Using Zipf's Law to Aid Language Learning

Introduction

Mathematics and linguistics, at the surface, seem to be on different sides of the educational system, but I find both to be fascinating. I've always admired people who could speak multiple languages so easily and pick up new ones like it was nothing. Being in a foreign language class myself for 4 years, I still find many aspects of language learning to be difficult and frustrating, particularly the problem of pacing and the sometimes overwhelming amount of new vocabulary.

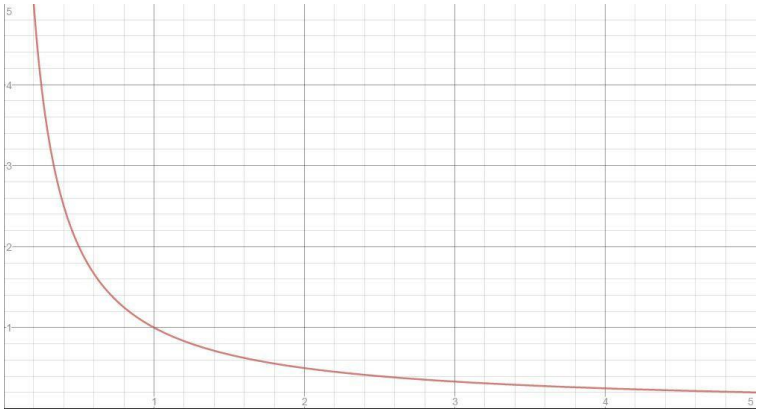
Since I plan to learn at least one more language in my lifetime, I wondered if there could be some sort of a mathematical pattern to language learning that could potentially make the process easier and lead to a greater comprehension of essential vocabulary. While watching a video, I was introduced to the concept of Zipf's law and its ability to model word frequency across languages. I thought that there may be some benefits in using this model and applying it to learning new languages. In this exploration I plan to use this law to analyze the word frequencies in an English language text both to show its benefits and limitations in the realm of language acquisition. For the sake of simplicity and to avoid any errors with languages, I chose only to use the law to model the English language. This may limit the scope of the investigation, but the possibilities of using this law to simplify English language learning is still wide reaching and could be beneficial both when teaching young children new vocabulary and introducing English as a second language.

Zipf's Law for Modeling Language

Zipf's law, discovered by American linguist George Kingsley Zipf, can be defined as "a pattern of distribution... by which the frequency of an item is inversely proportional to its ranking by frequency" ("Zipf's Law Dictionary Definition")and simple terms, it states that "few occur very often while many others occur rarely." (Black) This can be modeled by the power law equation in which ***f(r)*** is frequency***,r*** is the rank, ***a*** is a constant that differs between sources, and ***b*** > 0, usually with a value around 1[1] (Cancho).

$$f(r) \propto a r^{\frac{1}{b}}$$

This is equation can be modeled as



(Desmos Graphing Calculator)

Zipf's law is commonly associated with word frequency in language. In a given text the most common word occurs "x" amount of times (in the English language, this word is often "the"). The second most common word (typically "of" in English) will occur roughly "x/2" times, the third most common will occur "x/3" times and so on (Hosch). For this exploration in particular, the texts from which the word frequencies will be taken, will be from corpora,

[1] Differences in exponent values depend on the type of sample used. They are typically around 1, but can be greater when an atypical sample is used, such as speech patterns of young children or military combat texts (Cancho)

which are collections of written and/or spoken language taken from a variety of sources to represent the contents of a language ("What Is a Corpus?"). The 20 most common words in two English language corpora, along with their ranks and frequencies are listed in table 1 below:

Rank	**Word** (British National Corpus)	**Frequency** (British National Corpus)	**Word** (American National Corpus)	**Frequency** (American National Corpus)
1	the	61847	the	1204816
2	of	29391	of	616545
3	and	26817	and	595372
4	a	21626	to	533653
5	in	18214	a	490433
6	to	16284	in	409406
7	it	10875	it	255012
8	is	9982	is	227908
9	to	9343	for	204432
10	was	9236	I	188426
11	I	8875	you	179282
12	for	8412	that	168659
13	that	7308	was	158470
14	you	6954	I	157306
15	he	6810	with	150610
16	be	6644	on	141239
17	with	6575	‘s	127676
18	on	6475	‘s	111857
19	by	5096	but	111337

20	at	4790	be	108187

Accessing up-to-date word frequency lists for corpora proved to be slightly challenging, so the data used in these charts could be slightly out of date when compared to the current edition of the corpora, which are periodically updated. But, these lists were relatively easy to access compared to others and they either came directly from the source (American National Corpus) or from a university research center (British National Corpus). Using lists that were previously compiled may be a slight disadvantage to this exploration due to possible inaccuracies, however, there was no other way to reasonably gather this information efficiently, since the only other way would be by analyzing entire corpora. The frequencies given for the British National Corpus seemed low, however, when graphed along with the data from the American National Corpus, both follow Zipfian behavior and closely model the original graph of $f(r) \propto a\, r^{\frac{1}{b}}$ where $b=1$

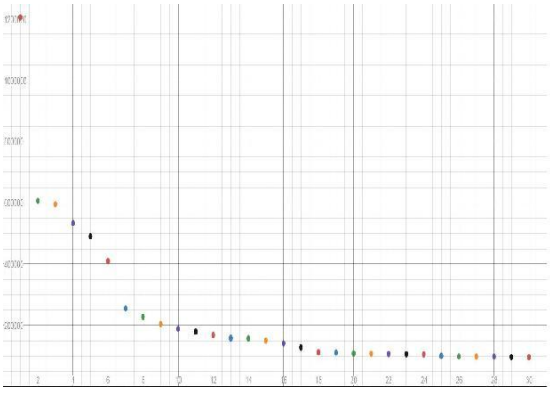


fig. 1 word frequencies in American National Corpus (***Desmos Graphing Calculator)***

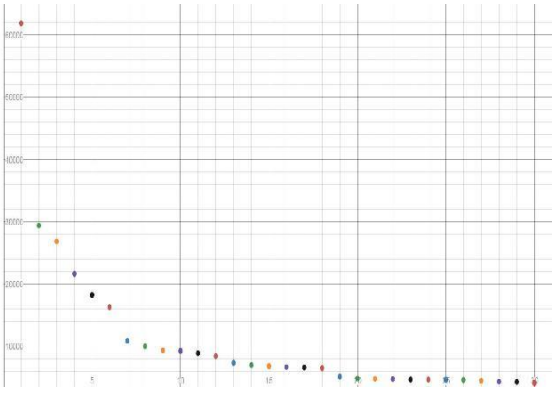


fig. 2 word frequencies in British National Corpus graphed in Desmos (***Desmos Graphing Calculator)***

Fig. 1 is graphical representation of the top 30 words by rank (x-axis) and frequency (y-axis) from the second release of the American National Corpus corpus (***Open American National Corpus***), which consists of 22 million American English words from varying written and spoken sources ("ANC Second Release"). Fig. 2 also graphically shows the rank and frequencies of the top 30 words (***UCREL***), but these words come from the British National Corpus which is composed of about 100 million words, of which 90% of them come from written materials and 10% from spoken materials ("Corpora").

Both graphs have different scaling since the corpora contain different total amounts of words. The most common word for both graphs is “the,” but in Fig. 1, the frequency is 1,204,816, while in Fig. 2 it is only 61,847. The actual words and their rankings also differ slightly, with some words being slightly lower or higher (though usually only by one or two ranks) due to the texts in the corpora being different and also possibly due to the slight variations between British and American English. However, despite the differing sources and scaling, both graphs roughly follow the same shape of the original.

The similarities are not just a coincidence. This pattern is not limited to just English language texts. It is instead thought to be “a universal property of world languages” (Cancho 249). Even ancient, extinct languages follow Zipf’s law. An example of this is the extinct Meroitic language. Two text were analyzed, both with and without bound morphemes, which are parts of a word that cannot stand alone such as the suffix -ly in English (Nordquist). The numbers gathered roughly exhibit Zipfian behavior, particularly when the morphemes are removed (see Table 1) (Smith).

Top 20 ranked words and rank-frequency count distributions for REM 1003

Regular			Bound Morpheme Removed		
Word	Count	Possible Meaning	Word	Count	Possible Meaning
seb	10	?	li	25	particle
qoleb	8	these??	seb	10	?
qor	7	king	qoleb	8	these??
tkk	7	?	lw	8	same as -lowi (is the, it is)??
amnp	5	Amun of Napata	qor	7	king
abrsel	4	Men	tkk	7	?
kdisel	4	Women	qes	6	Kush
abr	3	Man	amnp	5	Amun of Napata
adgite	3	?	abrsel	4	Men
arseli	3	?	kdisel	4	Women
grpgel	3	to command/commander?	lo	4	particle('is a/the')
grpglke	3	to command/commander?	te	4	locative particle
kdi	3	women	abr	3	Man
mno	3	Amun	adgite	3	?
ns	3	?	arse	3	?
qes	3	Kush	grpgel	3	to command/commander?
qesli	3	Kushite? (adj/noun)?	grpglke	3	to command/commander?
qesto	3	Kushite? (adj/noun)?	kdi	3	Woman
wwikewi	3		mno	3	Amun of Napata
100 (number)	2		ns	3	

Table 2. Source: Reginald Smith, "Investigation of the Zipf-plot of the Extinct Meroitic Language," ***Glottometrics*** 15, 2007, p. 56, table 2.

These findings could potentially help with further research and the eventual deciphering of unknown languages such as this one. This discovery is incredibly important, showing that the use of Zipf's law and other methods of computational linguistics can greatly expand the realms of language.

Cumulative Frequency

In a corpus, the top 150 words make up about half of the total number of words, though this can vary significantly depending on the corpus size (Powers). First, in order to test this, the cumulative frequency of the first 150 words in the corpus needs to be found. In order to do this more efficiently than adding, I used a program called CurveExpert to graph and find a best fit line for the data from the American National Corpus. After graphing the 150 points (done the same way as in Fig. 1 and Fig. 2), the graph resembled the original curve

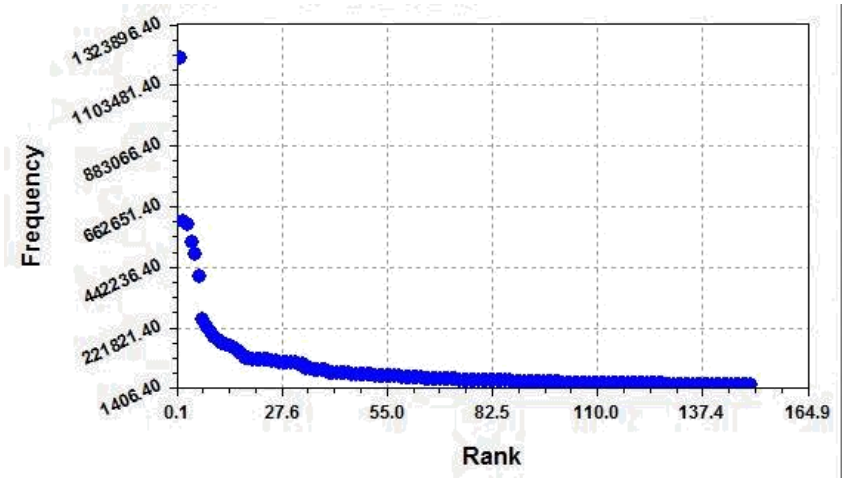


fig. 3 data for the word frequencies in the American National Corpus ("CurveExpert and GraphExpert Software")

Using this plot, a best fitting equation in the power law family was calculated

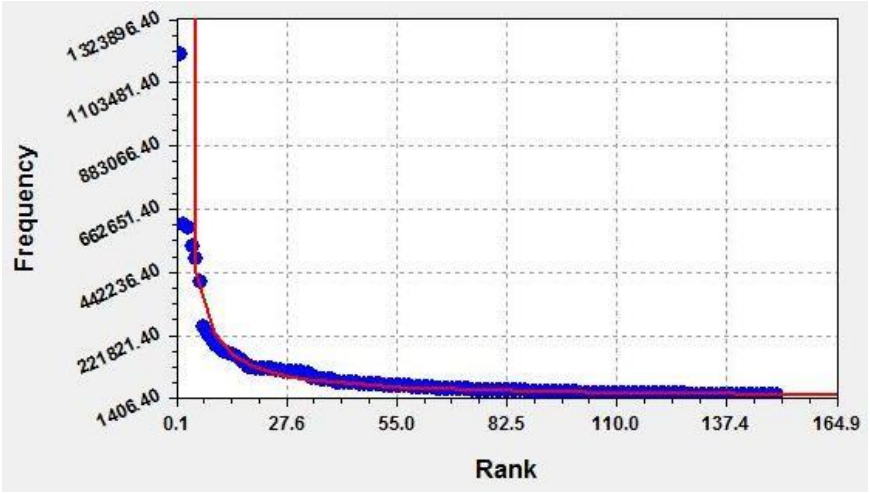


fig. 4 data and best fit line for the word frequencies in the American National Corpus ("CurveExpert and GraphExpert Software")

The best fitting equation for this line is

$$f(r) = 1946773 \cdot \frac{x}{0.9533591}^{1}$$

or

$$f(r) = 1946773x^{-0.9533591}$$

By taking the definite integral of this equation from 4 to 150, the cumulative frequency of the words in these ranks can be found. The reason this equation should be integrated from 4 to 150 is because in the graph, the best-fit line appears to not cross the first 3 points (see fig.4), so these will be added after finding the integral.

$$\int_{4}^{150} 1946773x^{-0.9533591}\ .$$
$$dx$$

$$[41739610.53x^{0.0466409}]_{4}^{150}$$

$$[41739610.53(150)^{0.0466409}] - [41739610.53(4)^{0.0466409}]$$

$$\approx 8200665$$

When I attempted to integrate this equation by hand the first time, I ended up with an answer that was much larger than the total number of words in the American National Corpus ($1.596483279 \times 10^{13}$)

After looking over my work again, I realized I made a calculation error. I then tried again and got a smaller number (820065.217). This was much closer, but it seemed too small in relation to the total word frequency in the entire corpus. Once I looked over my work again, I found that I had left off a number in the coefficient. After reworking it for a 3rd time, my answer came out to 8,200,665 after rounding the decimals. Then the frequencies of the first 3 ranking words, which totaled 2,416,733, were added to the integral value, making the cumulative frequency of the top 150 words in the American National Corpus around 10,617,398. To find the percentage I divided this number by 22,000,000 (the total word frequency in the corpus) and multiplied by 100.

$$22000000^{\frac{10617398}{}} \times 100 = 48.261\%$$

The best fit equation for the data from the frequency list for the British National Corpus was

$$\boldsymbol{f}(\boldsymbol{r}) = 61348.0689562 \cdot \frac{1}{\boldsymbol{x}^{0.833090723846}}$$

Or

$$\boldsymbol{f}(\boldsymbol{r}) = 61348.0689562\boldsymbol{x}^{-0.833090723846}$$

However, after finding the integral from 1 to 150 and it's percentage of the entire text, I found the percentage to be way too small, confirming my earlier suspicions of the numbers. The American National Corpus, however, proves the earlier claim that the top 150 words make up roughly 50% of the total word frequency. When applied to language learning, this particular idea can be incredibly valuable. This means that if one can understand around 150 to 200 words, they will be able to comprehend half of a text or spoken material in a given language. This also illustrates the importance of these top words in language acquisition and can provide an outline for how to structure vocabulary lessons; the words that make up the highest percentage of a given text should be taught first, so as to more quickly build comprehension skills.

Problems with Zipf's Law as a Language Learning Device

Though Zipf's law has the ability to accurately model a language, it does have limits as a language learning tool. When evaluating the improper integral from 1to infinity for the equation $\boldsymbol{f}(\boldsymbol{r}) \propto \boldsymbol{a}\, \boldsymbol{r}^{1}_{\boldsymbol{b}}$, $\boldsymbol{b} > 0$, and in the context of Zipf's law it is usually around 1. In the case of the American National Corpus, $\boldsymbol{b} = 0.9533591$. Since $\boldsymbol{b}$ is less than one, the the area under the curve of this equation (the integral) will be infinite.

$$\int_{1}^{\infty} 1946773x^{-0.9533591} \cdot dx$$

$$\lim_{b\to\infty} \int_{1}^{b} 1946773x^{-0.9533591} \cdot dx$$

$$\lim_{b\to\infty} [41739610.53x^{0.0466409}]_1^b$$

$$\lim_{b\to\infty} 41739610.53b^{0.0466409} - 41739610.53$$

$$\lim_{b\to\infty} \infty - 41739610.53$$

$$\lim_{b\to\infty} = \infty$$

Since evaluating the integral of the best fit equation for the graph of word frequencies in the American National corpus is used to find the cumulative frequency of all the words in the corpus, the fact that the area under this curve turns out to be infinite proves the equation cannot accurately model sources where $b \leq 1$. So, after a certain number of ranks, the accuracy of the model will break down and will no longer be a reliable way to determine the popularity of words in a source. So if someone was using a the American National Corpus as a guideline for the American English and was planning lessons based word importance (determined by rank), they would eventually not be able to determine which words are more important as they reached higher ranks due to the accuracy of the model breaking down.

The words that make up the top majority of words, though used most often, do not necessarily lead to a greater comprehension of a text. Not only does the model not account for the grammatical aspects of language learning, it also does not necessarily lead to comprehension based on vocabulary. The principle of least effort suggests that speakers communicate with the

minimum vocabulary required to introduce an idea, leading to less words being used more often. However, those wishing to gain information from a speaker desire a more specific explanation to comprehend the idea, meaning less common words would need to be utilized. But since speakers wish to use the least effort possible, these words are therefore less likely to be used. So if someone was to use Zipf's law as a model for what vocabulary to learn, after studying 150 words, they will theoretically know them, but their contexts and the meaning of the text as a whole will not be understood (Cancho and Sole).

Conclusions and Reflections

As shown in this exploration, Zipf's law is incredibly useful for modeling languages and is exhibited in a majority of human languages, both deciphered and undeciphered. However, as a tool to aid language learning, this model seems to have more disadvantages than advantages. Though it can serve as a baseline for creating lessons based on word popularity, that does not necessarily mean a language can be completely comprehended based on these top words, as explained with the principle of least effort. And along with this, depending on the source, Zipf's law begins to break down as the ranks increase and the accuracy of the model decreases, therefore making the law useless with less common words.

Before exploring this topic in depth, I thought that Zipf's law could be a successful and innovative way to efficiently learn a new language since it was based on mathematical models. But after my research, I've come to the conclusion that using this method alone would be very flawed and ineffective. Though I did not test the method myself for this exploration, I believe that the evidence presented based upon the data used is enough proof to show that using the

Zipf's law alone would not aid the process of language acquisition. However, perhaps it could be incorporated with traditional language learning methods. Instead of focusing on just the top words in general, one could instead split up these words based on parts of speech, such as verbs, pronouns, nouns, etc. This still gives the same effects as Zipf's law, where the top ranked words are used more often, but it also provides a larger variety of words since, as seen in table 1, the top words in a text do not tend to be very descriptive or detailed, which is essential for language comprehension.

Works Cited

"ANC Second Release." ***ANC Second Release | Open American National Corpus***. N.p., n.d. Web. 5 Jan. 2017. < http://www.anc.org/data/anc-second-release/>.

Black, Paul E. "Zipf's Law." ***Zipf's Law***. N.p., n.d. Web. 10 Sept. 2017. < https://xlinux.nist.gov/dads/HTML/zipfslaw.html>.

Cancho, R. F. I., and R. V. Sole. "Least Effort and the Origins of Scaling in Human Language." ***Proceedings of the National Academy of Sciences*** 100.3 (2003): 788-91. Web. 20 Feb. 2017. <http://www.pnas.org/content/100/3/788.full.pdf>.

Cancho, R. Ferrer I. "The Variation of Zipf's Law in Human Language." ***The European Physical Journal B*** 44.2 (2005): 249-57. Web. 25 Oct. 2016. < http://yaroslavvb.com/papers/cancho-variation.pdf>.

"Corpora." ***UCREL Corpus Holdings***. UCREL, n.d. Web. 4 Jan. 2017. < http://ucrel.lancs.ac.uk/corpora.html>.

"CurveExpert and GraphExpert Software." ***CurveExpert and GraphExpert Software RSS2***. N.p., n.d. Web. 10 Feb. 2017. <https://www.curveexpert.net/>.

Desmos Graphing Calculator. N.p., n.d. Web. 5 Jan. 2017. < https://www.desmos.com/calculator>.

Hosch, William L. "Zipf's Law." ***Encyclopædia Britannica***. Encyclopædia Britannica, Inc., 22 July 2013. Web. 10 Sept. 2016. < https://www.britannica.com/topic/Zipfs-law>.

Nordquist, Richard. "What Are Bound Morphemes in English?" ***About.com Education***. N.p., 13 Dec. 2016. Web. 26 Feb. 2017.

<http://grammar.about.com/od/ab/g/boundmorphterm.htm>.

Open American National Corpus. N.p., n.d. Web. 4 Jan. 2017. < http://www.anc.org/SecondRelease/data/ANC-all-count.txt>.

Powers, David M. W. "Applications and Explanations of Zipf's Law." ***Applications and Explanations of Zipf's Law*** (n.d.): 151-60. Web. 16 Dec. 2016. < http://www.clips.uantwerpen.be/conll98/pdf/151160po.pdf>.

Smith, Reginald. "Investigation of the Zipf-plot of the Extinct Meroitic Language."***Glottometrics*** 15 (2007): 53-61. Web. 19 Feb. 2017.

UCREL, Lancaster UK. UCREL, n.d. Web. 4 Jan. 2017. < http://ucrel.lancs.ac.uk/bncfreq/lists/1_2_all_freq.txt>.

"What Is a Corpus?" ***Oxford Dictionaries***. Oxford Dictionaries, 10 Sept. 2016. Web. 26 Feb. 2017. < https://en.oxforddictionaries.com/explore/what-is-a-corpus>.

"Zipf's Law Dictionary Definition." ***Zipf's Law Dictionary Definition | Zipf's Law Defined***. N.p., n.d. Web. 10 Sept. 2016. < http://www.yourdictionary.com/zipf-s-law>.

Assessment criteria:

Criterion		Comments *	Achievement level		
			Teacher	Moderator	Senior moderator
A	Communication	Sources were accurately cited, technology (Curve Expert) was used appropriately, and all information presented in the paper flowed perfectly from start to finish.	**0-4** 4		
B	Mathematical presentation	A variety of mathematical representations was used, including simple graphs, a table of word frequencies, and well formatted equations. Everything was clearly labeled, applicable, and well-placed.	**0-3** 3		
C	Personal engagement	An excellent, creative topic, where the writer took complex mathematical topics and linked them to non-typical content. The integral model for cumulative words was adapted perfectly.	**0-4** 4		
D	Reflection	Although exploration did not yield the original, desired results, the information gathered was still very relevant and on par with what was expected. Regular references to limitations and possible extensions.	**0-3** 3		
E	Use of mathematics	Although limited in scope, the application of integrals was perfectly placed. Most of the math was pulled directly from the Calc options topic and demonstrated sophistication and knowledge of the content.	**0-6** 5		
		Total:	**0-20** 19		