
La física del lenguaje¹

Román Orús

Donostia International Physics Center

roman.orus@dipc.org

Juan Uriagereka

University of Maryland

juan@umd.edu

Recibido: 16-2-2020

Aceptado: 14-1-2022

Resumen: en este artículo resumimos brevemente y de manera no técnica varios avances recientes en el estudio del lenguaje desde la perspectiva de la física. En particular, comentamos resultados referentes a la equivalencia entre MERGE y renormalización, así como de la teoría de Matrix Syntax. Proponemos también un experimento que pueden llevar a cabo estudiantes, de cara a jugar con una aproximación científica al lenguaje.

Abstract: in this paper we sketch very briefly and in a non-technical way several recent advances in the study of language from the perspective of physics. In particular, we comment on results concerning the equivalence between MERGE and renormalization, as well as the theory Matrix Syntax. We also propose an experiment for students in order to play with a scientific approach to language.

Palabras clave: física, lenguaje, renormalización, matriz, merge.

Key Words: physics, language, renormalization, matrix, merge

Índice

1. Introducción
2. Lenguaje y renormalización
3. Lenguaje, álgebra y física cuántica
4. Proyecto: el bloque fundamental del lenguaje

¹ Agradecimientos: mis sinceros agradecimientos a Ángel Gallego, Roger Martin, y Juan Uriagereka.

1. Introducción

Es de sobra conocido que el estudio del lenguaje se ha convertido en una disciplina pluridisciplinar. El lenguaje, es decir, el cómo hablamos, escribimos y, sobre todo, *pensamos*, es la ventana abierta al estudio del cerebro humano. Podemos entender nuestro cerebro sin abrirlo, tan solo con estudiar lo que produce. De hecho, el lenguaje es una de las características que nos hacen más humanos. Es un instrumento de transmisión de información bastante eficiente, pero es aún más eficiente como herramienta de razonamiento mental (parece estar hecho para pensar, y no para hablar). Es por ello que, en la actualidad, no es posible enfocar el estudio del lenguaje desde un punto de vista monolítico. Hace falta usar el conocimiento de la lingüística, claro está, pero también de la teoría de la información, de la neurociencia, de la biología, de la paleontología, y de muchas otras disciplinas.

En esta línea, durante los últimos años ha empezado a estar más y más claro que la física parece jugar un papel fundamental en la comprensión de algunos aspectos clave del lenguaje. Y no solo la física, sino también las matemáticas. Resulta fascinante que campos tan dispares como la física, que estudia el universo, y la lingüística, que estudia el lenguaje, al final compartan métodos y técnicas. Aun así, visto desde una perspectiva moderna, esto tampoco nos debería sorprender: el lenguaje nos explica en parte cómo procesa el *cerebro* la información, mientras que la física nos intenta explicar cómo el *universo* procesa la información. La base es muy similar y, si decimos que el cerebro es muy eficiente, imagínense entonces ustedes el universo. No debe sorprendernos pues que encontremos numerosas analogías y sinergias entre ambas disciplinas. E incluso descubrimientos paralelos pero con nombres distintos, como explicaremos a continuación.

Este artículo pretende ser un resumen relativamente poco técnico (en términos matemáticos) de dos grandes sinergias entre física y lenguaje. La primera es la analogía entre “renormalización” en física y “merge” en lingüística [1]. Viene a decir que el mecanismo por el que emergen las leyes físicas en sistemas complejos es similar, si no el mismo, al mecanismo por el que emergen las oraciones a partir de entidades más fundamentales como palabras, letras y fonemas. La segunda analogía tiene que ver con el uso del álgebra lineal para describir cadenas en el lenguaje, un enfoque que constituye las bases de una teoría llamada “Matrix Syntax” [2] que, sorprendentemente, recuerda en muchos aspectos a la física cuántica. Dirigimos al lector ávido de detalles técnicos a las dos referencias ya mencionadas. Finalmente, proponemos un experimento sencillo, que se puede llevar a cabo en un aula, y que trata de poner en práctica algunos de los aspectos referidos a las sinergias mencionadas arriba. En particular, proponemos calcular el bloque matemático fundamental del lenguaje humano, e identificar posibles aspectos universales.

Antes de continuar, un comentario: la ciencia, y en particular su belleza, se tiene que poder explicar sin ecuaciones, del mismo modo que la música se puede apreciar sin pentagramas, y que no hace falta saber pintar para admirar a Dalí. *Así que nuestro objetivo en lo que sigue, es intentar transmitir conceptos abstractos de manera intuitiva. En realidad, todo es sencillo.*

2. Lenguaje y renormalización

El concepto de “renormalización” es, históricamente, uno de los más importantes en física moderna. Conceptualmente, la renormalización tiene su origen en la idea básica de que para describir un sistema con muchísimas variables, no necesariamente todas estas variables son relevantes. Phil Anderson, premio Nóbel de física del 1977, lo resumió bien en un famoso artículo publicado en el 1972, titulado “More is different” [3]. En él, Anderson argumentaba que las leyes que gobiernan los sistemas compuestos no son meramente una “concatenación” de las leyes de sus constituyentes fundamentales, sino que nuevas leyes *emergen* como consecuencia de las interacciones a medida que el sistema se vuelve más complejo.

La renormalización es, pues, el arte de descifrar los grados de libertad relevantes para describir un sistema (físico) a una escala determinada. Pongamos por ejemplo un gas. Este está constituido por moléculas, y cada molécula individual se rige por las leyes de la mecánica cuántica. Conocemos su “estado cuántico”, y de ahí podemos predecir con mayor o menor probabilidad cómo será su comportamiento. El sistema se vuelve más complejo al considerar dos moléculas, las cuales interactúan entre ellas: ahí las ecuaciones ya no son tan sencillas. Pongamos que tenemos tres moléculas: el sistema es, probablemente, caótico e impredecible. Pues bien... una cantidad macroscópica de gas, como un litro del aire que respiramos, está constituido por aproximadamente 10^{23} moléculas, es decir, 1 seguido de 23 ceros. Imagínense las interacciones entre ellas. Y ahí la pregunta: para describir el gas, ¿necesitamos conocer con exactitud la solución de las ecuaciones de la mecánica cuántica para las 10^{23} moléculas? Afortunadamente la respuesta es *no*. Como bien sabemos, un gas se rige principalmente por variables termodinámicas, típicamente la Presión (P), el Volumen (V) y la Temperatura (T). Y obviamente, el comportamiento colectivo del gas, como tal, es muy distinto del que tendrían las moléculas que lo forman si estas existiesen independientemente las unas de las otras. Este es un ejemplo de *emergencia*: las leyes físicas que rigen el macro-mundo *emergen* de las que rigen el micro-mundo al añadir complejidad. A más complejidad, mayor número de fenómenos colectivos que nada tienen que ver con los constituyentes fundamentales.

El marco teórico apropiado para abordar el problema anterior es la renormalización. A pesar de que históricamente siempre se tiende a proceder a la inversa, es decir, a deducir las leyes del micro-mundo a partir de lo que vemos en el macro-mundo, ello no quita que un test de consistencia fundamental sea el intentar reproducir lo que vemos cada día a partir de las leyes de los bloques fundamentales como los átomos y moléculas. De hecho, es tal nuestra fe en la renormalización, que las nuevas teorías que se proponen en física de partículas y en gravedad cuántica deben ser *renormalizables*, es decir, deben permitir la descripción consistente de fenómenos emergentes a medida que consideramos sistemas más y más complejos. Por curioso que parezca, esto no es necesariamente así, y se ha convertido en un principio físico fundamental: la física nos tiene que permitir calcular.

En renormalización, la manera de proceder usualmente es mediante *cambios de escala* en la variable relevante (distancia, energía, tiempo... lo que sea). A cada cambio de escala (por ejemplo, a medida que hago un zoom-out del sistema), se consideran sistemas con menos detalles descritos por nuevas variables “promedio” o “renormalizadas”. Estas son las variables relevantes para describir el sistema a una escala concreta, y para encontrarlas es necesario *descartar la información que no es relevante* del sistema. Para un ejemplo práctico de lo que queremos decir, sin entrar en las tenebrosas pero bellas ecuaciones de la física teórica, observen la siguiente figura:

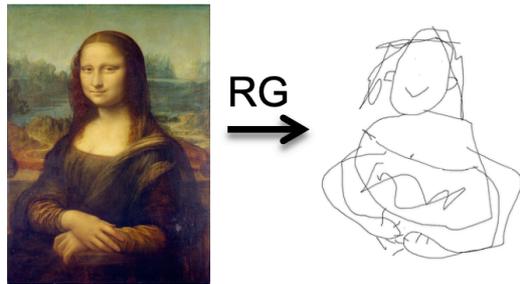


Fig.1: La Gioconda, renormalizada.

Por “RG” entendemos “Renormalization Group” (en Castellano “Grupo de Renormalización”), que el término técnico que los físicos utilizan para referirse a la renormalización. A la derecha está La Gioconda, con todo lujo de detalles. ¿Y a la derecha? Pues también, pero dejando muchos detalles de lado. Al pasar de la descripción de la izquierda a la de la derecha hemos descartado información que, por algún motivo, hemos considerado irrelevante. La descripción de la derecha nos ofrece una Gioconda con sus rasgos más fundamentales: una mujer, con los brazos cruzados, aparentemente sonriente, y punto. Para esta descripción, el nivel de detalle (es decir, de información) de la izquierda nos es irrelevante, y nos basta con el de la derecha. Algo similar ocurre en las teorías físicas, pero donde sustituimos cuadros de Leonardo por, pongamos, lagrangianos del modelo estándar de la física de partículas (objetos matemáticos a partir de los cuales se derivan las posibles interacciones fundamentales entre partículas). Los esbozos de la derecha se corresponderían con teorías emergentes (llamadas “teorías efectivas”) que permiten describir las moléculas, los gases, etc.

¿Y esto tiene que ver con el lenguaje? Pues sí, y gracias a Chomsky.

En 1993, Noam Chomsky propuso el *programa minimalista* [4]: un marco conceptual para las teorías del lenguaje basado en las ideas de perfección y economía del lenguaje. Chomsky argumenta que, de algún modo que aún no comprendemos por completo, el lenguaje ha de ser óptimo, asequible/simple, y eficiente. En definitiva, ha de ser un sistema óptimo o cuasi-óptimo para hacer lo-que-sea que se proponga. Este punto de vista diverge mucho de una aproximación filosófica al lenguaje (“yo creo que...”), y empieza a formalizar la lingüística como una ciencia en sí (“yo demuestro que...”). Si en física el objeto del estudio científico es el universo, en lingüística es el lenguaje.

Dentro del programa minimalista, Chomsky propone la operación MERGE como una de los bloques fundamentales de construcción del lenguaje. MERGE toma dos unidades lingüísticas, pongamos “perro” y “ladrador”, y las junta en

una, “perro ladrador”. En este caso, MERGE toma el núcleo “perro”, el sintagma adjetival “ladrador”, y lo transforma en una unidad diferente a un nivel superior, el sintagma nominal “perro ladrador”, que se comporta como sintagma nominal. Gráficamente, el MERGE entre unidades “alfa” y “beta” se puede representar así:

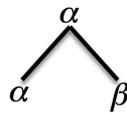


Fig.2: representación diagramática de MERGE.

Muchos lingüistas están acostumbrados a otra notación usando paréntesis, por ejemplo [NP [N perro] [A ladrador]]. Pero de cara a la analogía con la física, la notación diagramática nos es mucho más conveniente.

MERGE es una de las operaciones básicas del lenguaje. Mediante concatenación de distintos MERGE, podemos pasar de unidades fundamentales (palabras, letras, fonemas...) a unidades de orden superior (oraciones), las cuales *emergen*. Esto, básicamente, produce los bien conocidos árboles sintácticos. Por ejemplo:

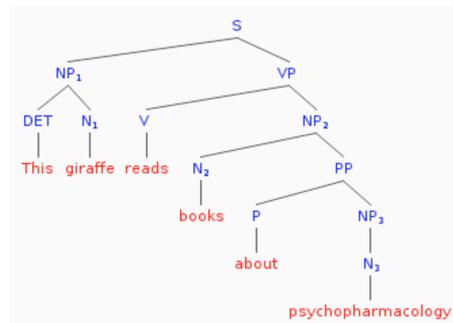


Fig.3: ejemplo de árbol sintáctico.

Y aquí es donde surge la analogía con la física. En lingüística, tenemos secuencias de palabras concatenadas mediante MERGE. La secuencia es temporal, y los sintagmas consecutivos así como la oración final emergen a escalas de tiempo largas. En definitiva, al árbol sintáctico de arriba un físico le añadiría estos dos ejes de coordenadas:

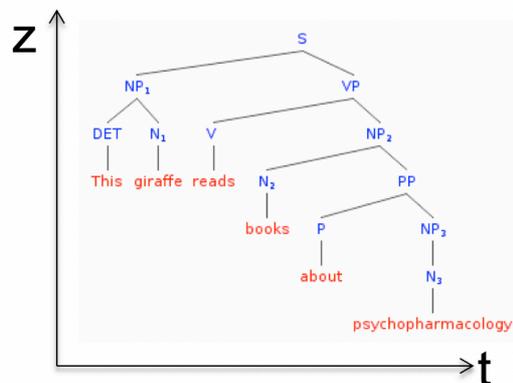


Fig.4: el mismo árbol sintáctico que en la Fig.3, pero con ejes físicos.

El eje horizontal “ t ” lo podemos entender como *tiempo*, ya que el lenguaje es (esencialmente) una secuencia de palabras que pronunciamos o pensamos en secuencia temporal. El eje vertical “ Z ” es la *escala* de tiempo. A mayor Z , mayor el “tiempo natural” relevante para los efectos que consideramos. “ Z ” es, básicamente, la separación temporal a la que consideramos posibles efectos dentro de la oración.

Esta descripción tan sencilla es *exactamente renormalización*, tal y como la entendemos en física. Para ser más precisos, MERGE es lo que en física llamamos “coarse-graining”: una operación mediante la cual unimos dos entidades fundamentales y producimos una nueva, que es más sencilla, y que está descrita por variables efectivas obtenidas a partir de descartar información irrelevante. En el caso del lenguaje, MERGE es el coarse-graining de unidades de información lingüística de acuerdo con diferentes escalas temporales. Los sintagmas y las oraciones son unidades lingüísticas complejas que *emergen* a escalas Z de tiempo grandes. El siguiente es un ejemplo más específico, donde la información lingüística se organiza en cuatro escalas de tiempo Z_1, \dots, Z_4 :

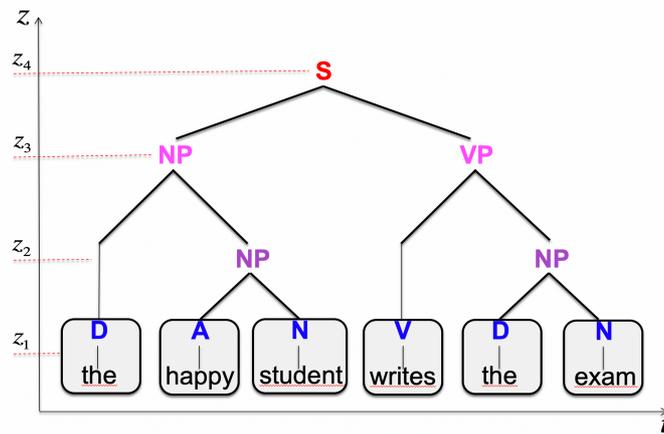


Fig.5: árbol sintáctico con ejes.

Esta aproximación al lenguaje ha resultado ser bastante práctica por diversos motivos. Pero el más importante es que ha permitido la aplicación de técnicas matemáticas y físicas al estudio del lenguaje. Por ejemplo, los llamados *modelos probabilísticos de lenguaje*, que son distribuciones de probabilidad de oraciones (como el que utiliza por ejemplo el algoritmo de predicción de Google), tienen una estructura interna en términos de las denominadas “redes de tensores” [5]. Usando técnicas matemáticas relacionadas con estas, se ha podido demostrar que el lenguaje es, en promedio, *crítico en el sentido físico*, es decir, que tiene correlaciones entre sus constituyentes que, en promedio, decaen muy lentamente (decaimiento polinómico y no exponencial). Este hecho remarcable concuerda a la perfección con las observaciones empíricas [6] y establece que el lenguaje es en cierto modo análogo a un sistema físico crítico unidimensional: crítico por su estructura de correlaciones y unidimensional porque dichas correlaciones se estructuran únicamente a lo largo de la variable temporal “ t ”. Entendemos pues que toda la artillería matemática de sistemas críticos en física, desarrollada durante los últimos cien años, tiene también aplicaciones en el estudio de las

propiedades del lenguaje. Asimismo, esta sinergia ha permitido también entender por qué determinados tipos de redes neuronales (como las “Deep Convolutional Networks” – redes convolucionales -) parecen ser particularmente buenas a la hora de procesar el lenguaje: da la casualidad de que este tipo de red neuronal organiza la información de acuerdo con una estructura de renormalización. Finalmente, es tentador pensar que esta es ciertamente la manera como el cerebro humano tiende a organizar la información. Y más aún si nos damos cuenta de que otros procesos mentales típicamente humanos se pueden entender también como renormalización: no solo el lenguaje, si no también la capacidad matemática, la música [7], el reconocimiento de imágenes (recuerden el ejemplo de La Gioconda)... nuestra hipótesis es que todas estas destrezas/capacidades humanas podrían ser diferentes manifestaciones de una capacidad fundamental del cerebro humano: *la capacidad de organizar y procesar información de acuerdo a distintas escalas físicas*. Tal capacidad podría ser debida a un único evento evolutivo específico. Y tal vez, al final del día, sea esto lo que más nos diferencia del resto de animales. Es decir: mi cerebro renormaliza muy bien la información, pero el de mi periquito pues no tanto.

3. Lenguaje, álgebra y física cuántica

Las sinergias entre física y lenguaje van más allá de lo expuesto arriba. En particular, recientemente propusimos una nueva teoría para algunos aspectos del lenguaje llamada “Matrix Syntax” (MS) [2]. MS utiliza muchos aspectos del álgebra lineal, la teoría de grupos, y la teoría de operadores para describir determinadas situaciones en sintaxis. El modelo resultante es similar en varios aspectos a la física cuántica, y funciona bien en la descripción de las correlaciones de largo alcance y las co-ocurrencias en el lenguaje.

MS tiene dos motivaciones. Una desde el punto de vista de la física, y otra desde el punto de vista del lenguaje. La motivación física es sencilla: el formalismo resultante es muy similar a algunos aspectos formales de la mecánica cuántica, lo cual es muy sorprendente. Solo por ello, MS merece ya ser considerada y estudiada con detalle. Desde el punto de vista del lenguaje, la motivación de MS es el deseo de entender las llamadas "cadenas", tal y como suceden en la oración 1(a):

- (a) Alicia parece saber mucha física.
- (b) Alicia sabe mucha física.

De algún modo, "Alicia" en 1(a) es el "sabedor" de físicos, exactamente igual que en 1(b). Al mismo tiempo, "Alicia" en 1(a) es también quien realiza la acción de “parecer” (“Alicia parece...”). Formalmente, en 1(a) podemos considerar que "Alicia" está en dos posiciones diferentes desde un punto de vista lógico: a la izquierda de "parece", y a la izquierda de "saber". Sin embargo, solo se externaliza en una de estas dos alternativas. Esta es la llamada "cadena". ¿Cómo sucede este fenómeno, y qué significa decir desde un punto de vista estructural e interpretativo? Estamos ante un problema abierto.

La intuición de por qué el álgebra lineal nos puede ayudar a entender las cadenas, proviene de la física cuántica. En las cadenas, como en 1(a), un elemento de la oración ("Alicia" en 1(a)) parece estar, formalmente, *en dos sitios a la vez a nivel mental, pese a que solo se externaliza en uno*. Aunque esto suene muy extraño, en mecánica cuántica este tipo de situaciones suceden a constantemente: una partícula subatómica puede estar en un estado cuántico que se corresponde con dos posiciones diferentes *a la vez*, siempre y cuando nadie la observe. Al observar (es decir, al *medir* su posición), el estado de la partícula *colapsa* y solo la observamos en una posición concreta. Pero su trayectoria y otras propiedades físicas son, hasta el mismo momento de la medida, compatibles con la superposición de dos posiciones simultáneas, y es así como debe tratarse matemáticamente.

No es nuestro deseo aquí adentrarnos en los innumerables detalles técnicos de MS, ya que nos harían perder el hilo. Lo que sí nos interesa recalcar es que, en MS, el tratamiento que se les da a los distintos elementos sintácticos (sintagmas, oraciones...) es el de *vectores matemáticos*. Para que nos entendamos, un vector viene a ser una flechita que apunta a alguna parte, como por ejemplo esta:

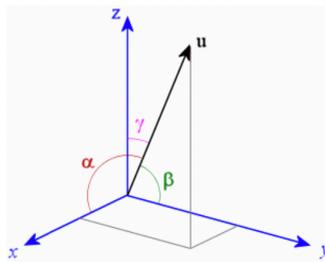


Fig.6: un vector en tres dimensiones

Los vectores son objetos matemáticos que, al sumarlos, producen nuevos vectores. Este es el tipo de matemática necesario para describir superposiciones en física cuántica, y es el mismo que usa MS para describir cadenas con elementos “en dos sitios a la vez”: estar en los dos sitios “a la vez” en la oración, corresponde a un vector que es “suma de los vectores de dos sitios”. Sin entrar en tecnicismos, el vector “u” de la figura de arriba podría corresponder a una oración, que en sí misma se forma por composición (“producto tensorial”) con vectores más sencillos correspondientes a sintagmas. Esta concatenación no es independiente de lo que decíamos en el punto anterior sobre la renormalización: debe obedecer sus reglas, y así lo hace MS [2]. En términos físicos, uno diría que MS es una teoría renormalizable del lenguaje, lo que hace todo el andamio que exponemos aquí consistente desde un punto de vista lógico.

La sorpresa con MS surge, sin embargo, al tirar de la manta. La pregunta fundamental aquí es: y una vez tenemos esto, ¿qué más podemos derivar además de las cadenas? Pues resulta que mucho, muchísimo. De hecho, muchísimo más que la mera solución al problema de las cadenas, que era nuestra motivación original. Las cadenas han sido *la excusa* para encontrar una construcción matemática del lenguaje que, no solo es tremendamente bella en sí misma, sino que además tiene un poder predictivo enorme. Aspectos tales como la *concordancia*, la relación de los complementos, las construcciones pasivas... y mucho más, surgen a partir de consideraciones puramente matemáticas. Son

teoremas. Por si esto fuera poco, la analogía natural que surge (¡sin buscarla!) con la física cuántica es, cuanto menos, intrigante, y merece ser considerada. ¿Estamos ante los resquicios lingüísticos de algún proceso cuántico del cerebro? No podemos más que hacer conjeturas en este aspecto, pero desde luego el tema da para pensar y debatir.

4. Proyecto: el bloque fundamental del lenguaje

Todo lo dicho arriba da para mucho debate y para mucha investigación. No hemos hecho más que ver la punta del iceberg, y serán otras mentes más jóvenes, brillantes y abiertas que las nuestras las que recojan el testigo y descifren la pregunta básica que aquí nos concierne: ¿qué narices es el lenguaje? No es moco de pavo, pero no nos asusta.

Los educadores tienen un papel clave en que lo anterior se pueda cristalizar. En opinión del que aquí escribe, la distinción tradicional entre “ciencias y letras” ha hecho muchísimo daño, y aquí hemos visto un ejemplo muy claro. ¿Cómo vamos a entender el lenguaje sin saber de matemáticas y de física? No tiene sentido que disciplinas que deberían ir de la mano ni se hablen, ni se interesen entre ellas. A veces también es útil asistir a las charlas “apócrifas”, ya que solo de la interdisciplinariedad surgen las ideas más fructíferas y disruptivas. Si no, corremos el riesgo de que la mayor parte de los avances sean solo incrementales. En definitiva: las ideas locas a veces hay que tomárselas en serio.

Es por ello que aquí proponemos un pequeño proyecto que se podría llevar a cabo en un aula de secundaria, por un grupo de estudiantes familiarizados con el análisis sintáctico y la probabilidad. Este consiste en calcular el bloque fundamental del lenguaje: la probabilidad de hacer MERGE de alfa y beta, y obtener gamma. La idea es calcular este objeto matemático en general para diferentes entornos, realizar una comparación, y obtener una serie de conclusiones sobre la universalidad del lenguaje. Más específicamente:

- Hacer grupos de 3-4 estudiantes
- Considerar aproximadamente 100 oraciones por grupo para un entorno determinado. Por ejemplo:
 - Personas con Español / Catalán / Inglés como lengua materna.
 - Perfiles influyentes de Twitter.
 - Niños de 2 años.
 - Gente con autismo, dislexia...
 - ...
- Hacer todos los árboles sintácticos de todas las oraciones, y calcular los porcentajes (%) de “alfa y beta Merge en gamma” (como por ejemplo, Sintagma Nominal y Sintagma Adjettival Merge en Sintagma Nominal, etc).
- Una vez se tengan todos los porcentajes, comparar los resultados para los distintos entornos.
- A partir de la comparación deberían poderse identificar rasgos similares independientes del entorno, lo que implicaría propiedades universales del lenguaje. *Estas propiedades universales están pues ancladas en*

nuestro cerebro, y no dependen de agentes externos. No pretendemos entender de donde vienen, sino identificarlas. Determinadas frecuencias, relaciones entre frecuencias...

- Discutir los resultados en grupo: analogía con la física, modelos de lenguaje, posible origen de tales propiedades universales...

Por supuesto, esta propuesta es muy extensa y puede ser acotada. Por ejemplo, nos podríamos restringir a un núcleo en concreto (como por ejemplo, un nombre), y calcular la probabilidad de que se seleccione cada tipo de sintagma como complemento. Aunque parezca mentira, hasta donde sabemos no se ha llevado a cabo nunca tal estudio, pese a ser muy sencillo y definitivamente relevante. Así que hay incluso la posibilidad de... ¡obtener un nuevo resultado científico en el estudio del lenguaje! Anímense, les aseguro que engancha.

- [1] Ángel Gállego, Román Orús, arXiv:1708.01525
- [2] Román Orús, Roger Martin, Juan Uriagereka, arXiv:1710.00372
- [3] Phil W. Anderson, Science, 177, 4047 (1972).
- [4] Noam Chomsky, MIT occasional papers in linguistics no. 1. Cambridge, Massachusetts (1993).
- [5] Román Orús, Nature Reviews Physics 1, 538-550 (2019).
- [6] E. Alvarez-Lacalle et al, PNAS 103(21):7956-7961 (2006); H. W. Lin, M. Tegmark, Entropy, 19, 299 (2017).
- [7] J. Katz, D. Pesetsky, <http://ling.auf.net/lingBuzz/000959>