Rank-frequency distributions of aesthetic units

Roman Rausch

first presented (remotely): *Omentielva Toltea*, Aug. 3rd 2019, Reykjavik, Iceland
published online: Aug. 18th 2024

Contents

1 Distribution of words

When dealing with objects of the same kind, we can analyze them statistically in terms of ranks and frequencies. A common linguistic application is to analyze the frequencies of words in a text or dictionary. Let us define the rank as r (starting with $r = 1$ for the most common one), and $f(r)$ as the corresponding relative frequency (so that $\sum_{r=1}^{N} f(r) = 1$ for N total ranks). The rank-frequency distribution of words is then well-described by the celebrated Zipf's law [Zipf, 1929, Zipf, 1949]:

$$f(r) \propto \frac{1}{r}. \tag{1}$$

Thus, if $f(r)$ is plotted doubly logarithmically the data become a straight line with slope -1. This is exemplified in Fig. 1 by the distribution of words in James Joyce's "Ulysses".

The Zipf distribution is famously ubiquitous and can be applied to data from seemingly unrelated fields – such as sizes of cities, GDPs of countries or the net worths of top billionaires [Simon, 1989, Yu et al., 2018]. Rigorous models to derive the power-law scaling are difficult to establish, but one observes that a necessary condition is a coherent, internally interacting and thus in some sense "equilibrated" system [Simon, 1989, Yu et al., 2018]. Thus, for example, cities over the whole European Union do not follow the scaling, while they do in any particular country where mutual economic interactions are stronger [Yu et al., 2018].

Most of the applications beyond word frequencies require a more general power law

$$f(r) \propto \frac{1}{r^k}, \tag{2}$$

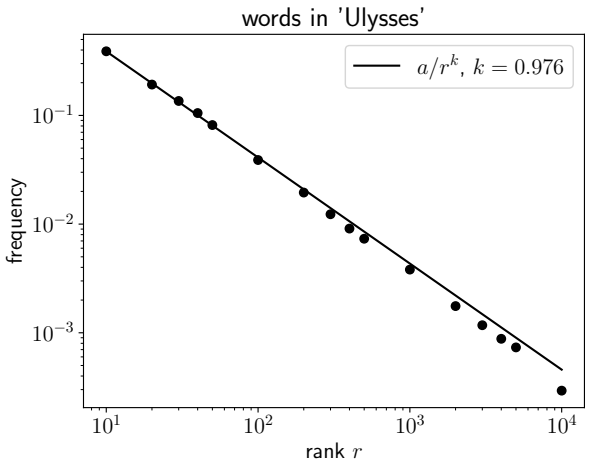

Figure 1: Rank-frequency distribution of the words in "Ulysses". Only selected ranks are used to avoid clutter. The fit is according to Eq. (2) and gives an almost perfect Zipf's law Eq. (1).

where k becomes a free parameter. One can test its robustness across various domains for oneself: In Fig. 2, I applied it to openly available data from Wikipedia at the time of *Omentielva Toltea* (2019), namely the largest German cities and the world's top billionaires. The 'relative frequency' of a city here means its share in the combined population of the cities; and correspondingly with billionaires and wealth. While cities shift in size and people move up and down in the ranks of billionaires, the distribution itself remains remarkably robust over time and a great deal has been written about it [Simon, 1989, Powers, 1998, Li, 2002, Chatterjee et al., 2007, Cristelli et al., 2012, Yu et al., 2018].

The continuous counterpart of Zipf's distribution is the "Pareto distribution", popularized as the "Pareto principle" ("∼20% of people in a team do ∼80% of the work") or as the "Matthew principle" ("the rich get richer and the poor get poorer") with sociological implications regarding wealth and credit distribution, discussing which is beyond the scope of this work [Merton, 1968].

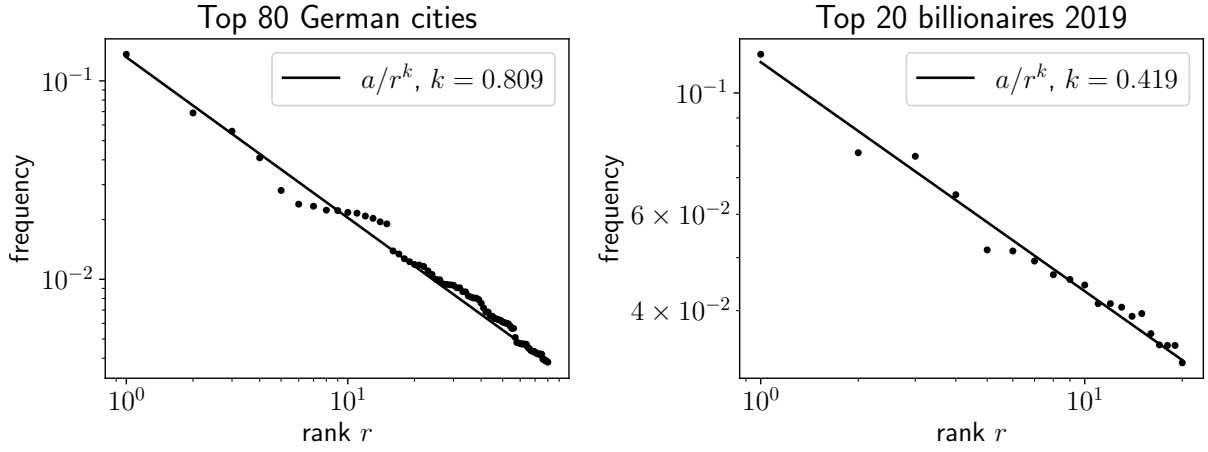

Figure 2: Rank-frequency distribution of the top 80 German cities (left) and of the top 20 billionaires (right) in 2019 with data from Wikipedia. The fit is according to Eq. (2).

Zipf's law is in fact an instance of Stigler's law [Merton and Gieryn, 1982], whereby a phenomenon is not named by its discoverer. While Zipf has not discovered the power-law distribution, he has proposed a kind of a model (or rather a metaphor) to explain it [Zipf, 1949, Powers, 1998]. According to him, words are like tools of a craftsman. In order to minimize effort, he has to keep frequently used tools close. The parameter k can be interpreted as the

dimensionality of the craftsman's space. For the word distribution we apparently have $k = 1$, i.e. the tools are aligned on a line.

2 Distribution of phonemes

The robustness of the Zipf distribution makes it all the more interesting to study cases where it falls short. One such case is the distribution of phonemes, either in a dictionary or in a text. The Russian mathematician Gusein-Zade (GZ) proposed a parameter-free model [Gusein-Zade, 1988], whereby rank-frequencies are random variables. Since the corresponding distribution is not known, the only reasonable assumption is that it is uniform. Because of the constraint $\sum_r f(r) = 1$, this can be interpreted geometrically as drawing uniform random variables from the surface of a simplex [Gusein-Zade, 1988]. Languages are expected to scatter along the expectation value, which is given by:

$$f(r) = \frac{1}{N} \sum_{m=r}^{N} \frac{1}{m}. \tag{3}$$

The formula is not intuitive, but can be approximated by an integral for large N using the Euler–Maclaurin formula:

$$\begin{aligned} f(r) &\approx \frac{1}{N} \int_r^N \frac{1}{x}\, dx + \frac{1}{2N} \left(\frac{1}{r} + \frac{1}{N} \right) \\ &= \frac{1}{N} \ln \left(\frac{N}{r} \right) + \frac{1}{2N} \left(\frac{1}{r} + \frac{1}{N} \right). \end{aligned} \tag{4}$$

This has a mixed logarithmic and power-law dependence on r. While the last term vanishes for large N and for large r at constant N, neglecting it will always result in the drastic setting $f(N) = 0$ instead of a small number $f(N) = 1/N^2$.

An improved approach is to rewrite the series using harmonic numbers $H_n = \sum_{m=1}^{n} 1/m = \ln n + \gamma + 1/2n + O(1/n^2)$:

$$\begin{aligned} f(r) &= \frac{1}{N} \left(H_N - H_{r-1} \right) \\ &= \frac{1}{N} \left(H_{N+1} - H_r + H_N - H_{N+1} + H_r - H_{r-1} \right) \\ &\approx \frac{1}{N} \left(H_{N+1} - H_r \right) \\ &\approx \frac{1}{N} \ln \left(\frac{N+1}{r} \right) \\ &\sim -\ln(r) \end{aligned} \tag{5}$$

Now the neglected terms $H_N - H_{N+1} + H_r - H_{r-1}$ become of order of $1/N^2$ for $r \approx N$, so that $f(N) = 1/N^2 + O(1/N^3)$ for large N. The error of this approximation is shown in Fig. 3.

Thus, the decrease with r is logarithmic $\sim -\ln(r)$ and slower than the power-law of the Zipf distribution. This implies that a semilogarithmic plot, whereby only the x-axis is plotted logarithmically, should yield a straight line. This is demonstrated in Fig. 4 for the distribution of phonemes in English. The quality of data description is commonly measured by the R^2 test, which becomes 1 when all data lie on top of the model curve. Remarkably, the GZ distribution achieves $R^2 \approx 0.98$ for the English data, compared to $R^2 < 0.2$ for the Zipf distribution with $k = 1$ and $R^2 < 0.9$ if k is fitted according to Eq. (2).

Gusein-Zade's paper actually dealt with the distribution of *letters* in Russian, but I would argue that it makes more sense to look at phonemes, since spelling can be arbitrarily obfuscate the pronunciation (e.g. when digraphs or trigraphs are involved to represent a single sound; or in the more extreme case of Chinese characters).

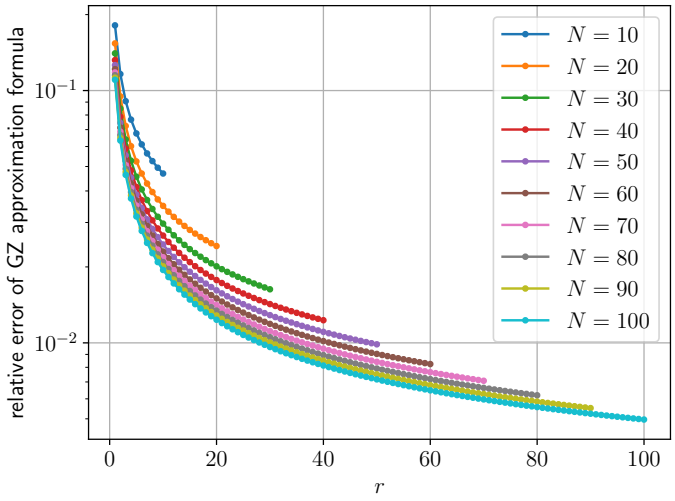

Figure 3: Relative error of the approximative formula in Eq. (5) compared to the exact Eq. (3).

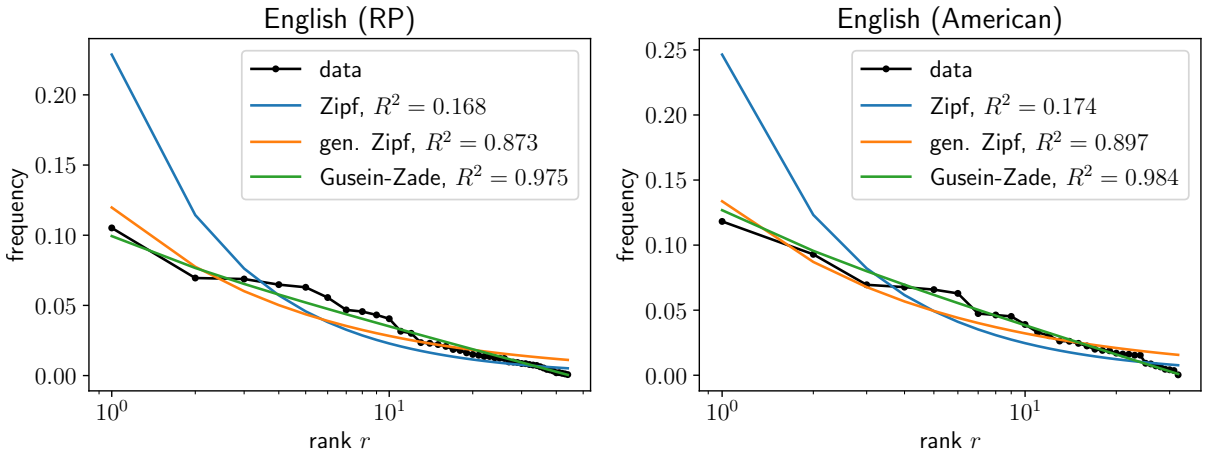

Figure 4: Rank-frequency distributions of English phonemes for the Received Pronounciation (RP) [Higgins, 2008] and the American pronunciation [Sigurd, 1968]. The generalized Zipf fit is according to Eq. (2). The Gusein-Zade distribution describes the data far better than Zipf's law.

In fact, I find that just as Zipf's law robustly describes the distribution of words, the GZ formula robustly describes the distribution of phonemes across various languages. Figure 5 shows this for openly available datasets on languages from different groups beyond English with similar results as in Fig. 4. Figure 6 shows rank-frequency distributions of consonants using ∼7000 wordlists from languages across the world [Everett, 2018], which also nicely follows the GZ distribution.

Tambovtsev and Martindale have examined different distributions for phonemes in various languages and, based on the R^2 values, conclude that the "phoneme frequencies follow a Yule distribution" [Tambovtsev and Martindale, 2007]. The Yule distribution is given by

$$f_r \propto \frac{c^r}{r^k}, \tag{6}$$

where an an additional free parameter c is introduced to further soften up the Zipf distribution. Thus, the statement is formally true, but I argue that this is a case of overfitting – the fit of a curve with 20-30 monotonically decreasing datapoints will be better the more free parameters one introduces, but descriptiveness is lost in the process. The much more interesting result hides in plain sight, namely that the GZ distribution is practically as good as Yule

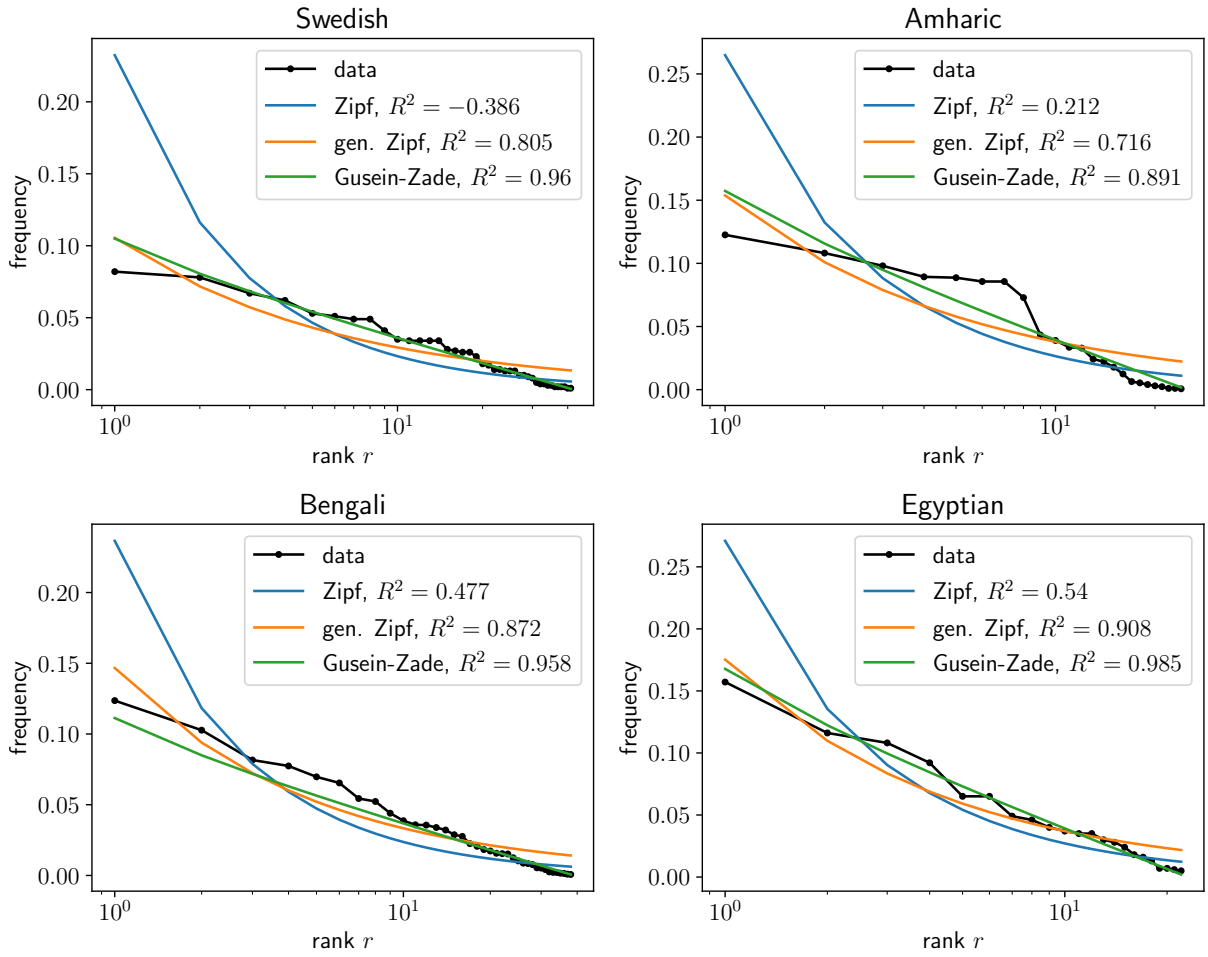

Figure 5: Same as in Fig. 4, but for phonemes in Swedish, Bengali [Sigurd, 1968], Amharic [Bender, 1974] and Egyptian [Peust, 2008] (only consonants).

(on average, $R^2 \sim 0.95$ compared to Yule's $R^2 \sim 0.97$ [Tambovtsev and Martindale, 2007]), but *without any free parameters at all*. This is despite the fact that the authors seem to have used the approximative log formula (4) instead of the exact formula (3).

3 Distribution of phonemes in constructed languages

Overall, natural languages seem to consistently achieve $R^2 = 0.94 \pm 0.05$ using the GZ formula to describe their phoneme distributions. Therefore, I propose that R^2 under the GZ distribution can be a measure of naturalness. In the following, I test how constructed languages behave in this respect.

3.1 Q(u)enya

I have taken Quenya data from the Eldamo database [Strack, 2019], cleaned it up by removing punctuation and standardizing the spelling. I phoneticized the spelling using the following character replacements: c→k, χ→h, þ→s, hl→ɬ, hr→ɽ, hw→ʍ, x→ks, q→kw, qu→kw. One can also argue for the need to replace ty→c (i.e. the palatal stop /c/ as a separate phoneme), but this does not change the result much either way due to its low occurrence. There is a similar uncertainty with regard to whether <ng> stands for /ŋg/ or /ŋ/ in various Tolkien's languages, but this

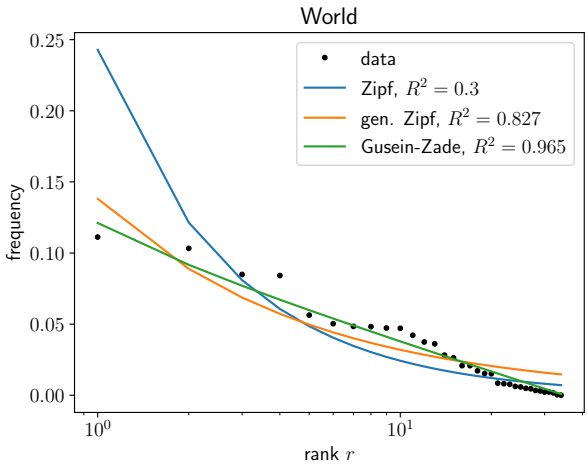

Figure 6: Rank-frequency distribution of the consonants in the world's languages [Everett, 2018].

has been also ignored for simplicity. The result is shown in Tab. 1 and we can see that the distribution is well described by the GZ formula irrespective of Early Qenya or late Quenya, or whether one takes the distribution from a dictionary or from texts.

language	*words*	*phonemes*	R^2
Quenya	2369	30	0.927
Quenya phrases	–	28	0.955
Middle Quenya	1858	30	0.927
Early Qenya	3755	27	0.968
Early Qenya phrases	–	27	0.924

Table 1: Values of R^2 of the Q(u)enya phoneme frequencies according to the Gusein-Zade formula.

3.2 Sindarin and Noldorin

I applied the same procedure to Goldogrin, Noldorin and Sindarin, with the phoneticization: th→þ, dh→ð, ch→χ, lh→ɬ, rh→ɋ, wh→ʍ, hw→ʍ, mh→v. The result is shown in Tab. 2 and we can see that the distribution is again well described by the GZ formula. Only the distribution taken from the the Sindarin phrases scores slightly worse.

language	*words*	*phonemes*	R^2
Sindarin	1265	33	0.952
Sindarin phrases	–	32	0.882
Noldorin	1327	32	0.975
Early Noldorin	815	27	0.928
Goldogrin	3218	30	0.966

Table 2: Values of R^2 of the Goldogrin/Noldorin/Sindarin phoneme frequencies according to the Gusein-Zade formula.

3.3 Non-Elvish languages by Tolkien

I applied the same procedure to Adûnaic, Black Speech and Khuzdul, whose corpora are much smaller. The phonemic replacements were: kh→κ, ph→φ, th→θ for Adûnaic, sh→ʃ, th→θ, gh→γ for Black Speech, and kh→κ for Khuzdul. The Black Speech data amounts to *aʃ nazg durbatulûk aʃ nazg gimbatul aʃ nazg þrakatulûk aɣ burzumiʃi krimpatul* and *uglúk u bagronk ʃa puʃdug saruman-glob búbhoʃ skai*, while the Khuzdul data amounts to an even shorter string of *balin fundinul uzbad κazaddûmu*, *baruk κazâd* and *κazâd aimênu*.

The result is shown in Tab. 3. Remarkably, the GZ model fits the data really well even for these small samples. I interpret this result as a formal expression of the fact that these phrases feel like excerpts from real languages, and it requires only a few phrases to achieve this effect.

language	*words*	*phonemes*	R^2
Adûnaic	160	28	0.920
Adûnaic phrases	–	28	0.918
Black Speech	34	24	0.897
Black Speech phrases	–	21	0.927
Khuzdul phrases	–	15	0.858

Table 3: Values of R^2 of the Adûnaic, Black Speech and Khuzdul phoneme frequencies according to the Gusein-Zade formula.

3.4 Obscure languages by Tolkien

One can push the approach further to such poorly attested early languages like Naffarin and Nebosh. Since little is known about their phonologies, one can only assume that the spelling is already mostly phonemic. The result is shown in Tab. 4, still with remarkably high R^2 scores.

language	*phonemes*	R^2
Naffarin phrase	22	0.973
Nevbosh limerick	23	0.897

Table 4: Values of R^2 of the frequencies of the (supposed) phonemes of Naffarin and Nebosh according to the Gusein-Zade formula.

3.5 Other constructed languages

I have applied the procedure to data from other famous constructed languages.

Esperanto: The following regularization was used: aj→ai, ej→ei, oj→oi, ŭ→u. I used the original wordlist in *Dr. Esperanto's International Language* (1889) [Zamenhof, 1889], where the words are given as roots, e.g. **patr'** for **patro** 'father' or **parol'** for **paroli** 'to speak' and a larger wordlist with endings [Makepeace, 2024]. Tab. 5 shows that adding the endings improves the R^2 score and that it is fully consistent with natural languages.

Dothraki: The following phoneticization was used: ch→č, sh→ʃ, th→θ, zh→ʒ, kh→χ, cch→čč, ssh→ʃʃ, tth→θθ, zzh→ʒʒ, kkh→χχ. The R^2 score was computed on a large wordlist [Henkel, 2024] and is fully consistent with natural languages.

Volapük: Only one replacement was applied: x→ks. The R^2 was computed on the Volapük Wikipedia article about Volapük itself and is consistent with natural languages, though on the low side.

Na'vi: The following phoneticization was used: px→P, tx→T, kx→K (ejective consonants), ng→ŋ, '→ʔ, é→e (explicit stress indication in one instance). I get $R^2 = 0.715$ computed on a large wordlist [Miller, 2024], which is the first instance of a clear deviation from natural languages.

Klingon: The following phoneticization was used: tlh→L, ch→č, gh→γ, '→ʔ, ng→ŋ. I get $R^2 = 0.484$ computed on a large wordlist [PanderMusubi, 2024], indicating an even stronger deviation from natural languages than in the case of Na'vi.

language	*words*	*phonemes*	R^2
Esperanto (1889), just roots	922	26	0.806
Esperanto, with endings	1805	27	0.989
Dothraki	1411	26	0.934
Volapük	–	26	0.824
Na'vi	2814	27	**0.715**
Klingon	4387	26	**0.484**

Table 5: Values of R^2 of the phoneme frequencies of various constructed languages according to the Gusein-Zade formula. Only Na'vi and Klingon get low scores (shown in bold).

The conclusion is that constructed languages mostly pass the naturalness test. Only alien languages (Na'vi and Klingon), whose explicit design goal is not to resemble any natural language, show perceptible deviations (whereby Na'vi seems more naturalistic than Klingon, in accordance with expectation). Note that the Yule distribution is not able to discriminate between these cases and fits R^2 for all languages very close to 1.

4 Extension to other aesthetic units

How can we understand that the GZ distribution performs so well for phonemes? Note that rather than dealing with several thousands of words, we have typically just 20-30 phonemes, so Zipf's craftsman metaphor does not seem appropriate – basically, all the phonemes are active, all the time. Judging in a utilitarian fashion, a completely flat distribution seems best, as words would become maximally distinguishable, but this is not found.

But phonology has an aesthetic component as well, the selection of phonemes makes up the style of the language which is created and modified by the collective action of its speakers. In this respect we may compare phonology to other art forms that evolve in real time and use a limited amount of "aesthetic units", namely dance and music, for which have a good idea where the aesthetic pleasure comes from.

A ballroom dance performance consists of more or less standardized moves or figures that have special names like "cross-body lead" (salsa) or "ocho" (Argentine tango). Music is complex and various things can be measured [Levitin et al., 2012, Mehr et al., 2019], but we can for example consider pitch classes as a relatively short set of units. Randomness now factors naturally into the performance: Aesthetic pleasure comes on the one hand from being surprised by an unexpected figure or note, on the other hand from being able to predict a great deal of them successfully, i.e. follow the melody. Neither too much randomness (white noise), nor too little (a constant beat) are perceived as very artistic.

In the case of music, data is readily available from the *music21* database [Cuthbert, 2024]. A typical example, Verdi's "La Donna è Mobile", is shown in Fig. 7. The number of pitches (30) is of the same order as the number

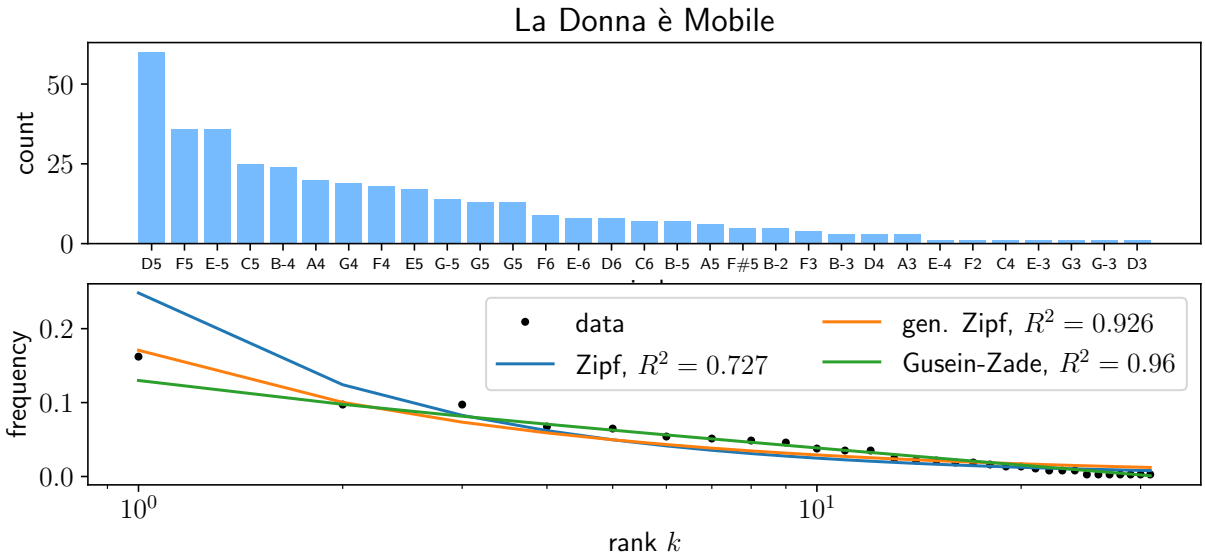

Figure 7: Rank-frequency distributions of pitches in Verdi's "La Donna è Mobile". The generalized Zipf fit is according to Eq. (2).

of phonemes in a language. We see that the pitches are indeed somewhat better described by the GZ distribution compared to Zipf's law. To get a wider picture, Tab. 6 presents averages of the pieces of several composers. The GZ consistently outperforms the generalized Zipf distribution despite being parameter-free, very similar to the case of phonemes.

composer	*pieces*	average R^2 (GZ)	average R^2 (gen. Zipf)	average k
Bach	433	0.927	0.826	0.628
Mozart	16	0.883	0.788	0.670
Beethoven	26	0.844	0.784	0.666
Palestrina	1318	0.889	0.799	0.626
Haydn	9	0.866	0.808	0.71
Monteverdi	97	0.924	0.838	0.722

Table 6: R^2 scores for the pitch class distributions averaged over many pieces for each composer available in the *music21* database [Cuthbert, 2024].

I note that a recent paper looked at another measure for music, namely rhythmic and melodic bigrams, which turn out to be universally power-law-distributed across various cultures [Mehr et al., 2019]. The linguistic analog would be probably phonetic bigrams, which is a higher-order correlation.

5 Summary

The proposition of this paper is that the rank-frequency distribution of aesthetic units in three domains (phonemes in language, pitch classes in music, figures in dancing) is governed by the Gusein-Zade distribution. The tail events are only logarithmically suppressed and there is less disparity between common and rare units than there would be under a Zipf distribution.

The robust validity of the GZ distribution for phonemes can be shown for data from natural languages, where it is a much better description than the Zipf distribution, even if the latter is generalized and the power-law exponent

k becomes a free parameter. Constructed languages which are made to be naturalistic also robustly follow GZ distribution, even when only short phrases are considered. It seems that it requires little effort to make a language sound natural under this measure. However, languages that are meant to be spoken by aliens and are unnatural for humans by design (Na'vi and Klingon) also fail the naturalness test under the GZ distribution.

For many classical music pieces, the distribution of pitch classes is better described by the GZ formula rather than Zipf's law, and they might be regarded as the analog of phonemes.

An application to dancing is yet to be made because useful data is lacking. An interesting possibility involves robotic dancing [Baillieul and Özcimder, 2012]. Humanoid robots can be taught the primitive elements of a dance. But these should be sticked together to a routine, and I propose that adherence to the GZ distribution will be a necessary condition for an aesthetic perception by human judges.

References

[Baillieul and Özcimder, 2012] Baillieul, J. and Özcimder, K. (2012). The control theory of motion-based communication: Problems in teaching robots to dance. In *2012 American Control Conference (ACC)*, pages 4319–4326.

[Bender, 1974] Bender, M. L. (1974). Phoneme frequencies in Amharic. *Journal of Ethiopian Studies*, 12(1):19–24.

[Chatterjee et al., 2007] Chatterjee, A., Yarlagadda, S., and Chakrabarti, B. K. (2007). *Econophysics of wealth distributions: Econophys-Kolkata I*. Springer Science & Business Media.

[Cristelli et al., 2012] Cristelli, M., Batty, M., and Pietronero, L. (2012). There is more than a power law in Zipf. *Scientific Reports*, 2(1):812.

[Cuthbert, 2024] Cuthbert, M. S. (2024). music21. retrieved from: `https://web.mit.edu/music21/`.

[Everett, 2018] Everett, C. (2018). The similar rates of occurrence of consonants across the world's languages: A quantitative analysis of phonetically transcribed word lists. *Language Sciences*, 69:125–135.

[Gusein-Zade, 1988] Gusein-Zade, S. M. (1988). Frequency distribution of letters in the Russian language. *Problemy Peredachi Informatsii*, 24(4):102–107.

[Henkel, 2024] Henkel, M. (2024). Dothraki resource grammar. retrieved from: `https://github.com/mariahenkel/dothraki_gf/blob/master/vocabulary/input_dot.txt`.

[Higgins, 2008] Higgins, J. (2008). RP phonemes in the advanced learner's dictionary. retrieved from: `http://www.minpairs.talktalk.net/phonfreq.html`.

[Levitin et al., 2012] Levitin, D. J., Chordia, P., and Menon, V. (2012). Musical rhythm spectra from Bach to Joplin obey a 1/f power law. *Proceedings of the National Academy of Sciences*, 109(10):3716–3720.

[Li, 2002] Li, W. (2002). Zipf's law everywhere. *Glottometrics*, 5(2002):14–21.

[Makepeace, 2024] Makepeace, P. (2024). Vortlisto. retrieved from: `https://github.com/paulmakepeace/vortlisto`.

[Mehr et al., 2019] Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., and Glowacki, L. (2019). Universality and diversity in human song. *Science*, 366(6468):eaax0868.

[Merton, 1968] Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63.

[Merton and Gieryn, 1982] Merton, R. K. and Gieryn, T. F. (1982). Science and social structure: A festschrift for Robert K. Merton. *(No Title)*.

[Miller, 2024] Miller, M. (2024). Na'vi / English dictionary 15.6.1. retrieved from: `https://learnnavi.org/navi-vocabulary/`.

[PanderMusubi, 2024] PanderMusubi (2024). Word lists and spell checking for Klingon. retrieved from: `https://github.com/PanderMusubi/klingon/tree/master`.

[Peust, 2008] Peust, C. (2008). On consonant frequency in Egyptian and other languages. *Lingua Aegyptia*, 16:105–134.

[Powers, 1998] Powers, D. M. W. (1998). Applications and explanations of Zipf's law. In *New Methods in Language Processing and Computational Natural Language Learning*.

[Sigurd, 1968] Sigurd, B. (1968). Rank-frequency distributions for phonemes. *Phonetica*, 18(1):1–15.

[Simon, 1989] Simon, H. A. (1989). The sizes of things. In *Statistics: A Guide to the unknown*. Duxbury Press.

[Strack, 2019] Strack, P. (2019). Eldamo - an Elvish Lexicon. retrieved from: `https://eldamo.org`.

[Tambovtsev and Martindale, 2007] Tambovtsev, Y. and Martindale, C. (2007). Phoneme frequencies follow a Yule distribution. *SKASE journal of theoretical linguistics*, 4(2):1–11.

[Yu et al., 2018] Yu, S., Xu, C., and Liu, H. (2018). Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. *arXiv preprint arXiv:1807.01855*.

[Zamenhof, 1889] Zamenhof, L. L. (1889). Dr. Esperanto's international language. retrieved from: `https://en.wikibooks.org/wiki/Esperanto/Appendix/English-Esperanto_word_list_(1889)`.

[Zipf, 1929] Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, 40:1–95.

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