1. Zipf's Law of Abbreviation

Zipf's Law of Abbreviation (also known as Zipf's Law of Brevity) is a statistical law in linguistics derived from the original Zipf's Law. While the classic law describes word frequency distributions, Zipf's Law of Abbreviation rather explains that the most frequent words are inclined to be shorter (Zipf, 1949). In fact, this has been empirically demonstrated for many languages, as Bentz and Ferrer-i-Cancho (2016) did for 986 languages. Likewise, this project aims to test this hypothesis by analyzing two brief song lyrics for Spanish and English.

2. Materials and Methods

In order to test Zipf's Law of Abbreviation, I used two songs in English and Spanish. As for the English text, I selected *It's too late* by Carole King (1971), whereas for Spanish I selected *Gasolina* by Daddy Yankee (2004). Both songs were obtained from internet webpages, which are available in the references. The lyrics of these songs are also available in the supplementary material section.

It is relevant to state that the texts were modified to avoid language interference. For instance, all the English words in the Spanish text were translated into their Spanish equivalent ('*todos los* weekenes *ella sale a vacilar*' was modified to '*todos los fines de semana ella sale a vacilar*'). Furthermore, all oral elisions such as *la' turbina'* or *stayed in bed all mornin'* were standardized to *las turbinas* or *stayed in bed all morning* to make tokenization and data processing easier. Lastly, all chorus and repeated verses were eliminated to avoid biases in repeated words, which would not be easily heard in casual speech. The adapted texts can be found in the GitHub repository. These two songs were chosen, since they belong to different genres, contexts, languages, and social registers. Thus, if Zipf's Law of Abbreviation is true, it should equally account for all texts.

Both texts were tokenized with the *spaCy* Python library, as it is simpler to use and standard practice in NLP, compared to other packages such as *regex*. The model *es_core_news_sm* was used for Spanish, and *en_core_web_sm* was used for English. From the resulting tokens, only those corresponding to words were retained, since Zipf's Law of Abbreviation concerns the relationship between word frequency and word length. To do so, *spaCy*'s default tokenizer was used. Then, all words were converted into lower cases by using *token.text.lower()*. Finally, all tokens were only selected if they were entirely composed of letters with *token.is_alpha*. This automatically rejected numbers, punctuation marks or alphanumeric symbols.

As for data processing, all words were added into a dictionary with their frequency in the text. Then, their length was measured with a second dictionary by adding a tuple as its value. Finally, Spearman Correlation was extracted from these measures with the *SciPy* library. Spearman Correlation was used because it allows us to quantify the degree of relatedness of two values by obtaining a rho value (ρ) ranking from -1 to 1. A result closer to -1 suggests that there is a strong negative correlation between two values, in this case, frequency and length.

3. Results and Discussion

The results of this study are summarized in Table 1. As we may observe, rho value is negative, suggesting that there is a negative correlation between frequency and length (the shorter a word is, the more likely it will be to appear, and vice-versa). P-value was lower than 0.01, which indicates that the probability that these results were due to the chance under null hypothesis are minimal.

It is important to note that, although values are negative, they are not closer to -1. This is due to the fact that Zipf's Law of Abbreviation is not systematic, but it is rather a statistical tendency. Furthermore, this is consistent with our findings, since the English text *It's too late* contains less token than the Spanish text *Gasolina*. Therefore, the model did not have that much data to process, leading to weaker estimates.

	Spanish text	**English text**
Spearman Rho Value (ρ)	-0.51	-0.36
P-Value	$p < 0.01$	$p < 0.01$

Table 1. Rho and p-values for Spanish and English texts.

To complement the quantitative analysis, a qualitative inspection was conducted. To do so, I looked at the raw data for frequencies and length in four words in both English (Table 2) and Spanish (Table 3). As it can be inferred from the tables below, shorter words (such as *and*, *I*, *no*, and *las*) have a higher frequency compared to longer words (such as *darling*, *breezing*, *janguea*, or *adrenalina*) for both texts. It is worth mentioning that Spanish and English present morphological differences, with Spanish words being often longer due to these morphological dissimilarities. Even so, these specific cases also support Zipf's Law of Abbreviation, since we may observe how bigger words occur less frequently compared to shorter words.

Word	**Frequency**	**Length**
And	8	3
I	6	1
Darling	1	7
Breezy	1	6

Table 2. Case study of four words in English, indicating Frequency and Length.

Word	**Frequency**	**Length**
No	7	2
Las	6	3
Janguea	1	7
Adrenalina	1	10

Table 3. Case study of four words in Spanish, indicating Frequency and Length.

4. References

Ayala, R. L., & Ávila, E. (2004). Gasolina [Song]. On *Barrio fino*. El Cartel Records; VI Music. Retrieved from https://videoele.com/CancionEle/gasolina-daddy-yankee.html.

Bentz, Christian & Ferrer-i-Cancho, Ramon (2016). Zipf's law of abbreviation as a language universal. In Bentz, Christian, Jager, Gerhard & Yanovich, Igor (eds.) Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. University of Tubingen, online publication system, https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.

Stern, T., & King, C. (1971). It's too late [Song]. On *Tapestry*. Ode. Retrieved from https://videoele.com/CancionEle/gasolina-daddy-yankee.html.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

5. Ressources used

- Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python (Version 3.7) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.1212303.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). *SciPy 1.0: Fundamental algorithms for scientific computing in Python* (Version 1.13.0) [Computer software]. https://doi.org/10.1038/s41592-019-0686-2
- Responses generated by ChatGPT (OpenAI, 2023) and Gemini (Google AI, 2024) were used to support methodological explanations and for coding assistance. I acknowledge that I am ultimately responsible for all code in the notebook submitted for the assignment.

6. Appendix

a. *Gasolina* by Daddy Yankee

Zúmbale mambo para que mis gatas prendan los motores,
que se preparen que lo que viene es para que le den ¡duro!

Mamita, yo sé que tú no te me vas a quitar, ¡Duro!
lo que me gusta es que tú te dejas llevar, ¡Duro!
todos los fines de semana ella sale a vacilar, ¡Duro!
mi gata no para de janguear, porque…

A ella le gusta la gasolina,
Dame más gasolina
cómo le encanta la gasolina.
Dame más gasolina

Ella prende las turbinas, no discrimina,
no se pierde ni una fiesta de marquesina,
se acicala hasta para la esquina,
luce tan bien que hasta la sombra le combina.
Asesina, me domina,
janguea en carros, motoras y limusinas,
llena su tanque de adrenalina,
cuando escucha reguetón en la cocina.

Aquí nosotros somos los mejores, no te me ajores,
en la pista nos llaman Los Matadores,
tú haces que cualquiera se enamore,
cuando bailas al ritmo de los tambores.

Esto va para las gatas de todos colores,
para las mayores, para las menores,

para las que son más zorras que los cazadores,
para las mujeres que no apagan sus motores.

Tenemos tú y yo algo pendiente,
tú me debes algo y lo sabes,
conmigo ella se pierde,
no le rinde cuentas a nadie.

b. *It's too late* by Carole King

Stayed in bed all morning just to pass the time
There's something wrong here, there can be no denying
One of us is changing, or maybe we just stopped trying

And it's too late, baby, now it's too late
Though we really did try to make it
Something inside has died
And I can't hide, and I just can't fake it
Oh, no, no, no

It used to be so easy living here with you
You were light and breezy, and I knew just what to do
Now you look so unhappy and I feel like a fool

There'll be good times again for me and you
But we just can't stay together
Don't you feel it too?
Still, I'm glad for what we had
And how I once loved you

Too late

Baby, it's too late

Now, darling, it's too late