Peer review (Part 1)

1. What is your general research question?

"¿Predice la ley de Zipf la frecuencia de palabras en contextos textuales diferentes?"

Queremos determinar si la ley de Zipf, la cual establece la frecuencia de aparición de las palabras, se puede comprobar con un análisis estadístico (con el lenguaje de programación R). El objeto del análisis será un conjunto de obras literarias escritas en inglés dirigidas a públicos diferentes (para tal propósito, utilizaremos obras de literatura juvenil y de literatura para adultos).

2. Why do you think this question is interesting? What does an answer to it tell us?

Gracias al análisis que realizaremos, podremos comprobar si la Ley de Zipf predice la frecuencia de aparición de las palabras en el idioma inglés. Realizar este análisis resultaría interesante, pues si se comprueba que la ley predice, en efecto, la frecuencia, ello nos permitiría generar vocabularios más útiles para el aprendizaje de lenguas extranjeras.

Asimismo, si se comprueba que la ley es cierta, ello podría conllevar consecuencias más profundas en el estudio del lenguaje, como por ejemplo, la existencia de una tendencia en las lenguas humanas para ahorrar contenido verbal no necesario; así pues, a la hora de descifrar mensajes (lingüísticos) encriptados, tal ley deberá ser considerada. Así pues, resultaría una herramienta interesante en el campo de la criptografía.

3. What is your specific research question?

¿Se cumple la ley de Abreviación de Zipf de la misma manera en textos de literatura juvenil inglesa que en los textos de literatura para adultos?

Se realizará un análisis estadístico de las palabras más usadas en novelas para un público juvenil y en novelas para un público adulto. Después, se procederá a identificar si los patrones de frecuencia se corresponden o no.

4. What kind of data would you use to address Question 2 if you had unlimited resources?

Antes que nada, se realizaría un estudio para determinar qué es lo que constituye una novela juvenil o para adultos. Una vez obtenida la definición de ambas categorías, se utilizaría un muy amplio corpus de novelas originales para un público juvenil y adulto. Después, se procedería al análisis estadístico con el lenguaje R.

5. What kind of data are you planning to use to address Question 2 within the scope of this class?

Transcripciones en PDF de novelas para adultos y juveniles y el lenguaje de programación R.

5.1. How will you obtain it?

Se procedería a buscar novelas en formato PDF, txt, rtf, odt, doc-docx, wps, html, etc. Por lo que respecta al lenguaje de programación R, utilizaremos la herramienta Google Collab proporcionada en el Aula Global de la asignatura Mètodes Empírics del Llenguatge 2, pues en ella se puede usar dicho lenguaje.

5.2. How much will you collect?

Entre 20 y 100 obras.

5.3. Do you think that is enough data to address Question 2? Why (not)?

Sí. Creemos que, aunque usemos el mínimo requerido, 20 obras, tal número será suficiente para realizar una estimación prudente sobre si la ley de Zipf se cumple o no. Esto es porque cada novela es una compilación larga de una lengua, lo que asegura que haya un sesgo mínimo en la fiabilidad de los datos. No obstante, en pos de asegurar, si cabe, mayor fiabilidad en los datos, creemos que debe existir una cierta heterogeneidad (dentro de los límites establecidos, esto es, novelas para adultos y juveniles) en las obras escogidas. Así pues, cada obra se escogerá de la forma susodicha.