Frequency and efficiency in Spanish fixed expressions

Ernesto R. Gutiérrez Topete
UC Berkeley

Zipf's law has long been held as a universal property in human language (Zipf, 1936, 1949). This law states that there is a negative correlation between a word's frequency and its length. In other words: "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences" (1936). Zipf claimed that this tendency stems from a need of communicative efficiency, reducing the effort in the production of more common words. This observation is the basis for his Principle of Least Effort—"the primary principle that governs our entire individual and collective behavior of all sorts, including the behavior of our language" (Zipf, 1949). Considerable empirical support for Zipf's law has been found concerning word frequency, data with variable sequence length, and neural data, all subsequently the focus of complex computational modeling (Aitchison, Corradi, & Latham, 2016; Piantadosi, Tily, & Gibson, 2011; Piantadosi, 2014; Lestrade, 2017).

Although Zipf's law has been vastly researched in natural language production, it has yet to be explored in linguistic structures beyond the word, such as fixed expressions. This type of research would further test the extent to which this law accounts for natural language, considering that fixed expressions have traditionally been regarded as representing a single unit in the mind of a speaker (Erker & Guy, 2012). This idea is notably encompassed in Sinclair's *idiom principle*, which asserts that "a language user has available to him or her a large number of semi-pre-constructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (Sinclair, 1991, p. 110). Thus, in the present study, I analyze the relationship between the length of a fixed expression and its frequency by testing whether or not highly frequent fixed expressions are more likely to be shortened than their low frequency counterparts.

To answer this question, I analyzed data from the Corpus del Español NOW Corpus, which included, at the time of analysis, over 7.2 billion naturally produced words taken from Spanish news outlets (Davies, 2016). Following Akbarian (2012), Simpson and Speake (1998), and Murar (2009), I examined proverbs (e.g. *ojos que no ven, corazón que no siente*). These fixed expressions refer to phrases or sayings that sum up situations and provide advice or a moral (Gramley & Pätzold, 1992). Only proverbs that include two clauses were considered for this study, as these syntactic compositions best facilitated the operationalization of fixed expression length; shortening was considered the omission of one of the clauses. For example, *más vale pájaro en mano* was treated as a shortening of *más vale pájaro en mano que cientos volando*.

A total of 30 fixed expressions—stratified by frequency (high, low, via median split)—was collected and analyzed, with each occurrence labeled as either *short* or *long*. The percent *shortened* was calculated and entered as a continuous dependent variable. The results of a linear regression applied to these data suggest that there is a difference in shortening rate between high and low frequency fixed expressions. High frequency is strongly correlated with shortening, notably supporting Zipf's law in this new domain: fixed expressions. Moreover, the results also provide support for Sinclair's idiom principle, showing that fixed expressions behave similarly to single words in regard to efficiency. In addition to cross–linguistic comparisons across Romance languages and others, further research should evaluate if fixed expressions are also sensitive to other factors linked to shortening such as informativeness, as Piantadosi et al. and Mahowald et al. have shown with single words, to see how context and frequency interact and influence the reduction of fixed expressions (2011; 2013).

References

Aitchison, L., Corradi, N., & Latham, P. E. (2016). Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Computational Biology*, *12*(12), e1005110.

Akbarian, I. (2012). What counts as a proverb? The case of NTC's dictionary of proverbs and clichés. *Lexikos*, *22*, 1–19.

Davies, M. (2016). The new 2.9 billion word NOW Corpus: Up-to-date as of . . . yesterday. The 20th Workshop on Linguistics and Language Processing. Kyung-Hee University.

Erker, D. & Guy, G. R. (2012). The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language*, 526–557.

Gramley, S. & Pätzold, K.-M. (1992). *A survey of modern English*. Routledge.

Lestrade, S. (2017). Unzipping Zipf's law. *PLoS One*, *12*(8), e0181987.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Murar, I. (2009). Pragmatic and functional uses of idioms. *Analele Universităţii din Craiova. Seria Ştiinţe Filologice. Lingvistică XXXI*, (1-2), 146–156.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Simpson, J. & Speake, J. (1998). *The concise Oxford Dictionary of Proverbs*. Oxford University Press.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Zipf, G. K. (1936). *The psychobiology of language*. London: Routledge.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. New York, NY: Addison-Wesley Press.