Chapter 15

Frequency and efficiency in Spanish proverbs

Ernesto R. Gutiérrez Topete[a]

[a]University of California, Berkeley

Zipf's law states that there is an inverse relationship between a word's length and its frequency; the more frequent words tend to be the shortest. Following this premise, I investigate if this universal property in language can extend to domains beyond the word. As such, the present study analyzes the use of proverbs, a specific type of fixed expressions, in a Spanish corpus of news language. The motivation for this study is to determine if more frequent proverbs are more likely to be shortened, relative to their lower-frequency counterparts. The results of this study indicate that there is a positive correlation between a proverb's general frequency in a corpus and its reduction rate. This paper argues that usage-based models of language representations are better equipped to account for the mechanisms used in the production of proverbs, compared to the traditional view of fixed expressions as enlarged single words.

1 Introduction

For interlocutors, communication is a trade-off between the speaker's production effort and the amount of information that the listener receives. That is, if the speaker provides less information to minimize the amount of effort required to articulate the signal, the listener will be faced with a bigger burden in trying to sort through the ambiguity. In trying to strike a balance, languages tend to reduce the articulation of more frequent words at the segment and syllable levels. For example, in Spanish we find that *grillo* → [ˈgri.o]; *después* → [de.ˈpue͡s]; and in English we find that *memory* → [ˈmɛm.ɹi]; and *information* → [ˈɪn.foʊ] (Lipski 1990, Brown 2008, Bybee 2002b, Mahowald et al. 2013). Such a phenomenon is

Ernesto R. Gutiérrez Topete. 2023. Frequency and efficiency in Spanish proverbs. In Barbara E. Bullock, Cinzia Russi & Almeida Jacqueline Toribio (eds.), *A half century of Romance linguistics: Selected proceedings of the 50th Linguistic Symposium on Romance Languages*, 323–345. Berlin: Language Science Press. DOI: 10.5281/zenodo.7525120

encompassed in Zipf's law, which states that there is a negative correlation between a word's frequency and its length in any given corpus of written language or naturally produced speech (Zipf 1936, 1949).

Usage-based models of language representation and use – for instance, as described by Bybee (2001) and Pierrehumbert (2001) – argue that there is a reciprocal relationship between the mental representation of language (i.e. the grammar) and its actual use. In other words, "usage feeds into the creation of grammar just as much as grammar determines the shape of usage" (Bybee 2006: 730). Furthermore, they postulate that language users store in exemplar clouds information related to the frequency and context in which the linguistic item was used, with more recent and/or frequent items being activated and retrieved more easily. Accordingly, research on frequency and context of use of linguistic structures finds that more frequent items are treated differently: more frequent words are learned sooner, recognized more quickly, articulated more easily, and perceived as more grammatical. These effects are seen in linguistic areas such as language change (Phillips 1984, 1999, Bybee 2000, 2001, 2002c), psycholinguistics (Vitevitch & Luce 1998, Vitevitch 2002), language variation (Ellis 2002a,b), language acquisition (Bybee 2002a, Erker 2011, File-Muriel & Brown 2010), and more.

When studying usage-based effects, linguists have primarily focused on the role of the speaker, paying particular attention to the phonetic/phonological level (i.e. sounds and clusters), as Bybee (2002b) has outlined. Other domains such as the intersection of frequency and fixed expressions have not been as rigorously studied. Fixed expressions are of particular interest because of their composition and representation. While fixed expressions have traditionally been regarded as representing a single unit in the mind of a speaker (see Sinclair 1991), the hypothesized exemplar clouds in the usage-based models "may in principle allow speakers to deduce what words are likely to cooccur, with what probability, formulating probabilistic generalizations well beyond the level of the fixed phrase …" (Erker & Guy 2012: 528). Thus, the stored information about language context, use, and co-occurrence of lexical items in collocations such as fixed expressions provides us with linguistic environments where Zipf's law can be tested – beyond the cluster, syllable, and lexical levels.

In this study, I analyze the effect that occurrence frequency has in a particular type of fixed expressions: the proverb. Proverbs are linguistic constructions that express, in a phrase or complete sentence, a specific idea – a metaphor that contains "a good dose of common sense, experience, wisdom, and above all truth" (Mieder 1989: 15). I focus the present study on the use of proverbs in a corpus of the language used in the news media. Admittedly, the language in the news industry may be considered less 'natural language' because of the formulaic crafting behind it. Nonetheless, it was selected for this study due to the fact that proverbs are very frequent in this genre owing to their economy: "[proverbs] let journalists rapidly explain something new via something known" (McLaughlin, personal communication, June 11, 2020). More precisely, this study investigates if more frequent fixed expressions such as proverbs are more likely to undergo articulatory reduction (e.g. *People who live in glass houses should not throw stones.* → *People who live in glass houses.*), as is the case with other linguistic domains (e.g. clusters, syllables, and lexical items). I hypothesize that more common phrases will experience higher rates of truncation in order to maximize efficiency in production. These results will help us further understand the extent to which Zipf's law applies to fixed expressions in (written) language, a novel linguistic domain.

The remainder of this paper is structured as follows: §2 explains frequency effects and its applicability to proverbs through the lens of usage-based models and Zipf's law. This section ends with a discussion of the use and prevalence of proverbs in the news media. §3 describes the present study, and §4 presents the results. Finally, this paper ends with a discussion of the findings of this study and a conclusion in §5 and §6, respectively.

2 Frequency, proverbs, and the news media

2.1 Zipf's law

Zipf's law (Zipf 1936, 1949), though not unchallenged (see for example Bentz & Ferrer-i-Cancho 2016), has long been held as a universal property in human language. In fact, it has been denominated "the most well known statement of quantitative linguistics" (Montemurro 2001: 1). This law states that there is a negative correlation between a word's frequency and its length in any given corpus of written language or naturally produced speech. In his own words: "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences" (1936: 25). The inverse relationship between a word's length and its relative frequency is robust and has been corroborated cross–linguistically by the countless studies that have examined this phenomenon (e.g. Bates et al. 2003, Ferrer-i-Cancho & Hernández-Fernández 2013, Piantadosi et al. 2011, Strauss et al. 2007, Wimmer et al. 1994, Zipf 1949). Furthermore, this pattern has also been found in other natural and/or non-human systems (e.g. data with variable sequence length, music, computer systems, etc., Aitchison et al. 2016).

The principle of abbreviation in human language, Zipf claimed, stems from a need of communicative efficiency, reducing the effort in the production of the more common words in language. 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: "Preface," paragraph 22). Ferrer-i-Cancho & Hernández-Fernández explain that the burden of communication falls on both the speaker and the hearer, which expands to many levels of the communication process. At the phonological level, the speaker wants to minimize the amount of articulated language, whereas the hearer wants to receive as much information in the signal as possible in order to decode the message (Pinker & Bloom 1990, Köhler 1986). At the lexical level, the hearer wants the most informative word (Köhler 1986, Zipf 1949). Yet, according to Köhler (1986), as cited in Ferrer-i-Cancho & Solé (2003), speakers tend to resort to the most common words to minimize communicative effort when speaking. Subsequently, the most frequent word acquire more uses (i.e., meanings), rendering these words more ambiguous than their less frequent counterparts (Ferrer-i-Cancho & Hernández-Fernández 2013: 788). What is easiest for the speaker yields a more complicated situation for the hearer and vice versa, thus prompting the need for a compromise between the two. Zipf refers to this efficiency trade-off as the principle of least effort.

2.2 Fixed expressions

Collocations or multi-word expressions are peculiar linguistic structures. On the one hand, they are composed of multiple words, creating whole phrases or even sentences. On the other hand, these chunks, as a whole, display similar frequency effects as lexical items; for instance, children can utter higher frequency multi-word phrases faster and "better" (Bannard & Matthews 2008), and adults are able to process these collocations more quickly (Arnon & Snider 2010). These phrases behave similarly to individual lexical items, and for that reason, they have traditionally been regarded as representing a single unit in the mind of a speaker. As Erker & Guy (2012) describe, traditional linguistics referred to fixed phrases or idioms such as *to kick the bucket* as enlarged lexical items, basically working as long words. 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: 110).

Gramley & Pätzold (1992) took Sinclair's principle further and created a systematic classification of fixed expressions. In their work, the authors categorized collocations, idioms, pragmatic idioms, and proverbs as separate sub-categories of fixed expressions. The authors attributed to the proverb: (a) the capability of expressing a speech act such as a promise, warning, request, etc.; (b) the nature of constituting a complete sentence; and (c) the characteristic of idiomaticity, where its actual meaning cannot be deduced by the denotation of its individual words (48–9). Nonetheless, Gramley & Pätzold (1992) note that proverbs are not necessarily 'fixed.' They elaborate that "proverb collections often list a number of variant forms, which shows that variability is a characteristic trait of proverbs" (1992: 60). In fact, proverbs are made up of complete sentences but are not always produced as such: "They are so well known that even fragments and mutations are easily associated with the full form," unlike most idioms, "which would become meaningless if changed in this way or allow only a literal reading" (1992: 60–61).

This leads to a conceptual conflict: a proverb has been considered *a single, enlarged word*, but unlike other types of fixed expressions, it can also have several variants – including fragmentations of the full form – while maintaining its meaning and still displaying similar lexical effects as individual lexical items. Now, the question that arises is: how can we conceptualize the mechanisms that allow for similar frequency effects of individual words and proverbs? In other words, if proverbs are treated as single words, should frequency effects of individual words be ignored or discarded? Alternatively, should lexical effects of individual words contained in the proverb be considered or prioritized in analyzing the frequency effects observed in this type of fixed expressions? As discussed below, the postulates regarding contextual and collocational information put forth by usage-based models allow for a representation of proverbs in a more uniform way.

2.3 Frequency in collocations

Usage-based and exemplar models (e.g. Johnson 1997) postulate that speakers and listeners store vast information about the linguistic items (e.g. sounds and words) they encounter in exemplar clouds (i.e. memory clusters). These exemplar clouds, then, help generalize these items into categories that can be used as production targets during speech articulation. As Erker & Guy (2012) explain, "[these] models further postulate that collocational information is retained in and deducible from memories of linguistic experience" (2012: 258). That is, exemplar clouds of sounds and words may be extended beyond this level, providing higher activation levels for words that co-occur frequently, where the articulation of a portion of a frequent collocation will evoke the remainder of this phrase from the listener's exemplar cloud. Additionally, more than one item may be activated by a previously activated/uttered item, giving the speaker more than one option to choose from, which may yield multiple variants for a particular phrase.

In their study of the use of subject pronoun + verb collocations in Spanish, Erker & Guy (2012) found that pronouns were omitted at a higher rate when they co-occurred with morphologically irregular and external activity verbs that had a higher general frequency. The opposite effect occurred with morphologically regular and mental activity verbs. From these results, the authors conclude that: (a) frequency does not affect language independently from other factors or constraints, and (b) frequency effects cannot be described simply as unidirectional.

2.4 Proverbs in news language

As described earlier, news language is not the quintessence of naturally produced language, but it provides us with an open field to explore the usage of proverbs. The language used in the news media industry has been characterized as a crafted language that employs certain devices or tools in a stylized manner (Mouriquand 1997, Conboy 2013). Among these tools we find anecdotes and proverbs, which are used with a similar purpose in mind: "[l]'objectif [des anecdotes] est à la fois de raconter une très courte histoire qui rende la lecture agréable et de donner à comprendre ce qui, conceptualisé, serait trop complexe" ('The purpose [of anecdotes] is both to tell a very short story that makes reading enjoyable, and to provide an understanding of what, conceptually, would be too complex') (Mouriquand 1997: 97). Similarly, "[l]e recours au proverbe, à la formule populaire présente l'intérêt de faire référence à un sens tout à fait clair dans l'esprit du lecteur et ainsi de faciliter la compréhension dans un passage difficile" ('The popular use of the proverb offers the consideration of referring to a meaning completely clear in the reader's mind, thereby facilitating comprehension of a difficult passage') (1997: 102).

The study of news language has traditionally focused on or, at the very least, started with the analysis of written newspapers. However, Conboy (2013) reports that news delivered in other types of media is often influenced by the style used in written newspapers. As such, all journalistic language, regardless of mode of delivery, may be attributed to the same genre, including spoken news on TV. Thus, the careful formulation of news language renders it non-natural, distinct from spontaneous speech. However, it is on a par with other forms of written language, such as books and magazines, two sources of written language to which Zipf's law also applies. Therefore, we can and should expect that this genre will render a form of language that resembles written language and that, as indicated above, contains proverb data that *is ripe for the picking*.

To my knowledge, proverbs in the Spanish news media have not been broadly studied from a linguistics perspective. However, it is well known that some of the most typical tools used in this industry, including the ones described by Mouriquand (1997), are widely used in this genre cross-linguistically. Thus, the present study works under the assumption that the characteristics about the news media described here will also apply to the news media industries in the Spanish-speaking world.

3 The present study

Although linguistic reduction has been well documented in many (non)linguistic domains and countless forms of communication, certain domains such as fixed expressions – and more specifically, proverbs – remain largely unexplored. The present study investigates whether or not more frequent proverbs are more likely to undergo linguistic reduction. Adhering to usage-based models and exemplar theory, I work under the premise that instances of linguistic items to which a speaker-listener is exposed are stored in the mind, in conjunction with all their relevant use and context information. As such, I assume that more frequent proverbs will benefit from the frequency effects observed in other domains such as lexical items; that is, they will be more easily recalled and/or evoked. Said differently, I hypothesize that proverbs will be governed by Zipf's law, demonstrating higher rates of shortening for more frequent items. Note that I will use the terms "reduction", "shortening", and "truncation" interchangeably in this paper.

3.1 Materials

The data used for this study came from the News on the Web (NOW) corpus, from the Corpus del Español, which at the time of data collection included over 7.2 billion words in Spanish (Davies 2016). The corpus is composed of web-based newspapers, magazines, and their respective comments sections, if available, thus rendering more formal (i.e. written news content) and less formal (i.e. naturally-produced user-comments) sources of language. The data was collected from 2012–present, and the corpus covers the press in all Spanish-speaking countries, including the United States (US). The news media was selected because of its high use of proverbs, as mentioned above, and because its parsimonious nature provides an environment where linguistic reduction is commonplace, compared to other corpora of written language. Additionally, the NOW corpus was selected because it provides a large corpus of data that spanned across Latin America, Spain, and the US.

A list of 30 target proverbs was selected from Pedicone de Parellada (2004), a collection of sayings commonly used in the press in Tucumán, Argentina (see Appendix A). I followed the definition of proverbs provided in Gramley & Pätzold (1992), where proverbs are described as expressions that (a) sum up to situations and provide advice or a moral, (b) are metaphorical, and (c) may have variable forms. The proverbs were stratified by frequency. Each proverb was labeled as either high frequency or low frequency, using a median split approach, based on their general frequency in the NOW corpus. The 30 expressions analyzed in this study are those from Pedicone de Parellada (2004), which had a frequency of at least 1 – in their long form – in the NOW corpus. It was important for the target proverbs to exist in their long form so that the process of shortening could be definitively demonstrated. Pedicone de Parellada's book was used to compile the materials of interest because it provided a list of proverbs that have been attested as commonplace in the contemporary Spanish-speaking media,[1] which aligned with the timeframe for the data collected in the NOW corpus. The present study, therefore, provides a synchronic observation of the behavior of proverbs in the news media.

3.2 Procedures

As the NOW corpus only allows for an interaction with the data through its interface, the data collection and analysis were conducted manually. In addition to the aforementioned criterion of one or more long-form occurrences in the corpus, the proverbs needed to be composed of two or more syntactic phrases (e.g. NP, VP, PP, etc.), as this facilitated the operationalization of proverb length, where shortening was considered present if at least one whole syntactic phrase had been omitted. Each occurrence of a proverb in the corpus was documented as an observation, and for each observation, the dependent variable was recorded as a decision between the binary SHORTENED or NOT SHORTENED outcome, according to the form in which the token appeared.

[1] Although this collection of proverbs comes from the Argentinian press, all proverbs are present in other varieties of Spanish; in fact, being a speaker of Mexican Spanish, I have a high level of familiarity with most of these proverbs. For that reason, these proverbs collected from the Argentinian press were analyzed in the NOW corpus, a multi-dialectal Spanish corpus.

Deletion, *not substitution*, was required for shortening to be counted as present. An example of reduction is a sentence with a PP and an NP which has the second phrase truncated, such as [*a buen entendedor* $_{\text{PP}}$][~~*pocas palabras*~~ $_{\text{NP}}$] ⟶[*a buen entendedor* $_{\text{PP}}$] 'a word to the wise'. Tokens in the dataset that had a syntactic phrase substitution or variants that differed from the standard form presented in Pedicone de Parellada (2004) were counted as long-form tokens (i.e. not shortened). For instance, [*a buen entendedor* $_{\text{PP}}$] [*sobran explicaciones* $_{\text{VP}}$] was considered a substitution, and thus not shortened. Moreover, *agua que no has de beber, déjala pasar* was considered a variant of *agua que no has de beber, déjala correr* 'don't be the dog in the manger', and thus not shortened. To ensure that all possible variants of a given proverb were included in the analysis, all its content words were searched in the corpus, and all data that had at least one complete syntactic phrase belonging to the proverb were collected.

3.3 Analysis

A total of 2,997 proverb occurrences were recorded in the corpus. Repeated tokens from articles that were published in multiple outlets (i.e. re-published articles) and tokens that made references to creative works (e.g. song or film titles) were excluded from the analysis. An analysis of frequency type showed that high frequency tokens had an average general frequency of 187.3 occurrences (SD=123.4) and a range of 54–503 occurrences, while lower frequency proverbs had a mean of 12.5 (SD=17.5), with a range of 1–51 occurrences. A *t*-test reported that the low and high frequency groups were significantly different from one another in regard to corpus occurrence ($df = 28$, $p < 0.001$). Furthermore, general corpus frequency and shortening frequency/rate per token type were calculated; the latter measurement refers to the rate of shortened over total number of occurrences for each proverb. A fixed-effects logistic regression model analyzed the dependent variable (i.e. whether an observation was shortened or not) as a function of general frequency in R (R Core Team 2018) using the `glm()` function. No other independent variables were analyzed in the statistical model.

4 Results

Across the board, proverbs in their long form were more common than shortened tokens: 2,388 to 609. Furthermore, the ratio for shortening ranged from 0% to 44.4%, indicating that, for any given proverb, there were more tokens it is long form that there were shortened tokens. Figure 1 shows the rate of shortened tokens per proverb for each of the 30 proverbs analyzed in this study over their general frequency in the NOW corpus.

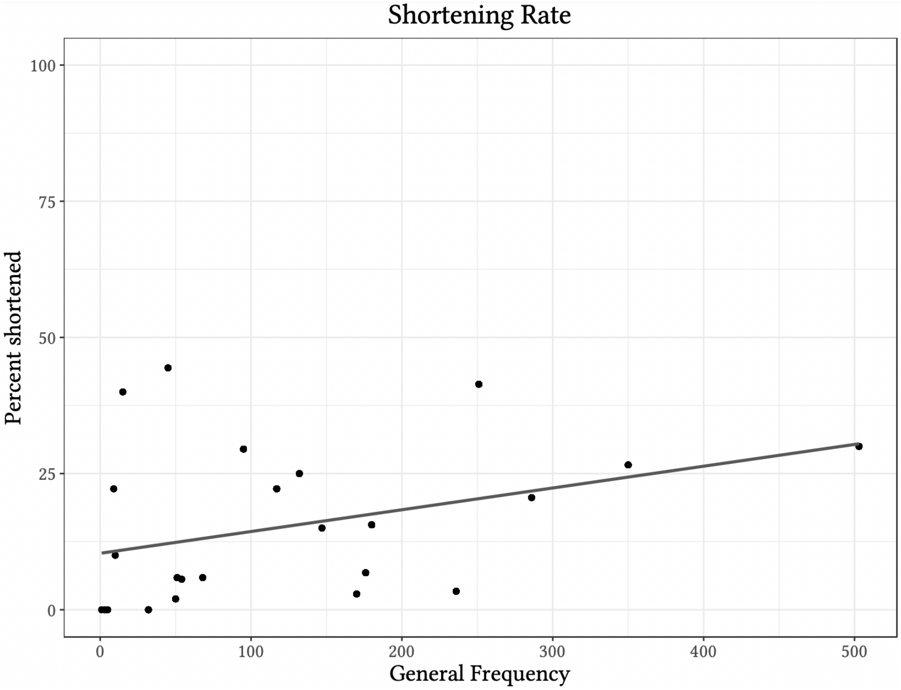

Figure 1: Rate of shortened tokens per proverb, given its general frequency in the NOW corpus. This graph includes the regression line.

The results of the logistic regression model, shown in Table 1, indicate that general frequency is a strong predictor of shortening rate ($p < 0.001$). The conversion of the model's log-odds coefficients into decimal values reveals the model's prediction that, at frequency 1, proverbs will display a reduction rate of 11.9%. Furthermore, as frequency increases by one unit, the probability of reduction is estimated to increase by 0.025%. That is, the model predicts that for a proverb that has a general frequency of 503 occurrences – the highest general frequency for the proverbs in this dataset – the shortening rate will increase to approximately 31.2%. These predictions are visualized in Figure 2, which displays the probability of reduction given proverb frequency.

Table 1: Logistic regression model of proverb shortening

Coefficients	Estimate	Std. Error	*z* value	*p* value	
(Intercept)	−1.9990626	0.0977211	−20.457	<2e-16	***
Gen-freq	0.0024009	0.0003107	7.728	1.09e-14	***

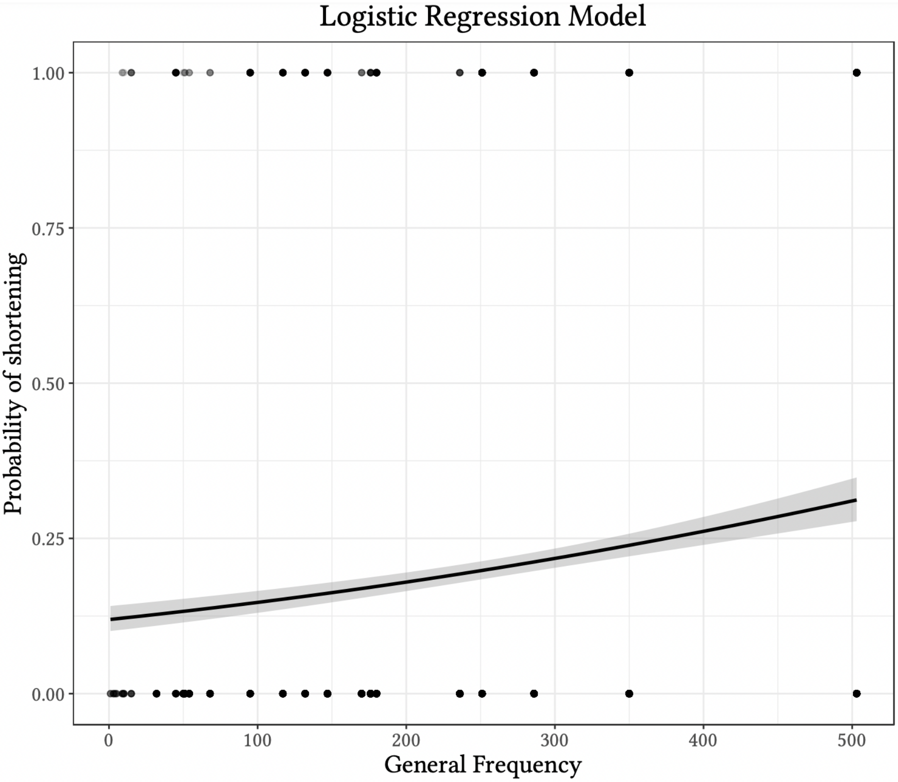

Figure 2: Probability of reduction over frequency increase for all 30 proverbs.

In sum, these results indicate that shortening shows a strong effect of frequency, adhering to Zipf's law. However, there were two particular outliers:

(1) Como la gata flora, cuando le ponen grita,
Like the cat flora when to-her something-is-given she-screams
cuando le sacan llora.
when from-her something-is-taken-away she-cries.
'There's no pleasing [someone]'

(2) Agua que no has de deber, déjala correr.
Water that no you-must to drink let-it flow
'Don't be the dog in the manger'

The former had a general frequency of 15 occurrences with a 40% reduction rate, whereas the latter had 45 occurrences and a reduction of 44.4%. While these proverbs were labeled as low frequency, they were in the top 3 most truncated proverbs. These outliers, which had an even higher shortening rate than the one reported for the most frequent proverbs, signal that frequency alone is not sufficient to account for this phenomenon. The first potential explanation for this peculiarity makes reference to syntactic length or complexity. At first glance, nonetheless, this explanation fails to hold water. The two outliers, repeated below with delineated syntactic structures, have comparable reduction rates but starkly distinct syntactic complexity.

(3) [Como la gata flora $_{\text{PP}}$], [cuando le ponen $_{\text{AdvP}}$]
[Like the cat flora $_{\text{PP}}$], [when to-her something-is-given $_{\text{AdvP}}$]
[grita $_{\text{VP}}$] [cuando le sacan $_{\text{AdvP}}$]
[she-screams $_{\text{VP}}$] [when from-her something-is-taken-away $_{\text{AdvP}}$]
[*llora* $_{\text{VP}}$]
[she-cries $_{\text{VP}}$]
'There's no pleasing [someone]'

(4) [Agua que no has de beber $_{\text{NP}}$], [déjala correr $_{\text{VP}}$]
[Water that no you-must to drink $_{\text{NP}}$] [let-it flow $_{\text{VP}}$]
'Don't be the dog in the manger'

The proverb in (3), evidently, is composed of a higher number of syntactic phrases relative to (4). Yet, it is also true that both proverbs contain complex structures (e.g. [*como la gata flora* $_{\text{PP}}$] for the former and [*agua que no has de beber* $_{\text{NP}}$] for the latter). It is possible that their more complex structures – even if only applicable to some of its phrases – may incentivize language users to truncate this phrase in order to reduce the production effort, be it speaking or writing. This would clearly fit within the expectations put forth by Zipf's law. As such, a post-hoc analysis that included (a) the total number of syntactic phrases and (b) the number of complex syntactic phrases per proverb was performed to identify those factors as potential indicators of reduction rate as a way to account for these two outliers. A statistical model that included the number of syntactic phrases and the number of complex syntactic phrases as non-interacting independent variables, shown in Table 2, found that these factors were not significant: ($p = 0.82$ and $p = 0.08$, respectively). The inclusion of these independent variables into the model does not change the effect observed for general frequency ($p < 0.001$). The next section provides a discussion of these results and elaborates on the factors considered in this study as potential predictors of proverb shortening rates.

Table 2: Logistic regression model of syntactic composition

Coefficients	Estimate	Std. Error	*z* value	*p* value	
(Intercept)	-2.6532178	0.7884364	-3.365	0.000765	***
Gen-freq	0.0032566	0.0003747	8.692	<2e-16	***
Syntactic phrases	0.1049060	0.3765724	0.279	0.780566	
Complex phrases	-0.6183682	0.3563933	-1.735	0.082728	

5 Discussion

To summarize the results in this study, proverbs appear to display higher rates of reduction when their general frequency increases in the NOW corpus. Therefore, proverbs seem to be governed by similar frequency effects that have been found in lexical items and other linguistic domains. This link suggests that, as proverbs become more frequent and thus easier to recall, speakers are more likely to shorten the production of these phrases. All in all, these results are a robust corroboration of Zipf's law, indicating that more frequent proverbs are more likely to be reduced.

The post-hoc statistical analysis presented above also reports that the number and complexity of syntactic phrases in these proverbs do not predict truncation. These results are corroborated by a further inspection of the rest of the data.

For instance, consider (5) and (6). The proverb in (5) had a general frequency of 286 occurrences with a shortening rate of 20.6%, while the proverb in (6) had 5 tokens with 0 shortened variants. Considering that these two sentences have a remarkably similar syntactic structure, if syntactic complexity was correlated with shortening rate, it would be expected that they display comparable reduction rates, but this is not the case. The pattern seen between (3) and (6) suggests that syntactic complexity does not favor a particular outcome when it comes to linguistic reduction, or at least not forthrightly.

(5) [Dime $_{\text{VP}}$] [con quién andas $_{\text{PP}}$] [y $_{\text{Conj}}$]
[Tell-me $_{\text{VP}}$] [with whom you-walk $_{\text{PP}}$] [and $_{\text{Conj}}$]
[te diré $_{\text{VP}}$] [quién eres $_{\text{PP}}$].
[to-you I-will-tell $_{\text{VP}}$] [who you-are $_{\text{PP}}$].
‘You will be judged by the company you keep’

(6) [Dime $_{\text{VP}}$] [cuanto tienes $_{\text{AdvP}}$] [y $_{\text{Conj}}$]
[Tell-me $_{\text{VP}}$] [how-much you-have $_{\text{AdvP}}$] [and $_{\text{Conj}}$]
[te diré $_{\text{VP}}$] [cuánto vales $_{\text{AdvP}}$].
[to-you I-will-tell $_{\text{VP}}$] [how-much you-are-worth $_{\text{AdvP}}$].
‘What you own determines your worth’

Considering that syntactic complexity does not appear to be directly correlated with linguistic shortening of proverbs, an additional analysis of variability seemed justified. Accordingly, I qualitatively analyzed the numbers of variants that were present for each of the thirty tokens in the dataset, starting with (3)–(6). However, from the onset, no clear pattern was detected. For instance, (3), a low frequency proverb with a high shortening rate, displayed a high degree of variability. The 15 occurrences found in the corpus were made up of a total of 8 variants (see some of these variants under (7)) other than the standard form shown in (Pedicone de Parellada 2004). Conversely, (5), a proverb with 286 tokens, only had 12 recorded variants.

(7) a. Como la gata flora: si se la ponen gritan y si se
Like the cat flora: if to-her it they-give they-scream and if to-her
la sacan lloran.
it they-take-away they-cry
There’s no pleasing [someone]

b. Como la gata flora, que si se la metes chilla y si
Like the cat flora: that if to-her it you-put-in she-cries and if
se la sacas llora.
to-her it you-put-out she-cries
There's no pleasing [someone]

c. Como la gata flora si te lo ponen gritas y si te
Like the cat flora: if to-you it they-give you-scream and if to-you
lo sacan lloras.
it they-take-away you-cry
There's no pleasing [someone]

While the rate of variability may appear to explain the outlier status of (3) and the compliance of (5), it fails to account for (6), which only had 5 tokens and 0% shortening; this example was composed of a total of three variants – more than 50% divergence from the standard form. Conversely, (4), with 45 occurrences and nearly 45% shortening, has only two variants present in the dataset. The shortening rates of (3) and (6), two low frequency proverbs, are in contrast with each other.

All in all, this study demonstrates that there is an efficiency trend in proverb production. The fact that this trend materializes at all in language use beyond the lexical level – i.e. in fixed expressions – signals that: (a) Zipf's law can be applied to linguistic items beyond the lexical item; (b) speakers and listeners can negotiate a balance between production effort and informativeness trade-off in proverbs, avoiding ambiguity; and (c) the language from the news media, albeit "not natural", may be used to study natural linguistic phenomena and general properties of human language, provided that this genre allow for the expression of these phenomena. Additionally, the present study also sheds lights on the traditional view of fixed expressions as enlarged lexical items, suggesting that usage-based models are better able to account for the frequency effects observed here.

First, these results are indicative of proverbs' adherence to Zipf's law in written news language. Zipf's law originated from an observation that the most common words in language also tend to be shorter. In fact, the more common the word, the shorter it will be. The general concept is applicable to proverbs; the more common proverbs were not shorter, *per se*, but they were shortened at a higher rate. This is suggestive of a general human tendency to operate efficiently in a way that extends beyond the phonological and lexical domains. Furthermore, the most common words tend to be so because they have the most uses, as indicated earlier. A similar parallel may be drawn between words and proverbs in this respect. The most common proverbs may be applicable to more conversational/social contexts; the more contexts it applies to (i.e. the more uses it has), the greater the need for efficiency.

The discussion of efficiency mentioned above leads us to the second observation made from the results of the present study: speakers are able to negotiate form length and meaning of proverbs while avoiding ambiguity. An advantage of proverbs is the metaphorical meaning that they have in addition to the literal one. This is an advantage that other fixed expressions such as idioms do not have. For example, "cut Barbara some slack" is not the same as "cut Barbara" or "Barbara some slack," where the former will have an idiomatic meaning, while the latter two lose their intended meaning; one only has a literal meaning, whereas the other becomes meaningless. In addition to maintaining their meaning, proverbs have a second advantage: intrinsic variability. Regardless of the number of variants or their rate if truncation, proverbs evoke a clear concept in the mind of the speaker-listener. As such, if one of multiple possible variants is produced, the listener (or reader) will arrive at the same concept or meaning being evoked. This last point should not be equated to *the number of variants will predict shortening*, because the results from this study indicate otherwise. Rather, the argument that I am making here is that the variable nature of proverbs provides flexibility in their expression by signifying to the speaker-listener that they are susceptible to being expressed with alternative forms – other than the standard, long form. The result of such an advantage is that speakers (or writers) are able to employ efficient ways of production (i.e. reduced forms), knowing that conceptual meaning remains intact for these proverbs.

Third, these results demonstrated that the language used in the news adheres to Zipf's law; put differently, more frequent proverbs in the news display higher reduction rates, in a way that adheres to Zipf's law. Although the news language genre is perceived as non-natural, it nonetheless follows efficiency conventions seen in other genres of written language as well as natural speech. Considering the parsimonious nature of language use in this industry, the tendency toward linguistic efficiency is not a surprise for the news media, especially if such a behavior also allows for a clear conveyance of specialized topics to the general population.

On the whole, the discussion of these results provides us with a better understanding of the mechanisms used for the shortening of proverbs. As described earlier, usage-based models of language representation argue that language use will impact the mental representation of language (i.e. the grammar). These models provide a better way to capture the frequency effects observed in this study, when compared to the more traditional depiction of proverbs. For instance, the variable nature of proverbs itself contradicts the notion that proverbs can be observed as "enlarged lexical items." While words may have their standard long form and a potential shortened version (e.g. *hippopotamus* → *hippo*), proverbs may have multiple long-form variants, in addition to multiple shortened versions (refer back to (7) for potential variants of the proverb '[*Como la gata flora*] [*cuando le ponen*] [*grita*][*cuando le sacan*] [*llora*]'). The inherent variability of proverbs makes them distinct from lexical items. Furthermore, the question of frequency becomes complicated when considering proverbs as a single word. For instance, it is not utterly clear if variants of a single proverb should be counted as separate words. If so, the next issue that arises is which variant to attribute a reduced token to.

Conversely, usage-based models consider these fixed expressions as strings of words linked by probabilistic generalizations. Exemplars of proverbs will be stored in the mind, each word linked to the next by probabilistic co-occurrence due to the contextual information stored in conjunction with each lexical item. As such, the activation and retrieval of one word in the string will activate the following word. The fact that one word may activate more than one lexical item accounts for the multi-variant proverbs, with higher frequency proverbs having an activation advantage compared to their lower-frequency counterparts. Lastly, the exposure to a shortened form of a proverb will thus activate the remainder of the phrase. This way, hearing or reading a shortened version of a proverb will still allow the listener/reader to retrieve the original conceptual meaning associated with the long-form proverb. Accordingly, usage-based models are able to account for the empirical observations surrounding proverb shortening trends in written news language.

6 Conclusion

All things considered, there is a clear relationship between general frequency and truncation rates in the production of proverbs in language in the news media. However, the presence of outliers indicates that general frequency is not the only factor that constrains this relationship. Syntactic complexity and variability as potential factors constraining truncation rates do not appear to be able to account for the shortening rates in these data. It is evident that, as reported in Erker & Guy (2012), frequency and shortening in collocations are connected but not independently of other constraints, and some of the constraints employed by some factors, such as variability, may sometimes manifest in more than one direction.

In sum, this study demonstrates that there is a clear correlation between the general frequency of a proverb in a given corpus and its reduction rate. These results corroborate Zipf's law in yet another linguistic domain: proverbs. Remarkably, this phenomenon seems to be very productive in proverbs, considering that altering the proverb's form allows for shortened proverbs to convey their original intended message. Lastly, the characteristics of proverbs discussed by the literature and the observations mentioned in the present study allow us to examine the effectiveness of usage-based models in accounting for the findings, compared to the traditional view of fixed expressions as single words.

While the corpus used in this study was large, the number of proverbs analyzed was small. As such, it became a challenge to identify other possible constraints that my be acting upon the proverbs. Future studies with more exhaustive proverb lists may be able to find stronger connections between the data and factors that may be influencing their production, in addition to general frequency.

Acknowledgements

First and foremost, I want to acknowledge Dr. Justin Davidson and Dr. Terry Regier for their help in developing and refining this research project. Also, I am grateful to Annie Helms for her initial help in conceptualizing this research idea and her extensive feedback on the study and this manuscript. Furthermore, I want to thank Madeline Bossi for her help with the syntactic analysis of these data. Finally, I want to show appreciation to the two anonymous reviewers and all the people who provided comments and feedback on previous presentations of this study to the Romance Linguistics community at UC Berkeley and to the LSRL 50 audience.

Appendix A List of proverbs

1. El que se va a Sevilla, pierde su silla.
   'Move your feet, lose your seat'
2. Entre bomberos no se pisan la manguera.
   'There is honor among thieves'
3. Más vale pájaro en mano que cientos volando.
   'A bird in the hand is worth two in the bush'
4. A buen entendedor, pocas palabras.
   'A word to the wise is enough'

5. En el país de los ciegos, el tuerto es rey.
   'In the land of the blind, the one-eyed man is king'
6. A rey muerto, rey puesto.
   'The king is dead; long live the king'
7. Más vale ser cabeza de ratón que cola de león.
   'Better to be a big fish in a small pond than a small fish in a big pond'
8. Dime cu''anto tienes y te diré cuánto vales.
   'What you own determines your worth'
9. Donde manda capitán, obedece marinero.
   'Where a captain rules, a sailor has no sway'
10. Aunque la mona se vista de seda, mona se queda.
    'You can't make a silk purse out of a sow's ear'
11. Casamiento y mortaja del cielo bajan.
    'Marriages are made in heaven'
12. Agua de mayo, pan para todo el año.
    '[something] couldn't have come at a better time'
13. El que con niño se acuesta amanece mojado.
    'If you lie down with dogs, you'll get up with fleas'
14. Como la gata flora, cuando le ponen grita, cuando le sacan llora.
    'There's no pleasing [someone]'
15. No por mucho madrugar amanece más temprano.
    'A watched pot never boils'
16. Al que madruga Dios lo ayuda.
    'The early bird catches the worm'
17. ¿A dónde va Vicente? A donde lo llame la gente.
    'Monkey see monkey do'
18. El que dice lo que no debe escucha lo que no quiere.
    'A word once spoken is past recalling'
19. Lo que mal comienza mal termina.
    'It was doomed from the start'
20. Quien mal anda mal acaba.
    'You get what you deserve'
21. Hombre prevenido vale por dos.
    'Forewarned is forearmed'
22. El que come y no convida tiene un sapo en la barriga.
    'He who eats and doesn't share, has got a tummy that's not fair'
23. Al pan pan y al vino vino.
    'Tell it like it is'
24. Agua que no has de beber déjala correr.
    'Don't be a dog in the manger'

25. El ojo del amo engorda al caballo.
    ‘It is the master’s eye that makes the mill go’
26. Pájaro que come vuela.
    ‘The early bird catches the worm’
27. Al mal tiempo buena cara.
    ‘Keep a stiff upper lip’
28. Sobre llovido mojado.
    ‘When it rains, it pours’
29. Dime con quién andas y te diré quién eres.
    ‘You will be judged by the company you keep’
30. El que no es conmigo contra mí es.
    ‘You are either with me or against me’

References

Aitchison, Laurence, Nicola Corradi & Peter E. Latham. 2016. Zipf’s law arises naturally when there are underlying, unobserved variables. *PLoS Computational Biology* 12(12). e1005110.

Arnon, Inbal & Neal Snider. 2010. More than words: Frequency effects for multiword phrases. *Journal of memory and language* 62(1). 67–82.

Bannard, Colin & Danielle Matthews. 2008. Stored word sequences in language learning: The effect of familiarity on children’s repetition of four-word combinations. *Psychological science* 19(3). 241–248.

Bates, Elizabeth, Simona D’Amico, Thomas Jacobsen, Anna Székely, Elena Andonova, Antonella Devescovi, Dan Herron, Ching Ching Lu, Thomas Pechmann, Csaba Pléh, Nicole Wicha, Kara Federmeier, Irini Gerdjikova, Gutierrez Gabriel, Daisy Hung, Hsu Jeanne, Gowri Iyer, Teodora Kohnert Katherine Mehotcheva, Araceli Orozco-Figueroa, Angela Tzeng & Ovid Tzeng. 2003. Timed picture naming in seven languages. *Psychonomic bulletin & review* 10(2). 344–380.

Bentz, Christian & Ramon Ferrer-i-Cancho. 2016. Zipf’s law of abbreviation as a language universal. In Christian Bentz, Gerhard Jäger & Igor Yanovich (eds.), *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, 1–4. Tübingen: Universität Tübingen. DOI: 10.15496/publikation-10057.

Brown, Earl Kjar. 2008. *A usage–based account of syllable- and word-final /s/ reduction in four dialects of Spanish*. Albuquerque, NM: The University of New Mexico. (Doctoral dissertation).

Bybee, Joan. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 65–85. Stanford, CA: Center for the Study of Language & Information.

Bybee, Joan. 2001. *Phonology and language use* (Cambridge Studies in Linguistics 94). Cambridge: Cambridge University Press.

Bybee, Joan. 2002a. Cognitive processes in grammaticalization. In Michael Tomasello (ed.), *The new psychology of language*, vol. 2, 151–174. New Jersey: Lawrence Erlbaum.

Bybee, Joan. 2002b. Phonological evidence for exemplar storage of multiword sequences. *Studies in second language acquisition* 24(2). 215–221.

Bybee, Joan. 2002c. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language variation and change* 14(3). 261–290.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82. 711–733.

Conboy, Martin. 2013. *The language of the news*. London: Routledge.

Davies, Mark. 2016. The new 2.9 billion word NOW Corpus: Up-to-date as of … yesterday. In Kyung-Hee University.

Ellis, Nick C. 2002a. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition* 24(2). 143–188.

Ellis, Nick C. 2002b. Reflections on frequency effects in language processing. *Studies in second language acquisition* 24(2). 297–339.

Erker, Daniel. 2011. *An acoustic sociolinguistic analysis of variable coda /s/ production in the Spanish of New York City*. New York: New York University. (Doctoral dissertation).

Erker, Daniel & Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88(3). 526–557.

Ferrer-i-Cancho, Ramon & Antoni Hernández-Fernández. 2013. The failure of the law of brevity in two new world primates: Statistical caveats. *Glottotheory International Journal of Theoretical Linguistics* 4(1). 45–55.

Ferrer-i-Cancho, Ramon & Ricard V Solé. 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences* 100(3). 788–791.

File-Muriel, Richard J. & Earl K. Brown. 2010. The gradient nature of s-lenition in Caleño Spanish. *University of Pennsylvania Working Papers in Linguistics* 16(2). 7.

Gramley, Stephan & Kurt-Michael Pätzold. 1992. *A survey of modern English*. London: Routledge.

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson & John W. Mullennix (eds.), *Talker variability in speech processing*, 145–165. San Diego, CA: Academic Press.

Köhler, Reinhard. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: N. Brockmeyer.

Lipski, John M. 1990. Elision of Spanish intervocalic /y/: Toward a theoretical account. *Hispania* 73(3). 797–804.

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

Mieder, Wolfgang. 1989. *American proverbs: A study of texts and contexts*. Frankfurt am Main: Herbert Lang.

Montemurro, Marcelo A. 2001. Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications* 300(3–4). 567–578.

Mouriquand, Jacques. 1997. *L'écriture journalistique*. Paris: Presses Universitaires de France.

Pedicone de Parellada, Elena Florencia. 2004. *El refranero hispánico: Pervivencia y circulación en la prensa gráfica, hoy*. San Miguel de Tucumán: Departamento de Publicaciones, Facultad de Filosofía y Letras, Universidad Nacional de Tucumán.

Phillips, Betty S. 1984. Word frequency and the actuation of sound change. *Language* 60(2). 320–342.

Phillips, Betty S. 1999. The mental lexicon: Evidence from lexical diffusion. *Brain and language* 68(1–2). 104–109.

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

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45), 137–157. Amsterdam: John Benjamins.

Pinker, Steven & Paul Bloom. 1990. Natural language and natural selection. *Behavioral and brain sciences* 13(4). 707–727.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna. https://www.R-project.org/.

Sinclair, John M. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Strauss, Udo, Peter Grzybek & Gabriel Altmann. 2007. Word length and word frequency. In Peter Grzybek (ed.), *Contributions to the science of text and language*, 277–294. Berlin: Springer.

Vitevitch, Michael S. 2002. Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear. *Language and speech* 45(4). 407–434.

Vitevitch, Michael S. & Paul A. Luce. 1998. When words compete: Levels of processing in perception of spoken words. *Psychological science* 9(4). 325–329.

Wimmer, Gejza, Reinhard Köhler, Rüdiger Grotjahn & Gabriel Altmann. 1994. Towards a theory of word length distribution. *Journal of Quantitative Linguistics* 1(1). 98–106.

Zipf, George Kingsley. 1936. *The psychobiology of language.* London: Routledge.

Zipf, George Kingsley. 1949. *Human behavior and the Principle of Least Effort.* New York: Addison-Wesley Press.