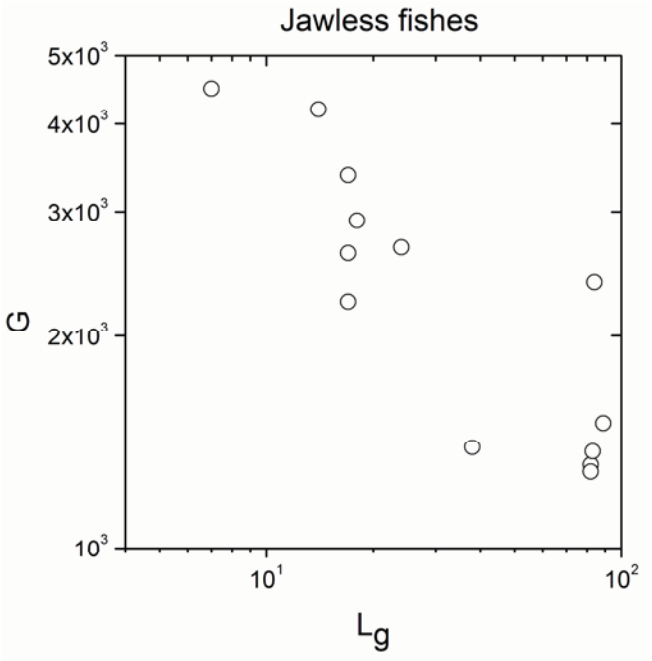


Figure 12. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for ray-finned fishes.

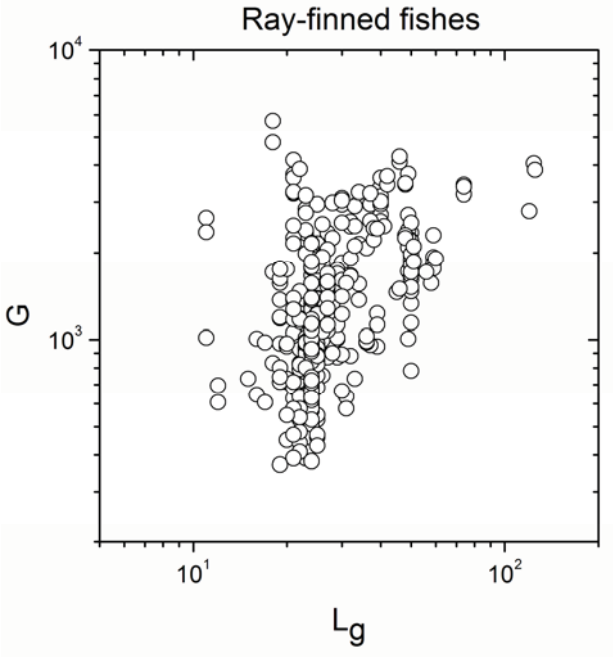
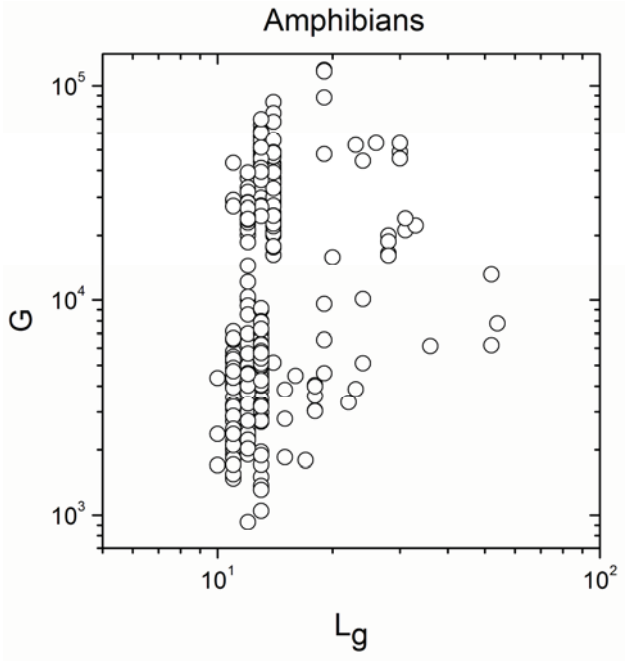


Figure 13. Genome size G (in Mb) *versus* the number of chromosomes L_g (in 1n) for amphibians.



Appendix B

Simpson's paradox [7,17] suggests that the conclusions about the correlations between G and L_g for a certain groups of organisms (Table 1) could change when these groups are subdivided using taxonomic information. Subdividing could yield paradoxical results such as (a) that a group of organisms shows no significant dependency but its subgroups do show a significant correlation or the opposite, that the significant correlation of the group is lost in the subgroups [7] or (b) that the sign of the significant correlation of the original group is the opposite of that of its subgroups [17].

When attempting to study how that correlation changes when taxonomic subgroups are considered, various serious problems were encountered. First, the necessary taxonomic information is not available for all species in public genome size databases. This is especially worrying for fungi, where the amount of missing information is massive. Second, due to the very limited coverage of the genome size databases, taxonomic subdivisions may contain only one subgroup or a few unless the taxonomic subgroup is low enough. Thirdly, at low taxonomic levels, subgroups turn out to have so little members that no significant correlations can be detected in the majority of them. The few significant correlations may not be representative of that scale of analysis due to the very limited coverage of genome size databases. Table 3 summarizes the results of the analysis of the dependency between the size of whole and the size of the parts at lower taxonomic levels within each original group. For simplicity, for each taxonomic sublevel, only those sublevels for which the group yielded more than one subgroup are considered.

Table 3. Summary of the correlations between genome size (G) and chromosome number (L_g) at different taxonomic levels. Boldface is used to indicate the taxonomic groups that are the target of our main analysis. +, −, ? are attached to the name of each target group to indicate, respectively, that the correlation between G and L_g was significant and positive, significant and negative, and none of them (at a significance level of 0.05). Below each target group of organisms, the total number of organisms in our dataset is shown. In each cell for which taxonomic data is available, a triple of numbers is shown above and a pair of numbers is shown below. The triple follows the format x,y,z, where x, y are respectively, the number of subgroups with significant positive and significant negative correlations, and z is the total number of subgroups. The pair follows the format x',y', where x' and y' are the number of organisms involved in significant positive and significant negative correlations, respectively.

Kingdom	**Phylum/Division**	**Class**	**Order**	**Family**	**Genus**
Fungi + 56	0,3,5 0,55		0,4,5 0,34		0,1,40 0,5
Plants	**Angiosperm** − 4706			22,7,194 2374,965	66,8,1114 1608,186
	Gymnosperm + 170			0,4,14 0,122	0,2,52 0,13
Animals	Arthropoda	**Insects** + 269	3,1,7 189,56	0,1,26 0,13	
	Chordata	**Reptiles** + 170	0,0,4 0,0	1,1,34 14,18	
		Birds ? 99	0,0,17 0,0	0,0,33 0,0	
		Mammals + 371	2,1,17 162,54	5,0,63 89,0	
		Cartilaginous fishes ? 52	0,1,9 0,24	1,0,20 7,0	
		Jawless fishes − 13	0,0,2 0,0	0,0,2 0,0	0,0,2 0,0
		Ray finned fishes + 647	4,0,30 262,0	3,0,115 214,0	
		Amphibians + 315	1,0,3 185,0	3,1,26 42,72	

To scrutinize the results of Table 3, we consider two definitions of Simpson's paradox: (a) the reversing of the sign of significant correlation between G and L_g when splitting a group into subgroups (b) the emergence or the loss of significant correlations between G and L_g when splitting a group into subgroups. Table 3 shows that, after splitting,

- The sign of the significant correlations was totally reversed, in full agreement with definition (a) of Simpson's paradox, only in fungi and gymnosperm plants.
- The sign of the significant correlation was totally maintained only in ray-finned fishes.

- The significant correlation was lost in jawless fishes, in agreement with definition (b) of the paradox.
- Significant correlations became a mixture of positive and negative correlations in angiosperm plants, insects, reptiles, mammals and amphibians.
- Non-significant correlations remained totally for birds.
- Significant correlations emerged only exceptionally in cartilaginous fishes (the number of significant correlations was very small with regard to the total number of subgroups), consistently with definition (b) of the paradox, but the sign of the correlation was not coherent.

This suggests that, with the currently available data, Simpson's paradox is only supported in some groups: Fungi, gymnosperm plants, jawless fishes and cartilaginous fishes. The limited coverage of genome sizes databases cannot exclude that the paradox appears in more groups when more organisms are added but also, the opposite effect could be found, namely, that the paradox disappears when more species are included.

4.5. Debate sobre la ley de Menzerath-Altmann en el genoma

En Hernández-Fernández y colaboradores (2011) se demostró que la relación entre el tamaño del genoma, *G*, y el número de cromosomas, L_g, no es lineal, pese a afirmaciones previas en esa dirección (Solé, 2010). La sencillez de un modelo –navaja de Ockham– no puede estar por encima de su capacidad explicativa y de concordancia con los datos. De hecho, nuestra perspectiva ya contemplaba que las relaciones entre *G* y L_g pudieran ser diversas y según los datos analizados (Hernández-Fernández *et al.*, 2011), se estudiaron todos los grupos de organismos propuestos en Ferrer-i-Cancho y Foros (2009) y se revisó la concordancia de la ley de Menzerath-Altmann, revisándose posibles efectos estadísticos como la paradoja de Simpson. Forma parte inherente de la ciencia el debate científico. Todo constructo debe ser revisable por la comunidad científica (Bunge, 2001).

La crítica de Solé (2010) se centra en Ferrer-i-Cancho y Forns (2009), y se resume en el inicio de su artículo (Solé, 2010, p.20):

> Chromosomes exhibit several features indicating that its spatiotemporal dynamics is self-organized. It has been recently suggested that a negative correlation between genome size and mean chromosome number would also be a fingerprint of selforganization, related to how human language is organized at the level of words and syllables. However, the vast dominance of non-coding DNA in eukaryotic genomes should prevent an interpretation of genome/chromosome size based on functional trade-offs related to information storage and transmission. Moreover, the reported negative correlation is shown to be an inevitable consequence of the definitions of chromosome and genome length and it is thus unrelated to any type of special generative process.

A tal efecto, tras la publicación crítica de Solé (2010), realizamos una revisión de nuestros argumentos y línea de investigación que se había originado en Ferrer-i-Cancho y Forns (2009), continuada en Hernández-Fernández y colaboradores (2011) y que se tradujo en varias respuestas y revisiones (Ferrer-i-Cancho *et al.*, 2012; Baixeries *et al.*, 2012; Ferrer-i-Cancho, Forns *et al*., 2013; Baixeries *et al.*, 2013) que recopilamos en el apartado siguiente y que han reforzado nuestra perspectiva más que debilitarla, como se argumentará.

Coincidimos con Solé (2010) en entender que la dinámica espaciotemporal de los cromosomas está autoorganizada, aunque las discrepancias de Solé (2010) apuntan a la trivialidad e irrelevancia de la ley de Menzerath-Altmann en el genoma, a la exageración de la metáfora comparativa entre el lenguaje y el nivel genómico estudiado

y a la dominancia del ADN no codificante en cualquier estudio global del genoma. Huelga decir que el espíritu crítico es uno de los pilares de la ciencia, y tanto la autocrítica como la duda razonable –y razonada– han estado presentes en nuestras publicaciones sobre la presencia de la ley de Menzerath-Altmann en el genoma, que no es todavía un tema cerrado, ni mucho menos. Hemos mostrado nuevas conexiones conceptuales entre el genoma y el lenguaje (Ferrer-i-Cancho, Forns *et al.*, 2013) y recopilado evidencia empírica y estadística al respecto (Hernández-Fernández *et al.*, 2011; Baixeries *et al.*, 2013). Nos hemos esforzado, además, en responder y en revisar nuestro trabajo (Ferrer-i-Cancho, Baixeries *et al.*, 2013). La cuestión es cómo se argumenta para defender las propias tesis y qué pruebas se aportan. Replicar experimentos o análisis estadísticos, a partir de datos, hace posible la ciencia moderna.

Para empezar, Solé (2010) sostiene que la preponderancia del ADN no codificante en el genoma –la visión clásica– invalida los estudios y la perspectiva de la lingüística genómica a escala cromosómica. Esta perspectiva tradicional, de alguna manera, en la que el ADN codificante es mínimo en comparación con el ADN basura (considerado a veces un *simple* esqueleto), y que conduce a la paradoja o puzle del valor C (Gregory, 2005; Petrov, 2001), como hemos visto, debe revisarse profundamente (Encode, 2012; Doolittle, 2013). De hecho, ahora más que nunca estudios como los defendidos aquí se antojan imprescindibles, precisamente para llegar a comprender hasta qué punto es relevante el que parecía ADN basura, pues puede ser que no sea ADN meramente estructural o inútil.

El debate sobre la redundancia en genómica se ha explorado desde perspectivas diversas (Krakauer y Plotkin, 2002) y en el lenguaje también hay un porcentaje elevado de redundancia e información aparentemente superflua que, no obstante, está presente en los sistemas lingüísticos y es relevante (Shannon, 1950). La redundancia puede poseer valor funcional, como sucede en el lenguaje, por lo que podría ser imprescindible también en el genoma. De hecho, hay dos líneas argumentales paralelas, entre el problema del ADN y la lingüística, que deberían quedar claras, que refutan los argumentos de Solé (2010) y que van en ambas direcciones de la investigación actual (Ferrer-i-Cancho, Forns *et al.*, 2013):

a) El mal llamado "ADN basura" parece que no es tal (Encode, 2012) y por tanto el genoma no está tan lejos del lenguaje en el sentido de contener redundancia y elementos sin "significado" directo.

b) En las lenguas encontramos también elementos sin significado, en el sentido léxico, asociados en general a funciones sintácticas, que podrían entenderse como "basura", bajo un análisis superficial, lo que reforzaría la conexión y los paralelismos entre el lenguaje y el genoma.

El hecho de que un fichero de texto sea comprimible es una demostración empírica del grado de redundancia del lenguaje. La detección y eliminación de la redundancia, por poner un ejemplo comparativo, es un problema clásico en los sistemas automáticos de generación de resúmenes y en la lingüística de corpus (Plaza, 2010), y en la actualidad se aplican técnicas automáticas de detección de la redundancia genética (Chen *et al.*, 2010). Chen y colaboradores (2010) concluyen en su estudio:

> Identifying redundancy is a complex problem in which gene pairs may be redundant in some phenotypes but not others. However, the results indicate that there is enough generality in the outcome of gene duplication to classify redundancy based on evidence from disparate phenotypes

En nuestra opinión, en la línea de Encode (2012), el mal llamado ADN basura plantea un reto y no podemos reducirnos a estudiar el ADN codificante si realmente queremos comprender el problema del ADN y el intrigante origen de la vida. La redundancia es esencial por ejemplo en la organización del genoma eucariota, especialmente en las secuencias del ADN telomérico que se repiten dando estabilidad e integridad de los cromosomas (Klug y Cummings, 1999, p.533).

En segundo lugar, está la cuestión de la relación de dependencia algebraica entre la longitud media de los cromosomas, L_c, y el tamaño del genoma, L_g, en pares de bases. Solé (2010) presenta solo el análisis de dos grupos de especies (mamíferos y un agrupamiento de todas las plantas) de los once grupos de organismos estudiados en Ferrer-i-Cancho y Forns (2009) y en Hernández-Fernández y colaboradores (2011), para concluir que las correlaciones negativas entre L_c y L_g son una consecuencia inevitable de las definiciones de cromosoma y genoma, es decir, trivial. Como exponemos en nuestro trabajo (Ferrer-i-Cancho, Forns *et al.*, 2013; Baixeries *et al.*, 2012), la ley de Menzerath-Altmann adaptada al caso del genoma se puede escribir como (Altmann, 1980):

$$L_c = aL_g^b e^{cL_g} \quad (1)$$

Donde *a*, *b* y *c* son constantes. Para Solé (2010) la ecuación anterior se reduce a $L_c = aL_g^{-1}$, es decir, *c*=0 y *b*=-1. Solé (2010) argumentó que la definición de L_c como un promedio, es decir $L_c = G/L_g$, con *G* la longitud total del genoma de una especie, implicaba una relación trivial de dependencia entre L_g y L_c, es decir $L_c = aL_g^{-1}$. Hemos demostrado que tal deducción es matemáticamente errónea (Ferrer-i-Cancho *et al.*, 2014) porque solo es válida para el caso particular en el que el tamaño del genoma no depende en promedio del número de cromosomas.

En nuestra revisión (Baixeries, Hernández-Fernández *et al.*, 2012) demostramos que, si se asumiera *c*=0, y se ajustase $L_c = aL_g^b$, más de la mitad de los grupos analizados (los mismos que en Hernández-Fernández *et al.*, 2011) no correlacionarían significativamente con *b*=-1 ($p<0.05$), y en general, excepto para pájaros, insectos, peces cartilaginosos y agnatos (cuatro de los once grupos), es significativamente mejor el ajuste de la ecuación 1 completa, con la constante *c* significativamente diferente de cero en cinco de los once grupos (Baixeries *et al.*, 2012). Un resumen de las reanálisis cuantitativas efectuadas se incluye en la tabla 4.1 (Ferrer-i-Cancho *et al.*, 2012; Baixeries, Hernández-Fernández *et al.*, 2012).

Ya en Hernández-Fernández y colaboradores (2011) vimos que en los pájaros y peces cartilaginosos no se obtuvo una correlación entre el número de cromosomas y el tamaño del genoma, o que las correlaciones de hecho eran negativas en plantas angiospermas y peces agnatos. En el caso de angiospermas la correlación negativa coincide con la encontrada por Vinogradov (2001), y podría estar relacionada con el carácter poliploide del genoma de las plantas (Gregory, 2005), y en el caso de los peces agnatos solo se incluyeron en el estudio las trece especies secuenciadas hasta la fecha en los corpus utilizados (Hernández-Fernández *et al.*, 2011), que son realmente pocas para este grupo.

Es obvio que se puede criticar fácilmente todo estudio fundamentado en bases de datos diciendo que ‘son pocos datos’, pues faltan muchísimas especies todavía por secuenciar, y ello influye sin duda en el resultado de los ajustes. No obstante, no es menos cierto que el hecho de que las bases de datos sean limitadas no es óbice para no explorar la presencia de regularidades estadísticas o leyes en las mismas: mal les habría ido, por ejemplo, a los primeros astrónomos si no hubiesen analizado los pocos datos de los cuerpos celestes que tenían, hasta que llegó Tycho Brahe, con su minucioso catálogo

estelar, para darle a Johannes Kepler una suculenta base de datos con la que deducir sus leyes del movimiento celeste.

Otra de las críticas de Solé (2010) se refiere al inapropiado uso de metáforas entre la lingüística y el estudio del genoma. Coincidimos con él, como no puede ser de otra manera, en que aunque las metáforas interdisciplinares han ayudado mucho al avance científico no deben usarse alegremente; precisamente por eso hemos reforzado nuestra postura, y la metáfora de partida, profundizado en la estadística y en análisis de corpus. La comparación entre el genoma formado por cromosomas y un texto formado por palabras, se ha acompañado de datos cuantitativos y de los pertinentes estudios y test estadísticos. No vemos, por otra parte, que Solé (2010) se haya aplicado lo que predica a sí mismo: su análisis presenta sesgos, con métodos de análisis superficiales y poco rigurosos, pues solo incluye los grupos que no contradicen su hipótesis de trivialidad de la ley de Menzerath-Altmann en el nivel cromosómico (Tabla 4.1).

Grupo analizado	Correlación entre L_c y L_g	$b \neq -1$ con $c=0$	$c \neq 0$	$b \neq -1$ con $c \neq 0$
Hongos	Sí	Sí		Sí
Angiospermas	Sí	*	Sí	Sí
Gimnospermas		Sí*	Sí	Sí
Insectos	Sí	Sí		Sí
Reptiles	Sí	Sí		Sí
Pájaros	Sí			
Mamíferos	Sí	*	Sí	
Peces cartilaginosos	Sí			
Peces agnatos	Sí	Sí		Sí
Peces con aletas		Sí	Sí	Sí
Anfibios	Sí	Sí	Sí	Sí

Tabla 4.1: Resumen de los estudios de la dependencia entre la longitud media de los cromosomas, L_c, y el tamaño del genoma, L_g (adaptada de Baixeries *et al.*, 2012) ajustando: $L_c = aL_g^b e^{cL_g}$. En la cuarta columna, a la derecha, se indica cuando es el mejor ajuste, pero con $b \neq -1$ y $c \neq 0$; la primera columna indica cuando hay correlación significativa entre L_c y L_g (Ferrer-i-Cancho y Forns, 2009; Hernández-Fernández *et al.*, 2011); en la segunda recogemos los casos en los que significativamente $b \neq -1$, indicando con * los grupos estudiados por Solé (2010); en la tercera columna se indican los grupos en los que significativamente $c \neq 0$ (Baixeries *et al.*, 2012).

Por otra parte, averiguar si existen leyes de escala en el genoma implica no renunciar a explorar ningún nivel de estudio bajo una perspectiva interdisciplinar (Bel-Enguix y Jiménez-López, 2011). Molina y van Nimwegen (2009) han explorado

diversos niveles funcionales del genoma en procariotas, nuestro trabajo (Hernández-Fernández *et al.*, 2011) se centró en el nivel cromosómico y el de Li (2012) en los exones únicamente del genoma humano, quizá más en concordancia con la crítica de Solé (2010) de apuntar a las partes del genoma "con significado". El estudio de Li (2012) apoya el uso de la lingüística genómica y sus conclusiones se centran en los exones del genoma humano, y nos cita como precursores de su trabajo, aunque a su juicio considera inapropiado el nivel de estudio cromosómico (Li, 2012):

> Previous discussion of a linguistic law called Menzerath's law (the longer a word, the shorter the syllables) in the genomic context was focused on the genome-chromosome-base level (the more number of chromosomes in a genome, the smaller the chromosome size). We apply this linguistic metaphor to more appropriate levels of gene, exon, and base. Using the human gene data, we found that the Menzerath's law at these levels holds true: the more number of exons in a gene, the shorted the averaged exon size.

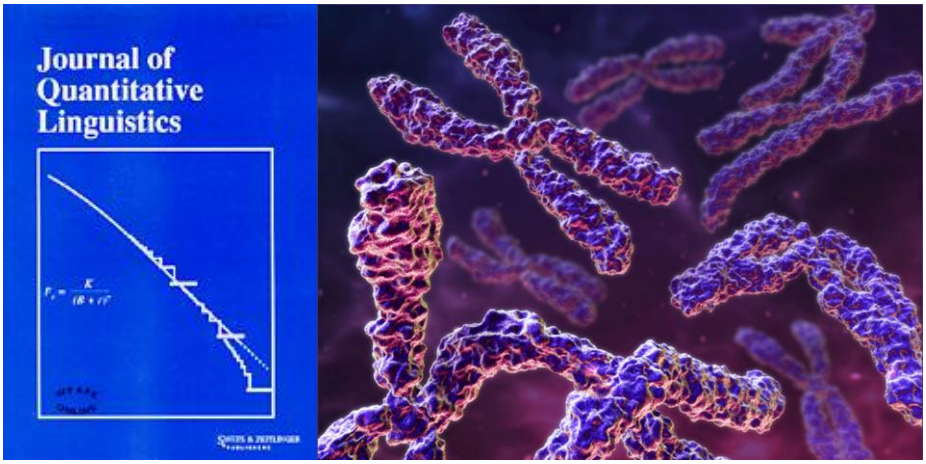
Sin embargo, las evidencias de Encode (2012) son incluso para los más críticos (Doolitle, 2013) un acicate para proseguir en una investigación global del genoma que no excluya ni algunas de sus partes (ADN no codificante), ni niveles de estudio. La aproximación de Li (2012), centrada en una única especie (humanos) y en los exones, no es incompatible para nada con la nuestra, que explora los genomas secuenciados conocidos en el nivel cariotípico. Li (2012) tampoco considera los paralelismos establecidos entre el ADN no codificante y los elementos lingüísticos sin significado léxico (Ferrer-i-Cancho, Forns *et al.*, 2013).

En el apartado siguiente se incluyen los dos artículos (Ferrer-i-Cancho, Forns *et al.*, 2013; Baixeries *et al.*, 2012b) que principalmente responden a las críticas directas de Solé (2010) e indirectas de Li (2012), que se han resumido aquí, dejando para más adelante (apartados 4.7. y 4.8.) una revisión de los modelos aleatorios en los que también se apoya Solé (2010) y que fueron revisados en Baixeries y colaboradores (2012) y en su *erratum* (Ferrer-i-Cancho, Baixeries *et al.*, 2013).

4.6. Respuestas a las críticas sobre la ley de Menzerath-Altmann en el genoma

Se incluyen aquí dos artículos:

- Baixeries, J., Hernández-Fernández, A., Forns, N. & Ferrer-i-Cancho, R. (2012b). The parameters of Menzerath-Altmann law in genomes. *Journal of Quantitative Linguistics* 20 (2), 94–104.
- Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G. & Baixeries, J. (2013b). The challenges of statistical patterns of language: the case of Menzerath's law in genomes. *Complexity* 18 (3), 11–17.



COMPLEXITY

Journal of Quantitative Linguistics, 2013
Vol. 20, No. 2, 94–104, http://dx.doi.org/10.1080/09296174.2013.773141

Routledge
Taylor & Francis Group

The Parameters of the Menzerath-Altmann Law in Genomes*

Jaume Baixeries[1], Antoni Hernández-Fernández[1,2], Núria Forns[3] and Ramon Ferrer-i-Cancho[1]
[1]TALP Research Center/LARCA, Universitat Politècnica de Catalunya, Barcelona, Spain; [2]Departament de Lingüística, General, Universitat de Barcelona, Spain; [3]Facultat de Biologia, Universitat de Barcelona, Spain

ABSTRACT

The relationship between the size of the whole and the size of the parts in language and music is known to follow the Menzerath-Altmann law at many levels of description (morphemes, words, sentences, …). Qualitatively, the law states that the larger the whole, the smaller its parts, e.g. the longer a word (in syllables) the shorter its syllables (in letters or phonemes). This patterning has also been found in genomes: the longer a genome (in chromosomes), the shorter its chromosomes (in base pairs). However, it has been argued recently that mean chromosome length is trivially a pure power function of chromosome number with an exponent of -1. The functional dependency between mean chromosome size and chromosome number in groups of organisms from three different kingdoms is studied. The fit of a pure power function yields exponents between -1.6 and 0.1. It is shown that an exponent of -1 is unlikely for fungi, gymnosperm plants, insects, reptiles, ray-finned fishes and amphibians. Even when the exponent is very close to -1, adding an exponential component is able to yield a better fit with regard to a pure power-law in plants, mammals, ray-finned fishes and amphibians. The parameters of the Menzerath-Altmann law in genomes deviate significantly from a power law with a -1 exponent with the exception of birds and cartilaginous fishes.

*Address correspondence to: Ramon Ferrer-i-Cancho, Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain. Tel: +34 934137870. Fax: +34 934137787. Email: rferrericancho@lsi.upc.edu

1. INTRODUCTION

The Menzerath-Altmann law is a linguistic law relating x, the size of a construct (e.g. a word), with y the size of its constituents (e.g. syllables). From a qualitative point of view, the law states that the larger the size of a construct the smaller the parts (Altmann, 1980). For instance, the longer a word (in syllables), the shorter its syllables (in letters or phonemes). The law is mathematically defined through the equation (Altmann, 1980)

$$y = ax^b e^{cx}, \tag{1}$$

here a, b and c are parameters. The law bears the name of the researcher who observed the qualitative dependence between the size of the whole and the size of the parts in language (P. Menzerath) and the researcher who put it into mathematical form (G. Altmann). The mathematical function in Equation (1) has been used not only in quantitative linguistics but also in other studies of scaling laws of genomes (Molina & van Nimwegen, 2009). Patterning consistent with the Menzerath-Altmann law in the genomes of various groups of organisms has been reported (Ferrer-i-Cancho & Forns, 2009) by taking y as the size of a genome in chromosomes (L_g) and taking x as the length of its chromosomes in million base pairs (L_c). These analyses made two major simplifications:

- x was taken as a mean length for consistency with previous linguistic research (e.g. Boroda & Altmann, 1991 and references therein) and due to the absence of information about the length of concrete chromosomes in public database for a sufficiently large number of organisms.
- Agreement with the Menzerath-Altmann law not through a fit of Equation (1) but through a statistical correlation test. A negative correlation between y and x was considered as consistent with the law but it does not imply that Equation (1) actually holds.

It has been argued that both simplifications could lead to trivial results (Solé, 2010). In Ferrer-i-Cancho & Forns (2009), chromosome lengths were defined as an average, i.e. $L_c = G/L_g$ for a given organism of genome length size G (in base pairs) and L_g chromosomes. It has been argued that the definition $L_c = G/L_g$ for a certain organism implies $L_c \sim 1/L_g$ for any organism and thus the negative dependency between L_c and L_g reported by

Ferrer-i-Cancho and Forns (2009) is unavoidable and therefore not relevant (Solé, 2010). By analysing only two groups of the eleven major groups of organisms that were considered in the original research on genomes (Ferrer-i-Cancho & Forns, 2009), it was concluded that the negative correlations between L_c and L_g reported in the study by Ferrer-i-Cancho and Forns (2009) are "an inevitable consequence of the definitions of chromosome and genome" (Solé, 2010).

In this article, the dependency between mean chromosome length and chromosome number is studied with statistical depth. Three possible nested mathematical definitions for the dependency between L_c, the mean chromosome length and L_g, the genome size (in base pairs), are considered:

$$L_c = aL_g^{-1} \tag{2}$$

$$L_c = aL_g^{b} \tag{3}$$

$$L_c = aL_g^{b}e^{cL_g} \tag{4}$$

where *a, b* and c are constants. Equation (2) is the one that, according to Solé (2010), genomes must obey trivially when chromosome length is defined as a mean. Equation (4) is an adaptation of Equation (1) to genomes. If the arguments by Solé (2010) were correct, the fit of Equation (3) to genome data should give $b \approx -1$ and the fit of Equation (4) should give $b \approx -1$ and $c \approx 0$. Incidentally, $b = -0.6$ is reported for the fit of Equation (3) to ants ($N = 105$) in the pioneering work by Wilde and Schwibbe (1989). Here the analysis of the Menzerath-Altmann law will be extended to the large groups of organisms employed in recent studies (Ferrer-i-Cancho & Forns, 2009; Hernández-Fernández et al., 2011) by means of the three nested models defined in Equations (2) to (4).

2. RESULTS

Table 1 shows that the fit of Equation (3) gives values of the parameter b that vary between -1.45 (jawless fishes) and 0.1 (amphibians). Table 1 also indicates that $b = -1$ is unlikely for fungi, gymnosperm plants, insects, reptiles, jawless fishes, ray-finned fishes and amphibians.

Table 1. Summary of the fit of $L_c = aL_g^{\,b}$. L_c is the mean chromosome length and L_g is the number of chromosomes in the major groups of organisms from the study by Hernández-Fernández et al. (2011). The number attached to the group name indicates the number of organisms for that group in our dataset. For the parameters a and b the notation "estimate ± standard error" is used. $b_{\min}$ and $b_{\max}$ are, respectively, the lower and the upper bound of b in a 97.5% confidence interval. F is the value of the F-statistic used to determine if parameter b contributes to decrease error significantly with regard to the error obtained by the fit of $L_c = aL_g^{-1}$. p is the p-value of the corresponding F-test. The values of F were rounded to leave only two significant digits. The values of p were rounded to leave a single significant decimal. An asterisk (*) is used to indicate the exponents that are inconsistent with $b = -1$, the prediction of Solé (2010), according to the F-test at a significance level of 0.05.

Group	a	b	$b_{\min}$	$b_{\max}$	F	p
Fungi (56)	11 ± 3	$-0.5^* \pm 0.2$	–0.8	–0.2	9.9	0.003
Angiosperm plants (4706)	$(51 \pm 5) \times 10^2$	-0.95 ± 0.05	–1.04	–0.86	1.5	0.2
Gymnosperm plants (170)	$(3 \pm 2) \times 10^3$	$-0.3^* \pm 0.2$	–0.8	0.1	10	0.001
Insects (269)	$(6 \pm 1) \times 10^2$	$-0.7^* \pm 0.1$	–0.9	–0.5	6.0	0.01
Reptiles (170)	$(8 \pm 3) \times 10^2$	$-0.6^* \pm 0.1$	–0.8	–0.4	9.2	0.003
Birds (99)	$(16 \pm 5) \times 10^2$	-1.0 ± 0.1	–1.3	–0.9	0.17	0.7
Mammals (371)	$(33 \pm 1) \times 10^2$	-0.99 ± 0.02	–1.02	–0.96	0.82	0.4
Cartilaginous fishes (52)	$(3 \pm 2) \times 10^2$	-0.8 ± 0.2	–1.3	–0.4	0.75	0.4
Jawless fishes (13)	$(11 \pm 2) \times 10^3$	-1.45 ± 0.08	–1.6	–1.32	41	6×10^{-5}
Ray–finned fishes (647)	$(30 \pm 9) \times 10$	$-0.54^* \pm 0.09$	–0.75	–0.36	14	2×10^{-4}
Amphibians (315)	$(10 \pm 6) \times 10^2$	$0.1^* \pm 0.2$	–0.4	0.5	21	6×10^{-6}

Table 2 indicates that Equation (4) gives a significantly better fit than Equation (3) for plants, mammals, ray-finned fishes and amphibians. Table 3 indicates that Equation (2) gives always a poorer fit except for birds, mammals and cartilaginous fishes. However, Table 3, shows that the fit of Equation (4) is always better than that of Equation (3) in all groups except insects, birds, cartilaginous fishes and jawless fishes, where the fit is worse. The latter can be interpreted as a failure of the fitting algorithm because Equation (3) is a particular case of Equation (4) with $c = 0$.

3. DISCUSSION

Table 4 summarizes all the qualitative results obtained so far on the dependency between L_g and L_c in this article and previous research in our major groups of organisms. The fact that the correlation between L_g and L_c is not significant but Equation (4) gives a better fit in gymnosperm plants and ray-finned fishes suggests that their dependency between L_g and L_c

Table 2. Summary of the fit of $L_c = aL_g{}^b e^{cLg}$. L_c is the mean chromosome length and L_g is the number of chromosomes in the major taxonomic groups from the study by Hernández-Fernández et al. (2011). For the parameters *a, b* and *c* the notation "estimate ± standard error" is used. F is the value of the F-statistic used to determine if parameter c contributes to decrease error significantly with regard to the error obtained by the fit of $L_c = aL_g{}^b$. p is the p-value of the corresponding F-test. The values of F were rounded to leave only two significant digits. The values of p were rounded to leave a single significant decimal. An asterisk (*) is used to mark the values of c that are significantly different from zero according to the F-test at a significance level of 0.05.

Group	a	b	c	F	p
Fungi	10 ± 9	–0.9 ± 0.5	–0.05 ± 0.06	0.76	0.4
Angiosperm plants	$(28 \pm 6)\times 10^2$	–0.36 ± 0.2	–0.07* ± 0.02	20	8×10^{-6}
Gymnosperm plants	$(0.3 \pm 7)\times 10^{12}$	23 ± 5	–1.9* ± 0.4	28	2×10^{-7}
Insects	$(6 \pm 3)\times 10^2$	–0.7 ± 0.5	0.01 ± 0.07	0.019	0.9
Reptiles	$(0.8 \pm 1)\times 10^4$	–1.9 ± 0.7	0.07 ± 0.04	2.8	0.1
Birds	$(4 \pm 5)\times 10^3$	–1.5 ± 0.6	0.03 ± 0.03	0.630	0.4
Mammals	$(30 \pm 2)\times 10^2$	–0.89 ± 0.05	–0.009* ± 0.004	4.1	0.04
Cartilaginous fishes	$(0.4 \pm 1)\times 10^3$	0.1 ± 1	–0.03 ± 0.04	0.78	0.4
Jawless fishes	$(9 \pm 4)\times 10^3$	–1.3 ± 0.3	–0.009 ± 0.02	0.20	0.7
Ray–finned fishes	$(2 \pm 1)\times 10^3$	–1.4 ± 0.2	0.023* ± 0.005	15	10^{-4}
Amphibians	$(0.6 \pm 6)\times 10^{16}$	26 ± 5	–1.6* ± 0.4	44	2×10^{-10}

may not be monotonic as the parameters reported for Equation (4) in Table 2 indicate.

It has been argued that the definition of L_c as a mean, namely $L_c = G/L_g$ implies an inverse proportionality dependency between L_g and L_c, i.e. $L_c \sim 1/L_g$ (Solé, 2010). We examined three nested models and have shown that Equations (3) and (4) are able to approximate the actual dependency between mean chromosome length and chromosome number better than the inverse proportionality relationship (Equation (2)), with the only exception of birds and cartilaginous fishes (Table 4). None of these two groups coincides with the two groups selected by Solé (2010), i.e. mammals and plants to support his claim that the Menzerath-Altmann law in genomes is a trivial power law with a -1 exponent. Notice that the results reported here do not imply that $L_c \sim 1/L_g$ gives a perfect fit for birds and cartilaginous fishes. Although more powerful mathematical and statistical arguments have been used to discard it (Hernández-Fernández et al., 2011), these two groups should be the subject of future research.

Our study of the dependency between mean chromosome length and chromosome number has been focused on a small family of nested mathematical models motivated by the proposal of an inverse proportionality

Table 3. The error of the fit versus the number of free parameters. A summary of the goodness of the fit of different equations for the relationship between L_c, the mean cromosome size and L_g, the number of chromosomes. The goodness of the fit is measured with s, the residual standard errors. Values were rounded to leave only two decimal digits. An asterisk (*) is used to indicate the cases where the non-linear regression technique failed when fitting the n-parameter equation model in the sense that yielded a value of s higher than that of the $n - 1$ parameter equation, with $n = 2$ or $n = 3$.

	s		
Group	$L_c = aL_g^{-1}$	$L_c = aL_g^{b}$	$L_c = aL_g^{b}e^{cLg}$
Fungi	1.91	1.77	1.77
Angiosperm plants	806.69	806.64	805.02
Gymnosperm plants	627.37	610.39	565.20
Insects	194.62	192.83	193.18*
Reptiles	38.35	37.45	37.25
Birds	9.08	9.12	9.14*
Mammals	49.76	49.77	49.56
Cartilaginous fishes	87.12	87.34	87.53*
Jawless fishes	55.13	26.48	27.51*
Ray-finned fishes	27.44	27.17	26.87
Amphibians	1418.08	1374.87	1289.92

Table 4. Summary of results on the dependency between L_c and L_g. The analysis of the dependency between L_c and L_g in all the groups of organisms through the fit of $L_c = aL_g^{b}e^{cLg}$. An asterisk (*) is used to indicate results borrowed from Ferrer-i-Cancho and Forns (2009) at a significance level of 0.05 (the significant and non-significant correlations are the same if the dataset of the present article is used).

Group	Correlation between L_c and L_g (*)	$b \neq -1$ when $c = 0$ (Table 1)	$c \neq 0$ (Table 2)	$b = -1$ and $c = 0$ yield maximum s (Table 3)
Fungi	Yes	Yes		Yes
Angiosperm plants	Yes		Yes	Yes
Gymnosperm plants		Yes	Yes	Yes
Insects	Yes	Yes		Yes
Reptiles	Yes	Yes		Yes
Birds	Yes			
Mammals	Yes		Yes	
Cartilaginous fishes	Yes			
Jawless fishes	Yes	Yes		Yes
Ray-finned fishes		Yes	Yes	Yes
Amphibians	Yes	Yes	Yes	Yes

relationship by Solé (2010), the mathematical definition of the Menzerath-Altmann law in quantitative linguistics research (Altmann, 1980; Teupenhayn & Altmann, 1984) and research on scaling laws in genomes (Molina & van Nimwegen, 2009). However, the issue of the mathematical function that would give a priori the best fit needs to be investigated further. Our analysis does not exclude the possibility that there are more appropriate functions to describe such dependency. With this regard, the motivation of the simple correlation analysis by Ferrer-i-Cancho and Forns (2009) was staying as much neutral as possible about the actual dependency between mean chromosome length and chromosome number. A trivial correlation can arise if genome size is statistically independent from the chromosome number; a property that has been rejected for the majority of groups (Hernández-Fernández, 2011). The combination of a simple correlation analysis, with further analyses to exclude trivial sources of correlations (Hernández-Fernández, 2011), results in a robust approach to the dependency between the mean size of the parts and the number of parts with lighter prior assumptions.

A relationship between the number of parts and the mean size of the parts consistent with the Menzerath-Altmann law has been reported in genomes, not only between chromosome number and mean chromosome size (here; Ferrer-i-Cancho & Forns, 2009), but also between the number of exons of a gene and the mean exon size in the human genome (Li, 2012). As for the second discovery, the hypothesis of the trivial power law with a -1 exponent (Solé, 2010) has also been excluded (Li, 2012). Indeed, the finding of the Menzerath-Altmann law in genomes is not surprising given the many parallels that have been investigated and established between human language and genomes (Bel-Enguix & Jiménez-Lopez, 2011 and references therein). However, the origins and the depth of this statistical coincidence between language, genomes and also music (Boroda & Altmann, 1991) should be investigated further.

5. METHODS

Data

The same dataset as in the study by Hernández-Fernández et al. (2011), which is an updated version of that of Ferrer-i-Cancho and Forns (2009), was used. Group sizes (in species) are shown in Table 1.

Non-linear regression

Throughout the article the fit of the functions defined in Equations (2) to (4) is studied. The goodness of the fit of these equations is evaluated by means of the residual standard error defined as

$$s = \sqrt{\frac{RSS}{N-p}} \tag{5}$$

where RSS is the residual sum of squares and $N - p$ is the degrees of freedom (N is the sample size and p is the number of parameters; $p = 1$ for Equation (2), $p = 2$ for Equation (3) and $p = 3$ for Equation (4)).

The general form of RSS for our functions is

$$RSS = \sum_{i=1}^{N} (y_i - a x_i^b e^{c x_i})^2 \tag{6}$$

where N is the number of organisms of the group and, x_i and y_i are, respectively, the value of L_g and L_c of the ith organism of the group.

The parameters that give the best fit minimizing s, which is equivalent to minimizing RSS, are obtained. First, a^*, the value of a that minimizes RSS given b and c, will be derived. The condition $dRSS/da = 0$ yields

$$a^* = \frac{\sum_{i=1}^{N} (y_i x_i^b e^{c x_i})}{\sum_{i=1}^{N} (x_i^{2b} e^{c x_i})} \tag{7}$$

a^* corresponds to a minimum of RSS (given b and c) if and only if $d^2RSS/d^2a > 0$. In our case, we have

$$\frac{d^2 RSS}{d^2 a} = \sum_{i=1}^{N} x_i^{2b} e^{c x_i} > 0 \tag{8}$$

because the number of chromosomes is a strictly positive number and thus a^* is a minimum.

When $b = 1$ and $c = 0$, one obtains the well-known estimator of the slope of a linear function through the origin (Sheather, 2010, p. 41), i.e.

$$a^* = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} x_i^2} \tag{9}$$

and the condition in Equation (8) holds trivially regardless of the sign of the x_i. The best fit of Equation (2) can be obtained exactly applying $b = -1$ and $c = 0$ to Equation (7), which gives

$$a^* = \frac{\sum_{i=1}^{N} (y_i/x_i)}{\sum_{i=1}^{N} (1/x_i^2)} \tag{10}$$

For the other functions (Equations (3) and (4)), a nonlinear regression algorithm (based on the Gauss-Newton method) that minimizes *RSS* numerically (Ritz & Streibig, 2008) was used. It is well-known that providing the appropriate initial values of the parameters of the ideal function is crucial for the success of the non-linear regression algorithm (Ritz & Streibig, 2008) because the algorithm could get trapped in local minima of *RSS*. It is customary to apply a transformation of the original curve to express it in a way that simple linear regression can be applied to obtain the initial values of the parameters for the non-linear regression technique (Ritz & Streibig, 2008). For the fit of Equation (3), the nonlinear regression algorithm was fed with initial values of a and b estimated through a linear regression of Equation 3 in logarithmic scale (Ritz & Streibig, 2008). Equation (3) is equivalent to

$$y' = bx' + a' \tag{11}$$

where $y' = \log L_c$, $x' = \log L_g$ and $a' = \log a$. A standard least squares linear regression gives the initial value of b and a'. The initial value of a' is obtained through $a = \mathrm{e}^{a'}$.

As for the fit of Equation (4), two different starting values of a, b and c were considered and the final values of a, b and c that yielded the smallest value of *RSS* where retained. The first initial set up was defined by a and b from the nonlinear regression best fit of Equation (3) and $c = 0$. The second initial set up was defined by $b = 0$ and the values of a and c estimated from linear regression on a logarithmic transformation of Equation (4) with $b = 0$. If logarithms are taken on both sides of Equation (4), one obtains

$$y' = cL_g + a' \tag{12}$$

where $y' = \log L_c$, and $a' = \log a$ as before. A standard least squares linear regression gives the initial value of c in the second initial setup and a'. The initial value of a' in the second initial setup is obtained through $a = e^{a'}$ as before.

The confidence intervals for b shown in Table 1 were computed using a non-parameteric bootstrap approach with 999 artificially generated datasets (Ritz & Streibig, 2008, pp. 96–99). A bootstrap technique was used instead of profile confidence intervals or Wald confidence intervals because of the greater robustness of the bootstrap approach (Ritz & Streibig, 2008).

When evaluating the power of the fit yielded by Equation (4) versus that of Equation (3), an extra-sums-of-squares F-test was used to determine if parameter c can be neglected and thus Equation (4) could be reduced to Equation (3) (Ritz & Streibig, 2008, pp. 103–105). We used an F-test instead of a t-test because the former is more robust than the latter (Ritz & Streibig, 2008). The same analysis was performed for evaluating the power of the fit of Equation (3) versus that of Equation (2).

ACKNOWLEDGEMENTS

This article is dedicated to G. Altmann and the late P. Menzerath. We thank G. Altmann for his clarifications and J. Perarnau for technical advice on the statistical analyses. This work was supported by the grant *Iniciació i reincorporació a la recerca* from the Universitat Politècnica de Catalunya and the grant BASMATI (TIN2011-27479-C04-03) from the Spanish Ministry of Science and Innovation (RFC and JB).

REFERENCES

Altmann, G. (1980). Prolegomena to Menzerath's law. In R. Grotjahn (Ed.), *Glottometrika 2* (pp. 1–10). Bochum: Brockmeyer.

Bel-Enguix, G., & Jiménez-López, M. D. (2011). Genetic code and verbal language: syntactic and semantic analogies. In G. Bel-Enguix, V. Dahl & M. D. Jiménez-López (Eds), *Biology, Computation and Linguistics. New Interdisciplinary Paradigms* (pp. 85–103). Amsterdam: IOS Press.

Boroda, M. G., & Altmann, G. (1991). *Menzerath's law in musical texts. Musikometrica, 3*, 1–13.

Ferrer-i-Cancho, R., & Forns, N. (2009). The self-organization of genomes. *Complexity, 15* (5), 34–36.

Hernández-Fernández, A., Baixeries, J., Forns, N., & Ferrer-i-Cancho, R. (2011). Size of the whole versus number of parts in genomes. *Entropy, 13*(8), 1465–1480.

Li, W. (2012). Menzerath's law at the gene-exon level in the human genome. *Complexity, 17* (4), 49–53.

Molina, N., & van Nimwegen, E. (2009). Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends in Genetics, 25*(6), 243–247.

Ritz, C., & Streibig, J. C. (2008). *Nonlinear Regression with R*. New York: Springer.

Sheather, S. J. (2010). *A Modern Approach to Regression with R*. New York: Springer.

Solé, R. V. (2010). Genome size, self-organization and DNA's dark matter. *Complexity, 16* (1), 20–23.

Teupenhayn, R., & Altmann, G. (1984). Clause length and Menzerath's law. In J. Boy & R. Köhler (Eds), *Glottometrika 6* (pp. 127–138). Bochum: Brockmeyer.

Wilde, J., & Schwibbe, M. H. (1989). Organizationsformen von Erbinformation im Hinblick auf die Menzerathsche Regel. In G. Altmann& M.H. Schwibbe (Eds), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen* (pp. 92–107). Hildesheim: Olms.

The Challenges of Statistical Patterns of Language: The Case of Menzerath's Law in Genomes

The importance of statistical patterns of language has been debated over decades. Although Zipf's law is perhaps the most popular case, recently, Menzerath's law has begun to be involved. Menzerath's law manifests in language, music and genomes as a tendency of the mean size of the parts to decrease as the number of parts increases in many situations. This statistical regularity emerges also in the context of genomes, for instance, as a tendency of species with more chromosomes to have a smaller mean chromosome size. It has been argued that the instantiation of this law in genomes is not indicative of any parallel between language and genomes because (a) the law is inevitable and (b) noncoding DNA dominates genomes. Here mathematical, statistical, and conceptual challenges of these criticisms are discussed. Two major conclusions are drawn: the law is not inevitable and languages also have a correlate of noncoding DNA. However, the wide range of manifestations of the law in and outside genomes suggests that the striking similarities between noncoding DNA and certain linguistics units could be anecdotal for understanding the recurrence of that statistical law. © *2012 Wiley Periodicals, Inc. Complexity 18: 11–17, 2013*

Key Words: statistical laws; language; genomes; music; non-coding DNA; Menzerath's law

RAMON FERRER-I-CANCHO[1]
NÚRIA FORNS[2]
ANTONI HERNÁNDEZ-FERNÁNDEZ[1,3]
GEMMA BEL-ENGUIX[4]
JAUME BAIXERIES[1]

[1]*Departament de Llenguatges i Sistemes Informàtics, Complexity and Quantitative Linguistics Lab, TALP Research Center/LARCA, Universitat Politècnica de Catalunya, Barcelona (Catalonia), Spain;* [2]*Departament de Microbiologia, Facultat de Biologia;* [3]*Departament de Lingüística General, Universitat de Barcelona, Barcelona (Catalonia), Spain; and* [4]*Laboratoire d'Informatique Fondamentale, University Aix-Marseille & CNRS, Marseille, France*
(email: rferrericancho@lsi.upc.edu)

1. INTRODUCTION

Attempts to demonstrate that statistical patterns of language have a trivial explanation have a long history that goes back at least to the research by G. A. Miller and collaborators questioning the relevance of Zipf's law for word frequencies around 1960 [1–3]. Zipf's law states that the curve that relates the frequency of a word f and its rank r (the most frequent word having rank 1, the second most frequent word having rank 2, and so on) should follow $f \sim r^{-\alpha}$ [4]. Miller argued that if monkeys were chained "*to typewriters until they had produced some very long and random sequence of characters*" one would find "*exactly the same 'Zipf curves' for the monkeys as for the human authors*" [3]. Under his view, Zipf's law would be an inevitable consequence of the fact that words are made of units, e.g., letters or phonemes. The typewriter argument has been revived many times since then [5–8]. However, rigorous analyses indicate that the curves do not really look the same and the parameters of this random typing model giving a good fit to real word frequencies are not forthcoming [9,10]. Here, we review a recent

 Vol. 18, No. 3
DOI 10.1002/cplx.21429
Published online 23 November 2012 in Wiley Online Library (wileyonlinelibrary.com)

claim that the finding of another statistical pattern of language, Menzerath's law, is also inevitable [11].

P. Menzerath hypothesized that "the greater the whole, the smaller its constituents" ("*Je größer das Ganze, desto kleiner die Teile*") in the context of language [12] (pp. 101). Converging research in music and genomes [13–16] suggests that Menzerath's law is a general law of natural and human-made systems. In this article, we leave the term Menzerath-Altmann law for referring to the exact mathematical dependency that has been proposed by the quantitative linguistics tradition for the relationship between x, the size of the whole (in parts) and y, the mean size of the parts, i.e. [17],

$$y = ax^b e^{cx} \quad (1)$$

where a, b, and c are the parameters of Menzerath-Altmann law.

In the pioneering research by Wilde and Schwibbe [14] and later work [15,20], Menzerath's law emerged as a negative correlation between L_c and L_g, where L_c is the mean chromosome length (the size of the constituents) and L_g is the chromosome number (the size of the construct measured in constituents). More recently, the law has been found in the dependency between mean exon size (the size of the constituents) and the number of exons of human genes (the size of the construct) [16].

However, it has been argued that this negative correlation is trivial [11]: the definition of L_c as a mean, i.e. $L_c = G/L_g$ leads (according to Ref. [11]) unavoidably to $L_c \sim {L_g}^b$ with $b = -1$, which is supported by the fact that mammals and plants give values of b that are very close to $b = -1$ ($b = -1.04$ for mammals and $b = -1.07$ for plants [11]). In the present article, $\sim$ is used to indicate proportionality. Furthermore, it has also been argued that a proper connection between human language and genomes cannot be established a priori using genomes as wholes and chromosomes as parts, due to the fluid nature of chromosomal arrangements and the vast dominance of noncoding DNA, which has no parallel in language [11].

Revising those arguments is critical for musicology, quantitative linguistics, and genomics. If they were correct, the relationship between the mean size of the constituents (y) and the number of constituents (x) which have been the subject of many studies [13,16–18] would be a trivial consequence of the definition of the size of the constituents as a mean. Following Miller's argument, producing Menzerath's law would be as easy as producing Zipf's law by monkeys chained to a typewriter. More precisely, the inevitability of $L_c \sim 1/L_g$ [11] predicts that Menzerath-Altmann law must always be Eq. (1) with $b = -1$ and $c = 0$ when defining the size of the parts as a mean. If such inevitability is correct, exponents deviating significantly from $b = -1$ should be the exception, not the rule in language, music and genomes.

Here we address the challenge of Menzerath's law in genomes [14–16] and beyond [13,17,18] by reviewing Solé's criticisms [11]: his mathematical and statistical arguments, essentially the inevitability of $L_c \sim 1/L_g$ (Section 2), as well as his conceptual arguments, mainly the mismatch between human language and genomes (Section 3). Finally, we will discuss some general questions that are crucial for understanding the recurrence of Menzerath's law (Section 4).

2. The mathematical and statistical debate.

2.1. Mixing Angiosperm and Gymnosperm Plants

Solé does not distinguish between angiosperm and gymnosperm plants [11]. However, our analyses have been revealing important differences between them: (1) concerning the relationship between L_g and L_c, Menzerath's law is only found in angiosperms [15], (2) G tends to increase as L_g increases in gymnosperms but G increases as L_g decreases in angiosperms [19] and, (3) the fit of $L_c \sim {L_g}^b$ yields $b = -0.95 \pm 0.05$ for angiosperms and $b = -0.3 \pm 0.2$ for gymnosperms [20], the latter being statistically inconsistent with $b = -1$ as Solé predicts [11]. As his division of plants differs from that of Ferrer-i-Cancho and Forns [15] and gymnosperms do not follow Menzerath's law, we proceed assuming that his notion of plant is equivalent or can be reduced to angiosperms.

2.2. $L_c = G/L_g$ does not Imply $L_c \sim 1/L_g$.

It has been argued that the definition of L_c as G/L_g unavoidably leads to an inverse proportionality dependency between L_g and L_c, i.e. $L_c \sim 1/L_g$ [11]. This can be refuted in two ways: empirically and mathematically.

2.2.1. Empirical Refutation

$L_c \sim 1/L_g$ is not inevitable because

- Amphibians exhibit a positive correlation between L_c and L_g that is incompatible with $L_c \sim 1/L_g$ [15].
- Menzerath's law (a significant negative correlation between L_c and L_g) was not found for gymnosperm plants and ray-finned fishes [15].
- Many empirical studies of Menzerath-Altmann law compute the size of the parts as an average as Ferrer-i-Cancho and Forns did [15] but the fit of Eq. (1) gives parameters that deviate from $b \approx -1$ (see Table 1 for a summary of research).
- $b = -0.6$ is reported for ants in the pioneering work by Wilde and Schwibbe [14] that is cited by Ferrer-i-Cancho and Forns [15].
- Solé reports estimates of b only for mammals and plants (according to his analysis $b = -1.04$ and $b = -1.07$, respectively) [11], whereas Ferrer-i-Cancho and Forns [15], con-

DOI 10.1002/cplx

TABLE 1

Some Parameters of Menzerath-Altmann Law

Type of Source	Size of the Whole (x)	Size of the Parts (y)	Languages	Samples	b	c	Ref.
Language	Morpheme length (in syllables)	Mean syllable length (in phomenes)	Indonesian	1	-0.37	0.048	[17]
	Word length (in syllables)	Mean syllable length (in phonemes)	English	1	0.15	-0.10	[17]
	Sentence length (in clauses)	Mean clause length (in words)	German, English, French, Swedish, Hungarian, Slovak, Czech, Indonesian	42	-0.27 ± 0.11[a]	N.A.	[18]
Music	mr-segment length (in F-motifs)	Mean F-motif length (in tones)	—	11	-0.44 ± 0.09[a]	N.A.	[13]

The summary is based upon the pioneering work of G. Altmann and collaborators. N.A. means that the two parameter version of Eq. (1), with $c = 0$, was fitted.

[a]This follows the notation $\mu \pm \sigma$, where μ is the mean value of b in all samples and σ is the corresponding standard deviation among samples.

sidered a total of 11 major groups [15] (see also Ref. [19]). Thus, nine groups have not been considered. $|b + 1|$ is a measure of the deviation from his prediction, i.e. $L_c \sim 1/L_g$. $|b + 1| = 0$ means a perfect matching with his prediction. $|b + 1|$ indicates that mammals and angiosperm plants are among the three groups with the smallest value of $|b + 1|$ (Table 2).

- A careful statistical analysis reveals that b deviates significantly from $b = -1$ in fungi, gymnosperm plants, insects, reptiles, jawless fishes, ray-finned fishes, and amphibians, groups for which Solé reports no result [11]. Furthermore, the parameter b of $L_c \sim L_g^{\,b}$ contributes significantly to improve the quality of the fit with regard to that of $L_c \sim 1/L_g$ for the same groups [20]. Put differently, if b is let free, then the error of the model is reduced significantly for these groups with regard to keeping it equal to -1.
- In a recent study of Menzerath-Altmann law in genomes at the gene-exon level, the relationship between the mean exon size in bases and the number of exons of a human gene yields $b \approx -0.5$ [16].

2.2.2. Mathematical Refutation

When G is a constant function ($G \sim 1$), we have that $L_c \sim L_g^{\,b}$ with $b = -1$ as it is argued by Solé [11]. Yet, if G is not constant, then $b = -1$ is not necessarily expected: (1) the exponent may change (e.g., if $G \sim L_g^{\,-2}$ then $b = -3$) and (2) the power-law $L_c \sim L_g^{\,b}$ could be lost (e.g., if $G \sim L_g e^{-Lg}$ then we would have $L_c \sim e^{-Lg}$).

A mathematical analysis indicates that $L_c \sim 1/L_g$ needs that G and L_g are uncorrelated [19]. Therefore, $L_c \sim 1/L_g$ is rejected if G and L_g are correlated. The empirical evidence for such correlation is the following: (1) G tends to increase as L_g increases in gymnosperm plants and mammals while G tends to decrease as L_g decreases in angiosperm plants [19] and (2), from the major taxonomic groups considered by [15], only birds and cartilaginous fishes show no significant correlation between G and L_g [19].

2.3. The Dependency Between L_c and L_g

So far we have been discussing the fit of $L_c \sim L_g^{\,b}$ with Solé's prediction of $b = -1$ to genomes. But we have never argued that the instantiation of Menzerath-Altmann law in genomes [recall Eq. (1)]:

$$L_c = aL_g^{\,b}e^{cL_g} \tag{2}$$

(with the possibility of $b = -1$ and/or $c = 0$, following Solé's arguments) is the best, or simply the most suitable for modeling the actual relationship between L_c and L_g in genomes. When preparing our original article [15], we were already aware of the challenge of designing biologically realistic equations and evaluating the goodness of their fit rigorously.

Therefore, we decided to use a simple correlation analysis between L_c and L_g to stay neutral about the actual dependency. While our original approach was nonparametric (based on a Spearman rank correlation test), Solé followed the parametric track with the assumption that genomes follow $L_c \sim L_g^{\,b}$ [11]. Our approach to test Menzerath's law [15] and our approach to reject $L_c \sim 1/L_g$ are both nonparametric [19]. In sum, our analysis requires fewer assumptions than his. However, we have had to follow a parametric approach in one of the branches of our genome research to show that even when strong assumptions are made about the actual dependency, his arguments do not stand, even for mammals and plants [20].

DOI 10.1002/cplx

TABLE 2

The Distance to $b = -1$

Group	$\lvert b + 1 \rvert$
Mammals[a]	0.014
Birds	0.042
Angiosperm plants[a]	0.051
Plants[a]	0.13
Cartilaginous fishes	0.18
Insects	0.31
Reptiles	0.39
Jawless fishes	0.45
Ray-finned fishes	0.46
Fungi	0.50
Gymnosperm plants	0.68
Amphibians	1.1

A summary of $\lvert b + 1 \rvert$, the difference between the exponent b obtained from the fit of $L_c \sim L_g^b$ and the exponent -1 that is expected from the arguments by Solé [11]. Groups are sorted increasingly by $\lvert b + 1 \rvert$. b was estimated using nonlinear regression as in Ref. [20]. The dataset is the same as that of Refs. [19] and [20]. The values of $\lvert b + 1 \rvert$ were rounded to leave only two significant digits.

[a]Is used for the only two groups used by Solé [11]. Two interpretations of Solé's notion of plant are offered: angiosperms and a mixture of angiosperms and gymnosperms.

3. The Conceptual Debate

3.1. The Unsupported Fluid Nature of Chromosomal Rearrangements

Solé states that "*the fluid nature of chromosomal rearrangements through time rules against any special multi-scale link between genome-level and chromosome-level patterns*" [11]. If the mathematical interpretation of this statement is that the genome and the chromosome level are statistically independent, then a large amount of research indicates that G and L_g are not independent in real genomes and that independence is in conflict with chromosome well-formedness (see Ref. [9] and references therein).

3.2. Languages also have "Dark Matter"

Solé argues that the dominance of noncoding DNA (what he also calls "*information-lacking DNA*", "*information-lacking DNA*" or "*junk DNA*"), should prevent us from using large-scale structures such as genomes as meaningful information-related units [11]. However, the view of non-coding DNA as "*dark matter*" or "*junk*" in a strict sense is outdated from the point of view of molecular biology [21–24]. Some researchers have suggested that "*there is in fact much less, if any, 'junk' in the genomes of the higher organisms than has previously been supposed*" [25].

Linguistic sequences and genomes are not so radically different concerning real or apparent "*junk*," "*dark matter*," or "*information-lacking DNA*". In general, words are classified into content, e.g., verbs, nouns, and function words, e.g., prepositions, conjunctions. While content words are said to have lexical meaning, function words are said to have grammatical meaning [26], i.e. function words lack lexical meaning [27] (pp. 55). For this reason they are called "*empty words*" by certain scholars [26]. Similarly, noncoding DNA is *empty*, in the sense that it does not code for specific proteins. The term "junk words" has also been used for referring to function words and particles in language sciences [28]. However, the closest analogy for the term "junk" in human language are the so-called filler words such as "um," "oh," "well" (Searls DB, Personal Communication, 2011).

Function words such as prepositions and conjunctions have an inherently relational meaning [29] and they are very important nodes in word networks: they are hubs or "authorities" in a network theory sense [30,31]. The logic structure of the sentence "Mary bought an apartment *in spite of* the economic crisis" is radically different from that of "Mary bought an apartment *thanks to* the economic crisis". The conjunctions "in spite of" and "thanks to" regulate the relationship between "Mary bought an apartment" and "the economic crisis" in the sentences above. In sum, *lexical meaning* and *protein coding* appear to be parallel terms, respectively, from the linguistic and genetic world. The same applies to *grammatical meaning* and *regulation*, the latter being a function served by noncoding DNA [22,24].

If we consider linguistic units with grammatical function as equivalent to noncoding DNA, then not only function words or particles parallel noncoding DNA, but also bound morphemes (e.g., the *–ed* ending of *walked*), as they also contain grammatical meaning. As linguistic sequences at many levels contain a mixture of elements with lexical and grammatical meaning (e.g., lexemes and bound morphemes in words), a DNA sequence may be a combination of coding and noncoding parts (e.g., exons and introns in genes). Words, phrases, clauses, sentences, i.e. units on which Menzerath's law has been reported (Ref. [33] and references therein), are "polluted" to some extent by "dark matter".

DOI 10.1002/cplx

TABLE 3

Percentage of Content, Function, and Filler Words in Two Registers: Conversation and News Report

	Conversation (%)	News (%)
Content words	41	63
Function words	44	37
Fillers	15	—

Adapted from Ref. [27], Table 2.4 (pp 61).

The statistics of the amount of function and content words provides us with an estimate of the amount of "dark matter" in language. Table 3 indicates that the proportion of a parallel of noncoding DNA in an English conversation is about 59%, which includes function and filler words, while it is about 37% in a news report. Therefore, languages also have a large proportion of elements reminiscent of non-coding DNA. But the true proportion of "noncoding" elements in languages could be higher if the grammatical morphemes that are attached to lexemes were included in the counts.

Interestingly, the evolution of the view of "fillers" in linguistics parallels the evolution of the view of noncoding regions in molecular biology. Progress in linguistic research indicates that "fillers" are more than mere "fillers" while progress in genomics indicates that "junk" DNA is more than mere "junk". As for linguistics, the understanding of filler words in linguistics has evolved from the term filler [32], as their meaning and their role in the sentence was gradually recognized, to particular kinds of discourse related particles or cue words (Ref. [34] and references therein). At present, the consensus is that "words" originally called fillers "*have no apparent grammatical relation to the sentences in which they appear*", and "*contrary to what prescriptivists' accusations, they do have a meaning, in that they seem to convey something about the speaker's relation to what is asserted in the sentence*" [34]. The view of other function words has also evolved similarly: function words believed to be empty contain indeed meaning [34,35]. As for molecular biology, the field is moving from the view of noncoding DNA as "junk" to that of functionally relevant material [21–24]. The view of repetitive segments in DNA sequences as mere "fillers" is being abandoned in molecular biology [36]. In both biology and linguistics, "dark matter" is becoming meaningful or functional matter, thanks to progress in core molecular biology and linguistics.

3.3. Misunderstanding of a Metaphor

Solé's focus on noncoding DNA as an obstacle for a proper connection between human language and genomes [11] shows that he has misunderstood the "*metaphor that genomes are words and chromosomes are syllables*" (abstract of Ref. [15]).

Patterning consistent with Menzerath's law is found at many linguistic levels: morphemes (in the seminal work by G. Altmann [17] that he cites) or sentences [18]; see Table 1. Probably the most radical example is music (see also Table 1), where the whole and the parts lack a "meaning" equivalent to that of content words. This suggests that Menzerath's law is a manifestation of abstract principles as many have proposed (see Ref. [13] and references therein [15]). In contrast, Solé shows a lack of abstraction when considering that language and genomes, in order to resemble statistically, must be practically identical [11]. Indeed, he interprets the linguistic metaphor that inspired our original article (genomes "are" words and chromosomes "are" syllables) not as a metaphor but as a narrow equivalence. We could have replaced words and syllables by other units: morphemes and syllables, sentences and clauses, or mr-segments, and F-motifs (Table 1). Words and syllables were probably the simplest metaphors for a general audience.

4. Discussion

We have seen that Menzerath's law is not inevitable in genomes and that it suffices that the number of parts (e.g., the number of chromosomes) and the size of the whole in the units of the parts (the size of chromosomes in bases) are correlated in order to reject a trivial case of the law [16,19]. However, we do not mean that the finding of a nontrivial Menzerath's law in the relationship between mean chromosome size and chromosome number [15,19] is due to the striking similarities between noncoding DNA and linguistic units with grammatical meaning that we have enlightened here but Solé neglected [11]. We have never argued that the finding of the law in genomes is indicative of meaning, syntax, or any other important property of language. The finding of Menzerath's law both when noncoding DNA is excluded [16] and when noncoding and coding-DNA are mixed [15], and beyond, i.e. in language (see Ref. [33] for a review) and music [13], suggests that a higher level of abstraction is necessary for understanding the recurrence of the law.

To our knowledge, it has not been investigated yet if noncoding DNA alone could lead to Menzerath's law, or more interestingly, a nontrivial Menzerath's law. Without this research, it is not possible either to have a clearer

DOI 10.1002/cplx

understanding of the role of noncoding DNA in the emergence of Menzerath's law in genomes or to question the relevance of the law in genomes. Perhaps, rather than precluding the emergence of the law or leading to a trivial law, noncoding DNA may contribute to the emergence of the law in a way that defies a trivial explanation.

Languages and genomes show a striking similarity at the semantic level: both possess units that have an arbitrary semantic reference of symbolic nature [37]. Our comparison goes further and suggests that genomes code for some abstract version of grammatical and lexical meaning, the former in noncoding regions and the latter in coding regions. However, the depth of the similarity and the possible DNA-specific properties must be investigated further. One of the challenges for language research is estimating the proportion of material with grammatical meaning including both free function words and bound morphemes.

Quantitative linguistics offers powerful tools for discovering and investigating nontrivial connections between human language and genomes [37,38]. However, the evolutionary mechanisms and the constraints that may underlie the recurrence of Menzerath's law still must be understood.

ACKNOWLEDGMENTS

The authors are grateful to M. D. Jiménez-López, D. Searls, F. Bartumeus, D. Alonso, and S. Caldeira for helpful discussions. This work was supported by the grant *Iniciació i reincorporació a la recerca* from the Universitat Politècnica de Catalunya and the grant BASMATI (TIN2011-27479-C04-03) from the Spanish Ministry of Science and Innovation (to R.F.C. and J.B.).

REFERENCES

1. Miller, G. A. Some effects of intermittent silence. Am J Psychol 1957, 10, 311–314.
2. Miller, G. A.; Chomsky, N.Finitary models of language users. In: Handbook of Mathematical Psychology II; Luce, R.; Bush, R.; Galanter, E., Eds.; Wiley: New York, 1963; pp 419–491.
3. Miller, G. A.Introduction. In: The Psycho-biology of Language; Zipf, G. K. MIT Press: Cambridge, MA, 1963; pp v–x.
4. Zipf, G. Human Behavior and the Principle of Least Effort; Addison-Wesley: Cambridge, MA, 1949.
5. Li, W. Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Trans Inf Theory 1992, 38, 1842–1845.
6. Suzuki, R.; Buck, J. R.; Tyack, P. L. The use of Zipf's law in animal communication analysis. Anim Behav 2005, 69, F9–F17.
7. Li, W. Zipf's law and the structure and evolution of languages, A.A.Tsonis, C.Schultz, and P.A.Tsonis, Complexity 1997, 2, 12–13 (Letter to the Editor). Complexity 1998, 3, 9–10.
8. Ferrer-i-Cancho R, Elvevåg B. Random texts do not exhibit the real Zipf's law-like rank distribution. PLoS ONE 2010, 5, e9411.
9. Baixeries, J.; Hernández-Fernández, A.; Ferrer-i-Cancho, R. Random models of Menzerath–Altmann law in genomes. Biosystems 2012, 107, 167–173.
10. Ferrer-i-Cancho, R.; Gavaldà, R. The frequency spectrum of finite samples from the intermittent silence process. J Am Soc Inf Sci Technol 2009, 60, 837–843.
11. Solé, R. V. Genome size, self-organization and DNA's dark matter. Complexity 2010, 16, 20–23.
12. Menzerath, P. Die Architektonik des deutschen Wortschatzes. Dümmler: Bonn, 1954.
13. Boroda, M. G.; Altmann, G. Menzerath's law in musical texts. Musikometrica 1991, 3, 1–13.
14. Wilde, J.; Schwibbe, M. H.Organizationsformen von Erbinformation Im Hinblick auf die Menzerathsche Regel. In: Das Menzerathsche Gesetz in informationsverarbeitenden Systemen; Altmann, G.; Schwibbe, M. H., Eds.; Olms: Hildesheim, 1989; pp 92–107.
15. Ferrer-i-Cancho, R.; Forns, N. The self-organization of genomes. Complexity 2010, 15, 34–36.
16. Li, W. Menzerath's law at the gene-exon level in the human genome. Complexity 2012, 17, 49–53.
17. Altmann, G. Prolegomena to Menzerath's law. Glottometrika 1980, 2, 1–10.
18. Teupenhayn, R.; Altmann, G. Clause length and Menzerath's law. Glottometrika 1984, 6, 127–138.
19. Hernández-Fernández, A.; Baixeries, J.; Forns, N.; Ferrer-i-Cancho, R. Size of the whole versus number of parts in genomes. Entropy 2011, 13, 1465–1480.
20. Baixeries, J.; Hernández-Fernández, A.; Forns, N.; Ferrer-i-Cancho, R. The parameters of Menzerath-Altmann law in genomes. J Quant Linguistics, in press.
21. Makalowski, W. Not junk after all. Science 2003, 300, 1246–1247.
22. Yazgan, O.; Krebs, J. E. Noncoding but nonexpendable: transcriptional regulation by large noncoding RNA in eukaryotes. Biochem Cell Biol 2007, 85, 484–496.
23. Carninci, P. RNA dust: Where are the Genes? DNA Res 2010, 17, 51–59.
24. Pennisi, E. ENCODE project writes eulogy for junk DNA. Science 2012, 337, 1159–1161.
25. Taft, R. J.; Pheasant, M.; Mattick, J. S. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 2007, 29, 288–299.
26. Lyons, J. Linguistic Semantics, An Introduction. Cambridge University Press.: Cambridge, 1995.

DOI 10.1002/cplx

27. Biber, D.; Johansson, S.; Leech, G.; Conrad, S.; Finegan, E.Chapter 1: Introduction: A corpus based approach to English grammar. In: Longman Grammar of Spoken and Written English; Pearson: Harlow, 1999.
28. Chung, C.; Pennebaker, J.The psychological functions of function words. In: Social Communication; Fiedler, K., Ed.; Psychology Press: New York, 2007; pp 343–359.
29. Bloom, L. One Word at a Time. Mouton: The Hague, 1973.
30. Ke, J.; Yao, Y. Analysing language development from a network approach. J Quant Linguistics 2008, 15, 70–99.
31. Ferrer-i-Cancho, R.; Riordan, O.; Bollobás, B. The consequences of Zipf's law for syntax and symbolic reference. Proc Biol Sci 2005, 272, 561–565.
32. Maclay, H.; Osgood, C. E. Hesitation phenomena in sponatenous English speech. Word 1959, 15, 19–44.
33. Cramer, I. The parameters of the Altmann-Menzerath law. J Quant Linguistics 2005, 12, 41–52.
34. Siegel, M. E. A. Like: The discourse particle and semantics. J Semantics 2002, 19, 35–71.
35. Ye, Z. When 'empty words' are not empty: Examples from the semantic analyses of some 'emotional adverbs' in Mandarin Chinese. Aust J Linguistics 2004, 24, 139–161.
36. Häsler, J.; Samuelsson, T.; Strub, K. Useful 'junk': Alu RNAs in the human transcriptome. Cell Mol Life Sci 2007, 64, 1793–1800.
37. Bel-Enguix, G.; Jiménez-López, M. D.Genetic code and verbal language: Syntactic and semantic analogies. In: Biology, Computation and Linguistics. New Interdisciplinary Paradigms; Bel-Enguix, G.; Dahl, V.; Jiménez-López, M. D., Eds.; IOS Press: Amsterdam, 2011; pp 85–103.
38. Searls, D. B. The language of genes. Nature 2002, 420, 211–217.

DOI 10.1002/cplx

4.7. Sobre los modelos de fragmentación aleatoria

Finalmente, tanto en la respuesta a Solé (2010) en Ferrer-i-Cancho y colaboradores (2013b) como en la revisión de Baixeries y colaboradores (2012) se analiza el modelo presentado por Solé (2010), basado en modelos previos desarrollados por otros investigadores (Sankoff y Ferreti, 1996): son los denominados modelos de fragmentación aleatoria.

Definido G como el tamaño total del genoma en pares de bases, con L_g el número de cromosomas haploide y $L_c = G/L_g$ la longitud media de cada cromosoma, en un modelo de fragmentación aleatoria (Sankoff y Ferreti, 1996; Solé, 2010) se realiza el siguiente proceso (revisado en Baixeries *et al.*, 2012) en dos estadios:

1.- En un primer estadio se generan los valores de G y L_g para un grupo de N especies, siguiendo el siguiente proceso:

- G se elige uniformemente al azar en el intervalo (G_m, G_M).
- L_g se escoge uniformemente al azar en el intervalo (L_m , L_M) y de forma independiente a G.
- Dados G y L_g, L_c se calcula $L_c = G/L_g$.

2.- En un segundo estadio se genera la longitud L_g de cada cromosoma aplicando un proceso de fragmentación aleatoria de la cadena de ADN (De *et al.*, 2001), por ejemplo seleccionando uniformemente al azar L_g puntos de rotura a lo largo de la secuencia de longitud G (Solé, 2010).

La propuesta de Sankoff y Ferreti (1996) tenía en consideración algunas evidencias empíricas, como el imponer una longitud máxima a cada cromosoma (de acuerdo con Schubert y Oud (1997)), y se cuestiona que se entiende por "aleatorio" explorando dos modelos, el uniforme y el proporcional, en los que respectivamente cada cromosoma se elige de forma equiprobable o proporcionalmente a su longitud. No obstante, Sankoff y Ferreti (1996) ignoran por simplicidad la existencia de centrómeros, fundamentales en la estructura cromosómica (Rago y Cheeseman, 2013), y contrastan con un número limitado de especies su modelo (De *et al.*, 2001).

Solé (2010) recoge el testigo, aunque ignorando las restricciones impuestas por sus predecesores para la formación de cromosomas, y plantea un modelo de fragmentación aleatoria en el que supone que G y L_g son magnitudes independientes, de manera que trivialmente la ley de Menzerath-Altmann (ecuación 1) se recupera con $b = -1$ y $c = 0$, luego $L_c = aL_g^{-1}$, y además se llega a cromosomas que no serían viables,

una consecuencia inevitable de los modelos de fragmentación aleatoria (Baixeries *et al.*, 2012). Nuestra línea de trabajo plantea por tanto dos contraargumentos a los modelos de fragmentación aleatoria del ADN (Baixeries *et al.*, 2012):

1. Como se ha visto, los datos actuales de especies secuenciadas no corroboran $L_c = aL_g^{-1}$ para todos los grupos de especies (Hernández-Fernández *et al.*, 2011). En todo caso, un buen modelo debería dar cuenta de todos los datos disponibles y no solo ajustar una parte de ellos.
2. En consecuencia, planteamos que G y L_g no son independientes, como no excluyen los precedentes de Sankoff y Ferreti (1996).

Sin haber todavía resuelto de forma definitiva el problema de la fragmentación del ADN en cromosomas, lo que sí hemos demostrado algebraicamente es que no puede descartarse la dependencia entre G y L_g y que, si un cromosoma bien formado debe poseer al menos un centrómero y dos telómeros (Sankoff y Ferreti, 1996), los modelos de fragmentación aleatoria como el de Solé (2010) generan un número nada desdeñable de organismos con cromosomas –o partes de cromosomas– vacíos (Baixeries *et al.*, 2012).

En nuestra opinión debe considerarse que un cromosoma posee unas partes ineludibles y cada una de ellas debe tener un número de bases mínimo, por definición diferente de cero (Baixeries *et al.*, 2012), y un máximo debido a límites biofísicos (Schubert y Oud, 1997). Schubert (2007) nos alerta de que los intervalos en los que es viable un cromosoma están relacionados con el tamaño total del genoma, el número de cromosomas, la estructura del cariotipo y los mecanismos de intercambio en la recombinación. Los teoremas y corolarios de Baixeries y colaboradores (2012) exploran la independencia entre G y L_g, y conducen a demostrar que en el modelo de fragmentación aleatoria de Solé (2010) se da $L_c = aL_g^{-1}$.

Imponiendo restricciones biológicas realistas se pierde la independencia entre G y L_g y se pierde entonces la relación $L_c = aL_g^{-1}$ tal como hemos demostrado matemáticamente (Baixeries *et al.*, 2012). El *erratum* presentado en la revista *Biosystems* (Ferrer-i-Cancho, Baixeries *et al.*, 2013) afina en los argumentos estadísticos pues no se debe confundir la independencia entre dos variables con su independencia en promedio.

Reflexiones sobre el significado de la presencia de la ley de Menzerath hallada en el cariotipo	**Artículos**	**Estado de la cuestión**
No hay correlación entre la longitud del genoma y el número de cromosomas	Solé (2010) la descarta como trivial o absurda (por la relevancia del ADN basura), admitida en pájaros y peces cartilaginosos (Hernández-Fernández *et al.*, 2011)	Abierta en el caso de pájaros y peces cartilaginosos (Hernández-Fernández et al., 2011): a estudiar corpus más completos.
Correlación negativa entre la longitud del genoma y el número de cromosomas, en algunos grupos taxonómicos	Hallada en peces agnatos y angiospermas (Vinogradov, 2001; Hernández-Fernández *et al.*, 2011)	Clara en angiospermas y agnatos.
La dominancia del ADN no codificante (ADN basura) invalida estudios de lingüística genómica en el nivel cromosómico	Defendida por Solé (2010), Encode (2012) pone en duda el concepto y la relevancia del ADN no codificante. Doolittle (2013) para una revisión crítica de Encode.	Encode (2012) invita a que el concepto de ADN basura se revise. El lenguaje posee "basura" (Ferrer-i-Cancho, Forns *et al.*, 2013).
Relación trivial entre la longitud del genoma y el número de cromosomas	Solé (2010) la sostiene presentando únicamente los datos de mamíferos y de todas las plantas juntas (angiospermas y gimnospermas) y apoyándose en un modelo de fragmentación aleatoria (Sankoff y Ferreti, 1996). Respondido en los artículos de este trabajo, y contrario a evidencias como las presentadas por Wilde y Schwibbe (1989)	Refutada en Baixeries *et al.* (2012, 2013), Hernández-Fernández *et al.* (2011) y Ferrer-i-Cancho, Forns *et al.* (2013).
Crítica sobre el nivel de estudio (cromosómico)	Realizada por Solé (2010), Li (2012) analiza en el nivel de exones la ley de Menzerath-Altmann, y nuestro trabajo el nivel cromosómico. Molina y van Nimwegen (2009) exploran diversos niveles en procariotas.	Averiguar si existen leyes de escala en el genoma implica no renunciar a explorar ningún nivel de estudio. Se ha mostrado que el nivel cromosómico es adecuado.
Metáforas inapropiadas entre el lenguaje y el genoma	Solé (2010) las critica, aunque tanto nuestro trabajo como el de otros (Li, 2012), y la perspectiva de Searls (2002) o Bel-Enguix y Jiménez-López (2011) animan a seguir en esta línea.	Hay semejanzas mutuas entre genoma y lenguaje, y los datos, test y artículos presentados aquí avalan la significancia y adecuación de las metáforas entre lenguaje y genoma.

Tabla 4.2: Resumen de algunas de las críticas planteadas a la presencia no trivial de la ley de Menzerath en el cariotipo, y el estado de la cuestión hasta la fecha.

Como puede comprobarse, las correcciones del *erratum* no alteran la conclusión de que el modelo de fragmentación aleatoria proporciona un ajuste insuficiente al estudio cuantitativo de la longitud de cada cromosoma, respecto el número de cromosomas, y a la no trivialidad de la ley de Menzerath-Altmann en el genoma (Baixeries *et al.*, 2012).

Por último, la tabla 4.2. resume algunas de las críticas o dudas sobre el significado de la presencia de la ley de Menzerath hallada en el cariotipo. Hay todavía muchas cuestiones abiertas pero es sin duda un tema apasionante en el que seguir explorando. Confiamos en que las analogías entre la lingüística y la genómica sigan dando sus frutos.

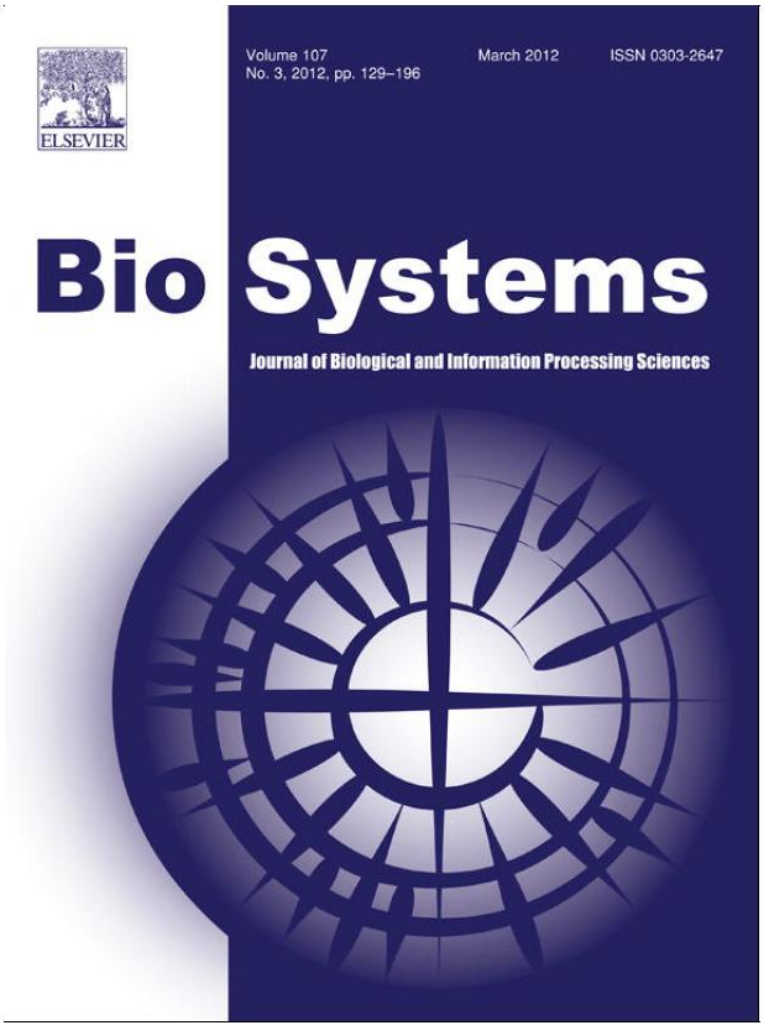
4.8. Baixeries, J., Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2012). *Random models of Menzerath-Altmann law in genomes.*

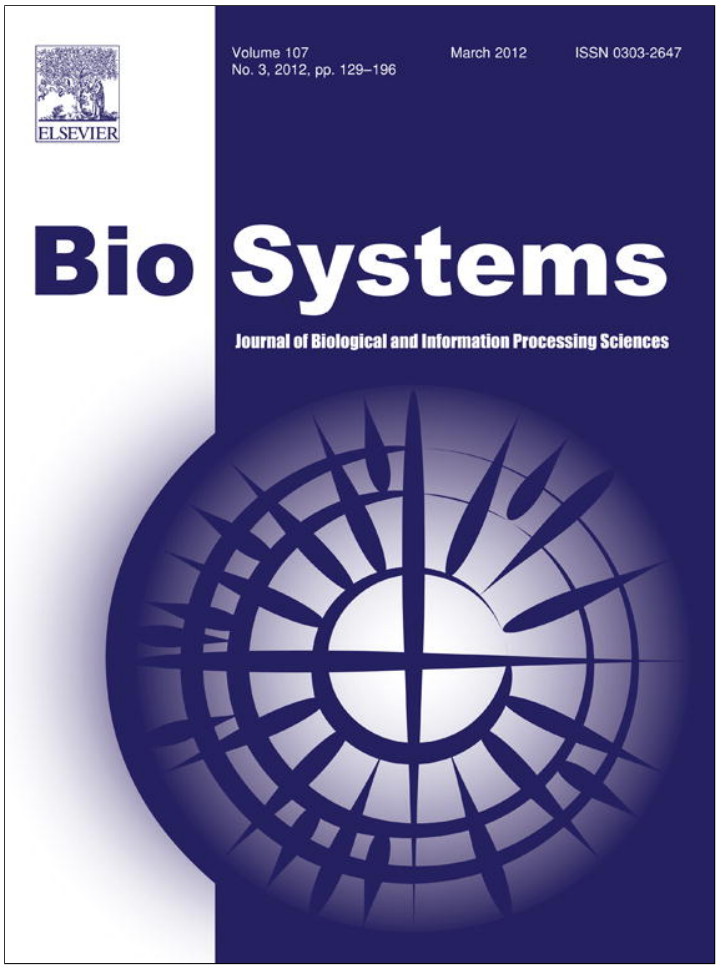
Se incluye en este apartado el artículo:

- Baixeries, J., Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2012). Random models of Menzerath-Altmann law in genomes. *BioSystems* 107 (3), 167–173.

Y seguidamente su *erratum*:

- Ferrer-i-Cancho, R., Baixeries, J. y Hernández-Fernández, A. (2013c). *Erratum* to "Random models of Menzerath-Altmann law in genomes" (*BioSystems* 107 (3), 167–173), *BioSystems* 111 (3), 216-217.





BioSystems 107 (2012) 167–173

Contents lists available at SciVerse ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Random models of Menzerath–Altmann law in genomes

Jaume Baixeries[a], Antoni Hernández-Fernández[b,c], Ramon Ferrer-i-Cancho[c,*]

[a] Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, LARCA Research Group, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain
[b] Departament de Lingüística General, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona (Catalonia), Spain
[c] Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain

ARTICLE INFO

Article history:
Received 13 October 2011
Received in revised form
16 November 2011
Accepted 28 November 2011

Keywords:
Random breakage
Power law
Genome size
Chromosome number
Menzerath–Altmann law

ABSTRACT

Recently, a random breakage model has been proposed to explain the negative correlation between mean chromosome length and chromosome number that is found in many groups of species and is consistent with Menzerath–Altmann law, a statistical law that defines the dependency between the mean size of the whole and the number of parts in quantitative linguistics. Here, the central assumption of the model, namely that genome size is independent from chromosome number is reviewed. This assumption is shown to be unrealistic from the perspective of chromosome structure and the statistical analysis of real genomes. A general class of random models, including that random breakage model, is analyzed. For any model within this class, a power law with an exponent of −1 is predicted for the expectation of the mean chromosome size as a function of chromosome length, a functional dependency that is not supported by real genomes. The random breakage and variants keeping genome size and chromosome number independent raise no serious objection to the relevance of correlations consistent with Menzerath–Altmann law across taxonomic groups and the possibility of a connection between human language and genomes through that law.

1. Introduction

Human language, music and genomes share a common statistical pattern: the mean size of the parts tends to decrease as the number of parts increases (Ferrer-i-Cancho and Forns, 2009). On the one hand, it is known in quantitative linguistics that the size of a construct (e.g., a sentence) tends to decrease as the number of constituents (e.g., clauses) increases (Altmann, 1980; Teupenhayn and Altmann, 1984). This behavior has been studied at many linguistic levels: e.g., morphemes, words, sentences (Altmann, 1980; Teupenhayn and Altmann, 1984) and also in music (Boroda and Altmann, 1991). This recurrent statistical pattern is presently known as Menzerath–Altmann law, the mathematical function that is used to define the relationship between the size of a construct and the size of its units (Cramer, 2005). If the mean size of the parts is S_P (e.g., the length of a clause in words), with the size of the whole S_W (e.g., the length of a sentence in clauses), the law states that.

$$S_P \sim {S_W}^b e^{cS_W} \quad (1)$$

where b and c are constants. On the other hand, the genomes of many groups of species exhibit a parallel qualitative behavior: the mean length of chromosomes tends to increase as the number of chromosomes of a species increases (Ferrer-i-Cancho and Forns, 2009; Wilde and Schwibbe, 1989). The significance of this statistical coincidence is a matter of current debate (Hernández-Fernández et al., 2011; Solé, 2010). One of the main arguments that has been raised against the relevance of the finding of patterning consistent with the law in genomes is a simple model of random breakage that can account for a negative correlation between mean chromosome length and chromosome number (Solé, 2010). Here the actual capacity of this model for explaining the real dependency between mean chromosome length and chromosome number will be analyzed. It will be argued that this random breakage model and all the variants within the same class of models are equivalent to a particular case of Menzerath–Altmann law, namely Eq. (1) with $b = -1$ and $c = 0$ and it will be shown that real genomes deviate significantly from Menzerath–Altmann law with precisely these concrete parameters.

* Corresponding author. Tel.: +34 934137870.
E-mail addresses: jbaixer@lsi.upc.edu (J. Baixeries), antonio.hernandez@upc.edu (A. Hernández-Fernández), rferrericancho@lsi.upc.edu (R. Ferrer-i-Cancho).

0303-2647/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2011.11.010

Let us define G as the size of a genome in base pairs, L_g as its number of chromosomes and $L_c = G/L_g$ as its mean chromosome length. It has been argued that Menzerath–Altmann law in genomes could be simply explained through a simple model of random breakage that generates the information about an organism in two stages. In the first stage, the values of G and L_g are generated for a group of N organisms through the following procedure:

- G is chosen uniformly at random within the interval (G^m, G^M).
- L_g is chosen uniformly at random within the interval (L_g^m, L_g^M) and independently from G.
- Given G and L_g, L_c is computed through $L_c = G/L_g$.

In the second stage, the length of each of the L_g chromosomes is generated applying a random breakage procedure (De et al., 2001), i.e. by selecting L_g breaking points along the sequence of length G uniformly at random (Solé, 2010). This procedure is also known as random fragmentation (Sankoff and Ferretti, 1996).

Notice that the independence between G and L_g assumed in the first stage is not a decision that is taken for simplicity but a property that it is claimed to actually hold in real genomes (Solé, 2010). Here it will be argued the opposite, namely that such independence is unrealistic from the perspective of chromosome structure and the statistical properties of actual genomes. It will be shown that the assumption of independence between G and L_g has negative consequences for the suitability of the model for actual genomes. Firstly, the assumption is no supported by real data (Hernández-Fernández et al., 2011) and, under certain parameter choices, can lead to organisms that are not well-formed or viable. Secondly, it will be shown that, due to such independence, $E[L_c|L_g]$, the expectation of L_c given L_g, follows a power law with a -1 exponent, i.e.

$$E[L_c|L_g] \sim L_g^{\beta} \tag{2}$$

with $\beta = -1$. This particular power-law (a particular case of Menzerath–Altmann law, i.e. Eq. (1) with $S_P = E[L_c|L_g]$, $S_W = L_g$, $b = -1$ and $c = 0$) will be shown to be insufficient to explain the actual dependency between L_c and L_g.

The remainder of the article is organized as follows. Section 2 evaluates the suitability of the assumption of independence between G and L_g of the random breakage model to real genomes from different perspectives. Section 2.1 argues that there is a structural dependency between G and L_g. A well-formed chromosome must have centromere and two telomeres and this imposes a lower bound on the minimum value of G given L_g. However, this is not a decisive argument against the random breakage model because chromosomes may be so large that this theoretical lower bound may have no observable effect. Section 2.2 reviews the statistical evidence of a dependency between G and L_g that the random breakage model denies. This is still not decisive evidence against the suitability of the random breakage model because the null hypothesis of independence between G and L_g cannot be rejected in two of the major groups of organisms reviewed, birds and cartilaginous fishes. A decisive rejection of the random breakage model (indeed a rejection of a wider class of models) for all the major groups of organisms reviewed arrives in Section 2.3. This subsection shows that

- All the major groups of organisms reviewed (including birds and cartilaginous fishes) deviate significantly from Menzerath–Altmann law with these particular parameters using complementary mathematical and statistical arguments. Indeed, Section 2.3 rejects a generalization of Menzerath–Altmann law with $b = -1$ and $c = 0$ extended by adding an additive term.
- Menzerath–Altmann law with $b = -1$ and $c = 0$ is equivalent to independence between G and L_g.

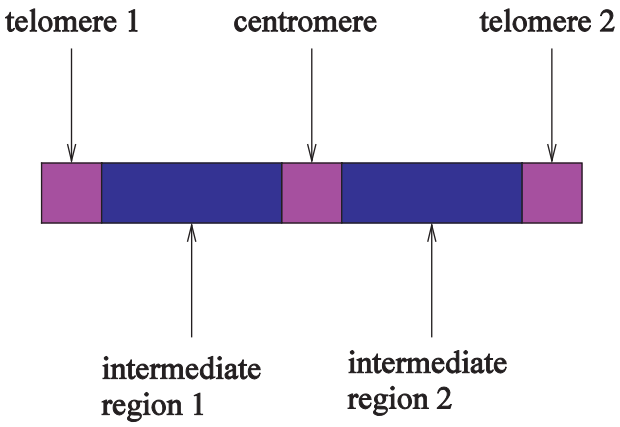


Fig. 1. Scheme of a metacentric chromosome, which has five parts, i.e. two telomeres, one centromere and two intermediate regions between a telomere and the centromere.

- The random breakage model and other models within the same class are equivalent to Menzerath–Altmann law with $b = -1$ and $c = 0$.

Section 2.3 also presents the exact mathematical dependency between L_g and L_c of the random breakage model with uniformly distributed G and L_g (Solé, 2010). Section 3 checks and clarifies the need of independence between G and L_g to obtain Menzerath–Altmann law with $b = -1$ and $c = 0$. This section shows that introducing a structural constraint (i.e. the centromere and the two telomeres must have a minimum size) in genomes that would be formed a priori by choosing chromosome number and genome size independently, leads to a dependency between L_g and L_c that deviates from Menzerath–Altmann law with $b = -1$ and $c = 0$ and whose exact mathematical form is presented. Section 4 summarizes all the problems surrounding Solé's (2010) assumption of independence between genome and chromosome level and explores other pitfalls of his random breakage model.

2. The Problems of Independence Between Genome Size and Chromosome Number

2.1. Well-formed or Viable Chromosomes

The random breakage model above does not impose any constraint on chromosome length but organisms with empty chromosomes are produced when $G < L_g$. In contrast, models of chromosome length evolution (De et al., 2001; Sankoff and Ferretti, 1996) consider that

> "...a viable and functional chromosome must minimally contain a centromere and two telomeres (and at least one gene whose function is not duplicated elsewhere in the genome). This imposes a lower bound on the size of a chromosome, on a purely structural basis. Finally, from the genetic viewpoint, there is reason to believe that for meiosis to be completed successfully, each chromosome must be of length sufficient for at least one crossover to be expected among the four aligned strands before they segregate into two pairs." (Sankoff and Ferretti, 1996, p. 8).

For simplicity, let us assume that a chromosome is made of three main parts: a centromere and two telomeres (Fig. 1). c^m, t_1^m, t_2^m, g_1^m, g_2^m and are defined, respectively as the minimum size of the centromere, the two telomeres and two intermediate regions

delimited by a telomere and the centromere. In general, a viable chromosome must satisfy, at least, the condition

$$(c^m + t_1^m + t_2^m + g_1^m + g_2^m)L_g \leq G \quad (3)$$

Following Sankoff and Ferretti (1996), let us consider that a viable genome must contain at least one centromere and two telomeres and that the intermediate regions could be empty. Our course, a further constraint could be added, namely that even the intermediate regions cannot be empty. Our choice of a general chromosome structure (Fig. 1) is motivated by the goal of showing the structural dependency between the genome and the chromosome scale from a mathematical point of view. Of course, more biologically realistic scenarios can be developed from it.

Assuming c^m, t_1^m, $t_2^m \geq 1$ and $g_1^m = g_2^m = 0$ as in Sankoff and Ferretti (1996), Eq. (3) yields $G \geq 3L_g$. Applying the parameters of Solé's random breakage model with $g_1^m = g_2^m = 0$, the viability condition in Eq. (3) is always satisfied if

$$(c^m + t_1^m + t_2^m)(L_g^M - 1) \leq G^m + 1 \quad (4)$$

holds. Next the parameter conditions that warrant that the random breakage model does not produce empty chromosomes or chromosomes with at least one empty part will be derived. Only the two telomeres and the centromere will be considered as main parts.

Empty chromosomes (chromosomes with no base pair) are avoided by imposing $c^m + t_1^m + t_2^m \geq 1$ in Eq. (4), which gives

$$L_g^M \leq G^m + 2, \quad (5)$$

in agreement with Hernández-Fernández et al. (2011). Interestingly imposing the condition that none of the main parts is empty, i.e. c^m, t_1^m, $t_2^m \geq 1$ on Eq. (4) it is obtained

$$3L_g^M \leq G^m + 4. \quad (6)$$

Notice that the independence between G and L_g needs Eq. (5) if genomes with chromosomes of length zero cannot be generated or Eq. (6) if genomes with empty main parts cannot be produced. Next the results above will be applied to the parameters chosen by Solé (2010) to simulate the random breakage model. It will be shown that in one of the parameter settings, chromosomes with non-viable structure are generated.

The random breakage model was simulated with two parameter settings (Solé, 2010). In both, $N = 500$, $G^M = 10{,}000$, $L_g^m = 10$ and $L_g^M = 200$ were used. In simulation (a) $G^m = 9000$ was used (for Fig. 2(a) of Solé (2010)) and in simulation (b) $G^m = 50$ was used (for Fig. 2(a) of Solé (2010)). While parameter setting (a) satisfies the necessary and sufficient condition for absence of empty chromosomes (Eq. (5)) and the necessary and sufficient condition for absence of chromosomes with empty main parts (Eq. (6)), parameter setting (b) violates both conditions (in this case $L_g^M > G^m + 2$, which implies $3L_g^M > G^m + 4$ as $G^m > 0$). Next we will focus on parameter setting (b), which was used to explain qualitatively the high dispersion in the actual relationship between L_c and L_g in angiosperm plants. The issue of whether $G^m = 50$ (or $G^m = 9000$ in parameter setting (a)) is realistic enough, knowing that a minimum complete living organism has been argued to require about $G^m = 6 \times 10^5$ (Luisi, 2006 and references therein), will be left aside and we will focus on the implications for viability of such as small minimum genome size from a purely structural perspective.

The random breakage model was simulated varying N and using the remainder of parameters from setting (b). Our Fig. 2 shows the growth of the number of organisms with empty chromosomes as N grows. For $N = 500$, 2.9 ± 1.7 organisms with empty chromosomes are produced on average while 13.3 ± 3.6 organisms with at least one chromosome with empty main parts are produced on average. However, if the simulations had taken the value of N of angiosperm plants from the sample studied by Hernández-Fernández et al. (2011), i.e. $N = 4706$, then 27.6 ± 5.2 organisms with empty chromosomes and 125.1 ± 11.0 organisms with at least one empty main chromosome part would be obtained on average.

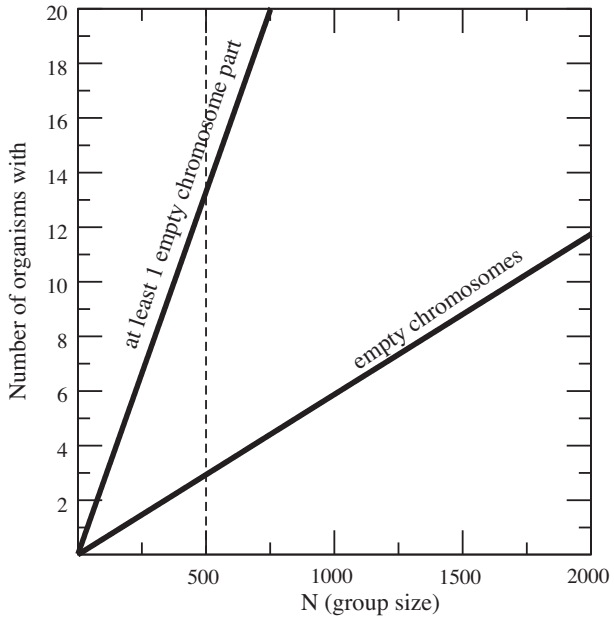


Fig. 2. Visual analysis of the viability of the organisms generated by the random breakage model as a function of N, the group size. Two solid lines are shown, one for the linear growth of the number of organisms with empty chromosomes (organisms such that $L_g > G$) that are produced by the random breakage model as the group size N increases and another one for the linear growth of the number of organisms with at least one empty main chromosome part (organisms such that $3L_g > G$; only the two telomeres and one centromere are considered as main parts). As a guide to the eye, a vertical line with $N = 500$ (dashed line) is included. The solid curves were obtained by averaging the results of simulating the random breakage model over 10^6 replicas. All the parameters of the model were taken from Fig. 2(b) of Solé (2010) except N.

We have studied the problem of viability by just requiring chromosomes or main chromosome parts to not be empty, which both define necessary but not sufficient conditions for actual chromosome viability as parts of size one may still not be large enough to warrant viability. These simple requirements for chromosome viability have been chosen to illustrate the problem of denying the dependency between G and L_g. A deeper analysis of viability could be performed with more realistic values of c^m, t_1^m, t_2^m, g_1^m, g_2^m or G^m (Luisi, 2006).

2.2. Independence Between Chromosome Number and Genome Size in Well-formed Chromosomes

In most major groups of organisms examined by Ferrer-i-Cancho and Forns (2009), a significant correlation between genome size and chromosome number has been found (Table 1). Genome size tends to decrease as the number of chromosomes increases in angiosperm plants and jawless fishes while it tends to increase as chromosome number increases in many other groups of species: fungi, gymnosperm plants, insects, reptiles, mammals, ray-finned fishes and amphibians (Table 1). The assumption of independence between genome size and chromosome number made by the random breakage is only valid for birds and cartilaginous fishes according to Table 1. From the biological perspective, it has been suggested that the negative correlation between G and L_g in angiosperms would result from a trade-off between recombination mechanisms (Vinogradov, 2001).

Table 1
Summary of results on the dependency between L_c (mean length of chromosomes) and L_g (chromosome number) and the dependency between G (genome size) and L_g. *Yes* is used to indicate statistically significant correlations at a significant level of 0.05. + and − are used to indicate the sign of the correlation. (*) is used to indicate results borrowed from Ferrer-i-Cancho and Forns (2009). The significant and non-significant correlations remain if the updated dataset in Hernández-Fernández et al. (2011) is used. (**) is used to indicate results borrowed from Hernández-Fernández et al. (2011). The results on the correlation between G and L_g in angiosperm plants are supported independently by the pioneering research by Vinogradov (2001). The hypothesis of linearity between G and L_g was rejected by means of a non-parametric linearity test (Hernández-Fernández et al., 2011).

Group	Correlation between L_c and L_g (*)	Correlation between G and L_g (**)	Non-linear dependency between G and L_g (**)
Fungi	Yes −	Yes +	Yes
Angiosperm plants	Yes −	Yes −	Yes
Gymnosperm plants		Yes +	Yes
Insects	Yes −	Yes +	Yes
Reptiles	Yes −	Yes +	Yes
Birds	Yes −		Yes
Mammals	Yes −	Yes +	Yes
Cartilaginous fishes	Yes −		Yes
Jawless fishes	Yes −	Yes −	Yes
Ray-finned fishes		Yes +	Yes
Amphibians	Yes +	Yes +	Yes

2.3. The Dependency Between Mean Chromosome Length and Chromosome Number in the Random Breakage Model

Firstly, it will argued that a power law with a $\beta = -1$ exponent and an additive constant, i.e.

$$E[L_c|L_g] = \frac{a}{L_g} + d, \tag{7}$$

where a is a proportionality constant and d is an additive constant and $L_g > 0$, is not supported by the major groups of organisms considered by Ferrer-i-Cancho and Forns (2009) and Hernández-Fernández et al. (2011). To see it, notice that the fact that $L_c = G/L_g$ means that Eq. (7) is equivalent to

$$E[G|L_g] = a + dL_g, \tag{8}$$

namely that $E[G|L_g]$ is a linear function of L_g. Interestingly, the linear dependency between genome size and chromosome number that Eq. (8) defines has been rejected by means of a non-parametric linearity test (Table 1), which indirectly rejects Eq. (7). Further support for the non-linear dependency between G and L_g can be obtained assuming that G is exactly the sum of all chromosome sizes, i.e. (Solé, 2010)

$$G = \sum_{i=1}^{L_g} g_i, \tag{9}$$

where g_i is the size in base pairs of the ith chromosome. The definition in Eq. (9) implies that if $L_g = 0$ was reached then $E[G|L_g] = 0$, which gives $a = 0$ in Eq. (8). Notice that $a = 0$ implies $d \neq 0$ as a genome cannot be empty when $L_g > 0$. Two predictions of $a = 0$ on Eqs. (7) and (8), $E[L_c|L_g] = d$, i.e. no significant correlation between L_c and L_g, and $E[G|L_g] = dL_g$ with $d \neq 0$, i.e. a significant correlation between G and L_g, are not found in any major group of organisms except in gymnosperm plants and ray-finned fishes (Table 1). The fact that this alternative non-parametric linearity test is not able to reject linearity for these two groups of organisms is not a problem for our arguments: Eq. (7) can be discarded because Eq. (8) is rejected or because a significant negative correlation between L_c and L_g is not found as in gymnosperm plants and ray-finned fishes (Table 1).

Secondly, a general class of random models, of which the random breakage model is a particular case, will be presented. It will be shown that the models of this class obey Eq. (7) with $d = 0$ and a being determined solely by the distribution of G. For the random breakage model above it will be shown that

$$a = \frac{G^M(G^M - 1) - (G^m + 1)G^m}{2(G^M - G^m - 1)}. \tag{10}$$

Consider an extended random model that generates the information about an organism in two stages. It is assumed that $G, L_g \geq 1$. In the first stage, the values of G and L_g for a group of N organisms through the following procedure:

- G is chosen at random from a certain distribution A.
- L_g is chosen at random from a certain distribution B and independently from G.
- Given G and L_g, L_c is computed through $L_c = G/L_g$.

In the second stage, the length of each of the L_g chromosomes is generated following a certain procedure C.

Uniform distributions were selected for distributions A and B in the random breakage model used to explain correlations consistent with Menzerath–Altmann law in genomes (Solé, 2010). Reciprocal translocation and random breakage are examples of mechanisms that have been considered for procedure C (De et al., 2001; Li et al., 2011; Sankoff and Ferretti, 1996; Solé, 2010). However, the distribution B and procedure C are irrelevant for $E[L_c|L_g]$ in this general class of random models. In all the calculations that follow, true independence between G and L_g is assumed. This means that we assume that chromosomes of length zero or chromosome with empty main parts are not discarded. The following general theorem will be used to derive the exact mathematical form of $E[L_c|L_g]$ in this general class of random models.

Theorem 1. *Two random natural numbers X and Y, such that $X > 0$, are independent if and only if $Z = Y/X$ satisfies $E[Z|X] = E[Y]/X$.*

Proof. By definition of independence between X and Y,

$$p(Y = y|X = x) = p(X = x) \tag{11}$$

for any x and y. Multiplying by $z = y/x$ on both sides of the previous equality, it is obtained

$$p(Y = y|X = x)z = p(X = x)\frac{y}{x} \tag{12}$$

for any x and y. Applying the fact that $Y = y$ is equivalent to $Y/X = y/x$, Eq. (12) leads to

$$p(Z = z|X = x)z = p(X = x)\frac{y}{x} \tag{13}$$

for any x and y thanks to the definitions $Z = X/Y$ and $z = y/x$.

Summing over y on both sides of the equality of Eq. (13) yields

$$\sum_y p(Z = z|X = x)z = \frac{1}{x}\sum_y p(Y = y)y \tag{14}$$

for any x. By the definition of expectation and conditional expectation and $z = y/x$, Eq. (14) is transformed into

$$E[Z|X = x] = \frac{E[Y]}{x} \tag{15}$$

for any x, which finally gives $E[Z|X] = E[Y]/X$ as we wanted to prove.

For the general class of random models above, the following corollary shows that $E[L_c|L_g]$ is a power law whose exact form is determined solely by distribution A (distribution B and procedure C are irrelevant) and that the exponent of the law does not depend on A.

Corollary 1. *The general class of random models above is equivalent to the class of models yielding*

$$E[L_c|L_g] = aL_g^{\beta}, \tag{16}$$

with $a = E[G]$ *and* $\beta = -1$.

Proof. Straightforward from Theorem 1 taking X as L_g, Y as G, Z as L_c and $a = E[G]$.

The exact form of $E[L_c|L_g]$ for the random breakage model (Solé, 2010) will be derived applying some elementary results on uniformly distributed random numbers that will be obtained first.

Theorem 2. *Let X be a uniformly distributed integer random variable within the interval (x^m, x^M), i.e. $p(X=x) = 1/(x^M - x^m - 1)$ if $x \in (x^m, x^M)$ and $p(X=x) = 0$ otherwise.*

The expectation of X is

$$E[X] = \frac{x^M(x^M-1) - (x^m+1)x^m}{2(x^M - x^m - 1)}. \tag{17}$$

Proof.

$$E[X] = \sum_{x=x^m+1}^{x^M-1} p(X=x)x = p(X=x)\left(\sum_{x=1}^{x^M-1} x - \sum_{x=1}^{x^M} x\right)$$

$$= \frac{x^M(x^M-1) - (x^m+1)x^m}{2(x^M - x^m - 1)}. \tag{18}$$

Corollary 2. *The expectation of G and L_c in the random breakage model (Solé, 2010) are respectively*

$$E[G] = \frac{G^M(G^M-1) - (G^m+1)G^m}{2(G^M - G^m - 1)} \tag{19}$$

$$E[L_g] = \frac{L_g^M(L_g^M-1) - (L_g^m+1)L_g^M}{2(L_g^M - L_g^m - 1)}. \tag{20}$$

Proof. Applying Theorem 2 for G with $x^m = G^m$ and $x^M = G^M$ gives Eq. (19). Applying Theorem 2 for L_g with $x^m = L_g^m$ and $x^M = L_g^M$ gives Eq. (20).

Corollary 3. *In the random breakage model (Solé, 2010), the expectation of L_c when L_g is given is $E[L_c|L_g] = a/L_g$ with a defined as in* Eq. (10).

Proof. Straightforward applying Corollaries 1 and 2.

3. The Effect of Viability on the Dependency Between the Size of the Parts and the Size of the Whole

Corollary 1 states that $E[L_c|L_g] \sim 1/L_g$ can only be achieved by independence between G and L_g. Next the effect of a particular viability concern on the shape of $E[L_c|L_g]$ will be studied in the simple random breakage model (Solé, 2010). As expected from Corollary 1, the new $E[L_c|L_g]$ deviates from a/L_g.

For simplicity, let us assume that the main three parts, i.e. two telomeres and the centromere, must have of least length $k \geq 0$ each, i.e. c^m, t_1^m, $t_2^m = k$ for an organism to be viable (we assume that the remainder of the parts could be empty). Thus a viable organism needs $3kL_g \leq G$ (when $k = 0$ no viability constraint is imposed). Under these constraints, the next theorem describes the expected mean chromosome length as a function chromosome number.

Theorem 3. *Consider that the chromosome number L_g ($L_g > 0$) and the genome size G are natural numbers. Consider also that L_g is given and that then G must be generated. If a chromosome has three main parts that each must have length equal or greater than k (with $k \geq 0$) in order to be viable (the remainder of the parts can be empty) it follows that*

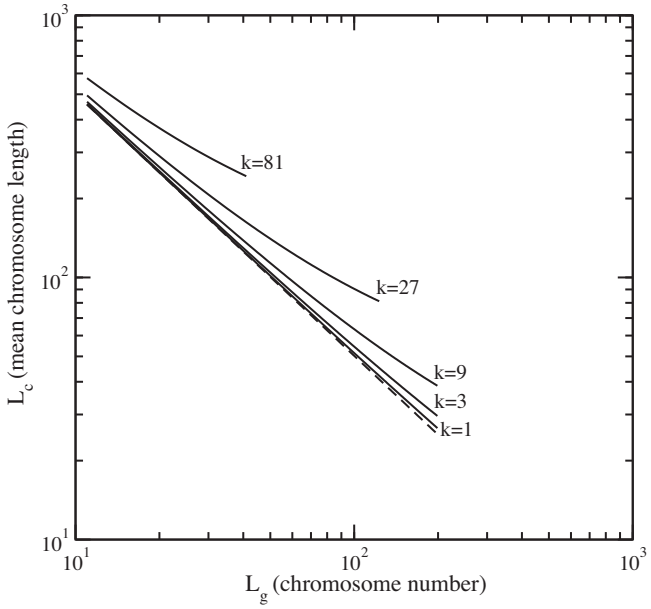


Fig. 3. A comparison of the expected value of L_c given L_g as a function of the viability factor k in the extended random breakage model. In this extension, an organism is not viable unless $3kL_g \leq G$, where L_g is the chromosome number of the organism and G is its genome size. According to this criterion, all non-viable organisms are removed. Dashed line is used for $k = 0$ (no viability constraints) while solid line is used for $k > 0$. Parameter setting (b) from Solé (2010), i.e. $G^m = 50$, $G^M = 10{,}000$, $L_g^m = 10$ and $L_g^M = 200$, was used. Theoretical expectations for the mean chromosome length for different values of k were obtained applying Theorem 3.

(1) *If $3kL_g \leq G^m + 1$ then all organisms of L_g chromosomes are viable, i.e. $E[L_c|L_g] = a/L_g$, with a defined as in* Eq. (10).
(2) *If $3kL_g \geq G^M$ then no organism of L_g chromosomes is viable, i.e. $E[L_c|L_g]$ is undefined.*
(3) *If $G^m + 1 < 3kL_g < G^M$ then some organisms of L_g chromosomes are not viable and*

$$E[L_c|L_g] = \frac{G^M(G^M-1) - 3kL_g(3kL_g-1)}{2L_g(G^M - 3kL_g)}. \tag{21}$$

Proof. The proof of (1) follows from Corollary 3. The proof of (2) is straightforward as no viable organism can be generated in that parameter setting. As for the proof of (3), consider the value of G of a viable organism follows a uniform distribution within the open interval $(3kL_g - 1, G^M)$. That is, the distribution keeps being uniform as in the original random breakage model (Solé, 2010) when viability constraints are taken into account but with a left-truncation introduced by viability. By Theorem 2 with $x^m = 3kL_g - 1$ and $x^M = G^M$, it is obtained

$$E[G|L_g] = \frac{G^M(G^M-1) - 3kL_g(3kL_g-1)}{2(G^M - 3kL_g)}. \tag{22}$$

Knowing that $E[L_c|L_g] = E[G/L_c|L_g] = E[G|L_g]/L_g$, it is finally obtained

$$E[L_c|L_g] = \frac{G^M(G^M-1) - 5kL_g(5kL_g-1)}{2L_g(G^M - 5kL_g)} \tag{23}$$

as we wanted to prove.

Fig. 3 shows $E[L_c|L_g]$ for different values of k employing parameter setting (b) of (Solé, 2010). It can be seen that k introduces a dependency between G and L_g that leads to a deviation between $E[L_c|L_g]$ and a power-law with -1 exponent (a straight line in double logarithmic scale with -1 slope) that is expected from independence between G and L_g. The exact dependency between G and L_g that was chosen for Theorem 3 was used simply to illustrate the consequences of breaking the independence that has been assumed

and defended (Solé, 2010). More realistic dependencies between G and L_g could be defined.

4. Discussion

There is a serious design flaw in the random breakage model: the explicit assumption that genome size and chromosome number are and must be independent. Referring to the models by Sankoff and Ferretti (1996) and De et al. (2001), it has been stated that

> "The good fit obtained by these theoretical approaches and the current knowledge on the fluid nature of chromosomal rearrangements through time rule against any special multiscale link between genome-level and chromosome-level patterns" (Solé, 2010).

This argument is problematic for many reasons:

- In these models (Sankoff and Ferretti, 1996; De et al., 2001), chromosomes lengths are generated from a constant genome size and a constant chromosome number that are borrowed from a target species. These numbers are kept constant by the dynamical rules of the models. Therefore, these models constitute no evidence against "multiscale links".
- As we have discussed in the present article, a multiscale link is determined by the kind of constraints on chromosome structure (e.g., non-empty chromosome parts) that models of chromosome evolution have taken into account (Sankoff and Ferretti, 1996; De et al., 2001).
- Statistical studies have unraveled significant dependencies between genome size and chromosome number (Table 1; see also Trivers et al., 2004). It is possible to predict, for a given species, chromosome sizes by chromosome number, and furthermore, given either genome size or average chromosome length it is possible to predict the size range of all chromosomes of that species (Li et al., 2011).
- Experiments indicate that "upper and lower tolerance limits for chromosome size are apparently determined by the genome size, chromosome number and karyotype structure of a given species" (see Schubert (2007) and references therein).
- It is has been argued that multiscale links could result from the interplay between opposite biological forces (Schubert, 2007 and references therein), e.g., trade-offs between recombination mechanisms.
- It has been shown here that the independence assumption of the random breakage model leads to a "power-law" dependency between mean chromosome length and chromosome number (a particular case of Menzerath–Altmann law, Eq. (1) with $b = -1$ and $c = 0$) that is not supported by actual genome data.

Any model is a simplification of reality and many models are only focused on a few statistical properties or simply one target statistical property as the random breakage model. The key is not making simplifications that do not let a model reproduce one of its target statistical properties. In the simple random breakage model (Solé, 2010), the target distribution is the dependency between L_c and L_g. Unfortunately, the independence between G and L_g defended in the random breakage model leads to a power-law dependency between L_c and L_g with a -1 exponent that has been rejected (recall Section 2.3). The only two taxonomic groups examined to support the random breakage model, plants and mammals (Solé, 2010), are among the groups for which that power-law provides an insufficient fit.

There are still more aspects of the random breakage model (Solé, 2010) that need to be revised. Firstly, the model generates chromosome lengths that are produced by selecting break points in the genome sequence uniformly at random and independently. Random breakage is known to be a suitable model in extremal conditions, e.g., the fragmentation of DNA induced by high dosages of ^{7}Li radiation (Fuquan et al., 2005). The idea that a genome can break at any point in normal circumstances has been refuted by independent computational analyses of genomes (e.g., Peng et al., 2006; Hinsch and Hannenhalli, 2006; Ruiz-Herrera et al., 2006). The conclusion of these studies is that the breaks are found in "fragile regions" or "hotspots" (Becker and Lenhard, 2007). Besides, random breakage produces chromosome lengths that follow the broken stick distribution (Smart, 1976). A recent study shows that the gamma distribution gives the best fit among various candidates (not including the broken stick distribution) to actual chromosome lengths (Li et al., 2011). The quality of the fit a broken stick distribution to actual chromosome lengths should be evaluated and compared to that of a gamma distribution. Obtaining a sufficiently good fit to actual chromosome lengths with a broken stick distribution is a serious challenge for the random breakage model (Solé, 2010) because random breakage or fragmentation is not used as a true model but as a control for more realistic models of chromosome length evolution (Sankoff and Ferretti, 1996; De et al., 2001) and, in general, these more realistic models yield a better fit to actual data than random breakage. This challenge cannot be obviated since, at present, the actual statistical pattern of genomes that the random breakage model (Solé, 2010) is able to reproduce to a sufficient degree is not known. Besides, it should also be checked rigorously if genome sizes and chromosome number are uniformly distributed as the random breakage model assumes. Uniformity for genomes sizes is clearly not the case in general. The genome size of eukaryotes varies over five orders of magnitude but the distribution is skewed towards small values (Oliver et al., 2007). The skewness of the distribution of genome sizes has been demonstrated within angiosperm plants but the direction of the skewness (towards small values or towards large values) varies according to the subgroup under consideration (Vinogradov, 2001). It has been argued that eukaryotic genome sizes might evolve in a stochastically proportionate manner, which necessarily produces far more small genomes than large genomes, even in the absence of selection against large genomes (Oliver et al., 2007). Actual genome sizes at the level of all eukaryotes are consistent with a log-normal distribution (Oliver et al., 2007), which is predicted by proportional evolution after sufficiently long periods of time (Lewontin and Cohen, 1969).

In sum, the random breakage model and other models within the same class cannot trivially explain the dependency between mean chromosome length and chromosome number that is found in genomes. These models should be extended with realistic genome-chromosome dependencies in order to provide a satisfactory fit to actual genome data. At present, these models raise no serious objection to the possibility of a connection between linguistic systems and genomes through Menzerath–Altmann law.

Role of the Funding Source

The funder had not role in the study, design, collection, analysis and interpretation of data, the writing of the report and the decision to submit the paper for publication.

Acknowledgements

This article is dedicated to the memory of A. Hernández Carpio. We are grateful to N. Forns for a careful review and D. Menacho for helpful discussions. This work was supported by the project SESAAME-BAR (TIN2008-06582-C03-01) of the Spanish Ministry of Science and Innovation (JB and RFC).

References

Altmann, G., 1980. Prolegomena to Menzerath's law. Glottometrika 2, 1–10.

Becker, T.S., Lenhard, B., 2007. The random versus fragile breakage models of chromosome evolution: a matter of resolution. Molecular Genetics and Genomics 278, 487–491.

Boroda, M.G., Altmann, G., 1991. Menzerath's law in musical texts. Musikometrica 3, 1–13.

Cramer, I., 2005. The parameters of the Altmann–Menzerath law. Journal of Quantitative Linguistics 12 (1), 41–52.

De, A., Ferguson, M., Sindi, S., Durrett, R., 2001. The equilibrium distribution for a generalized Sankoff–Ferretti model accurately predicts chromosome size distribution in a wide variety of species. Journal of Applied Probability 38, 324–334.

Ferrer-i-Cancho, R., Forns, N., 2009. The self-organization of genomes. Complexity 15 (5), 34–36.

Fuquan, K., Kui, Z., Yong, Z., Tianguang, C., Meinan, N., Li, S., Minghui, C., Yizhong, Z., 2005. Analysis of length distribution of short DNA fragments induced by ^{7}Li ions using the random-breakage model. Chinese Science Bulletin 50 (9), 841–844.

Hernández-Fernández, A., Baixeries, J., Forns, N., Ferrer-i-Cancho, R., 2011. Size of the whole versus number of parts in genomes. Entropy 13 (8), 1465–1480, doi:10.3390/e13081465.

Hinsch, H., Hannenhalli, S., 2006. Recurring genomic breaks in independent lineages support genomic fragility. BMC Evolutionary Biology 6, 90.

Lewontin, R.C., Cohen, D., 1969. On population growth in a randomly varying environment. Proceedings of the National Academy of Sciences of the United States of America 62, 1056–1060.

Li, X., Zhu, C., Lin, Z., Wu, Y., Zhang, D., Bai, G., Song, W., Ma, J., Muehlbauer, G.J., Scalon, M.J., Zhang, M., Yu, J., 2011. Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. Molecular Biology and Evolution 28 (6), 1901–1911, doi:10.1093/molbev/msr011.

Luisi, P.L., 2006. Chapter 11: approaches to the minimal cell. In: The Emergence of Live. From Chemical Origins to Synthetic Biology. Cambridge University Press, Cambridge, pp. 242–267.

Oliver, M.J., Petrov, D., Ackerly, D., Falkowski, P., Schofield, O.M., 2007. The mode and tempo of genome size evolution in eukaryotes. Genome Research 17, 594–601.

Peng, Q., Pevzner, P.A., Tesler, G., 2006. The fragile breakage versus random breakage models of chromosome evolution. PLoS Computational Biology 2, e14.

Ruiz-Herrera, A., Castresana, J., Robinson, T.J., 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? Genome Biology 7, R115.

Sankoff, D., Ferretti, V., 1996. Karyotype distributions in a stochastic model of reciprocal translocation. Genome Research 6, 1–9.

Schubert, I., 2007. Chromosome evolution. Current Opinion in Plant Biology 10, 109–115.

Smart, J.S., 1976. Statistical tests of the broken-stick model of species-abundance relations. Journal of Theoretical Biology 59 (1), 127–139.

Solé, R.V., 2010. Genome size, self-organization and DNA's dark matter. Complexity 16 (1), 20–23.

Teupenhayn, R., Altmann, G., 1984. Clause length and Menzerath's law. Glottometrika 6, 127–138.

Trivers, R., Burt, A., Palestis, B.G., 2004. B chromosomes and genome size in flowering plants. Genome 47 (1), 1–8.

Vinogradov, A.E., 2001. Mirrored genome size distributions in monocot and dicot plants. Acta Biotheoretica 49, 43–51.

Wilde, J., Schwibbe, M.H., 1989. Organizationsformen von Erbinformation Im Hinblick auf die Menzerathsche Regel. In: Altmann, G., Schwibbe, M.H. (Eds.), Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Olms, Hildesheim, pp. 92–107.

BioSystems 111 (2013) 216–217

Contents lists available at SciVerse ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Erratum

Erratum to "Random models of Menzerath–Altmann law in genomes" (BioSystems 107(3) (2012) 167–173)

Ramon Ferrer-i-Cancho[a,*], Jaume Baixeries[a], Antoni Hernández-Fernández[a,b]

[a] Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center/LARCA, Universitat Politècnica de Catalunya, Barcelona (Catalonia), Spain
[b] Departament de Lingüística General, Universitat de Barcelona, Barcelona (Catalonia), Spain

ARTICLE INFO

Article history:
Received 14 December 2012
Accepted 7 January 2013

Keywords:
Menzerath–Altmann law
Scaling laws
Genomes
Independence
Mean independence

ABSTRACT

Here we improve the mathematical arguments of Baixeries et al (BioSystems 107(3) (2012) 167–173). The corrections do not alter the conclusion that the random breakage model yields an insufficient fit to the scaling of mean chromosome length as a function of chromosome number in real genomes.

1. Introduction

Consider that X is the number of chromosomes of a species, Y its genome size in bases, $Z = Y/X$ is the mean chromosome size and $E[Z|X = x]$ is the conditional expectation of Z given x (a concrete value of X). We have argued that

$$E[Z|X] = aX^{-1}, \tag{1}$$

is equivalent to independence between X and Y (Baixeries et al., 2012). The claim is wrong because Eq. (1) holds if and only if Y is mean independent of X. Here we explain why and correct related problems. All these corrections improve the accuracy of our mathematical arguments but do not alter the major conclusion of Baixeries et al. (2012), namely that Eq. (1) and the corresponding random breakage model yield an insufficient fit to real genomes.

2. Mathematical preliminaries

The next statement indicates Eq. (1) is equivalent to constant $E[Y|X = x]$.

Lemma 1. *Consider a constant a, two random natural variables, X and Y, and a third random number Z, such that $X > 0$ and $Z = Y/X$. Then, $E[Z|X = x] = a/x$ if and only if $E[Y|X = x] = a$ for any x.*

Proof. $E[Z|X = x] = E[(1/x)Y|X = x] = E[Y|X = x]/x$. □

So far, we have not cared about the value of a. The following statement indicates that a is not arbitrary, showing the equivalence between the constancy of the conditional expectation and mean independence (Poirier, 1995, p. 67).

Lemma 2. *Consider a constant a and two random natural variables, X and Y. $E[Y|X = x] = a$ for any x if and only if Y is mean independent of X, i.e. $E[Y|X = x] = E[Y]$ for any x.*

DOI of original article: http://dx.doi.org/10.1016/j.biosystems.2011.11.010.
* Corresponding author.
E-mail addresses: rferrericancho@lsi.upc.edu (R. Ferrer-i-Cancho), jbaixer@lsi.upc.edu (J. Baixeries), antonio.hernandez@upc.edu (A. Hernández-Fernández).

0303-2647/$ – see front matter
http://dx.doi.org/10.1016/j.biosystems.2013.01.004

Proof. Showing that $E[Y|X=x]=E[Y]$ for any x implies $E[Y|X=x]=a$ for any x is trivial because $E[Y]$ does not depend on x. Now we aim to show that $E[Y|X=x]=a$ for any x implies $E[Y|X=x]=E[Y]$ for any x. By the law of total probability for expectations (DeGroot and Schervish, 2012, p. 258), $E[Y]=E[E[Y|X=x]]$ and the substitution $E[Y|X=x]=a$, $E[Y]=E[a]=a$. □

Eq. (1) is equivalent to Y being mean independent of X, as shown next.

Theorem 1. *Consider a constant a and two random natural variables, X and Y, and a third random number Z, such that $X>0$ and $Z=Y/X$. Then, $E[Z|X=x]=a/x$ if and only if Y is mean independent of X, i.e. $E[Y|X=x]=E[Y]$ for any x.*

Proof. Chaining Lemma 1 and Lemma 2, we have $E[Z|X=x]=a/x$ if and only if $E[Z|X=x]=E[Y]$ and by Lemma 2 we have $a=E[Y]$. □

3. Corrections

We have claimed (Theorem 1 by Baixeries et al. (2012)) that
Two random natural numbers X and Y, such that $X>0$, are independent if and only if $Z=Y/X$ satisfies $E[Z|X]=E[Y]/X$.
In order to be correct, that theorem should state
Given *two random natural* **variables** *X and Y, such that $X>0$,* ***Y*** is mean independent of ***X*** *if and only if $Z=Y/X$ satisfies $E[Z|X]=E[Y]/X$.*
or
If *two random natural* **variables** *X and Y, such that $X>0$, are independent* **then** *$Z=Y/X$ satisfies $E[Z|X]=E[Y]/X$.*
The point is Theorem 1 and the fact that independence implies mean independence (Kolmogorov, 1956; Poirier, 1995) but the reverse does not hold. To see it, consider that

$$p(X=x,Y=y)=\begin{cases}1/2 & \text{if } x=0 \text{ and } y=0\\ 1/4 & \text{if } x=1 \text{ and } y=1\\ 1/4 & \text{if } x=1 \text{ and } y=-1.\end{cases} \quad (2)$$

Thus $E(Y|X=1)=(1/2)(-1)+(1/2)1=0$ and $E(Y|X=0)=0$. Therefore Y is mean independent of X but X and Y are not independent because $p(X=1,Y=0)=0 \neq p(X=1)p(Y=0)$ as $p(X=1)=1/2$ and $p(Y=0)=1/2$.

The proof that we provided for our incorrect theorem (Baixeries et al., 2012) indeed only demonstrates that independence implies $E[Z|X]=E[Y]/X$. Theorem 1 and the fact that independence is more restrictive than mean independence (as shown above) indicate that the converse is impossible to prove.

Baixeries et al. (2012) studied a general class of random breakage models where X and Y are independent (with $Z=Y/X$) and presented the following corollary (Corollary 1 by Baixeries et al. (2012)):

The general class of random models above is equivalent to the class of models yielding $E[L_c|L_g]=aL_g^{-\beta}$, with $a=E[G]$ and $\beta=1$.
(the meaning of the notation is the following: $X=L_g$, $Y=G$, $Z=L_c=G/L_g$ and $b=\beta$).

That corollary is incorrect as these random models are only a subclass of all the models yielding Eq. (1). In order to be correct, it should state

The general class of random models above **is a strict subset of** *the class of models yielding $E[L_c|L_g]=aL_g^{-\beta}$, with $a=E[G]$ and $\beta=1$.*

Again, the point is that independence implies mean independence but not the reverse. Therefore, the generalized class of random models where X and Y are independent covers only a fraction of the models that according to Theorem 1 lead to Eq. (1).

Besides the theorem and the corollary revised above, Baixeries et al. (2012) needs corrections in other places:

- p. 168 *Menzerath-Altmann law with $b=1$ and $c=0$ is equivalent to independence between G and L_g.*
  should read
  Menzerath-Altmann law with $b=1$ and $c=0$ is equivalent to G **being mean independent of** *L_g.*
- p. 168 *the need of independence between G and L_g to obtain Menzerath-Altmann law with $b=1$ and $c=0$.*
  should read
  the need **that *G* is mean independent of L_g** *to obtain Menzerath-Altmann law with $b=1$ and $c=0$.*
- p. 170 *In all the calculations that follow, true independence between G and L_g is assumed.*
  should read
  In all the calculations that follow, **mean independence between *G* and L_g (which includes independence between *G* and L_g as a particular case)** *is assumed.*
- p. 171 *Corollary 1 states that $E[L_c|L_g]\sim 1/L_g$ can only be achieved by independence between G and L_g*
  should read
  Corollary 1 states that $E[L_c|L_g]\sim 1/L_g$ can only be achieved **if *G* is mean independent of L_g**.

Acknowledgements

We are grateful to Ł. Debowski, P. Delicado, R. Gavaldà, J. Mačutek and E. Pons for their valuable mathematical insights. This work was supported by the grant *Iniciació i reincorporació a la recerca* from the Universitat Politècnica de Catalunya and the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation.

References

Baixeries, J., Hernández-Fernández, A., Ferrer-i-Cancho, R., 2012]. Random models of Menzerath-Altmann law in genomes. Biosystems 107, 167–173.
DeGroot, M.H., Schervish, M.J., 2012]. Probability and Statistics, 4th edition. Wiley, Boston.
Kolmogorov, A.N., 1956]. Foundations of the Theory of Probability, 2nd edition. Chelsea Publising Company, New York.
Poirier, D.J., 1995]. Intermediate Statistics and Econometrics: A Comparative Approach. MIT Press, Cambridge.

CAPÍTULO 5

Otros niveles de comunicación química

En los capítulos anteriores hemos visto que las leyes y principios de la lingüística cuantitativa están presentes desde la comunicación acústica hasta la química del ADN. Por otra parte podemos preguntarnos, ¿operan las mismas leyes en la comunicación química que en la comunicación acústica o visual? ¿Aparece por ejemplo el principio de compresión en la comunicación mediante feromonas? Presentamos aquí la última línea de investigación abierta en este trabajo.

La comunicación química ya existía mucho antes de que la evolución dotara a muchas especies de complejos sistemas de comunicación acústica o visual. Para entender la comunicación es imprescindible entender la comunicación química. Ajena en buena medida a nuestros limitados sentidos, con sólo observar la Naturaleza descubrimos sus maravillas gracias a la tecnología actual. Pero, ¿qué conexiones existen entre los sistemas de comunicación acústicos y la comunicación química? ¿Hay patrones universales en la comunicación independientemente de la modalidad?

La comunicación química no se limita a secuencias de ADN, como vimos en el capítulo anterior. En las células se dan múltiples mecanismos de transducción química (Alberts *et al.*, 2008). En general, la transducción de señales en la célula sucede cuando una molécula de señalización (intracelular o extracelular) activa un receptor (interno o en la superficie de la célula), creándose una respuesta que puede ser amplificada posteriormente y generar una cadena de respuestas celulares. Todas las células poseen además complejos mecanismos de comunicación química, tanto intracelulares como extracelulares e intercelulares (figura 5.1), que todavía se están descubriendo e investigando, también bajo la perspectiva de la teoría de la información (Kerszberg, 2003, para una revisión teórica sobre el asunto).

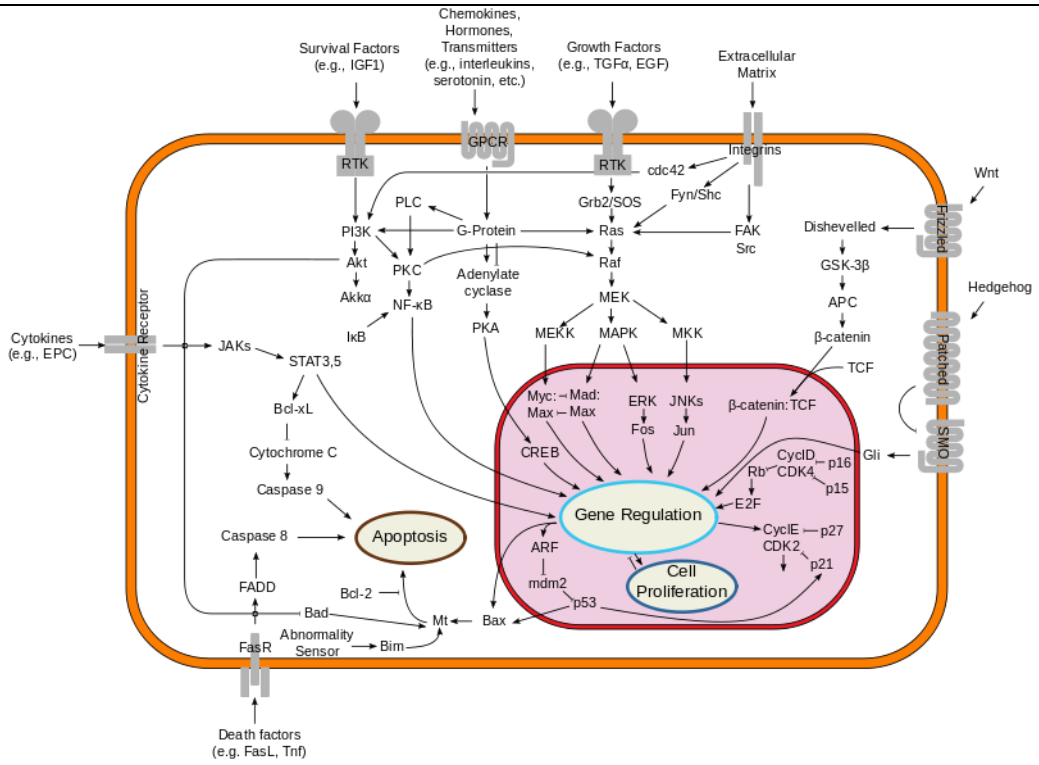


Figura 5.1: Esquema general de los diversos patrones de comunicación celular, entre los que destacan la apoptosis (o muerte celular programada), la regulación genética y la reproducción. Se indican asimismo las moléculas que intervienen en los diferentes procesos.
Fuente: http://commons.wikimedia.org/wiki/File:Signal_transduction_pathways.svg

La complejidad de la comunicación química es un problema que está siendo atacado recientemente desde perspectivas muy diversas como, por citar algunos enfoques, la mecánica estadística aplicada al problema de la senescencia y la apoptosis (muerte celular), relacionada con el acortamiento de los telómeros (Dao Duc y Holcman, 2013), la teoría de la información en la comunicación entre plantas e insectos (Doyle, 2009) o el análisis lingüístico de la química de la comunicación en bacterias (Schauder y Bassler, 2001).

La aproximación cuantitativa requiere de buenos datos. En nuestro caso los hallamos en la base de datos Pherobase (El Sayed, 2012), que contiene valiosa información recopilada sobre la comunicación mediante infoquímicos. Gracias a esta base de datos pudimos realizar la que, hasta donde sabemos, es la primera aproximación zipfiana y desde la lingüística cuantitativa al estudio general de la distribución de los infoquímicos en la Naturaleza (Hernández-Fernández y Ferrer-i-Cancho, 2014).

5.1. Infoquímicos y feromonas

Las feromonas son las sustancias químicas involucradas en la comunicación semioquímica (Law y Regnier, 1971) que se da entre muchos seres vivos, no únicamente insectos (Chapman, 1998). Casi todas las especies, de diversos géneros y filos, que van de los mamíferos a las plantas o las algas, poseen comunicación química, y muchas de ellas comparten incluso las mismas feromonas en sus sistemas de comunicación (Wyatt, 2003).

Para no empantanarnos en la terminología, aceptaremos seguir la tradición actual, en la que el étimo infoquímico, más general que el de feromona, se refiere a las sustancias involucradas en la comunicación química, capaces por definición de transmitir información (Shannon, 1948), aunque a menudo se emplee como sinónimo también semioquímico (Wyatt, 2010). La reciente división de los infoquímicos según su función (Wyatt, 2010), distingue varios tipos de infoquímicos, más o menos apropiados según el área de estudio (Dicke y Sabelis, 1988). Así, la clasificación de El-Sayed (2012) para Pherobase, diferencia entre cinco funciones (para otras clasificaciones véase Nordlund y Lewis, 1976; Dicke y Sabelis, 1988; Wyatt, 2010):

- Feromonas, involucradas en la comunicación intraespecífica.
- Atrayentes, o infoquímicos que causan la agregación de individuos, segregados por la propia especie o sintetizados en laboratorios.

- Alomonas, o aleloquímicos[24] que benefician al emisor, como sucede con las sustancias defensivas.
- Kairomonas, o aleloquímicos beneficiosos para el receptor, como pasa con algunas sustancias que emiten plantas tras ser atacadas por insectos y que para su desgracia atraen a más.
- Sinomonas, o aleloquímicos que benefician tanto al emisor como al receptor en interacciones.

La comunicación química ocurre en tres pasos fundamentales según Okubo y colaboradores (2001), siguiendo el modelo de Shannon (1948): primero un individuo libera una señal química (olfativa o gustativa), seguidamente se transmite dicha señal por el ambiente y finalmente es percibida por un segundo individuo. Esta es, de nuevo, una perspectiva externalista de la comunicación, pues cabe considerar, no obstante, que existen además mecanismos internos que conducen al emisor a liberar la sustancia (a iniciar la comunicación) y al receptor por su parte a procesar la información percibida.

Las feromonas se comportan como un medio de transmisión externo de señales, cuyas principales ventajas son el alcance a distancia y su capacidad de difusión y persistencia en el medio en el que se expanden (Wyatt, 2010). Su uso está condicionado al hábitat predominante de la especie y a su actividad, como ya concluyeron los estudios pioneros de Wilson (1958) y Bossert y Wilson (1963). La palabra "feromona" fue introducida por Karlson y Lüscher (1959). Etimológicamente, procede del griego "*pherein*" (transportar o llevar) y "*hormon*" (que excita o estimula) y es por tanto, por definición, una partícula que comunica individuos distintos, en contraposición a las hormonas que cada individuo utiliza en su comunicación interna (Wyatt, 2003).

La evolución de la comunicación tiene un claro componente social en todas las especies, pues existen siempre conexiones entre el lenguaje, o sistema de comunicación, la cognición social y, en el caso de algunas especies, la cultura (ver para una revisión Fitch *et al*., 2010). Tanto los sistemas de comunicación acústicos o visuales como los químicos basados en feromonas comparten esta conexión social, fundamental en la comunicación de las especies (Riba, 1990), siendo quizá los insectos un abundante

[24] En la alelopatía una especie produce uno o más compuestos (aleloquímicos) que influyen en el crecimiento, supervivencia o reproducción de otras especies. Estos compuestos son conocidos como aleloquímicos y regulan las relaciones mutualistas que se dan generalmente entre plantas, bacterias, hongos, corales, y otros organismos (Willis, 2007).

ejemplo en nuestro planeta (Chapman, 1998). Pese a las múltiples diferencias entre unos sistemas y otros según la modalidad, si existen universales en los sistemas de comunicación éstos deberían encontrarse en todos ellos (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013).

Diversas presiones evolutivas han influido en la comunicación química, empezando por la influencia del clima de los ecosistemas. Así por ejemplo, las altas temperaturas y la humedad elevada aumentan la velocidad de evaporación, reduciendo la persistencia de las feromonas, por lo que habitualmente las feromonas que utilizan las especies que viven en los trópicos poseen un mayor peso molecular medio (Alberts, 1992). Los mamíferos y los pájaros suelen asociar las feromonas a proteínas de tipo ureico, ya sean procedentes de la orina o la sudoración, para aumentar su persistencia (Wyatt, 2003).

Por tanto, las características químicas de las feromonas están directamente relacionadas con su función comunicativa y el ecosistema en el que son segregadas. Las feromonas se difunden en un medio en el que persisten durante un tiempo, de forma que no deben ser percibidas de forma inmediata. En el medio aéreo predominan las cadenas de carbono, y ya Wilson y Bossert (1963) acotaron que las feromonas empleadas en el aire debían tener cadenas de carbono de entre 5 y 20 carbonos, y un peso molecular (PM, o en inglés *molecular weight*, MW) de entre 80 y 300. Al predominar las moléculas lineales de carbono hay una trivial relación lineal entre PM y el número de carbonos de la cadena, tal y como se aprecia en la figura 5.2, en la que se representan las feromonas obtenidas por síntesis, todas ellas empleadas por al menos una especie en su sistema de comunicación, para 1686 compuestos de la base de datos Pherobase (El-Sayed, 2012), actualizada a fecha 22 de junio de 2012. Se confirma así la tendencia a la linealidad de las cadenas de carbono que suelen conformar las feromonas, con un límite superior marcado por la conexión de carbonos con hidrógenos y la posibilidad de dobles o triples enlaces del carbono (Hernández-Fernández y Ferrer-i-Cancho, 2014b).

Para feromonas con pocos carbonos la combinatoria reduce mucho las posibilidades de sustancias nuevas, mientras que al aumentar el número de carbonos se dispara la diversidad molecular (Wilson, 1970; Chapman, 1998; Wyatt, 2003). Como se aprecia en la figura 5.3, las feromonas que se utilizan como señales de alarma suelen tener pesos moleculares pequeños, para facilitar así su secreción y difusión rápida, mientras que las feromonas sexuales suelen aumentar el tamaño para mejorar su persistencia y detectabilidad (Chapman, 1998; Wyatt, 2003). Curiosamente los gritos o

las llamadas de atención o alarma en el lenguaje humano también suelen hacerse mediante palabras cortas, emitidas a intensidades acústicas más altas de lo normal, y que tendemos a alargar para asegurarnos de que son oídas, actuando por un lado la tendencia a la compresión y por otro lado la necesidad del éxito comunicativo (Ferrer-i-Cancho y Hernández-Fernández, 2013; Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013).

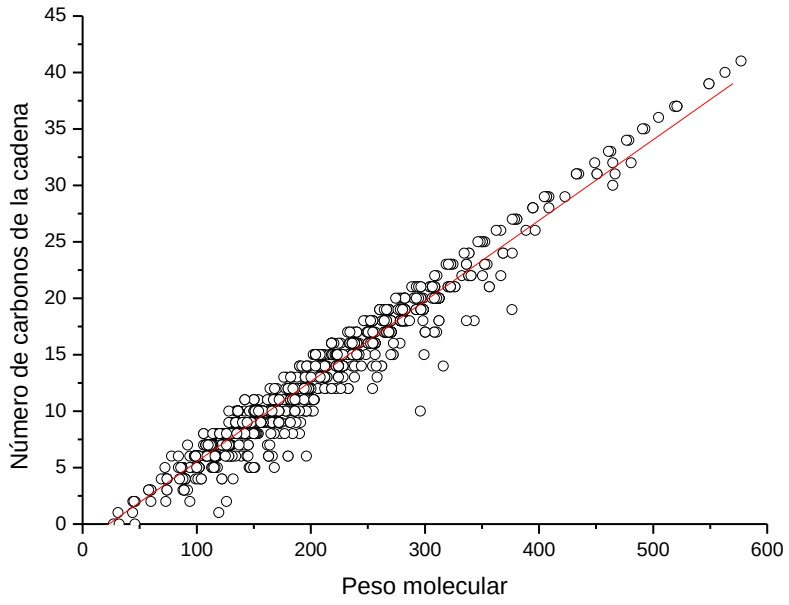


Figura 5.2.: Relación lineal entre el peso molecular de los infoquímicos de síntesis (N=1686) y el número de carbonos de cada una. Según un ajuste de regresión lineal simple se tiene que el número de carbonos C = 4.12·PM -1.29, con un coeficiente de correlación ρ = 0.97 y una desviación estándar para cada coeficiente de σ_a = ±0.02 y σ_b= ± 0.01, respectivamente. Datos obtenidos de www.pherobase.com, de Hernández-Fernández y Ferrer-i-Cancho (2014b).

La especificidad de una feromona, para por ejemplo atender a un estímulo específico como la presencia de un depredador concreto (en el caso de las feromonas de alarma), se puede dar a través de dos mecanismos químicos sencillos (Wyatt, 2003). Uno de ellos sería la evolución de una única molécula que logre una gran combinatoria, como sucede en las feromonas peptídicas, que a través de los veinte aminoácidos disponibles (en organismos eucariotas) permiten millones de posibilidades, simplemente escogiendo unos pocos aminoácidos (Browne *et al.*,1998), de forma que tendríamos 20^N posibilidades, con N el número de aminoácidos utilizados en la molécula. El otro

mecanismo es el uso de una única mezcla de compuestos simples, de bajo peso molecular, como se da en las feromonas de multicomponente (Wyatt, 2003).

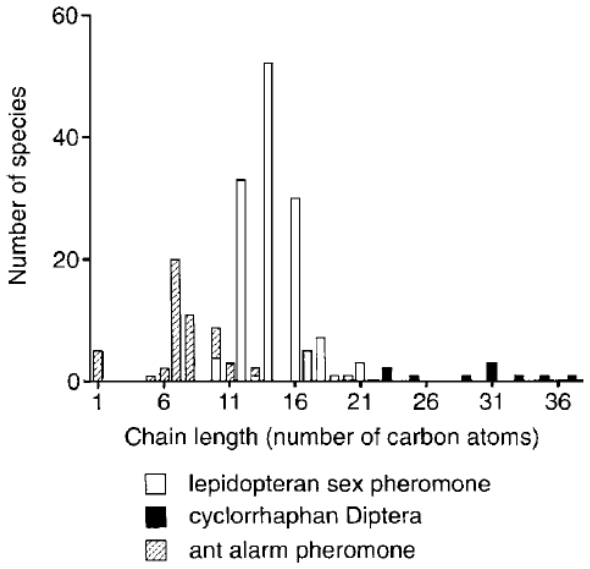


Figura 5.3.: Número de átomos de carbono de feromonas de alarma de las hormigas, de bajo peso molecular, comparadas con feromonas sexuales de lepidópteros y feromonas de contacto de dípteros. Figura extraída de Chapman (1998) y citada también en Wyatt (2003). Se aprecia como las feromonas que se utilizan como señales de alarma suelen tener pesos moleculares pequeños, para facilitar así su secreción y difusión rápida, mientras que las feromonas sexuales suelen aumentar el tamaño para mejorar su persistencia

En el medio acuático, no obstante, aunque el tamaño de las moléculas sigue siendo relevante pasa a ser primordial la solubilidad de las feromonas, dada la relevancia tanto de la distribución temporal como el espectro de señales para la comunicación efectiva, y la mayor facilidad que permite el agua para la difusión de moléculas de mayor tamaño (Atema, 1995). Así pues, en las especies acuáticas hay feromonas de tamaño similar a las del medio aéreo, como las feromonas sexuales de muchos peces, o moléculas mucho mayores que a pesar de su tamaño pueden ser altamente solubles (Wyatt, 2003). La comunicación química en el medio acuático afecta a la estructura de las poblaciones, su organización y función en el ecosistema (Hay, 2009).

5.2. El estudio cuantitativo de los infoquímicos

Los infoquímicos proporcionan información o actúan como precursores comunicativos (señales) que probablemente evolucionaron filogenéticamente a partir de

compuestos que originariamente no poseían una función comunicativa (Symonds y Elgar, 2004; Steiger *et al.*, 2011). Estudiar cuantitativamente las distribuciones de los infoquímicos en los ecosistemas ha sido terreno de la biogeografía y la ecología (Mateu, 1993) y es especialmente relevante en la dinámica de poblaciones (Turchin, 2003). El estudio de las poblaciones se centra en el análisis detallado de cada ecosistema, y las preguntas principales, tal y como Turchin (2003, p.8) plantea en la introducción de su obra, podrían ser:

> First, are dynamics of the studied population characterized by a stationary distribution of densities? (This is the issue of population regulation.) If yes, there is some characteristic mean level around which the population fluctuates, and fluctuations are characterized by a certain (finite) variance. What ecological mechanisms are responsible for setting this mean level? (This is the focus of population statics.) What mechanisms set the amplitude of fluctuations? Finally, are there detectable statistical periodicities, and what is the order and trajectory stability characterizing dynamics?

En nuestro caso, los datos disponibles en Pherobase (El-Sayed, 2012), nos permiten tener una visión global, más que centrada en cada ecosistema (Hernández-Fernández y Ferrer-i-Cancho, 2014). A las preguntas cruciales que plantea Turchin (2003) podemos aportar la hipótesis de que los canales de comunicación química influyan en la distribución y dinámica de poblaciones.

Si definimos el grado de un infoquímico como el número de especies asociadas al mismo, porque lo producen o son sensibles a dicha sustancia, y entonces ordenamos por rango los infoquímicos, de manera que el de mayor grado, utilizado por más especies, posee rango uno, el siguiente rango dos y así sucesivamente, independientemente de la función del infoquímico (El-Sayed, 2012), entonces podemos estudiar la distribución del uso de infoquímicos en la naturaleza (Hernández-Fernández y Ferrer-i-Cancho, 2014).

Tal y como se vio en el capítulo 2, podemos ajustar diversas funciones, otorgando pesos estadísticos o no a la distribución (Li *et al.*, 2010) para, siguiendo criterios como el de Akaike (1974), determinar qué función de las analizadas (véase tabla 2.1, página 34) proporciona el mejor ajuste respecto el número de parámetros independientes. En Hernández-Fernández y Ferrer-i-Cancho (2014) comprobamos que para Pherobase (El-Sayed, 2012) el ajuste de Zipf de dos regímenes era el mejor, tanto si asignábamos pesos estadísticos como si excluíamos los infoquímicos sintetizados en el laboratorio (Figura 5.4.).

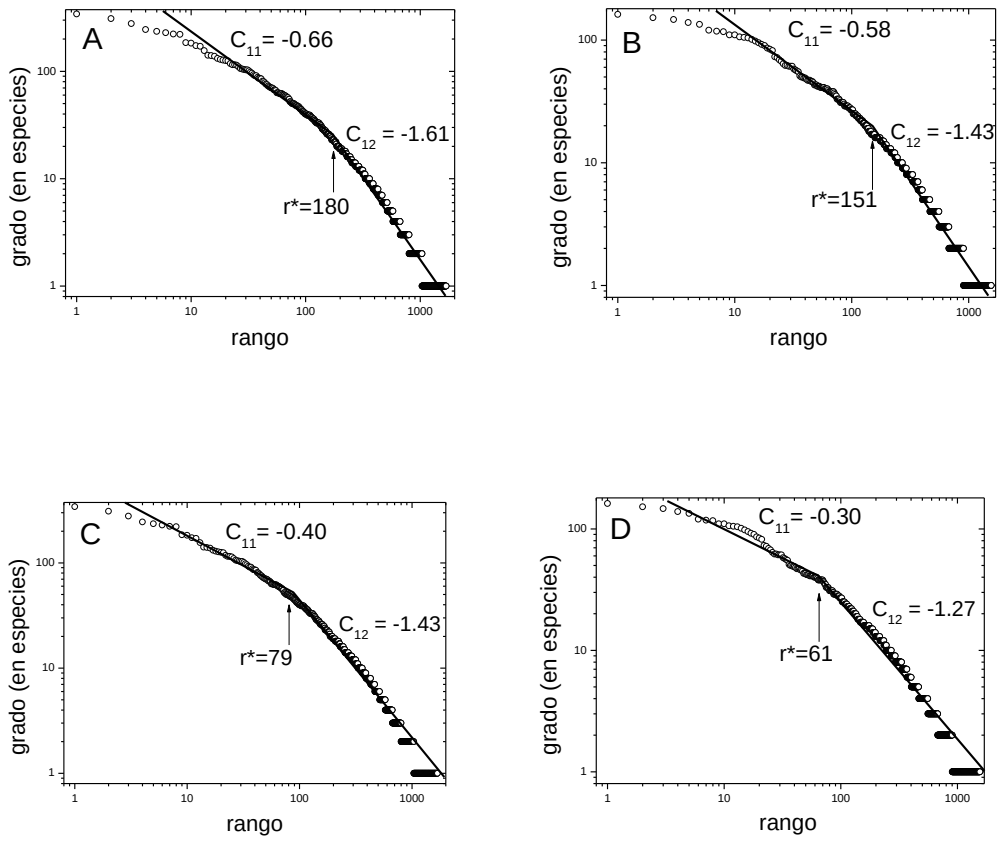


Figura 5.4.: Grado (número de especies asociadas a cada infoquímico) respecto su rango en doble escala logarítmica, con el mejor ajuste encontrado de dos regímenes de Zipf (línea), indicando en cada figura el punto de cambio de régimen (r*) y el exponente de Zipf de cada zona. A y B se corresponden a todos los datos estudiados mientras que en C y D se excluyeron los atrayentes. A y C incluyen los ajustes sin pesos estadísticos, mientras que B y D son el caso con pesos estadísticos (para más detalles véase Hernández-Fernández y Ferrer-i-Cancho, 2014).

En los datos analizados hay un total de 1686 compuestos, que según la clasificación de El-Sayed (2012) son utilizados como atrayentes por 6253 especies, alomonas por 2094 especies, kairomonas por 712 especies, feromonas por 8568 especies y, finalmente, como sinomonas solo por 6 especies de toda la base de datos. Por tanto, hay un claro predominio de los atrayentes y las feromonas comunicativas en la base de datos (tabla 5.1), así como del canal aéreo y de las poblaciones de insectos (El-Sayed, 2012), lo que produce sesgos en el análisis que se deben considerar (Hernández-Fernández y Ferrer-i-Cancho, 2014).

	Atrayentes	Alomonas	Kairomonas	Feromonas	Sinomonas	TOTAL
Repertorio nuclear	4833 (41.11%)	1506 (12.81%)	410 (3.49%)	5005 (42.57%)	1 (<0.01%)	11755 (100%)
Repertorio periférico	1420 (24.16%)	588 (10.00%)	302 (5.14%)	3563 (60.62%)	5 (0.09%)	5878 (100%)
TOTAL	6253 (35.46%)	2094 (11.88%)	712 (4.04%)	8568 (48.59%)	6 (0.03%)	17633 (100%)

Tabla 5.1.: Distribución de los infoquímicos de Pherobase (El-Sayed, 2012) analizados en Hernández-Fernández y Ferrer-i-Cancho (2014), por función, y según su distribución en el repertorio nuclear (primer régimen de Zipf) y el periférico (segundo régimen de Zipf). Se incluyen entre paréntesis los datos en porcentaje.

La distribución de dos regímenes de Zipf encontrada en la distribución de rangos del grado de cada infoquímicos, fue también hallada en la distribución de frecuencias de palabras (tokens) respecto al rango (Ferrer-i-Cancho y Solé, 2001) y ha sido confirmada recientemente (Petersen *et al*., 2012). En el caso de las palabras, el punto de transición entre zonas definía el límite entre el núcleo del vocabulario, en el que predominan las palabras con mayor versatilidad sintáctica, y el vocabulario periférico, potencialmente infinito, en el que la precisión semántica impera (Ferrer-i-Cancho y Solé, 2001).

La ubicuidad y variedad de las feromonas e infoquímicos se explica por la selección natural (Wyatt, 2009; Hauser, 1996). En el caso de los infoquímicos, el repertorio nuclear está formado por un conjunto de sustancias utilizadas por muchas especies y que cuenta, entre las sustancias de mayor grado, con varios isómeros del tetradecenil acetato, una conocida feromona sexual de los insectos (Shorey, 1976; Chapman, 1998). Manejamos la hipótesis de que este núcleo de compuestos responde tanto a la presión de la eficiencia comunicativa, en la que algunas sustancias son más efectivas en el medio aéreo, como a la de minimización del coste de producción, siendo las moléculas más cortas las que se tienden a producir con mayor facilidad (Hernández-Fernández y Ferrer-i-Cancho, 2014b), mientras que para evitar las interferencias en los canales químicos, que en algunos ecosistemas densamente poblados pueden estar muy solicitados (caso de los bosques tropicales, véase Basset *et al.*, 2012), hay un repertorio periférico potencialmente infinito, utilizado por menos especies, y que permite evitar la confusión (Hernández-Fernández y Ferrer-i-Cancho, 2014).

Bajo la perspectiva zipfiana la existencia de un repertorio periférico es consecuencia de la necesidad de diversificar las posibilidades comunicativas de las especies (Zipf, 1949), otorgando especificidad a la comunicación, frente a la tendencia a unificar o homogeneizar los repertorios al existir compuestos preferibles por su

disponibilidad, fácil difusión, producción y alta persistencia, según el canal y el ecosistema (Okubo *et al.*, 2001), que entrarían pues, estos últimos, dentro del repertorio nuclear encontrado (Hernández-Fernández y Ferrer-i-Cancho, 2014). En definitiva, la diversificación se opone a la unificación (Ferrer-i-Cancho y Solé, 2003; Zipf, 1949) también en la escala ecológica y en la comunicación química, y bajo un enfoque darwinista es de esperar que, con el tiempo, el canal químico se ocupe de forma óptima por las diversas especies (Hernández-Fernández y Ferrer-i-Cancho, 2014), afectando a la dinámica de las poblaciones (Turchin, 2003). Lógicamente, el problema está abierto, ya que nuestro estudio es general, y son necesarios más estudios particulares que se centren exclusivamente en datos de cada ecosistema concreto, para corroborar si la hipótesis, y la existencia de un repertorio químico nuclear, se recupera o no a menor escala, por ejemplo dentro de grupos taxonómicos específicos.

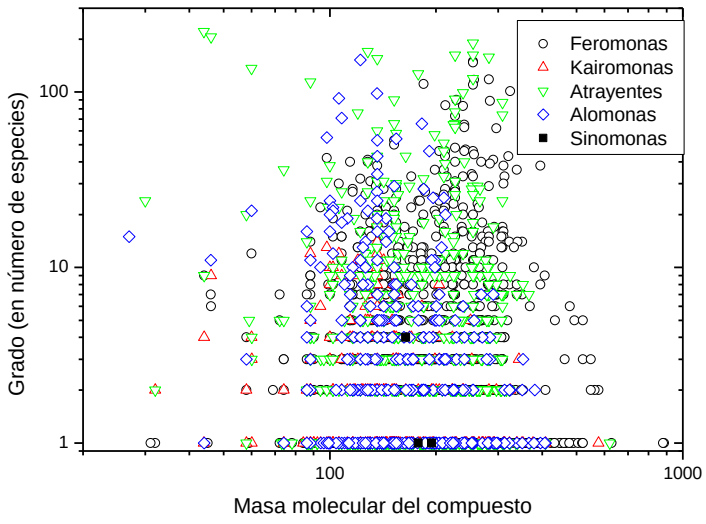


Figura 5.5.: Distribución del grado de utilización de infoquímicos según la masa molecular de los compuestos y su función (Hernández-Fernández y Ferrer-i-Cancho, 2014b, en preparación), en doble escala logarítmica.

Otro de los problemas abiertos en el que estamos trabajando (Hernández-Fernández y Ferrer-i-Cancho, 2014b) es el análisis del principio de compresión de la distribución de infoquímicos (figura 5.5.). Las feromonas son una solución comunicativa para los organismos más pequeños que no pueden efectuar emisiones acústicas de frecuencia óptima para la comunicación en el medio en el que viven (Dusenbery, 1992; Dusenbery y Snell, 1995). El coste metabólico de una feromona es

bajo comparado con otro tipo de señales (Thornhill y Alcock, 1983). Además, la comunicación mediante feromonas tiene costes energéticos y químicos mayores en los mamíferos que en los insectos, pues los primeros deben segregar proteínas para aumentar la persistencia y duración de sus señales, y repetir más a menudo las secreciones para mantener sus zonas de influencia (Wyatt, 2009).

Las feromonas suelen emitirse en paralelo a otros canales de comunicación (tabla 5.2), de manera que en general la comunicación suele tener componentes multisensoriales que deben considerarse. Según Wyatt (2003) este hecho puede suponer la redundancia de la información transmitida por la feromona (cuando se informa de lo mismo a través de modalidades diferentes), la modulación de la señal (cuando se añade a otros canales de comunicación), o que sea indispensable para la comunicación efectiva (puesto que las diversas modalidades aportan informaciones complementarias pero necesarias todas ellas). La problemática nos parece apasionante, plantea como se ve varios frentes abiertos, y es nuestro objeto de estudio en la actualidad (Hernández-Fernández y Ferrer-i-Cancho, 2014b) analizar la presencia o no de principios generales de la comunicación (Ferrer-i-Cancho, Hernández-Fernández *et al*., 2013) en el canal químico (tabla 5.2.). Queda, sin duda, mucho trabajo por delante.

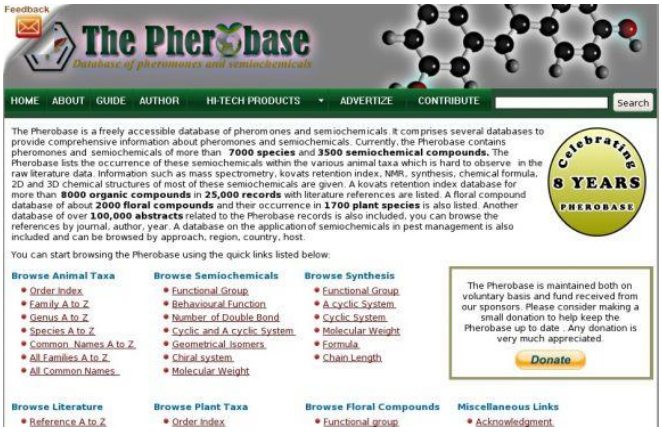
Característica del canal	**Tipo de señal**			
	Química	**Acústica**	**Visual**	**Táctil**
Alcance (distancia)	Largo	Largo	Medio	Muy corto
Velocidad de transmisión	Lenta	Rápida	Rápida	Rápida
Coste energético (emisor)	Bajo	Alto	Bajo a moderado	Bajo
Especificidad potencial (*)	Muy alta	Alta	Moderada	Limitada
Duración temporal	Breve a muy alta	Breve	Breve	Muy breve
Compresión (evidencias)	A investigar	Existen	Existen	A investigar

Tabla 5.2.: Características de los diversos canales comunicativos según el tipo de señal involucrada. Adaptada y modificada de Wyatt (2003) que a su vez se inspiró en Alcock (1989). (*) La especificad potencial se relaciona con la diversidad potencial del repertorio pero sin combinaciones de unidades discretas.

5.3. Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2014). *The infochemical core. (enviado)*

Se incluye aquí el artículo:

- Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2014). The infochemical core. *Journal of Quantitative Linguistics*, enviado el 30 de julio de 2013, pendiente de publicación.



The infochemical core

Short title: The infochemical core

Antoni Hernández-Fernández[1,2,*], Ramon Ferrer-i-Cancho[1]

(1) Complexity and Quantitative Linguistics Lab. Departament de Llenguatges i Sistemes Informàtics. TALP Research Center. Universitat Politècnica de Catalunya, Barcelona (Catalonia), Spain.

(2) Departament de Lingüística General. Universitat de Barcelona, Barcelona (Catalonia), Spain.

*Author for correspondence (antonio.hernandez@upc.edu).

Keywords: chemical communication, pheromones, infochemicals, semiochemicals, Zipf's law.

Total length of the manuscript: 7008 words.

*Address correspondence to: Antoni Hernández-Fernández, Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya, Campus Nord, Edifici Vèrtex, Plaça Eusebi Güell, 6, 08034 Barcelona (Catalonia), Spain. Tel: +34 649 00 51 53. Email: antonio.hernandez@upc.edu

ABSTRACT

Vocalizations and less often gestures have been the object of linguistics research over decades. However, the development of a general theory of communication with human language as a particular case requires a clear understanding the organization of communication through other means. Infochemicals are chemical compounds that carry information and are employed by small organisms that cannot emit acoustic signals of optimal frequency to achieve successful communication. Here the distribution of infochemicals across species is investigated when they are ranked by their degree or the number of species with which it is associated (because they produce or they are sensitive to it). The quality of the fit of different functions to the dependency between degree and rank is evaluated with a penalty for the number of parameters of the function. Surprisingly, a double Zipf (a Zipf distribution with two regimes

with a different exponent each) is the model yielding the best fit although it is the function with the largest number of parameters. This suggests that the world wide repertoire of infochemicals contains a chemical nucleus shared by many species and reminiscent of the core vocabularies found for human language in dictionaries or large corpora.

1. INTRODUCTION

Quantitative linguistics is a discipline with a tremendous capacity to explore connections between human language and other natural systems. The key is that a potential connection between language and a certain natural systems can be investigated simply by looking at certain statistical properties for which only minimal assumptions are required. For instance, the tendency of more frequent words to be shorter (Zipf, 1935, 1949), is also found in DNA sequences (Naranan & Balasubrahmanyan, 2000) and the behavior of non-human species (see Ferrer-i-Cancho et al., 2013b for a review). A non-trivial version of Menzerath's law (the tendency of the size of the parts of a linguistic construct to decrease as its number of parts increases (Menzerath, 1954) is found in genomes at different levels of analysis (Wilde & Schwibbe, 1989; Li, 2012; Ferrer-i-Cancho et al., 2013a). The depth of the connection depends on various factors. One factor is the existence of conceptual similarities between both at some level of abstraction (Ferrer-i-Cancho et al., 2013a). For instance, the striking conceptual similarities between human words and codons (Bel-Enguix & Giménez-López, 2011) are reinforced by other similarities arising from statistical analyses (Naranan & Balasubrahmanyan, 2000; Balasubrahmanyan & Naranan, 2000; Naranan, 2011).

Another factor is the number of statistical properties that coincide simultaneously: the likelihood of a non-trivial connection cannot decrease as the number of shared statistical regularities increases (Rao et al., 2012; Ferrer-i-Cancho & McCowan, 2012). For instance, the various statistical similarities between dolphin whistles and human words (McCowan et al., 1999; Ferrer-i-Cancho & McCowan, 2009; Ferrer-i-Cancho & McCowan, 2012) suggest that

dolphin whistles may have a communicative function resembling that of human words. However, this is not the only reason why quantitative linguistics has a tremendous capacity to bridge the gap between different fields. Due to its minimal assumptions, quantitative linguistics is at a privileged position for selecting candidates for extraterrestrial forms of intelligence (Doyle et al., 2011) or shedding light on the linguistic nature of undeciphered scripts (Rao, 2010).

Contrary to what many believe, quantitative linguistics is more than data analysis or the art of collecting statistical curiosities. A less known contribution of quantitative linguistics going back at least to the foundation of modern quantitative linguistics by G. K. Zipf during the 1st half of the XX century (Zipf, 1949), are abstract principles that can explain the recurrence of certain statistical patterns, not only in language but beyond, to finally establish laws of language or human behavior. G. K. Zipf's idea of a minimum equation defining the cost of a set of tools has recently been interpreted as a precursor of the notion of mean code length in information theory (Ferrer-i-Cancho et al., 2013b). Compression, i.e. the minimization of that mean code length has been used to shed light on the origins of the law of brevity not only in human language but in other species (Ferrer-i-Cancho et al., 2013b) and DNA sequences (Naranan & Balasubrahmanyan, 2000). A compromise between entropy minimization and mutual information minimization, inspired by Zipf's conflict between unification and diversification (Zipf, 1949) has been used to explain the recurrence of "Zipf's law" patterning in languages (Ferrer-i-Cancho, 2005; Prokopenko et al., 2010; Dickman et al., 2011). That explanation is abstract to enough to be valid for DNA (Ji, 1999) and other natural systems where Zipf's law-like patterning has been found (e.g., Searls, 2002; McCowan et al., 1999).

Human vocalizations, under the form of speech or written language (Akmajian et al., 1997), and less intensively gestures (Sandler & Lillo-Martin, 2006) have been objects of research from standard linguistics and neighboring disciplines over decades. However, the development of a

general theory of communication with human language as a particular case requires a clear understanding of the organization of information transfer through other means. While language is believed to be uniquely human (Hauser et al., 2002; Pinker, 2003) lacking a statistically rigorous testing of the statement and researchers struggle to date the origins of language in modern humans or ancestors, bacteria started linguistic communication millions and millions of years before a vocalizing multicellular organism appeared on Earth (Ben Jacob et al., 2004). Thus, chemical communication might be the oldest communication system produced in our planet. The same applies to writing, a milestone in the development of communicative scales for being a persistent way of communication allowing on one to detach from the here are and now. However, bacteria pioneered the exchange of documents, i.e. plasmids, in our planet (Head 1999/2000).

Language research is anthropocentric: it is based on the assumption that putting human language and human biology at the center provides a sufficient level of abstraction for solving the puzzle of the origins of language (Hurford, 2012; Fitch, 2010). However, this human centered vision has certain flexibility: it allows one to include existing or extinct species that are close in phylogeny or jumping further to species exhibiting complex vocal behavior like us, e.g., songs. This view is challenged again by non-vocal and non-gestural linguistic communication of brainless unicellular organisms through chemical compounds: bacteria (Ben Jacob et al., 2004). This adds further support for the proposal of a new paradigm for language research including biology and computer science (Bel-Enguix & Jiménez-López, 2012).

Here we consider the particular case of chemical communication in multicellular organisms, a domain that has received, to our knowledge, little attention by quantitative linguistics research, with some exceptions (Doyle, 2009). The goal of this article is applying concepts and tools from quantitative linguistics to shed light on the organization of infochemicals, chemical compounds that carry information (Wyatt, 2003).

2. INFOCHEMICALS

Although the complexity of chemical communication probably requires a zoosemiotic approach (Maran, Martinelli & Turovski, 2011; Riba, 1990), currently we must settle for analyzing signals that can be detected and describe the communicative contexts in which they are emitted. Chemicals provide information (cues) or act as precursors of communication (signals) or as central elements of communication systems that probably evolved from non-communicative compounds with a phylogenetic pattern (Symonds & Elgar, 2004; Steiger et al., 2011, for a review). Chemical signals constitute much of the language of life in the sea (Hay, 2009) and also in dry land (Wyatt, 2003). Humans have a poor capacity to understand chemical interactions and have a rather stunted sense of smell, but our current technology allows us to explore the fascinating world of infochemicals.

Infochemicals are usually divided in two groups: pheromones and allelochemicals (Wyatt, 2010; Dicke & Sabelis, 1988; Nordlund & Lewis, 1976). Pheromones are defined classically as semiochemicals involved in intraspecific communication (Law & Regnier, 1971; Regnier & Law, 1968), substances secreted to the outside by an organism (in contrast with hormones, secreted inside an organism) and perceived by a second individual of the same species in which they release a specific reaction (Karlson & Lüscher, 1959; Wyatt, 2003), in opposition to allelochemicals that mediate interspecies communications (Nordlund & Lewis, 1976) like happens between plants and insects.

The modern division of infochemicals by function (Wyatt, 2010, for a general review) distinguishes various classes of infochemicals, more or less appropriate according to the area of research concerned (Dicke & Sabelis, 1988). Thus, for example, the classification chosen by El-Sayed for the Pherobase, a free database of infochemicals (El-Sayed, 2012) distinguishes between five behavioral functions (for alternative classifications see Nordlund & Lewis, 1976;

Dicke & Sabelis, 1988; Wyatt, 2010): pheromones, involved in intraspecific communication; attractants, or infochemicals that cause aggregation of individuals, secreted by species or synthesized by humans; allomones, or allelochemicals that benefit the sender; kairomones or allelochemicals beneficial for the receiver; and finally synomones, that are allelochemicals benefiting both signaler and receiver in mutualistic interactions.

Pherobase (El-Sayed, 2012) is a wide coverage database that incorporates the list of species that produce, or are sensitive to, each infochemical, as well as other biochemical characteristics. Pheromones and allelochemicals are a way of transmitting information whose main advantages are their ability to spread and persistence in the environment in which they expand. Their diffusion is conditioned to the predominant species habitat and its activity (Okubo et al., 2001), as already concluded by pioneer studies of infochemicals (Wilson, 1958; Wilson & Bossert, 1963; Wilson, 1970).

TABLE 1 NEAR HERE

Here the degree of an infochemical is defined as the number of species that are associated to it, because they produce or are sensitive to it, according to Pherobase (El-Sayed, 2012). The infochemical with higher degree has rank 1, the 2^{nd} one has rank 2,… and so on (Table 1). The number of functions that an infochemical serves for a given species is irrelevant for degree. For instance, if an infochemical is associated to only one species, the infochemical will have degree one regardless of the number of functions served.

In this article, the fit of different functions to the rank distribution of infochemical associations in Nature is studied. The list of functions considered is summarized in Table 2. It is found that the function providing the best balance between the goodness of the fit and the number of parameters is a double Zipf (a power law with two different exponents) although it is the functions with the largest number of parameters. This suggests that infochemicals have a core

repertoire analogous to the core vocabulary found in human language (Ferrer-i-Cancho & Solé, 2001; Petersen et al., 2012).

TABLE 2 NEAR HERE

3. MATERIALS AND METHODS

f(r) is defined as the degree of rank *r*, *n* as the maximum rank (*r* = 1,2,…,*n*) and *T* as the sum of all the degrees, i.e.

$$T = \sum_{r=1}^{n} f(r) \tag{1}$$

with *n* the size of the repertoire.

Diverse two-parameter models such as Zipf's function, Beta function, Yule function or Menzerath-Altmann function (see Table 2) can be fitted to rank-frequency data (Li et al., 2010). Here Li et al's (2010) model selection methodology is adopted to study rank-degree data in infochemicals. A two regime distribution Zipf distribution (Ferrer i Cancho & Solé, 2001) is added to the list of functions explored by Li et al.(2010).

3.1. Materials

The degree and the classification of each infochemical comes from the Pherobase (El-Sayed 2012) which is freely available at http://www.pherobase.com/. A study primarily concerned about compounds that exists in nature and that regulate the communication without human intervention should discard attractants synthesized in human laboratories. Therefore, two kinds of analyses are considered: one concentrated on pheromones and allelochemicals of Pherobase (El-Sayed, 2012) and another of the whole database, including attractants synthesized by humans. The whole database comprises a repertoire of *n*=1686 chemical

compounds and T=17633 species-infochemical associations. A summary of the elementary features of the more frequent elements of the dataset is shown in Table 1.

3.2. Methods

The methodology for fitting functions to the dependency between rank and degree is borrowed from Li et al's for the dependency between the frequency of a word and its rank (Li et al., 2010). Baayen's (2008) methodology is used to compute the breakpoint parameter of the double Zipf (parameter r^* in Table 2). Li et al's (2010) methodology consists of a linear regression of the target function on a double logarithmic transformation, i.e. log(r) versus log (f_r), and then using Akaike's information criterion, a combination of likelihood to evaluate the quality of the fit and a penalty for the number of parameters used. Table 2 summarizes the functions that are fitted to the rank distribution of infochemicals and the corresponding double logarithmic transformation that is used for linear regression.

SEE is defined as the sum of the squared differences between the logarithm of $f_{r,}$ the observed degree of rank r, and the logarithm of F_r, expected degree of rank r, i.e.

$$SSE = \sum_{r=1}^{n} w_r (\log(f_r) - \log(F_r))^2 \tag{2}$$

being w_r a weight that is w_r = 1 in unweighted regression and w_r =1/r for weighted regression following Li et al's (2010) methodology. The goal of the weighted regression is giving more importance to low ranks.

The log-likelihood L is defined as (Venables & Ripley 1999; Li et al., 2010) as

$$L = C - \frac{n}{2}\log\left(\frac{SSE}{n}\right) - \frac{n}{2} \tag{3}$$

where *n* is the maximum rank and *C* is an additive constant that depends of the model (see Table 2). For model selection, Akaike Information Criterion (AIC) is used as in Li et al. (2010). As Li et al. (2010) point out, any constant term of Eq. 3 will be canceled when two models are compared and then the AIC defined by Akaike (1974) is:

$$AIC = -2L + 2K = n\log\left(\frac{SSE}{n}\right) + 2K\,, \tag{4}$$

where *K* is the number of free parameters (Table 2) in the model under consideration. So, the AIC difference of a model, Δ, is defined as the difference between the AIC of the model and that of the model with smallest AIC (Li et al., 2010). Then, trivially Δ=0 for the best model. The relationship between AIC and SSE defined in equation 3 is still true for weighted regression as Li et al. (2010) demonstrate, but the correlation-based of R^2 for weighted regression in logarithmic has a specific definition (see Li et al.(2010) for further details).

To fit the double Zipf, a function not considered by Li et al. 2010, Baayen's (2008, p.234-239) technique is used to compute the breakpoint automatically. That breakpoint is the rank *r** that defines the boundary between two consecutive simple Zipf distributions. The optimal breakpoint *r** is the rank that minimizes the *SSE*, which is obtained by exploring exhaustively all the possible values of *r**, that is fitting the double Zipf with a given *r** that is varied between 1 and *n* and keeping the *r** yielding the smallest *SSE*.

FIGURE 1 NEAR HERE

Fig. 1 indicates that the breakpoint obtained for the Pherobase is not a local optimum. R statistical package is used to analyze all the data (see R online manuals in http://www.r-project.org/ and Baayen (2008)). Table 2 summarizes the list of functions fitted to the rank distribution of degree.

4. RESULTS

Table 3 and Table 4 summarize the fitting of different functions to the empirical relationship between the degree of an infochemical and its rank according to Pherobase (El-Sayed 2012), including or excluding attractants, respectively. Both the correlation coefficient (ρ) and AIC differences (Δ) suggest that the double Zipf is the function providing the best fit, both in the weighted and unweighted regression, regardless of whether attractants are included or not.. All the other functions considered (simple Zipf, Beta, Yule and Menzerath-Altmann) are far from yielding the best fit: the 2nd best function, both in unweighted and weighted regression for the whole database, is Yule function with AIC differences Δ=967.5 and Δ=1450, respectively (Table 3) and similar results for the analysis excluding attractants (Table 4). AIC differences of this order of thousand are normally considered sufficient to discard the model under consideration (Burham & Anderson 2002).

TABLE 3 AND TABLE 4 NEAR HERE

Figure 2 shows the best fit of the double Zipf equation with and without weights, for the whole database (Fig 2 A and C) and without attractants (Fig 2 B and D), respectively. When all infochemicals are considered, the breakpoint is r^*= 180 for unweighted, and r^*=79 for weighted regression. When attractants are excluded, the breakpoint is r^*=151 for unweighted, and r^*=61 for weighted regression (see Table 5). The value of the breakpoint parameter of the double Zipf suggests that infochemicals are divided into two groups: a core of the order of one hundred infochemicals and the remainder.

TABLE 5 AND FIGURE 2 NEAR HERE

5. DISCUSSION

This double Zipf distribution of ranks (in the degree of an infochemical as a function of its rank rank) is also found in the rank distribution of words, i.e. the frequency of occurrence of a word

(in tokens) as a function of its rank (Ferrer-i-Cancho & Solé, 2001; Petersen et al., 2012), where the breakpoint defines the boundary between a vocabulary core, i.e. a finite vocabulary of semantically versatile word types, and a potentially infinite peripheral vocabulary.

The connection may not be obvious at first glance: our target has not been the frequency of occurrence of an infochemical in nature but its degree, namely, the number of species that produce it or are sensitive to each infochemical. However, word frequency is connected lawfully with another linguistic variable, namely word polytextuality, which can be defined as the number of texts of a corpus where the target word appears at least once (Köhler, 1986). The metaphor that words types are infochemicals, texts are the infochemicals from the environment that the members of a species has produced or been sensitive to and the positive correlation between frequency and polytextuality (Köhler, 1986), suggests that the infochemicals in the high degree regime (the low ranks) form a core repertoire analogous to the core lexicon found in the high frequency regime of human language (Ferrer-i-Cancho & Solé, 2001), while the infochemicals in the low degree regime would form a peripheral repertoire analogous to the peripheral lexicon found in the lower frequency domain of words (Ferrer-i-Cancho & Solé, 2001).

Cores have also been investigated in cognitive networks (Baronchelli et al., 2013). A network core is defined by Baronchelli et al. (2013) as "*a powerful subset of the network because of the high frequency of occurrence of its nodes, their importance for the existence of remainder of nodes, or the fact that is both densely connected and central (in terms of graph distance)"*. A network analysis of cross-referencing between dictionary entries has shown that dictionaries have a core consisting of about 10% of words (typically with a concrete meaning and acquired early), from which other words can be defined (Picard et al., 2009). In our case, we are analyzing a bipartite graph of infochemical- species associations (one partition for infochemical and another partition for species). Table 5 summarizes the proportion of infochemicals types

within the chemical core taking the breakpoint of the double Zipf as the boundary. The percentage of infochemicals that belong to the core repertoire is about 10% according to unweighted regression, about the same percentage of word types in the grounding kernel of a dictionary (Picard et al., 2009). The 10% of the core both in dictionaries and infochemicals might be simply anecdotal and should be explored further.The percentage of infochemical-species associations within the chemical kernel is two-thirds (11755 infochemicals in the core repertoire over 17633 in total, just 66.66%) of the total but this percentage varies according to the kind of function, and increases to more than seventy percent for attractants and allomones (Table 6). Similarly, the core vocabulary of words identified by means of a double Zipf by Ferrer-i-Cancho & Solé (2001) is responsible for 85% of word tokens in a large English multiauthor corpus. Although infochemical degrees and word frequencies are not fully comparable, the high percentages of associations/tokens in the core highlights the importance of cores.

TABLE 6 NEAR HERE

Pheromones and kairomones are just a little below sixty percent over infochemicals associations in the chemical kernel (Figure 3) and both have a greater presence than attractants and allomones (in percentage) in the peripheral chemical repertoire, with synomones not appearing because their number is very low (Figure 3). Pheromones increases their presence in the peripheral chemical repertoire probably as a response of a major necessity of communication specificity and to reducing communication interferences with other species. The information carried by general attractants of the infochemical core may need to be completed and detailed by peripheral, perhaps species-specific pheromones to avoid confusion.

FIGURE 3 NEAR HERE

The chemical kernel constitutes a core of infochemicals shared by many species (Table 5). We hypothesize that the design of such core could be driven by the economy of its compounds from the perspective of the sender (e.g., production ease of the compound) and communicative efficiency in a given environment from the perspective of the receiver. Communicative efficiency is determined by various factors such as detection ease or persistence of the chemical compound. The heterogeneity of the ecosystems of the species included in the database probably limits the cases of communicative interference between species. Furthermore, to avoid the confusion that might arise in a ecosystem with different species using the same infochemicals, species may adopt different diffusion strategies and exploit the zoosemiotic communication context to reduce confusion (Maran, Martinelli & Turovsky, 2011; Riba, 1990).

The most frequently analyzed infochemicals (Table 2), by definition on the core chemical repertoire, includes for example some isomers of Tetradecenyl Acetate that are well-known sex pheromones (Shorey, 1976; Chapman, 1998) and are attractive for all of the males in a large group of different near species. As already pointed out by Shorey (1976) in his classical review, other chemicals present in relatively small quantities enhance the attractiveness of the pheromone for males of the correct species and at the same time inhibit attraction of males of wrong species.

Each species has its own chemical communication system, with a finite set of associated infochemicals that can emit, V_e, and another set to which they respond, V_r. Actually, $V_e \neq V_r$ since there are semiochemicals (e.g. synthesized) that a species will not issue and those who feel attracted. Using set theory, the chemicals that constitute a chemical communication system is $V_e \cup V_r$, while the elements emitted and also detectable are $V_e \cap V_r$. In this Universe of chemical communication there are species that are sensitive to a shared chemical repertoire,

emitted or not by themselves. It is a complex Universe, because there are substances that, within the same species, can attract females and not males or vice versa.

The analysis of non-human communication systems may be able to provide insights into the efficiency of signaling systems that might otherwise be inaccessible (Doyle, 2009). Keeping a certain distance, parallels between our study and other linguistic phenomena, e.g., shared elements between different languages can be established. Above we have considered the metaphor of a species as a text of related infochemicals. Here we look at the infochemicals to which a species is related as a language. There are infochemicals shared in the communication systems of some species so, arguably, species may speak "different chemical languages" but do have some communicative elements in common (not necessarily with the same meaning). Similarly, different human languages can share word elements such as phonemes or syllables in their linguistic systems. Some languages not only share those building blocks of words but also word themselves. In a more complex way, languages share sometimes elements like words of basic vocabulary, e.g. Swadesh list, especially if they are related languages. From an evolutionary perspective, related languages tend to maintain diachronically basic words in the same or practically identical word form (Wang & Wang, 2004) just like different species can be sensitive to very similar -or the same - semiochemicals when they come from the same phylum (Symonds & Elgar, 2004).

The ubiquity and variety of pheromones can be explained by natural selection (Wyatt, 2009; Hauser, 1996). Under a Zipfian perspective (Zipf, 1949), the existence of a peripheral chemical repertoire (the infochemicals out of the core) could be a consequence of the need of diversifying the communication possibilities of species, which would be particularly useful in a noisy channel. If the principle of least effort leads all species to emit similar chemicals compounds from the core because of their high availability, ease of production or utility in different environments (Okubo et al., 2001), then interference and confusion emerges as

inevitable, especially in very rich terrestrial ecosystems like tropical forests (Basset et al., 2012). Thus, diversification opposes to unification (Ferrer-i-Cancho & Solé, 2003; Zipf, 1949) also in chemical communication systems. From a Darwinian evolution perspective, it is expected that an optimal occupation of the chemical channel in ecosystems arises over time. In sum, the existence of both a chemical core and a periphery could be a natural consequence that communication has to solve two different problems: efficient coding of information and transmission (Ferrer-i-Cancho et al., 2013).

Although linguistics is mainly concerned about the vocal modality, interest in gestural language through conventional sign language or spontaneous gestures has been increasing over time (Goldin-Meadow et al., 2008; Sandler & Lillo-Martin, 2006). The chemical modality might be the next frontier to be explored intensively. Quantitative linguistics now has the challenge of understanding the mechanisms underlying the emergence of two regimes in words and infochemicals and identifying the distinctive features of those cores. Throughout this long research track we hope to answer a very important question: what are the principles of organization shared between human language and chemical communication? Understanding both as evolutionary and complex adaptative systems might be crucial (Beckner et al., 2009; Bel-Enguix & Jiménez-López, 2012).

The quantitative exploration of the Pherobase is just at the beginning and our analysis has focused on the large scale. Future research should pay attention to specific ecological niches or concrete phylum. The generosity of those who share data on infochemicals is helping to explore connections between apparently distant domains, what will become increasingly more common in 21st century science.

ACKNOWLEDGEMENTS

We thank G. Bel-Enguix for helpful comments. We are grateful to A.M.El-Sayed and all the people who have made the Pherobase® possible and to Jordi Las for his help with the programming of the data acquisition. RFC thanks the support of the grant "*Iniciació i reincorporació a la recerca*" from the Universitat Politècnica de Catalunya and the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation.

REFERENCES

Akmajian, A., Demers, R, A,, Farmer, A. K. & Harnish, R. (1997). *Linguistics. An introduction to language and communication.* Cambridge, MA: MIT Press, 2nd edition.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control,* 19, 716–722.

Baayen, R.H.(2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press.*

Balasubrahmanyan, V.K. & Naranan, S. (2000). Information Theory and Algorithmic Complexity: Applications to Linguistic Discourses and DNA Sequences as Complex Systems Part II: Complexity of DNA sequences, analogy with linguistic discourses. *Journal of Quantitative Linguistics,* 7(2), 153-183.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N. & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, 17 (7) 348-360.

Basset, Y., et al. (2012). Arthropod diversity in a tropical forest. *Science*, 338, pp. 1481-1484.

Beckner, C., et al. (2009). Language is a complex adaptive system. *Language Learning*, 59 (s1), 1-26.Bel-Enguix, G. & Jiménez-López, M. D. (2011). Genetic code and verbal language: syntactic and semantic analogies. In: *Biology, computation and linguistics. New interdisciplinary paradigms.* Bel-Enguix, G.; Dahl, V.; Jiménez-López, M. D., Eds.; IOS Press: Amsterdam. pp. 85-103.

Bel-Enguix, G. & Jiménez-López, M. D. (2012). Biocomputing: an insight from linguistics. *Natural Computing,* 11, 131-139.

Ben Jacob, E., Becker, I., Shapira, Y. & Levine, H. (2004). Bacterial linguistic communication and social intelligence, Trends in Microbiology, 12 (8), 366–372.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference. A practical information-theoretic approach*. 2nd edition. New York: Springer.

Byers, J.A. (2006). Pheromone component patterns of moth evolution revealed by computer analysis of the Pherolist. *Journal of Animal Ecology* **75**, 399–407

Chapman, R. F. (1998). *The insects. Structure and function*, 4th edition. Cambridge: Cambridge University Press.

Dicke, M. & Sabelis, M.W. (1988). Infochemicals terminology: Based on cost-benefit analysis rather than origin of compounds? *Functional Ecology,* 2, 131-139.

Dickman, R., Moloney, N. R. & Altmann, E. G. (2012). Analysis of an information-theoretic model for communication. *Journal of Statistical Mechanics*: *Theory and Experiment*, P12022.

Doyle, L. R., McCowan, B., Johnston, S., and Hanser, S. F. (2011). Information theory, animal communications, and the search for extraterrestrial intelligence. *Acta Astronomica,* 68, 406–417.

El-Sayed, A.M. (2012). *The Pherobase: Database of Pheromones and Semiochemicals.* http://www.pherobase.com.

Ferrer i Cancho, R. (2005). Zipf's law from a communicative phase transition. *European Physical Journal B,* 47, 449-457.

Ferrer-i-Cancho, R. & 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-i-Cancho, R. & Solé R.V. (2001). Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8, 165-173.

Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G. & Baixeries, J. (2013a). The challenges of statistical patterns of language: the case of Menzerath's law in genomes. Complexity, 18 (3), 11–17.

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J. & Semple, S. (2013b). Compression as a universal principle of animal behavior. Cognitive Science, in press.

Ferrer-i-Cancho, R. & McCowan, B. (2009). A law of word meaning in dolphin whistle types. *Entropy,* 11 (4), 688-701.

Ferrer-i-Cancho, R. & McCowan, B. (2012). The span of correlations in dolphin whistle sequences. Journal of Statistical Mechanics, P06002.

Fitch, W. T. (2010). *The evolution of language*. Cambridge: Cambridge University Press.

Greenhill, S.J., Blust, R. & Gray, R.D. (2008). The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics Online*. 4**,** 271–283.

Goldin-Meadow, S., So, W.-C., Ozyurek, A., & Mylander, C. (2008). The natural order of events: how speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences,* 105 (27), 9163-9168.

Hauser, M. D. (1996). *The Evolution of Communication*. Cambridge, Massachusetts: MIT Press.

Hauser, M. D., Chomsky, N. Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how dit it evolve? *Science,* 298, 1569–1579.

Hay, M. (2009). Marine chemical ecology: chemical signals and cues, structure marine populations, communities, and ecosystems. *Annual Review of Marine Sciences*, 1, 193–212.

Head, T. (1999/2000). Communication by documents in communities of organisms. *Millenium III (4)*, 33-42

Hurford, J. (2011). *The origins of grammar. Language in the light of evolution.* Oxford: Oxford University Press.

Ji, H. (1999). The Linguistics of DNA: Word, Sentences, Grammar, Phonetics and Semantics. *Annals of the New York Academy of Scie*nces, 870, 411-417.

Karlson, P. & Lüscher, M.(1959). 'Pheromones': a new term for a class of biologically active substances. *Nature*, 183, 55–56.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Law, R. H. & Regnier, F. E. (1971). Pheromones. *Annual Review of Biochemistry*, 40, 533–548.

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

Li, W. (2012). Menzerath's law at the gene-exon level in the human genome. *Complexity,* 17, 49-53.

Pinker, S. (2003). *Language as an adaptation to the cognitive niche*. In: *Language evolution: states of the art*, M. H. Christiansen and S. Kirby (eds), pp. 16-37. New York: Oxford University Press.

Maran,T., Martinelli, D., Turovsky, A.(Eds.) (2011). *Readings in zoosemiotics*. Göttingen: De Gruyter-Mouton.

McCowan, B., Hanser, S. F. and Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour,* 57, 409-419.

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes*. Dümmler: Bonn.

Naranan, S. & Balasubrahmanyan, V.K. (2000). Information Theory and Algorithmic Complexity: Applications to Linguistic Discourses and DNA Sequences as Complex Systems Part I: Efficiency of the Genetic Code of DNA. *Journal of Quantitative Linguistics,* 7 (2), 129-151.

Naranan, S. (2011). Historical Linguistics and Evolutionary Genetics. Based on Symbol Frequencies in Tamil Texts and DNA Sequences. *Journal of Quantitative Linguistics,* 18(4), 359-380.

Nordlund, D.A. & Lewis, W.J. (1976). Terminology of chemical releasing stimuli in intraspecific and interspecific interactions. *Journal of Chemical Ecology,* 2, 211–220.

Okubo, A., Armstrong, R.A., & Yen,J. (2001). *Diffusion of "Smell" and "Taste": Chemical Communication*. In *Diffusion and Ecological Problems*, 2nd edition, ed. A.Okubo & S.A. Levin, pp. 107–126. New-York: Springer-Verlag.

Petersen, A.M., Tenenbaum, J.N., Havlin, S., Stanley, H.E. & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports,* 2, 943.

Picard, O., Blondin-Masse, A., Harnad, S., Marcotte, O., Chicoisne, G. & Gargouri, Y. (2009). Hierarchies in Dictionary Definition Space. *23rd Annual Conference on Neural Information Processing Systems (NIPS): Workshop on Analyzing Networks and Learning*.

Prokopenko, M., Ay, N., Obst, O., and Polani, D., (2010). Phase transitions in least-effort communications. *Journal of Statistical Mechanics: Theory and Experiment*, 11025.

Rao, R. (2010). Probabilistic analysis of an ancient undeciphered script. *IEEE Computer,* 43 (4), 76-80.

Rao, R. P .N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R. & Mahadevan, I. (2012). Entropy, the Indus script, and language: a reply to R. Sproat. *Computational Linguistics*, 36 (4), 795-805.

Regnier, F. E. & Law, R. H. (1968). Insect Pheromones. *Journal of Lipid Research*, 9, 541–551.

Riba, C. (1990). *La comunicación animal. Un enfoque zoosemiótico*. Barcelona: Anthropos.

Sandler, W. & Lillo-Martin, D. (2006). Sign language and linguistic universals. Cambridge: Cambridge Universty Press.

Searls, D. B. (2002). The language of genes. *Nature,* 420, 211-217.

Semple, S., Hsu, M. J. & Agoramoorthy, G. (2010). Efficiency of coding in macaque vocal communication. *Biology Letters, 6*, 469-471.

Shorey, H.H.(1976). *Animal Communication by Pheromones*. New York: Academic Press.

Steiger, S., Schmitt, T. & Schaefer, H.M. (2011). The origin and dynamic evolution of chemical information transfer. *Proceedings of the Royal Society of London B,* 278, 970–979. Symonds M.R.E. & Elgar M.A. (2004). The mode of pheromone evolution: evidence from bark beteles. *Proceedings of the Royal Society of London* B, 271, 839–846.

Wang, F. & Wang, S.Y. (2004). Basic Words and Language Evolution. *Language and Linguistics,* 5(3),643-662.

Wilde, J. & Schwibbe, M. H. (1989). Organizationsformen von Erbinformation Im Hinblick auf die Menzerathsche Regel. In *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Altmann, G.; Schwibbe, M. H., Eds.; Olms: Hildesheim, 1989. pp. 92-107.

Wilson, E. O. & Bossert, W. H. (1963). Chemical communication among animals. *Recent Progress in Hormone Research*, 19, 673–716.

Wilson, E. O. (1970). Chemical communication within animal species. In *Chemical ecology,* 9, ed. E. Sondheimer & J. B. Simeone, pp. 133–155. New York: Academic Press.

Wilson, E.O. (1958). A chemical release of alarm and digging behavior in the ant Pogonomyrmex badius (Latreille). *Psyche* 65, 41-51.

Wyatt, T.D. (2009). Fifty years of pheromones. *Nature*, 457, 262–263.

Wyatt, T.D. (2010). Pheromones and signature mixtures: defining species-wide signals and variable cues for individuality in both invertebrates and vertebrates. *Journal of Comparative Physiology A*, 196,685–700.

Wyatt, T.D.(2003). *Pheromones and Animal Behaviour*. Cambridge: Cambridge University Press.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology.* Cambridge, MA: MIT Press.

Zipf, G.K. (1949). *Human behaviour and the principle of least effort.* Cambridge, MA: Addison-Wesley.

Table 1: List of the ten infochemicals with the highest degree (in number of species) according to Pherobase (El-Sayed 2012). f_r is the degree of the infochemical with the the r-th largest degree. Each chemical compound can serve different behavioral functions.

Infochemicals	r	f_r
(Z)-9-Tetradecenyl acetate	1	342
(Z)-11-Tetradecenyl acetate	2	309
(Z)-11-Hexadecenyl acetate	3	278
2,6,6-Trimethylbicyclo[3,1,1]hept-2-ene	4	245
(Z)-7-Dodecenyl acetate	5	235
(E)-11-Tetradecenyl acetate	6	229
Carbon dioxide	7	222
Ethanol	8	221
(Z)-11-Hexadecenal	9	185
1-Octen-3-ol	10	183

Table 2: List of functions used in this article (simplified and adapted from Li et al. 2010 and enriched with the double Zipf) and the parameters fitted, i.e. a , b and normalization constants (C,C'). In the double logarithmic transformation for linear regression, C_0 is the independent term, C_1 the coefficient that multiplies log r. r^* is the rank of the breakpoint in the double Zipf function and defines two regimes, a 1st regime from r=1 to r=r^* and a 2nd regime for r>r^*.

MODEL	**FREE PARAMETERS** (***K***)	**ORIGINAL FUNCTION** (***r***=rank, F_r=degree, ***n***=repertoire size)	**DOUBLE LOGARITHMIC TRANSFORMATION**
Zipf	1	$F_r = \frac{C}{r^a}$	**log F_r = C_0 + C_1 log r**
Beta	2	$F_r = \frac{C(n+1-r)^b}{r^a}$	**log F_r = C_0 + C_1 log r +C_2log(n+1-r)**
Yule	2	$F_r = \frac{C \cdot b^r}{r^a}$	**log F_r = C_0 + C_1 log r +$C_4 r$**
Menzerath-Altmann	2	$F_r = C \cdot r^b \cdot e^{-a/r}$	**log F_r = C_0 + C_1 log r + C_3/r**
Double Zipf	3	1st regime: $F_r = \frac{C}{r^a}$; 2nd regime: $F_r = \frac{C'}{r^{a'}}$	**1st regime: log F_r = C_{01} + C_{11} log r** **2nd regime: log F_r = C_{02} + C_{12} log r**

Figure 1: The sum of squared errors of the double Zipf, *SSE*, versus the breakpoint rank, *r**. A and B are the plots for the whole Pherobase, unweighted and weighted respectively. C and D are analogous to A and B excluding attractants synthesized by humans, unweighted and weighted respectively. It can be seen that the breakpoint *r** obtained corresponds to a global minimum of deviance in all cases. Only in C we find another local minimum for very large ranks that is not relevant due to its comparatively large value of *SSE* with regard to the global minimum.

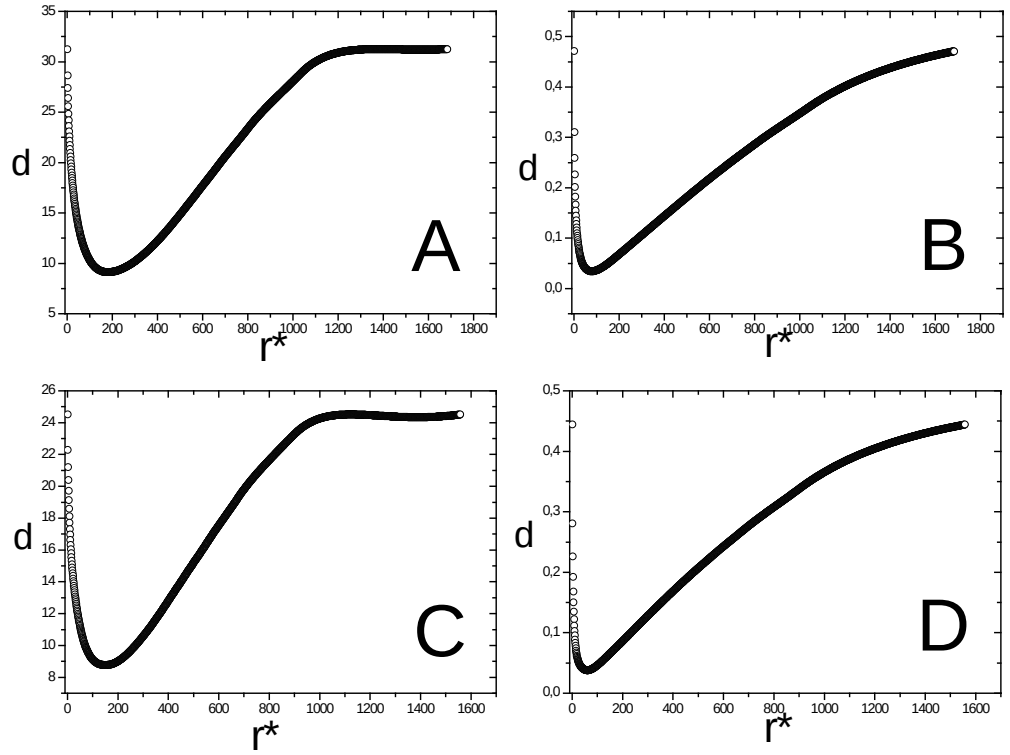


Figure 2: Degree (number of species that that are associated to each infochemical) versus the infochemical rank in double logarithmic scale (white circles) versus the best fit of the double Zipf model (solid line). A and B correspond to the whole database while attractants are excluded for C and D. A and C are unweighted while B and D are weighted. In each subplot, the rank breakpoint (r^*), the Zipf's exponent for the first and second regime (C_{11} and C_{12}, respectively) are shown.

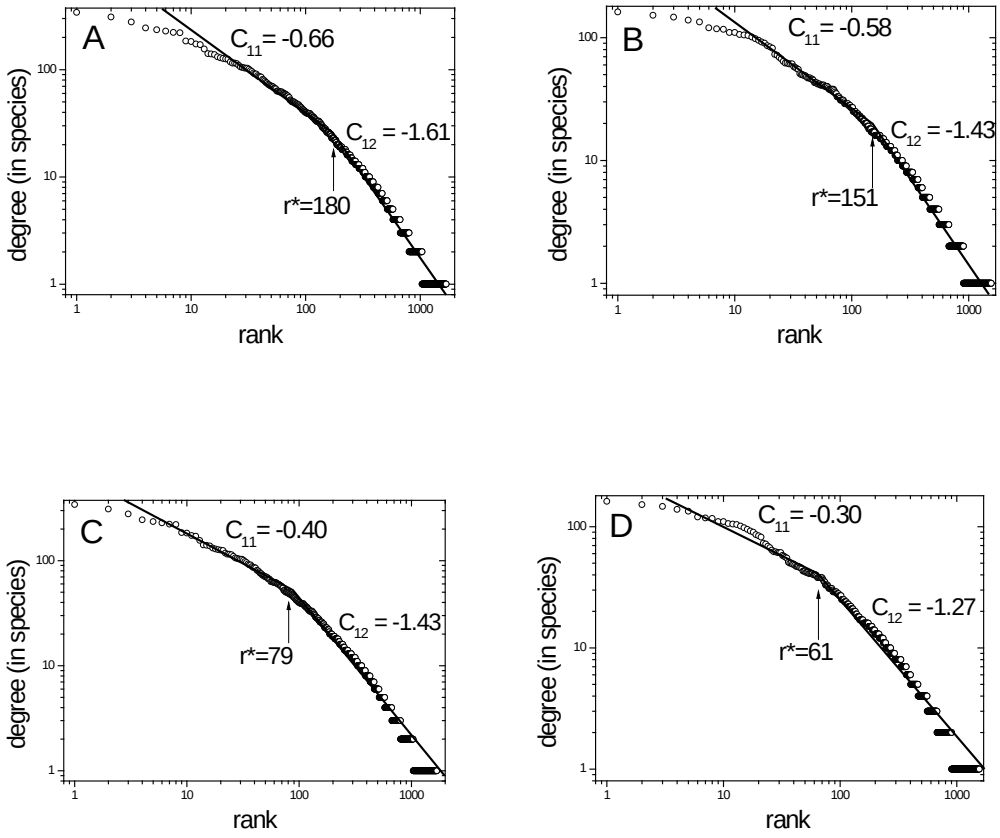


Table 3: Summary of the results of functions fitted for the whole database according to Pherobase (El-Sayed, 2012). For every target function, the coefficients giving the best fit are shown (the meaning of each coefficient is explained in Table 1). *Δ* is the difference between the AIC of the target function and that of the function giving the lowest AIC), *SSE* is the sum of squared errors and *ρ* is the correlation coefficient ($R^2_{corr,log}$ and $R^2_{corr,log,weight}$, for unweighted and weighted fit, respectively, following Li et al's (2010) notation).

Function	**Unweighted**				**Weighted**			
	Coefficients	Δ	***SSE***	***ρ***	**Coefficients**	Δ	***SSE***	***ρ***
Zipf	C_0 =4.12± 0.02 C_1 = -1.29±0.01	2074.4	31.24	0.94	C_0 =2.84± 0.01 C_1 = -0.74±0.01	4425.2	0.47	0.90
Beta	C_0 =3.59 ± 0.05 C_1= -1.21± 0.01 C_2 = 0.051±0.004	1939.9	28.81	0.95	C_0 =-0.17± 0.08 C_1= -0.625±0.005 C_2 = 0.39±0.01	3426.4	0.26	0.95
Yule	C_0 =3.36± 0.02 C_1 = -0.90± 0.01 C_4=-39·10^{-5} ±1·10^{-5}	967.5	16.18	0.97	C_0 =2.66± 0.01 C_1 = -0.49± 0.03 C_4 =-92·10^{-5} ±1·10^{-5}	1450.5	0.08	0.98
Menzerath -Altmann	C_0 =4.41± 0.02 C_1 = -1.39± 0.01 C_3 = -2.9±0.1	1354.7	20.36	0.96	C_0 = 3.40± 0.01 C_1 = -0.98± 0.01 C_3 = -0.95±0.02	2756.8	0.17	0.96
Double Zipf	**C_{01} =1.43± 0.01 C_{11} = -0.66± 0.01 C_{12} = -1.61±0.01**	**0**	**9.11**	**0.98**	**C_{01} =1.83± 0.01 C_{11} = -0.40± 0.01 C_{12} = -1.43±0.01**	**0**	**0.034**	**0.99**

Table 4: Summary of the results of functions fitted for the Pherobase (El-Sayed, 2012) excluding attractants. The format is the same as in Table 3.

Function	Unweighted				Weighted			
	Coefficients	Δ	SSE	ρ	Coefficients	Δ	SSE	ρ
Zipf	C_0 =3.68± 0.02 C_1 = -1.17±0.01	1601.5	24.50	0.94	C_0 =2.52± 0.01 C_1 = -0.66±0.01	3845.3	0.44	0.88
Beta	C_0 =3.34 ± 0.05 C_1= -1.12± 0.01 C_2 = 0.033±0.004	1543.6	23.58	0.94	C_0 =-0.26± 0.08 C_1= -0.55±0.01 C_2 = 0.37±0.01	3030.3	0.26	0.93
Yule	C_0 =3.08± 0.02 C_1 = -0.85± 0.01 C_4=-35·10^{-5} ±1·10^{-5}	881.3	15.41	0.96	C_0 =2.34± 0.01 C_1 = -0.42± 0.01 C_4 =-95·10^{-5} ±1·10^{-5}	1448.2	0.1	0.98
Menzerath -Altmann	C_0 =3.95± 0.02 C_1 = -1.26± 0.01 C_3 = -2.6±0.1	924.9	15.85	0.96	C_0 = 3.09± 0.01 C_1 = -0.98± 0.01 C_3 = -0.95±0.02	2107.5	0.15	0.96
Double Zipf	**C_{01} =1.32± 0.01 C_{11} = -0.58± 0.01 C_{12} = -1.43±0.01**	**0**	**8.76**	**0.98**	**C_{01} =1.73± 0.01 C_{11} = -0.30± 0.01 C_{12} = -1.27±0.01**	**0**	**0.038**	**0.99**

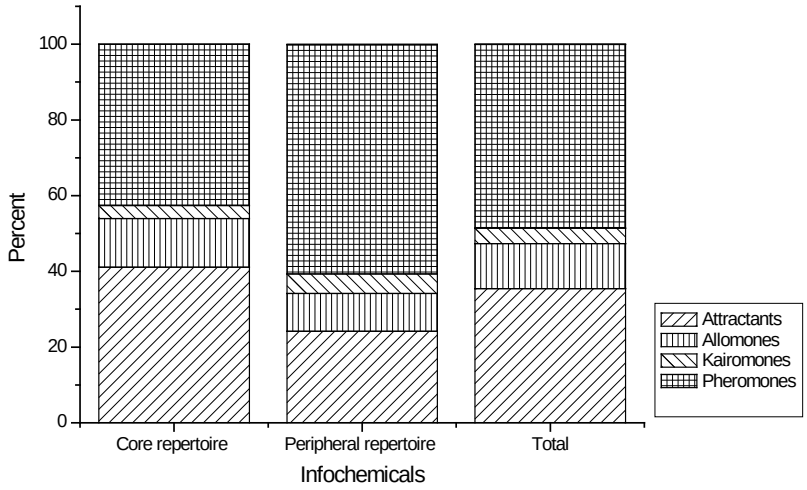
Table 5: Percentage of infochemical types both in the core and peripheral chemical repertoire considering the breakpoint *r** of the two regime distribution as the point of separation between both (El-Sayed, 2012). *T* is the number of infochemical-species associations (the total sum of degrees) and *n* is the repertoire size (in infochemical types).

	All infochemicals		**All infochemicals without attractants**	
	Unweighted	**Weighted**	**Unweighted**	**Weighted**
Core chemical repertoire (1st regime)	10.68%	4.69%	9.69%	3.91%
Peripheral chemical repertoire (2nd regime)	89.32%	95.31%	90.31%	96.09%
Breakpoint (*r)**	180	79	151	61
Total infochemical-species associations (*T*)	17633	17633	11380	11380
Repertoire (*n*)	1686	1686	1560	1560

Table 6: Number of infochemical-species associations that are inside the "Core Chemical Repertoire", and in the "Peripheral Chemical Repertoire", by kind of infochemical and for the whole database (percentages are shown in parentheses and are relative to the total in the right-most column). The breakpoint r^* of the double Zipf function in unweighted regression defines the boundary between both repertoires.

	Attractants	**Allomones**	**Kairomones**	**Pheromones**	**Synomones**	**TOTAL**
Core repertoire	4833 (41.11%)	1506 (12.81%)	410 (3.49%)	5005 (42.57%)	1 (<0.01%)	11755 (100%)
Peripheral repertoire	1420 (24.16%)	588 (10.00%)	302 (5.14%)	3563 (60.62%)	5 (0.09%)	5878 (100%)
TOTAL	6253 (35.46%)	2094 (11.88%)	712 (4.04%)	8568 (48.59%)	6 (0.03%)	17633 (100%)

Figure 3: Percentage of infochemical associations inside the "Core Chemical Repertoire", and in the "Peripheral Chemical Repertoire" by kind of infochemical. Synomones cannot be seen due to their very low proportion. Data is borrowed from Table 6.



DISCUSIÓN Y CONCLUSIONES GENERALES

Las fronteras de la lingüística

Vertebran este trabajo campos muy diversos en los que la lingüística cuantitativa se ha mostrado válida para afrontar problemas propios no solo de la lingüística tradicional sino también de la neurolingüística, la comunicación animal y vegetal, o la genómica. Nuestro enfoque permite replantear y extender las leyes de la lingüística más allá del lenguaje humano.

La lingüística, como la cinemática en el estudio de los movimientos de los cuerpos, empezó siendo fundamentalmente descriptiva; las lenguas se analizaron creando los conceptos que se creyeron necesarios. La dicotomía establecida entre enfoques diacrónicos y sincrónicos fue un primer paso hacia el planteamiento de la dinámica de las lenguas, es decir, hacia el cuestionamiento de las causas del fenómeno lingüístico. Al escudriñar las causas emergieron teorías y modelos, y no bastó con descomponer, desmenuzar y pormenorizar lo observado: la matemática formalizó la lingüística, a la vez que la teoría de la información maduraba en el siglo XX, y la estadística se erigió entonces en esencial para aproximarse a los datos que suministraba la experiencia.

Somos conscientes de que los temas abordados en este trabajo han sido muy diversos, de la genómica a la comunicación en los seres vivos, pero todos ellos poseen dos puntos en común: la información y la comunicación. Sobre la información, el filósofo Jesús Mosterín nos recuerda que los animales superiores poseen dos órganos procesadores de información, el genoma y el cerebro, y ambos son capaces de detectar, asimilar, almacenar, modificar, usar y transmitir información, además de hallarse materialmente en un soporte físico, ya sea el ADN o las neuronas (Mosterín, 1993). En lo que respecta a la comunicación, al hecho comunicativo, desde el modelo de Shannon (1948) es evidente que es necesario un emisor y un receptor para que la comunicación tenga lugar, lo que es obvio en la comunicación entre los seres vivos.

En la genómica el receptor es la propia especie en su camino evolutivo. La reproducción implica un mensaje a la generación siguiente, un mensaje en una *botella* celular que es el material genético. Dicho mensaje ha llegado a nuestra especie que, gracias a sus capacidades cognitivas, empieza ahora a ser capaz de discernir y comprender una ínfima parte de su contenido. En paralelo a cómo empezábamos a comprender nuestros sistemas lingüísticos nos dimos cuenta de que, en nuestro interior, existía un mensaje que tal vez sea mucho más importante que nuestra comunicación externa, en cuanto afecta a todos los seres vivos. A tal efecto, valga como reflexión final la cita de Mosterín (1993):

> Determinar unidades de información genética equivale a segmentar los cromosomas. Un gen es una unidad de información genética correspondiente a un cierto segmento cromosómico. Y esa segmentación no es unívoca, sino que puede efectuarse de diversos modos. (…) El léxico mismo puede

analizarse en las palabras que lo componen. Cada palabra es un meme, pero también lo es cada fonema, o cada una de las acepciones significativas de la palabra. Según el contexto de investigación, será útil elegir como unidades de cultura trozos lingüísticos mayores o menores: la lengua entera, el dialecto, el léxico, la palabra, la acepción, el fonema, etc. (…) Los cromosomas son unidades naturales, existentes con independencia de nuestras convenciones, y su número y estructura están ya dados en cada una de las células eucariotas. Las dimensiones culturales, sin embargo, son meros constructos nuestros, que nos sacamos de la manga para organizar de un modo manejable la complejidad inextricable de los datos culturales.

Mosterín (1993) apunta al problema crucial de la segmentación, al hecho de que se puede dividir de forma diferente tanto un texto lingüístico como el genoma. No obstante, como bien indica, los cromosomas son unidades con fronteras físicas claras, bien delimitados y existentes con independencia a constructos cerebrales humanos, constructos como podrían ser las palabras o los fonemas, o la propia división en genes que, como indica, no es unívoca (Alberts *et al.*, 2008, para una revisión). Debemos, pues, considerar de entrada como superiores las divisiones que la naturaleza nos presenta de forma diáfana, más allá de los constructos humanos, lo que refuerza más si cabe el análisis del nivel cromosómico que se ha llevado a cabo en nuestro trabajo (artículos del capítulo 4).

Mosterín (1993) alude también a los memes que, introducidos por Richard Dawkins, son unidades de información análogas a los genes que se transmiten de generación en generación, y nos recuerda que las unidades lingüísticas son memes, sobreentendiendo que la información está inextricablemente unida a la comunicación. Sostenemos –y hemos comprobado– que hay leyes generales que gobiernan la comunicación y, conceptualmente, deberían ser extensibles al nivel *memético*. Si las leyes de la dinámica de cuerpos se aplican con éxito tanto en la escala atómica como en las galaxias, pasando por todos los niveles mesoscópicos (intermedios), con sus correspondientes correcciones cuánticas y relativistas, probablemente las leyes de la comunicación, para constituirse como tales leyes de pleno derecho, que operan en la lingüística, en la comunicación de los seres vivos (en sus diversas modalidades, química, táctil, visual o acústica) y en la genética, deberían adaptarse a cada una de las escalas de estudio, y se deberían analizar también las desviaciones que se den para ajustarse cada vez más a la evidencia empírica existente.

6.1. Resultados, conclusiones y trabajo futuro de cada artículo

Sobre las leyes de la lingüística cuantitativa, para empezar, vimos en Ferrer-i-Cancho y Hernández-Fernández (2008) una generalización matemática sobre la ley de Zipf, a considerar en el futuro, dada la relevancia de la ley de Zipf en el lenguaje: el exponente de la distribución de frecuencias de palabras y el exponente de la relación potencial entre la frecuencia y su rango, coinciden únicamente cuando su valor es el número de oro (Ferrer-i-Cancho y Hernández-Fernández, 2008). Asimismo, se exploró por vez primera la desviación del exponente de Zipf en enfermos de Alzheimer, detectando la evolución verbal de la enfermedad (Hernández-Fernández y Diéguez-Vide, 2013), hecho que en el futuro se debería estudiar en más profundidad, con corpus más completos y longitudinales (en el caso de Alzheimer de GDS1 a GDS5). En otras palabras, las variaciones en la ley de Zipf podrían predecir la evolución sintáctica de estos pacientes, halladas en pacientes GDS5 pero no en GDS4. Mediante futuros sistemas de detección automática se pretende conseguir describir la evolución de ciertas enfermedades con alteraciones verbales (Hernández-Fernández y Diéguez-Vide, 2013), como podrían ser sujetos adultos con deterioro cognitivo leve (DCL). Es esencial definir mejor las desviaciones de las leyes generales en los casos de patología del lenguaje. La intención es clara: mejorar la detección y ayudar en los tratamientos de las diversas patologías.

La ley de Menzerath-Altmann, bien conocida en el lenguaje, se ha revisado y fundamentado desde una perspectiva teórica y estadística (Ferrer-i-Cancho *et al.*, 2014). Posteriormente se ha explorado su presencia en el nivel cromosómico (Hernández-Fernández *et al.*, 2011), interesante como se ha visto por su independencia de modelos cognitivos (Mosterín, 1993), y se han rebatido de forma sólida las críticas directas (Solé, 2010) e indirectas (Li, 2012) recibidas, revisando cada una de las inconsistencias tanto conceptuales como estadísticas que se habían esgrimido en nuestra contra (Ferrer-i-Cancho, Forns *et al.*, 2013).

En Baixeries y colaboradores (2013) se analizó y determinó de forma cuantitativa cada uno de los posibles ajustes de la ley de Menzerath-Altmann, $Z = aX^b e^{cX}$, en el nivel cromosómico del genoma. En concreto, se obtuvo que el ajuste de una ley potencial pura, sin el término exponencial, daba exponentes *b* de oscilaban entre -1.6 y 0.1, siendo *b=-1* no significativo estadísticamente para hongos, plantas gimnospermas, insectos, reptiles, peces con aletas o anfibios, en contra de Solé (2010).

El añadido del término exponencial de la ley de Menzerath-Altmann suponía una mejora en plantas (gimnospermas y angiospermas), mamíferos, peces con aletas y anfibios, y con la única excepción de pájaros y peces cartilaginosos, los parámetros de la ley de Menzerath-Altmann se desviaban significativamente de la simple ley potencial con exponente b=-1 propuesta como trivial, refutándose así de forma cuantitativa las criticas recibidas y mejorándose la exploración de la ley de Menzerath-Altmann en el nivel cromosómico.

En Hernández-Fernández y colaboradores (2011) además se concluyó la capacidad del espacio de fases G/Lg, es decir el espacio formado por el número total de pares de bases y el número de cromosomas haploide, de distinguir a grupos de organismos con un pasado evolutivo diferenciado. Entendemos, por supuesto, que todo el trabajo realizado en genómica está a merced de actualizaciones como consecuencia de nuevas secuenciaciones o aproximaciones biofísicas (Daban, 2014).

A la vez, en diversos artículos que han ido más allá de la mera refutación, y que apuntalaban nuestra posición, se han revisado también los modelos genéticos de fragmentación aleatoria (Baixeries *et al*., 2012; Ferrer-i-Cancho, Baixeries, *et al.*, 2013), estudiando sus límites y aportando, como crucial, el hecho de que cualquier modelo que se aplique a la división del genoma debe, cuanto menos, concordar con los datos disponibles y ser realista (por ejemplo no generando cromosomas vacíos). Nuestras revisiones matemáticas posteriores, que matizaban algunos corolarios y teoremas, mejoraron el artículo original (Baixeries *et al*., 2012), sin perder un ápice de rigor, ni de relevancia, los resultados y conclusiones establecidas con anterioridad (Ferrer-i-Cancho, Baixeries, *et al.*, 2013).

Se han establecido nuevos paralelismos entre la lingüística y la genómica, en lo referente a la presencia del mal llamado ADN basura (no codificante) que tiene su análogo en los elementos sin significado léxico del lenguaje (Ferrer-i-Cancho, Forns et al., 2013), y que en el proceso de investigación han salido reforzados gracias a las nuevas evidencias empíricas que han suministrado trabajos como Encode (2012). Lejos de estar cerrado, el nuevo campo que abre Encode (2012) debería promover el trabajo futuro conjunto de lingüistas y genetistas, con aproximaciones que desde la lingüística cuantitativa exploren nuevas maneras de resolver el misterio de la vida.

Por otra parte, se llevó a cabo también la comprobación de la ley de brevedad en los corpus de siete lenguas diferentes y su exploración, no siempre exitosa, en la comunicación animal, tanto en el repertorio de delfines como en las emisiones de

primates no humanos y cuervos (Ferrer-i-Cancho y Hernández-Fernández, 2013). Es sin duda un terreno a seguir explorando en los sistemas de comunicación de otras especies: ¿es la ley de brevedad universal o se restringe a sistemas de comunicación sofisticados? Este estudio nos condujo de nuevo a plantear el porqué de las excepciones de la ley de brevedad, y ulteriormente a proponer a la comunidad científica la extensión y aplicación del principio de compresión, originario de la teoría de la información, a la ciencia cognitiva y la comunicación (Ferrer-i-Cancho, Hernández-Fernández *et al.*, 2013). El principio de compresión es sin duda un principio más de los que operan en los sistemas de comunicación, aunque su relación precisa con los principios de minimización de la energía y de maximización de la información mutua todavía deba establecerse.

Para concluir, el estudio cuantitativo de la distribución de infoquímicos de Pherobase (El-Sayed, 2012) nos ha dejado el descubrimiento de la existencia de dos regímenes de Zipf en la distribución del grado o número de especies que utilizan dichos compuestos en la comunicación (Hernández-Fernández y Ferrer-i-Cancho, 2014) con la existencia de un repertorio químico nuclear y otro periférico en la comunicación química, en analogía a lo que sucede en el lenguaje (Ferrer-i-Cancho y Solé, 2001). El repertorio nuclear podría incluir los compuestos más eficientes y óptimos en un ecosistema y canal concreto, además de ser compuestos sin un coste energético excesivo para las especies. Los organismos, no obstante, deben recurrir al repertorio periférico, con sustancias de menor frecuencia de uso, para dar mayor especificidad a sus mensajes y evitar interferencias comunicativas, pues sería imposible que todas las especies empleasen el mismo infoquímico sin generar la más absoluta confusión (Hernández-Fernández y Ferrer-i-Cancho, 2014). Análogamente, en el lenguaje la carga léxica viene dada por el vocabulario formado por las palabras menos frecuentes, mientras que las palabras más frecuentes destacan por su conectividad sintáctica y menor capacidad semántica. Por otra parte, nuestro trabajo actual –y futuro inmediato– apunta al hecho de que los infoquímicos también estarían sujetos al principio de compresión.

6.2. Conclusiones generales

La tabla 6.1 (página siguiente) resume los principales resultados y conclusiones de cada artículo incluido en este trabajo, vistos de forma sucinta en el apartado anterior. Puede cotejarse con la tabla 0.1 (ver página 10) en la que se presentaban inicialmente los objetivos de cada uno. En general, hemos logrado realizar los objetivos planteados en cada uno de los artículos presentados.

Artículos	Principales resultados	Conclusiones
1. Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2014). The infochemical core. Journal of Quantitative Linguistics, pendiente de publicación.	El análisis de diversas funciones al grado o distribución de uso de infoquímicos (Zipf, Beta, Yule, Menzerath-Altmann y doble Zipf), da como resultado que los dos regímenes de Zipf son el mejor de los ajustes, según el AIC.	Hay un repertorio nuclear de infoquímicos compartido por múltiples especies, y uno periférico más específico, al que recurren para evitar las interferencias en el canal químico. Los infoquímos se organizan de forma parecida a las palabras.
2. Hernández-Fernández, A. y Diéguez-Vide, F. (2013). La ley de Zipf y la detección de la evolución verbal en la enfermedad de Alzheimer, *Anuario de Psicología/The UB Journal of Psychology*, 43 (1), 67-82.	Se han observado desviaciones del exponente de Zipf en las palabras de frecuencia media para pacientes GDS5, pero no en GDS4. La desviación del exponente de Zipf en los pacientes GDS5 muestra que es posible predecir la evolución de un estadio a otro en el Alzheimer.	El estudio de desviaciones del exponente de Zipf se puede aplicar en la neurolingüística y podría permitir deducir cuándo existe una alteración en la sintaxis a partir de la simple producción oral del enfermo. Posible relación entre el exponente de Zipf y el GDS.
3. Hernández-Fernández, A., Baixeries, J., Forns, N. y Ferrer-i-Cancho, R. (2011). Size of the whole versus number of parts in genomes. *Entropy*, 13 (8), 1465–1480, doi:10.3390/e13081465.	El tamaño de los cromosomas y el genoma se correlacionan en 9 de 11 grupos estudiados (todos salvo pájaros y peces cartilaginosos) y la presencia de la ley de Menzerath se da en 7 de los 9 grupos en los que hay correlación.	Se refutan estadísticamente los argumentos dados sobre la trivialidad de la presencia de la ley de Menzerath-Altmann en el nivel cromosómico del genoma. Se distinguen grupos de organismos en el espacio de fases G/L_g.
4. Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2013). The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, 4 (1), 45-55.	Se demuestra la presencia de la ley de brevedad en siete lenguas diferentes, así como en otros sistemas de comunicación animal. Se justifican las excepciones halladas hasta la fecha, especialmente en los corpus de dos pequeños primates.	Se pueden dar errores estadísticos si los repertorios son pequeños: se imposibilita así comprobar si la ley de brevedad está presente o no en un corpus. Otros factores, como la presencia de llamadas a larga distancia, podrían falsear los resultados obtenidos.
5. Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J. y Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*. DOI: 10.1111/cogs.12061.	La minimización de la longitud esperada de un código implica que la longitud esperada de una palabra o vocalización no puede aumentar, cuando aumenta su frecuencia de uso. La reducción temporal de una señal implica ahorro energético.	El principio de compresión opera en los sistemas de comunicación y en el comportamiento animal. La ley de brevedad es una consecuencia del principio de compresión. No es el único principio que actúa en los sistemas de comunicación.
6. Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G. y Baixeries, J. (2013). The challenges of statistical patterns of language: the case of Menzerath's law in genomes, *Complexity*, 18, 11–17.	Se demuestra que la ley de Menzerath en el cariotipo no es inevitable ni trivial. Las lenguas humanas poseen el equivalente al ADN no codificante, es decir, elementos sin aparente significado léxico.	Se han refutado todos y cada uno de los argumentos que defendían la trivialidad de la ley de Menzerath en el nivel cromosómico. Tanto el lenguaje como el genoma poseen unidades con referencia semántica arbitraria o simbólica.
7. Ferrer-i-Cancho, R., Baixeries, J., Hernández-Fernández, A., Debowski, L. y Macutek, J. (2014). *When is Menzerath-Altmann law mathematically trivial? A new approach.* Pendiente de publicación. Disponible en: http://arxiv.org/abs/1210.6599	Los test no paramétricos entre el número de pares de bases y el número de cromosomas establecen que su relación no es trivial en diez de once grupos taxonómicos analizados. La ley de Menzerath-Altmann no es inevitable ni trivial.	Revisión de los test estadísticos de correlación entre el número de pares de bases y el número de cromosomas, que permiten valorar la presencia de la ley lingüística de Menzerath-Altmann en el genoma.
8. Ferrer-i-Cancho, R., Baixeries, J. y Hernández-Fernández, A. (2013). Erratum to "Random models of Menzerath-Altmann law in genomes" (*BioSystems* 107 (3), 167–173). *BioSystems* 111 (3), 216-217.	Los teoremas y corolarios del artículo original (Baixeries et al., 2012) se actualizan especialmente al distinguir entre 'independencia' e 'independencia promedio'.	Las puntualizaciones estadísticas dadas a Baixeries *et al.* (2012), mejoran el artículo original en la revisión de los modelos de fragmentación aleatoria en el genoma, sin alterar sus resultados.
9. Baixeries, J., Hernández-Fernández, A., Forns, N., y Ferrer-i-Cancho, R. (2013). The parameters of Menzerath-Altmann law in genomes," *Journal of Quantitative Linguistics*, 20, 94–104.	El ajuste de una ley potencial pura da exponentes entre -1.6 y 0.1, y un exponente trivial de -1 en la ley de Menzerath-Altmann no ajusta a la mayoría de grupos taxonómicos, a excepción de pájaros y peces cartilaginosos.	Refutación cuantitativa de la trivialidad e inevitabilidad de la ley de Menzerath-Altmann en el nivel cromosómico. Potencialidad de los modelos de ajuste mediante regresión no lineal, por encima de la regresión lineal simple.
10. Baixeries, J., Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2012). Random models of Menzerath-Altmann law in genomes. *BioSystems* 107 (3), 167–173.	Algunos modelos de fragmentación aleatoria generan cromosomas vacíos. Los modelos de fragmentación aleatoria no pueden explicar trivialmente la dependencia entre el tamaño del genoma y el número de cromosomas.	La independencia del tamaño del genoma respecto al número de cromosomas, asumida en los modelos de fragmentación aleatoria, da pie a inconsistencias y a no concordar con la evidencia empírica.
11. Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2008). Power laws and the golden number. G. Altmann, I. Zadorozhna i Y. Matskulyak (eds.), *Problems of general, germanic and slavic linguistics* (pp. 518-523). Chernivtsi: Books– XXI.	Determinación de la relación matemática existente entre el exponente de Zipf de la distribución de frecuencias de palabras y el exponente de la distribución de rangos.	El exponente de la distribución de frecuencias de una magnitud y el exponente de la distribución de rangos son iguales cuando su valor es igual al número de oro o sección áurea.

Tabla 6.1.: Resumen de los principales resultados y conclusiones de los trabajos que conforman esta tesis.

Más allá de cada artículo, llegado a este punto, esperamos que el lector haya comprendido la dimensión de nuestro trabajo. La patología del lenguaje, la comunicación animal y la genómica son algunas de las fronteras de la lingüística que hemos transitado. Decíamos en el primer capítulo que el lenguaje puede ser entendido como el dominio de la lingüística, y por tanto como una realidad cognitiva exclusivamente humana con una realidad palpable (las lenguas) y además como un sistema de comunicación más de la Naturaleza. El desarrollo de la lingüística ha sido paralelo al estudio de la comunicación animal y, aunque hay cientos de estudios comparativos, pocos han intentado plantear un marco teórico común sobre el que trabajar. Creemos que, en parte, lo hemos hecho aquí.

El método científico es aplicable al estudio del lenguaje y la comunicación. Las matemáticas, y la estadística especialmente, son la herramienta esencial para contrastar y validar –o falsar– hipótesis. Así, hemos podido estudiar la presencia de la ley de Menzerath-Altmann en el nivel cromosómico del genoma, haciendo hincapié en los paralelismos entre el lenguaje y el ADN. También hemos comprobado que la ley de brevedad se sigue en las lenguas humanas y en algunos sistemas de comunicación animal. Además, hemos revisado también la ley de Zipf, en especial las desviaciones de su exponente en la enfermedad de Alzheimer, analizando sin dogmatismos cómo otros modelos podrían dar mejores ajustes, siguiendo criterios de teoría de la información (como el AIC), modelos que se han utilizado, para concluir, en el estudio de la distribución del uso de infoquímicos, que siguen la ley de Zipf en dos regímenes.

Son solo algunas leyes, bien conocidas en la lingüística cuantitativa, que se han explorado en las fronteras de la lingüística y que, sin embargo, nos han llevado a una profunda reflexión sobre la ciencia cognitiva. De esas cavilaciones en el piélago de la lingüística emergió la propuesta de extender el principio de compresión de la teoría de la información a los sistemas de comunicación.

Las distancias cualitativas y cuantitativas del lenguaje y la cognición humana con otros sistemas de comunicación han dado, históricamente, un carácter antropo-céntrico al estudio de la comunicación. La lingüística cuantitativa es capaz de aprender de la biodiversidad del planeta, de las miríadas de sistemas de comunicación que nos envuelven en un simple paseo por el bosque, a la vez que aportar sus leyes para ver si son una casualidad humana, y estamos entonces realmente solos en el Cosmos, o, por el contrario, son universales de la comunicación, de la comunicación de un mundo vivo que apenas empezamos a comprender.

Referencias

En negrita los artículos incluidos en este trabajo.

Akaike, H. (1974). A new look at statistical model identification. IEEE Trans. Automat. Control 19, 716–722.

Alberts, A. C. (1992). Constraints on the design of chemical communication-systems in terrestrial vertebrates. *American Naturalist*, **139**, 62–89.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. y Walter, P.(2002). *Molecular Biology of the Cell*. New York: Garland Publishing; quinta edición, 2008.

Altmann, G. (1980). Prolegomena to Menzerath's law. En R. Grotjahn (Ed.), *Glottometrika*, pp.1–10. Bochum: Brockmeyer.

Andersen, H. (2001). *On Kuhn*. Belmont: Wadsworth.

Andres, J. Kubáček, L., Machalová, J. y Tučková, M. (2012). Optimization of parameters in the Menzerath–Altmann law. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica*, Vol. 51 (2012), 1, 5-27.

Ash, R. B. (1965). *Information Theory*. New York: Wiley.

Atema, J. (1995). Chemical signals in the marine environment: Dispersal, detection and temporal signal analysis. *PNAS*, 92(l), 62-63.

Ay, N., Flack, J. y Krakauer, D.C. (2007). Robustness and complexity co-constructed in multimodal signalling networks. *Philosophical Transactions of the Royal Society of London B* 362 (1479), 441–447.

Aylett, M. y Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, 2006. 119(5): p. 3048-3058.

Baayen, H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, H. (2007). *Analyzing linguistic data. A practical introduction to statistics.* Nijmegen: Cambridge University Press.

Baixeries, J., Elvevåg, B. y Ferrer-i-Cancho, R. (2013). The evolution of the exponent of Zipf's law in language ontogeny. *PLoS ONE* 8 (3), e53227. [doi: 10.1371/journal.pone.0053227]

Baixeries, J., Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2012). Random models of Menzerath-Altmann law in genomes. *BioSystems* 107 (3), 167–173.

Baixeries, J., Hernández-Fernández, A., Forns, N., y Ferrer-i-Cancho, R. (2013). The parameters of Menzerath-Altmann law in genomes," *Journal of Quantitative Linguistics*, 20, 94–104.

Bannard, C. y Lieven, E. (2008). Repetition and reuse in child language learning. En: R.Corrigan et al. (Eds), *Formulaic Language*, Amsterdam: John Benjamins Pub., 299-321.

Basset, Y., *et al.* (2012). Arthropod diversity in a tropical forest. *Science*, 338, 1481-1484.

Bel-Enguix, G., Dahl, V., y Jiménez-López, M.D. (eds.) (2011). *Biology, Computation and Linguistics. New Interdisciplinary Paradigms.* Amsterdam: IOS Press.

Bennet-Clark, H. C. (1998). Size and scale effects as constraints in insect sound communication. *Philosophical Transactions of the Royal Society of London B*, 353(1367), 407–419.

Bezerra, B. M., Souto, A. S., Radford, A.N. y Jones, G. (2011). Brevity is not always a virtue in primate communication. *Biology Letters* 7, 23-25.

Boeckx, C. (2006). *Linguistic Minimalism*. New York: Oxford University Press.

Boroda, M.G. y Altmann, G. (1991). Menzerath's law in musical texts. *Musikometrica* 3, 1–13.

Bossert, W.H. y Wilson, E.O.(1963). The analysis of olfactory communication among animals. *Journal of Theoretical Biology*, 5, 443-469.

Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, Massachusetts: MIT Press.

Brown, C. H. y Sinnot, J. M. (2006). Cross-species comparisons of vocal perception. En: S. Greenberg, y W. A. Ainsworth (Eds.), *Listening to speech: An auditory perspective* (pp. 183–201). Londres: Routledge.

Browne, K. A., Tamburri, M. N. y Zimmer-Faust, R. K. (1998). Modelling quantitative structure-activity relationships between animal behaviour and environmental signal molecules. *Journal of Experimental Biology*, 201, 245–258.

Brumm, H. y Slabbekoorn, H. (2005). Acoustic communication in noise. *Advances in the Study of Behavior*, 35, 151–209.

Brumm, H. y Zöllinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* 148, 1173-1198.

Bunge, M. (1978). *Phylosophy of physics.* Dordrecht: Reidel.

Bunge, M. (2001). *La investigación científica.* México: Siglo XXI.

Bunge, M. (2010). *Las pseudociencias, ¡vaya timo!* Pamplona: Editorial Laetoli.

Chalmers, A.F. (2000). *¿Qué es esa cosa llamada ciencia?* Madrid: Siglo XXI Editores, tercera edición revisada y aumentada. http://ulagos.files.wordpress.com/2012/03/libro-que-es-esa-cosa-llamada-ciencia.pdf

Chapman, R. F. (1998). *The insects. Structure and function*, 4th edn. Cambridge: Cambridge University Press.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology*, *52A*, 273-302.

Chater, N. y Brown, G. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science* 32 (1), 36-67.

Chater, N. y Manning, C.D. (2006). Probabilistic models of language processing and acquisition, *TRENDS in Cognitive Sciences*, 10 (7), 335-345.

Chater, N. y Vitányi, P. M. B. (2002). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19-22.

Chen, H-W., Bandyopadhyay, S., Shasha, D.E. y Birnbaum, K.D. (2010). Predicting genome-wide redundancy using machine learning. *BMC Evolutionary Biology*, 10 (357), 1-15.

Chomsky, N. (1986). *Knowledge of language: Its nature, origins, and use*. New York: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge: The MIT Press.

Clemins, P. y Johnson, M.T. (2006). Generalized perceptual linear prediction (gPLP) features for animal vocalization analysis, *Journal of the Acoustical Society of America*, 120 (1), 527-534.

Corominas-Murtra, B. y Solé, R. (2010). Universality of Zipf's law. *Physical Review E*, *82* (1), 011102.

Cover, T. M. y Thomas, J. A. (2006). *Elements of information theory* (segunda edición). Hoboken, NJ: Wiley.

Cramer, I. (2005). The parameters of the Altmann–Menzerath law. *Journal of Quantitative Linguistics* 12 (1), 41–52.

Cuetos, F. (2012). *Neurociencia del lenguaje*. Barcelona: Ed. Panamericana.

Daban, J.R. (2014). The energy components of stacked chromatin layers explain the morphology, dimensions mechanical pe chromosomes, *Interface*, 11, 20131043.

Dao Duc, K. y Holcman, D. (2013) Computing the length of the shortest telomere in the nucleus, *Physical Review Letters*, 111, 228104 (2013).

De Leon, M.J. y Reisberg B. (1999). *An atlas of Alzheimer's Disease. The encyclopedia of visual medicine series.* Carnforth: Parthenon Publishing. Disponible en: http://www.alzinfo.org/clinical-stages-of-alzheimers

De, A., Ferguson, M., Sindi, S. y Durrett, R. (2001). The equilibrium distribution for a

Dóminich, S. y Horváth, M.S. (2008). *"Golden" Properties of the World Wide Web*. Trabajo póstumo, en: http://cir.dcs.uni-pannon.hu/cikkek/Dominich_Golden_Properties_WWW.pdf

Doolittle, W.F. (2013). Is junk DNA bunk? A critique of Encode. *PNAS*, doi: 10.1073/pnas.1221376110, 1-7.

Doyle, L.R., McCowan, B., Hanser, S.F., Chyba, C., Bucci, T. y Blue, J.E. (2008). Applicability of Information Theory to the quantification of responses to anthropogenic noise by southeast alaskan humpback whales. *Entropy*, 10, 33-46; DOI: 10.3390/entropy-e10020033

Dunham, I. *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.

Dusenbery, D. B. y Snell, T. W. (1995). A critical body size for use of pheromones in mate location. *Journal of Chemical Ecology*, **21**, 427–438.

El-Sayed, A.M. (2012). *The Pherobase: Database of Pheromones and Semiochemicals.* http://www.pherobase.com.

Encode Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the Encode pilot project. *Nature*, 447 (7146), 799–816.

Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 85-96.

Endler, J. A. (1993). Some general comments on the evolution and design of animal communication systems. *Philosophical Transactions of the Royal Society of London B, 340*, 215-225.

Eroglu, S. (2013). Parameters of the Menzerath-Altmann law: Statistical mechanical interpretation as applied to a linguistic organization**.** http://arxiv.org/ftp/arxiv/papers/1307/1307.7140.pdf

Eroglu, S. (2013b). Menzerath–Altmann law for distinct word distribution analysis in a large text, Physica A, vol. 392, 12, 2775-2780.

Evans, N., y Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429–492.

Fenk-Oczlon,G. y Fenk, A.(2008). Complexity trade-offs between the subsystems of language. En: M. Miestamo, K.Sinnemäki y F.Karlsson (eds.), *Language Complexity. Typology, contact, change.* Vol. 94, 43-66. Amsterdam: John Benjamins.

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

Ferrer-i-Cancho, R. (2005b). Hidden communication aspects inside the exponent of Zipf's law. *Glottometrics*, *11*, 98–119.

Ferrer-i-Cancho, R. (2005c). Decoding least effort and scaling in signal frequency distributions. *Physica A: Statistical Mechanics and its Applications*, *345* (1), 275–284.

Ferrer-i-Cancho, R. (2005d). Zipf's law from a communicative phase transition. *European Physical Journal B* 47, 449-457.

Ferrer-i-Cancho, R. (2013). The optimality of attaching unlinked labels to unlinked objects. Arxiv: http://arxiv.org/ftp/arxiv/papers/1310/1310.5884.pdf

Ferrer-i-Cancho, R. y Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics*, P06009.

Ferrer-i-Cancho, R. y Elvevåg, B. (2010) Random texts do not exhibit the real Zipf's Law-like rank distribution. *PLoS ONE* 5(3):e9411. doi:10.1371/journal.pone.0009411

Ferrer-i-Cancho, R. y Forns, N. (2009). The self-organization of genomes. *Complexity* 15 (5), 34–36.

Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2008). Power laws and the golden number. In: "*Problems of general, germanic and slavic linguistics*", Altmann, G., Zadorozhna, I. y Matskulyak, Y. (eds.), Chernivtsi: Books - XXI. pp. 518-523.

Ferrer-i-Cancho, R. y Hernández-Fernández, A. (2013). The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, 4 (1), 45-55.

Ferrer-i-Cancho, R. y Moscoso del Prado Martín, F. (2011). Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, L12002.

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

Ferrer-i-Cancho, R., Baixeries, J. y Hernández-Fernández, A. (2013c). Erratum to "Random models of Menzerath-Altmann law in genomes" (*BioSystems* 107 (3), 167–173). *BioSystems* 111 (3), 216-217.

Ferrer-i-Cancho, R., Baixeries, J., Hernández-Fernández, A., Debowski, L. y Macutek, J. (2014). ***When is Menzerath-Altmann law mathematically trivial? A new approach.*** **Pendiente de publicación. Disponible en: http://arxiv.org/abs/1210.6599**

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

Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G., y Baixeries, J. (2013b). The challenges of statistical patterns of language: the case of Menzerath's law in genomes, ***Complexity*****, 18, 11–17.**

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J. y Semple, S. (2013). Compression as a universal principle of animal behavior. ***Cognitive Science*****, 2013, 1-14. DOI: 10.1111/cogs.12061.**

Ferrer-i-Cancho, R., y Lusseau, D. (2009). Efficient coding in dolphin surface behavioral patterns. *Complexity*, 14(5), 23–25.

Ferrer-i-Cancho, R. y Solé R.V. (2001). Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8, 165-173.

Feyerabend, P. (1993). *Contra el método*. Barcelona: Planeta. Traducción de: Feyerabend, P. (1975). *Against method*. New York: New Left Books.

Fitch, W. T. (2000) The evolution of speech: a comparative view. *Trends in Cognitive Science* 4, 258.

Fitch, W. T., Huber, L. y Bugnyar, T. (2010) Social cognition and the evolution of language: Constructing cognitive phylogenies. *Neuron* **65**, 795-814.

Fitch, W.T. (2009). Prolegomena to a future science of biolinguistics. *Biolinguistics*, Vol. 3 (4), 283-320.

Fletcher, N. H. (2004). A simple frequency-scaling rule for animal communication. *Journal of the Acoustical Society of America*, 115(5), 2334–2338.

Font-Clos, F., Boleda, G. y Corral, A. (2013). A scaling law beyond Zipf's law and its relation to Heaps' law. *New Journal of Physics,* **15,** 093033. doi:10.1088/1367-2630/15/9/093033

Gabaix, X. (1999). Zipf's law for cities: an explanation. *The Quarterly Journal of Economics*, 114(3), 739-767.

Galtier, N., Nabholz, B., Glemin, S. y Hurst, G.D.D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal, *Molecular Ecology*, 18, 4541–4550

generalized Sankoff–Ferretti model accurately predicts chromosome size distribution

Gillooly, J. F. y Ophir, A. G. (2010). The energetic basis of acoustic communication. *Proceedings of the Royal Society B*, 277, 1325–1331.

Gleick, J. (1989). *Caos.* Barcelona: Crítica, colección Drakontos, ed. revisada, 2012.

Gleick, J. (2011). *The information.* New York: Pantheon Books, Random House.

Gregory, R.T. (Ed.) (2005). *The Evolution of the Genome*. Londres: Academic Press.

Grice, H.P. (1989). *Studies in the Way of Words*. New York: Harvard University Press.

Guiraud, P. (1968). The semic matrices of meaning. *Social Science Information*, 7(2):131-139.

Ha, L.Q., Sicilia-García, I., Ming, J. y Smith, F.J. (2002). Extension of Zipf's Law to words and Phrases, *Journal of Computational Linguistics and Chinese Language Processing, Vol. 8, No. 1*, 77-102, Febrero de 2003.

Haken, H. (1983). *Synergetics, an introduction: nonequilibrium phase transitions and self-organization in physics, chemistry, and biology*. New York: Springer-Verlag.

Hauser, M.D., Newport, E.L. y Aslin, R.N. (2001). Segmentation of the speech stream in a nonhuman primate: Statistical learning in cotton-top tamarins. *Cognition*, 78, B53–B64.

Hay, M. (2009). Marine chemical ecology: chemical signals and cues, structure marine populations, communities, and ecosystems. *Annu. Rev. Marine Sciences*, **1**, 193–212.

Heisenberg, W. (1956). Física y filosofía. Buenos Aires: Ediciones la isla.

Herdan, G. (1964). *Quantitative linguistics*. Belfast: Butterworth &Co. Publishers.

Hernández-Fernández, A. (2006). *La ley de Zipf en el método comparativo*. Tesina para la obtención del DEA, Universidad de Barcelona, enero de 2006, no publicada.

Hernández-Fernández, A. y Diéguez-Vide, F. (2013). La ley de Zipf y la detección de la evolución verbal en la enfermedad de Alzheimer, *Anuario de Psicología/The UB Journal of Psychology*, 43 (1), 67-82.

Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2014). The infochemical core. *Journal of Quantitative Linguistics*, enviado el 30 de julio de 2013.

Hernández-Fernández, A. y Ferrer-i-Cancho, R. (2014b). Compression in infochemicals. En preparación.

Hernández-Fernández, A., Baixeries, J., Forns, N. y Ferrer-i-Cancho, R. (2011). Size of the whole versus number of parts in genomes. *Entropy*, 13 (8), 1465–1480, doi:10.3390/e13081465.

Hirsch, J.E. (2005). *An index to quantify an individual´s scientific research output*. *http://arxiv.org/abs/0911.3144*

Hockett, C.F. (1963): *The problem of universals in language.* En: *Universals of Language* (Editado por J.H.Greenberg), pp.1-29. Cambridge: MIT Press.

Howes, D. (1968). Zipf's law and Miller's Random-Monkey model. *The American Journal of Psychology*, Vol. 81, No. 2, 269-272.

Hřebíček, L. (1995). *Text levels. Language constructs, constituents and the Menzerath-Altmann law*. Trier: Wissenschaftlicher Verlag.

Hrncir, M., Barth, F.G. y Tautz, J. (2006). Vibratory and airborne-sound signals in bee communication (*Hymenoptera*). En: Drosopoulos, S. y Claridge, M.F. (eds), *Insect sounds and communication*. Boca Raton: CRC Press Taylor & Francis Group, p.421-439.

Jayaram, B.D. y Vidya, M.N. (2009). The relationship between word length and word frequency in Indian languages. *Glottotheory, 2*, 62-69.

Ji, A., Johnson, M.T., Walsh, E.J. McGee, J. y Armstrong, D.L. (2013). Discrimination of individual tigers (*Panthera tigris*) from long distance roars, *Journal of the Acoustical Society of America,* 133(3), 1762-1769.

Karlson P., Lüscher M. (1959). "Pheromones: a new term for a class of biologically active substances". *Nature* **183** (4653): 55–56.

Kello, C.T., Brown, G.D., Ferrer-i Cancho, R., Holden, J.G., Linkenkaer-Hansen, K., Rhodes, T. y Van Orden, G.C. (2010). Scaling laws in cognitive sciences. *Trends in cognitive sciences*, *14* (5), 223–232.

Kelso, J.S. (2000). *Dynamic patterns: the self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Kerszberg, M. (2003). Genes, neurons and codes: remarks on biological communication. *BioEssays*, 25, 699–708.

King, P. (2000): Internalismo, externalismo y autoconocimiento, *Crítica. Revista Hispanoamericana de Filosofía*, Vol. XXXII (96), 99–119.

Klug, W.S. y Cummings, M.R. (Eds) (1999). *Conceptos de genética*. Madrid: Prentice-Hall.

Köhler, R. (1990). Elemente der synergetischen linguistik. *Glottometrika*, 12: 179-188.

Köhler, R. (2005). *Gegenstand und arbeitsweise der quantitativen linguistik.* En: Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski (Eds.): *Quantitative*

linguistik - quantitative linguistics. Ein internationales handbuch. De Gruyter: Berlin/ New York, pp. 1-16.

Köhler, R. (2005b). *Synergetic linguistics*. En: Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski (Eds.): Quantitative linguistik - Quantitative linguistics. *Ein internationales handbuch.* De Gruyter: Berlin/ New York, pp. 760-775.

Kornai, A. (2008). *Mathematical linguistics*. London: Springer.

Kraft, V. (1986). *El círculo de Viena*. Barcelona: Taurus Ediciones.

Krakauer, D.C. y Plotkin, J.B. (2002). Redundancy, antiredundancy, and the robustness of genomes. *PNAS*, 99(3), 1405-1409.

Law, R.H. y Regnier, F.E. (1971). Pheromones. *Annual Review of Biochemistry*, **40**, 533–548.

Lewicki, M.S. (2002). Efficient coding of natural sounds. *Nature Neuroscience, 5*, 356-363.

Li, W. (2012). "Menzerath's law at the gene-exon level in the human genome". *Complexity* 17 (4), 49–53.

Li, W., Miramontes, P. y Cocho, G. (2010). Fitting ranked linguistic data with two-parameter functions. *Entropy* 2010, ***12***, 1743-1764. doi:10.3390/e12071743

Liu, H. y Huang, W. (2012). Quantitative linguistics : State of the art, theories and methods. *Journal of Zhejiang University*, 43(2): 178-192.

Luisi, P.L. (2006). Chapter 11: approaches to the minimal cell. En: *The Emergence of live. Chemical origins to synthetic biology*. Cambridge: Cambridge University Press, 242–267.

Luscombe, N.M., Qian, J., Zhang, Z., Johnson, T. y Gerstein, M. (2002). The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biology* 2002, 3(8):research0040.1–0040.7.

Mach, E. (1959). *The analysis of sensations and the relation of the physical to the psychical*, New York: Dover Publications.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. 3ª edición. Volumen 2: *The Database*. Mahwah: Lawrence Erlbaum Associates.

Magurran, A.E. (2004). *Measuring biological diversity*. Oxford: Blackwell.

Mandelbrot, B. (1952). *An information theory of the statistical structure of language. In Proceedings of Symposium on Application Communication Theory*, Butterworth, London, 22–26 September 1952; 486–500.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication theory*, 486–502.

Mandelbrot, B. (1953b) Contribution à la théorie mathématique des jeux de communication, *Publ. Inst. Stat. Univer. de Paris*, 2, 1-124.

Mandelbrot, B. (1962). On the theory of word frequencies and on related markovian models of discourse. *Structure of language and its mathematical aspects*, 190–219.

Mandelbrot, B. (1966). Information theory and psycholinguistics: a theory of words frequencies, En: P. Lazafeld, N. Henry (Eds.), *Readings in Mathematical Social Science*, MIT Press, Cambridge, MA, 1966.

Manin, D.Y. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, 32 (7), 1075–1098.

Manin, D.Y. (2009). Mandelbrot's model for Zipf's law: Can Mandelbrot's model explain Zipf's law for language? *Journal of Quantitative Linguistics*, 16(3), 274-285.

Margulis, L. (1970). *The origin of Eucaryotic cells*. New Haven(USA): Yale University Press.

Margulis, L. (1996). Archaeal-eubacterial mergers in the origin of Eukarya: Phylogenetic classification of life. *Proceedings of the National Academy of Sciences*, 93, 1071-1076.

Marsden, M.P.F. y Laemmli, U.K. (1979). Metaphase chromosome structure: Evidence for a radial loop model. *Cell*, 17(4), 849–858.

Mateu Bellés, J. (1993). Biogeografía. En: *Geografía General I. Introducción y Geografía física*. Madrid: Taurus.

McConnell, M. J., Lindberg, K. J., Brennand, J. C., Piper, T., Voet, C., Cowing-Zitron, S., Shumilina, R. S., Lasken, J. R., Vermeesch, J., Hall, I.M. y Gage, F.H. (2013). Mosaic copy number variation in human neurons. *Science*, 342 (6158): 632. DOI: 10.1126/science.1243472

McCowan, B., Hanser, S.F. y Doyle, L. R. (1999): Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires, *Animal Behaviour*, **57,** 409–419.

Menzerath, P.(1954). *Die architektonik des deutschen wortschatzes*. Bonn: Dümmler.

Miller, G.A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, Vol. 70, No. 2, 311-314.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, *1* (2), 226–251.

Molina, N. y van Nimwegen, E. (2009) Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends in Genetics*, 25 (6), 243-247.

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

Mosterín, J. (1993). Filosofía de la cultura. Madrid: Alianza Editorial.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, *46* (5), 323–351.

Novo Villaverde, F.J. (2007). *Genética Humana*. Madrid: Pearson.

Okubo, A., Armstrong, R.A., y Yen, J. (2001). Diffusion of "smell" and "taste": chemical communication. En A. Okubo y S.A. Levin (eds.), *Diffusion and ecological problems*, pp. 107–126. New-York: Springer-Verlag.

Otero, E. (2004). La distinción kuhniana entre tipos de ciencia y la inconsistencia fundacional de los estudios sociales de la ciencia. *Ciencias Sociales Online*, 2004, Vol. I, 1, 1-7.

Pennisi, E. (2012). Genomics. Encode project writes eulogy for junk DNA. *Science*, 337(6099), 1159–1161.

Petersen, A.M., Tenenbaum, J.N., Havlin, S., Stanley, H.E. y Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words, *Scientific Reports*, *2*,943.

Petrov, D.A. (2001) Evolution of genome size: New approaches to an old problem. *Trends in Genetics*, 17, 23–28.

Piantadosi, S., Tily, H. y Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences,* doi: 10.1073/pnas.1012551108 .

Plaza Morales, L. (2010). *Uso de grafos semánticos en la generación automática de resúmenes y estudio de su aplicación en distintos dominios: biomedicina, periodismo y turismo.* Tesis no publicada, en: http://nlp.uned.es/~lplaza/papers/PhDThesis.pdf

Polikarpov, A.A. (2006). Towards the foundations of Menzerath's law. En: P. Grzybek (Ed.), *Contributions to the science of text and language text, speech and language technology* 31, pp. 215-240.

Popescu, I.I. (*ed*) (2009). *Word frequency studies*. Berlín: de Gruyter.

Popescu, I.I., Mačutek, J. y Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

Powers, D.M.W. (1998). Applications and explanations of Zipfs law. En D.M.W. Powers (ed.) NeMLaP3/CoNLL98: *New methods in language processing and computational natural language learning*, ACL, 151-160. http://acl.ldc.upenn.edu/W/W98/W98-1218.pdf

Prestwich, K.N. (1994). The energetics of acoustic signalling on anurans and insects. *American Zoologist*, 34, 625-643.

Pulvermüller, F. (1999*).* Words in the brain's language. *Behavioral and Brain Sciences* (1999) **22,** 253–336.

Pulvermüller, F. (2002). *The neuroscience of language*. Cambridge: Cambridge Universtity Press.

Rago, F. y Cheeseman, I.M. (2013). The functions and consequences of force at kinetochores. *Journal of Cell Biology*, 200 (5), 557–565.

Regnier, F.E. y Law, R.H. (1968). Insect pheromones. *Journal of Lipid Research*, **9**, 541–551.

Reisberg B *et al.* (1982). The Global Deterioration Scale for assessment of primary degenerative dementia. *American Journal of Psychiatry* 1, 139(9), 1136-1139.

Riba, C. (1990). *La comunicación animal. Un enfoque zoosemiótico*. Barcelona: Anthropos.

Robinson, P.J., Fairall, L., Huynh, V.A.T. y Rhodes, D. (2006). EM measurements define the dimensions of the ''30-nm'' chromatin fiber: Evidence for a compact, interdigitated structure. *PNAS*, 103(17), 6506-6511.

Rojido, G.M. (2001). One hundred years of anaphylaxis. *Alergol Inmunol Clin 2001; 16: 364-368.*

Rosch, E. (1978). Principles of categorization. En: Rosch, E. y Lloyd, B.B. (eds), *Cognition and categorization* (pp. 27-48). Hillsdale: Lawrence Erlbaum. http://commonweb.unifr.ch/artsdean/pub/gestens/f/as/files/4610/9778_083247.pdf

Roy, D. *et al.* (2006). *The Human Speechome Project*. Twenty-eighth Annual Meeting of the Cognitive Science Society. http://web.media.mit.edu/~dkroy/papers/pdf/cogsci06.pdf

Rumelhart, D.E. y McClelland, J.L. y el PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, Cambridge: MIT Press

Saffran, J.R., Aslin, R.N. y Newport, E.L. (1996). Statistical learning by 8-month old infants. *Science*, 274, 1926–1928.

Saichev, A., Malevergne, Y., y Sornette, D. (2010). *Theory of Zipf's law and beyond* (Vol. 632). Dordrecht: Springer.

Sankoff, D. y Ferretti, V. (1996). Karyotype distributions in a stochastic model of reciprocal translocation. *Genome Research* 6, 1–9.

Sapp, J. (1994). Evolution by association. New York, N.Y.: Oxford University Press.

Schauder, S. y Bassler, B.L. (2001). The language of bacteria. *Genes and Development*. 15, 1468-1480.

Schubert, I. (2007). Chromosome evolution. *Current Opinion in Plant Biology*, 10, 109–115.

Schubert, I. y Oud, J.L. (1997). There is an upper limit of chromosome size for normal development of an organism. *Cell*, 88, 515–520.

Schwann, T. y Schleyden, M.J. (1847). *Microscopical researches into the accordance in the structure and growth of animals and plants*. London: Printed for the Sydenham Society.

Searls, D.B. (2002). The language of genes. *Nature*, 420, 211.

Semple, S., Hsu, M. J., y Agoramoorthy, G. (2010). Efficiency of coding in macaque vocal communication. *Biology Letters*, 6(4), 469–471.

Seyfarth R.M., y Cheney D.L. (2003). Signalers and receivers in animal communication. *Annual Review of Psycholo*gy, 54, 145–173.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423, 623–656, July, October, 1948.

Shannon, C. (1950). The redundancy of English. En *Cybernetics*: *Transactions of the 7th Conference*. New York: Josiah Macy, Jr. Foundation, pp. 248–272. Reimpreso en 2003, *Cybernetics*. Berlin: Diaphanes Verlag, pp.248-272.

Shannon, C. y Weaver, W. (1949). *The mathematical theory of communication*. Chicago: University of Illinois Press.

Shepard, R.N. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-345.

Shorey, H.H. (1976). *Animal communication by pheromones*. New York: Academic Press.

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika*, 425–440.

Simon, H.A. (1960). Some further notes on a class of skew distribution functions. *Information and Control*, *3* (1), 80–88.

Smith Churchland, P. (1985). *Neurophilosophy*. Cambridge, MA: MIT Press.

Smith, F.J. y Devine, K. (1985). Storing and retrieving word phrases, *Information Processing & Management*, **21**, 3, 215-224.

Sokal, R.R. y Rohlf, F.J. (1995). *Biometry: The principles and practice of statistics in biological research*. New York: Freeman.

Solé, R.V. (2010). Genome size, self-organization and DNA's dark matter. *Complexity* 16 (1), 20-23.

Sovilj, M. (2011). *Prenatal bases of development of speech and language and prenatal stimulation*. In *QIM 2011 Round Table Knowledge Federation Proceedings*, Karabeg, D. (*ed*.), Belgrade: QIM 2011 Secreatariaat Editions.

Stanley, H.E., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K. y Simons, M. (1999). Scaling features of noncoding DNA. *Physica A*, 273, 1-18.

Steiger, S., Schmitt, T. y Schaefer, H.M. (2011). The origin and dynamic evolution of chemical information transfer. *Proceedings of the Royal Society of London B*, 278, 970–979.

Stemmer, B. y Whitaker, H.A. (2008). *Handbook of the neuroscience of language*. Londres: Elsevier.

Strauss, U., Grzybek, P. y Altmann, G. (2007). Word length and word frequency. En: *Contributions to the science of text and language*, (P. Grzybek (Ed.)), pp. 277–294. Dordrecht: Springer.

Sueur, J. (2006). Insect species and their songs. En: Drosopoulos, S. y Claridge, M.F. (eds), *Insect sounds and communication*. Boca Raton: CRC Press Taylor & Francis Group, pp. 207-217.

Symonds M.R.E. y Elgar M.A. (2004). The mode of pheromone evolution: evidence from bark beteles. *Proceedings of the Royal Society of London* B, 271, 839–846.

Teupenhayn, R. y Altmann, G. (1984). Clause length and Menzerath's law. *Glottometrika* 6, 127–138.

Thom, R. (1993). *Parábolas y catástrofes*. Barcelona: Tusquets.

Thornhill, R. y Alcock, J. (1983). *The evolution of insect mating systems*. Cambridge, Massachusets: Harvard University Press.

Tomaschek, F., Wieling, M., Arnold, D. y Baayen, R. H. (2013). Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience. *Interspeech* 2013 (en prensa).

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

Turchin, P. (2003), *Complex population dynamics: a theoretical/empirical synthesis*, Nueva York: Princeton University Press.

Van Ewijk, L. (2011). *Word retrieval in acquired and developmental language disorders: a bit more on processing*. Amsterdam: Netherlands Graduate School of Linguistics.

Venables, W.N. y Ripley, B.D. (1999). *Modern applied statistics with S-PLUS.* New York: Springer-Verlag.

Vinogradov, A.E. (2001) Mirrored genome size distributions in monocot and dicot plants. *Acta Biotheoretica*, *49*, 43–51.

Von Newmann, J. (1955). *'Method in the physical sciences'*, in *The unity of knowledge*, editted by L. Leary, p.158. Reimpreso en J. Von Neumann, F. Bródy (ed.) y T. Vámos (*ed*.), *The Neumann compendium* (2000), p.628.

Von Newmann, J. y Morgenstern, O. (1944). *Theory of games and economic behavior*. *Chapter 1: formulation of the economic problem.* Princeton: Princeton University Press, 7-8.

VV.AA. (2002). To honor G.K. Zipf. *Glottometrics*, 3, 2002, Special Issue. Lüdenscheid: RAM-Verlag.

Waggensberg, J. (1994). *Ideas sobre la complejidad del mundo.* Barcelona: Tusquets.

Waggensberg, J. (2007). *El gozo intelectual*. Barcelona: Tusquets.

Wallace, C.S. y Freeman, P.R. (1987). Estimation and inference by compact coding, *Proceedings of the Royal Statistical Society B, 49*, 240-265.

West, G.B. y Brown, J.H. (2005). The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *Journal of Experimental Biology*, *208*, 1575-1592.

Wilde, J. y Schwibbe, M.H. (1989). Organizationsformen von erbinformation im hinblick auf die menzerathsche regel. En G. Altmann y M.H. Schwibbe (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen* (pp. 92–107). Hildesheim: Olms.

Willis, R.J. (2007). *The history of allelopathy*, Dordrecht: Springer.

Wilson, E.O. (1970). Chemical communication within animal species. En E. Sondheimer y J.B. Simeone (eds.), *Chemical ecology,* **9**, pp. 133–155. New York: Academic Press.

Wilson, E.O. y Bossert, W.H. (1963). Chemical communication among animals. *Recent Progress in Hormone Research*, **19**, 673–716.

Wilson, E.O. (1958). A chemical release of alarm and digging behavior in the ant Pogonomyrmex badius (Latreille). *Psyche* **65**, 41-51.

Winkler, H. (1920). *Verbreitung und ursache der parthenogenesis im pflanzen - und Tierreiche*. Jena: Verlag Fischer.

Wolfsberg, T., McEntyre, J. y Schuler, G. (2001). Guide to the draft human genome. *Nature,* 409 (6822), 824–826.

Wyatt, T.D. (2003). *Pheromones and animal behaviour*. Cambridge: Cambridge University Press.

Wyatt, T.D. (2009). Fifty years of pheromones. *Nature,* 457, 262–263.

Wyatt, T.D. (2010). Pheromones and signature mixtures: defining species-wide signals and variable cues for individuality in both invertebrates and vertebrates. *Journal of Comparative Physiology A*, 196, 685–700.

Yule, G.U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J.C.Willis, F.R.S. *Philosophical Transactions of the Royal Society of London B*, *213*, 21–87.

Zimmermann, E., Newman J.D. y Jürgens U. (Eds.) (1995) *Current topics in primate vocal communication.* New York: Plenum Press.

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Cambridge, MA: MIT Press.

Zipf, G.K. (1949). *Human behaviour and the principle of least effort.* Cambridge, MA: Addison-Wesley.

Zöllinger, S.A. y Brumm, H. (2011). The Lombard effect. *Current Biology,* 21, R614-615.

Zöllinger, S.A., Podos, J., Nemeth, E., Goller, F. y Brumm, H. (2012). On the relationship between, and measurement of, amplitude and frequency in birdsong, *Animal Behaviour,* http://dx.doi.org/10.1016/j.anbehav.2012.04.026