Sindarin dictionary statistics

Roman Rausch

Sep. 2nd 2013

Contents

Introduction

This is a statistical evaluation of the Sindarin dictionary hosted at `http://www.sindarin.de`.

1 All phonemes

The frequencies of all Sindarin phonemes are found to be:

rank	*phoneme*	*frequency*
1	**a**	0.145
2	**n**	0.11
3	**e**	0.094
4	**r**	0.087
5	**i**	0.075
6	**l**	0.071
7	**o**	0.055
8	**g**	0.044
9	**d**	0.043
10	**þ**	0.041
11	**u**	0.036
12	**m**	0.03
13	**s**	0.027
14	**t**	0.023
15	**b**	0.019
16	**k**	0.015
17	**w**	0.015
18	**f**	0.012
19	**h**	0.012
20	**v**	0.011
21	**ð**	0.011
22	**p**	0.009
23	**χ**	0.007
24	**j**	0.002

25	**y**	0.002
26	**R**	0.002
27	**L**	0.002
28	**W**	0.001

Notation (for easiness of counting, digraphs were converted to unigraphs):

- **k** for /k/, pronounced [k], spelled <c> by Tolkien
- **þ** for /þ/, pronounced [þ], spelled <th> and sometimes <þ> by Tolkien
- **ð** for /ð/, pronounced [ð], spelled <dh> and sometimes <ð> by Tolkien
- **χ** for /x/, pronounced [x] or [χ], spelled <ch> by Tolkien
- **j** for /j/, pronounced [j], spelled <i> by Tolkien
- **R** for /r̥/, pronounced [r̥], spelled <rh> by Tolkien
- **L** for /ɬ/, pronounced [ɬ], spelled <lh> by Tolkien
- **W** for /ʍ/, pronounced [ʍ], spelled <hw> or <wh> by Tolkien

Assumptions for simplicity:

- The difference between long and short vowels is neglected.
- Diphthongs are counted as two vowels.
- It is not always clear how <ng> is supposed to be pronounced (either /ŋ/ or /ŋg/). It was treated as /n/ + /g/.

1.1 Discussion

For the rank-frequency distribution $p(r)$ (where r is a phoneme's rank), an ad-hoc formula was first proposed by Zipf in 1929 [1]:

$$p(r) \sim \frac{1}{r}$$

with the normalization $s(N) = \sum_{k=1}^{N} 1/k$, where N is the total amount of phonemes.

Several authors noticed since then that it does not fit the data across languages too well and have proposed other ad-hoc fitting functions [3, 4]. In 1988, Gusein-Zade proposed a formula [2] based on a sensible assumption, namely that rank-frequencies are drawn from a uniform probability density and that $p(r)$ can be approximated by the corresponding expectation value for any given language. This leads to:

$$p(r) = \frac{1}{N} \sum_{k=0}^{N-r} \frac{1}{r+k}$$

For large N and for large r at fixed N this can be approximated by:

$$p(r) \approx \frac{1}{N} \log \frac{N+1}{r}$$

It turns out that this formula describes real-language data rather well and no wild fitting is required (see below). The fact that a model assumption enters the calculation seems to have been overlooked or misunderstood by other authors, probably because Gusein-Zade's paper was published in Russian. One can see that it makes no sense to generalize the Zipf distribution by adding fittable parameters, like $r^{-\beta}$ (as it often seems to be

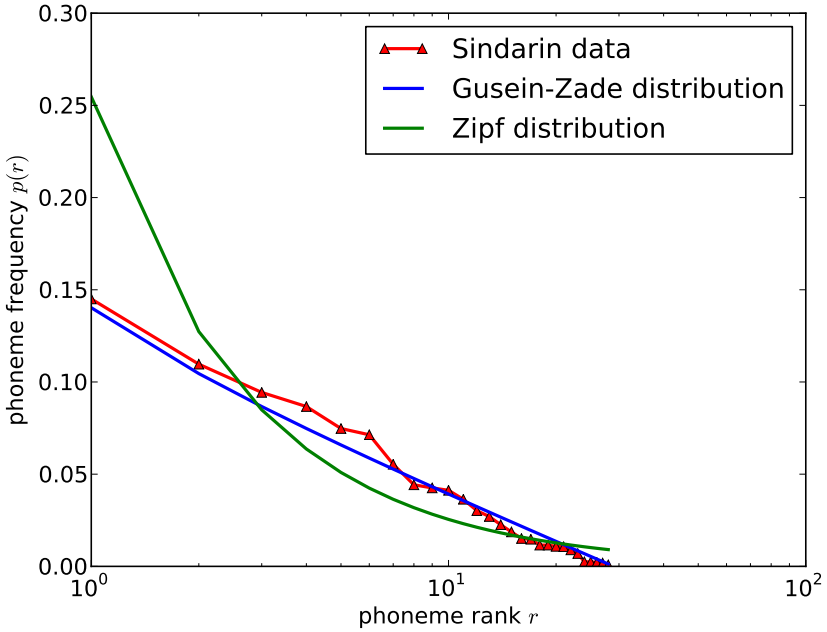

Figure 1: Rank-frequency distribution of Sindarin phonemes

done) because the dependency is different, approximately $\log 1/r$ rather than a power law[1]. This means that a semilogarithmic plot of $p(r)$ should produce a straight line. This is indeed the case for the Sindarin data, as seen in fig. 1.

Comparing it with data from natural languages (fig. 2) one finds a similarly good agreement for English and Swedish, somewhat worse for Bengali. Except for Bengali, deviations are spread both above and below the Gusein-Zade function which suggests a statistical rather than a systematic error. I do not know how reliable the Bengali data are.

Note that the formula does not predict how common a certain sound is, but rather how frequent the phoneme ranked r is (whatever the phoneme itself may be). It turns out that this value is completely determined by the total amount of phonemes N.

Note also that it matters for the individual frequencies whether one considers a dictionary or a text. In the latter case, English [ð] obviously becomes much more common [5] due to the *the*s and *that*s (in Sindarin texts, the frequency of **i** is expected to go up for the same reason). However, the distribution seems to stay the same: The RP data in figure 2 are from a dictionary, the American English data from a text.

Finally one should note that the RP data for English include diphthongs as separate phonemes, while the American English data do not; but again, this does not seem to affect the distribution itself.

We can thus conclude that the rank-frequency distribution of the Sindarin phonemes is indistinguishable from that of a natural language.

2 Vowels & consonants

Rank frequencies for vowels only:

rank	*phoneme*	*frequency*
1	**a**	0.355
2	**e**	0.231
3	**i**	0.183
4	**o**	0.136
5	**u**	0.089
6	**y**	0.006

Rank frequencies for consonants only:

[1]This does not mean that the Zipf distribution cannot be applicable somewhere else. It does seem to describe the distribution of words in a text [7].

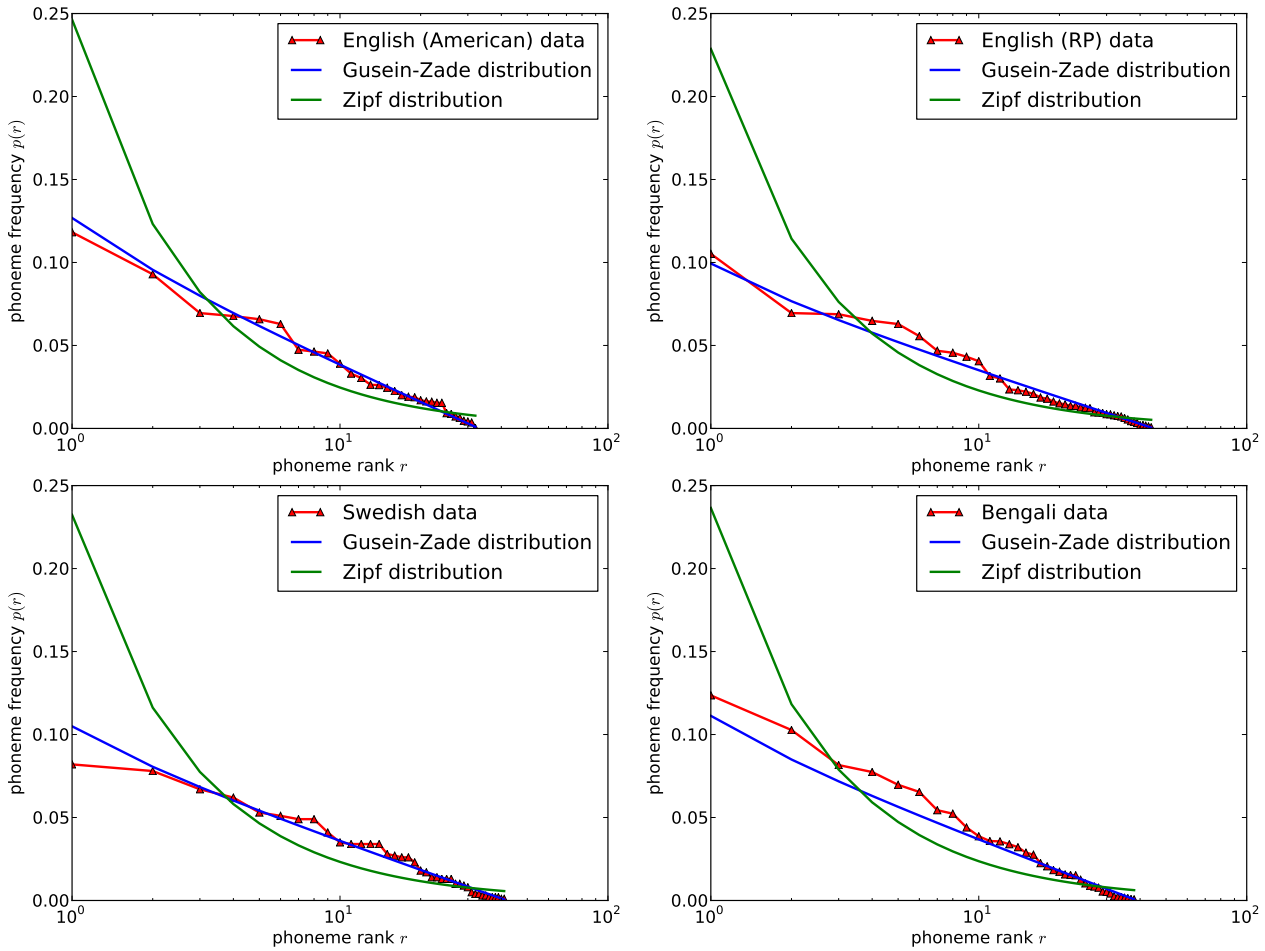

Figure 2: Rank-frequency distributions of phonemes for various natural languages. The American English, Swedish and Bengali data are from the references in [3], the RP data are from [5].

rank	*phoneme*	*frequency*
1	**n**	0.185
2	**r**	0.146
3	**l**	0.121
4	**g**	0.075
5	**d**	0.072
6	**þ**	0.07
7	**m**	0.051
8	**s**	0.046
9	**t**	0.038
10	**b**	0.032
11	**k**	0.026
12	**w**	0.025
13	**f**	0.019
14	**h**	0.019
15	**v**	0.018
16	**ð**	0.018
17	**p**	0.015
18	**χ**	0.012
19	**j**	0.004
20	**R**	0.004
21	**L**	0.003
22	**W**	0.001

Vowel-to-consonant ratio:

consonants	0.592
vowels	0.408

3 Place & manner of articulation

Place of articulation:

dentals	0.567
labials	0.184
velars	0.15
interdentals	0.1

Manner of articulation:

sonorants/semivowels	0.541
stops and fricatives	0.459

Distribution among stops:

rank	*phoneme*	*frequency*
1	**g**	0.291
2	**d**	0.28
3	**t**	0.148
4	**b**	0.123
5	**k**	0.099
6	**p**	0.059

Distribution among fricatives:

rank	*phoneme*	*frequency*
1	**þ**	0.344
2	**s**	0.226
3	**f**	0.096
4	**h**	0.096
5	**v**	0.09
6	**ð**	0.089
7	**χ**	0.058

Distribution among sonorants/semivowels:

rank	*phoneme*	*frequency*
1	**n**	0.343
2	**r**	0.271
3	**l**	0.223
4	**m**	0.094
5	**w**	0.046
6	**j**	0.008
7	**R**	0.007
8	**L**	0.006
9	**W**	0.002

4 Bigrams and entropy

A bigram is a cluster of two letters, or, in this case, two phonemes. One can introduce the conditional probability $p_i(j)$ to find the phoneme j if the preceding phoneme is i. It forms a matrix with normalized rows: $\sum_j p_i(j) = 1$. If one weighs the rows with the frequencies $p(i)$, one obtains the probability to get the phonemes i and j in two sequential draws: $p(i,j) = p(i)p_i(j)$. This is now of course normalized with respect to the total sum: $\sum_{ij} p(i,j) = 1$. The procedure is readily generalized to n-grams.

Linguistically, the matrix shows us a language's phonotacics and the restriciveness of its phonology. (Probably, one can also use it to write a ruthlessly efficient hangman algorithm.) Obviously, the higher the spread of values across the bigram matrix, the freer the phonology. This is exactly what is measured by the n-gram entropy[2]:

$$H_n = -\sum_{i_1 i_2 \ldots i_n}^{N} p(i_1, i_2, \ldots, i_n) \log_2(p(i_1, i_2, \ldots, i_n))$$

H_n can already be computed for the unigram frequencies $p(i)$, but as discussed above, their distribution is mostly determined by the total amount of phonemes N, so that the same goes for the entropy. It seems more interesting to look at the bigram entropy H_2: The smaller it is, the more restrictive the phonology. Note that for any value of n, H_n has the maximum value of $H_{max} = log_2(N)$ which corresponds to the case that all n-grams are equiprobable, which would make the phonology absolutely free and all the phonemes uncorrelated.

The following three tables show $p_i(j)$, computed for vowels only, consonants only, and for all phonemes. Colors are used as a visual guide to highlight values from 0.1 to 0.2 (blue); 0.2 to 0.3 (green); 0.3 to 0.4 (purple); 0.4 to 0.5 (orange); and finally above 0.5 (red).

Vowels only:

[2]The logarithm to base 2 is a convention and one says then that the entropy is measured in "bits". Of course, this sets the scale rather than the unit – H is dimensionless.
The interpretation of H in information theory is as (the average) uncertainty: H is zero if a probability is equal to one (a completely certain event), increases with N (the more outcomes, the higher the uncertainty), and is maximal at fixed N if all probabilities are equal (all outcomes equiprobable, hence maximal uncertainty). Finally, the uncertainty of two independent events is the sum of the individual uncertainties.

	a	**e**	**i**	**o**	**u**	**y**
a		0.51	0.233	0.004	0.253	
e	0.2		0.733	0.033	0.033	
i	0.696	0.174		0.13		
o		1				
u	0.048		0.952			
y						

Consonants only:

	p	t	k	b	d	g	s	f	þ	χ	h	v	ð	n	l	r	L	R	m	w	W	j
p															0.5	0.5						
t															0.2	0.8						
k									0.154						0.154	0.692						
b					0.032										0.065	0.903						
d				0.068	0.017							0.017			0.068	0.695				0.136		
g					0.007									0.014	0.435	0.116				0.428		
s	0.021	0.604	0.031	0.021	0.01	0.063	0.25															
f								0.833							0.056	0.111						
þ						0.079									0.095	0.825						
χ				1																		
h																						
v														0.542	0.042	0.417						
ð				0.063												0.938						
n		0.072	0.038	0.004	0.102	0.148		0.008	0.015		0.011			0.557	0.008	0.004				0.034		
l		0.046		0.037	0.055	0.009		0.037	0.083	0.101		0.101	0.009		0.45	0.028				0.046		
r				0.056	0.033	0.056		0.011	0.239	0.106		0.056	0.056	0.278	0.017	0.033			0.006	0.056		
L																						
R																						
m	0.159			0.136	0.023						0.023				0.159	0.136			0.364			
w					1																	
W																						
j																						

All phonemes:

	a	e	i	o	u	y	p	t	k	b	d	g	s	f	þ	χ	h	v	ð	n	l	r	L	R	m	w	W	j
a		0.104	0.048	0.001	0.052			0.002	0.001	0.015	0.102	0.025	0.067	0.006	0.079	0.015	0.002	0.026	0.018	0.151	0.082	0.143			0.044	0.021		
e	0.007		0.025	0.001	0.001				0.002	0.027	0.076	0.077	0.047	0.001	0.065	0.012		0.013	0.045	0.246	0.204	0.114	0.001		0.013	0.024		
i	0.097	0.024		0.018						0.015	0.019	0.021	0.021	0.004	0.105	0.004	0.001	0.013	0.019	0.244	0.121	0.16			0.1	0.01		
o		0.024								0.015	0.068	0.041	0.068		0.083	0.013	0.004	0.041	0.02	0.218	0.124	0.274			0.007	0.002		
u	0.019		0.374							0.016	0.031	0.097	0.022		0.04			0.025	0.022	0.14	0.062	0.125		0.003	0.025	0.003		
y												0.087						0.043		0.348	0.261	0.261						
p	0.341	0.451	0.073	0.037	0.024																0.043	0.043						
t	0.418	0.136	0.13	0.153	0.079																0.018	0.072						
k	0.41	0.158	0.072	0.137	0.094	0.036									0.015						0.015	0.07						
b	0.331	0.357	0.013	0.076	0.019	0.006					0.007										0.013	0.184						
d	0.293	0.107	0.132	0.118	0.121	0.018				0.015	0.004							0.004			0.015	0.149				0.029		
g	0.23	0.055	0.037	0.241	0.04						0.003									0.006	0.174	0.046				0.171		
s	0.17	0.073	0.085	0.03	0.061		0.012	0.355	0.018	0.012	0.006	0.037	0.147															
f	0.387	0.153	0.189	0.081	0.027									0.143							0.01	0.019						
þ	0.325	0.134	0.134	0.081	0.069							0.021									0.025	0.215						
χ	0.514	0.081	0.081	0.162	0.135					0.054																		
h	0.47	0.243	0.13	0.096	0.061																							
v	0.307	0.173	0.067	0.12	0.013															0.181	0.014	0.139						
ð	0.121	0.379	0.034	0.155	0.034					0.018												0.275						
n	0.212	0.161	0.099	0.077	0.047	0.003		0.029	0.015	0.002	0.041	0.059		0.003	0.006		0.005			0.224	0.003	0.002				0.014		
l	0.303	0.16	0.132	0.124	0.071	0.002		0.01		0.008	0.012	0.002		0.008	0.017	0.021		0.021	0.002		0.094	0.006				0.01		
r	0.2	0.128	0.185	0.136	0.069	0.005				0.015	0.009	0.015		0.003	0.067	0.029		0.015	0.015	0.077	0.005	0.009			0.002	0.015		
L	0.222	0.167	0.222	0.278	0.111																							
R	0.318		0.136	0.182	0.364																							
m	0.282	0.167	0.216	0.088	0.048	0.004	0.032			0.027	0.005						0.005				0.032	0.027			0.072			
w	0.578	0.321	0.092								0.018																	
W	0.375	0.125	0.5																									
j	0.72			0.16	0.12																							

For the unigram and bigram entropies, one obtains:

	H_1	H_1/H_{max}	H_2	H_2/H_{max}
Sindarin data	4.111	0.855	3.051	0.635
Gusein-Zade $N = 28$	4.251	0.884		

Unfortunately, data from natural languages are hard to come by. For English, Shannon gives $H_1 = 4.14$, $H_1/H_{max} = 0.88$ and $H_2 = 3.56$, $H_2/H_{max} = 0.76$. However, this was calculated for the $N = 26$ Latin letters rather than for phonemes. Making a comparison nevertheless, one can see that the phonology of Sindarin is much more restricted, which makes sense.
H_2 is expected to be smaller than H_1 for any language (which is equivalent to the existence of phonotactics). To find a lower bound, languages like Japanese or Hawaiian are promising candidates.

5 Sources

To get a distribution by source, only unique entries were counted. Because of the ubiquitous conceptual changes by Tolkien, an editorial decision has to be made regarding what to count as unique.

For example, N. **naith** 'gore' (Ety:387), S. **neith, naith** 'angle' (PE17:55) and S. **naith** 'spearhead, gore, wedge, narrow promontory' (UT:282) were regarded as the same (polysemous) word, with various possible translations into English, and a joined reference (Ety:387, PE17:55, UT:282).
On the other hand, S. **eitha-** '1. prick with a sharp point, stab 2. treat with scorn, insult' (HEK-, WJ:365) and S. **eitha-** 'to ease, assist' (ATHA-, PE17:148) are clearly two different (homophonous) words, and are therefore kept separate. In this case it is obvious from their different etymologies.
There is a grey zone, however: For example, EN **baran** 'brown, swart, dark-brown' and S. **baran** 'brown, yellow-brown' suggest a conceptual change, albeit a small one, so that they were counted as separate entries, and thus also as different words for the statistics.

This gives the following absolute and relative counts (compare also the Hiswelóke charts [8]):

source	*count*	*rel. count*
Ety	1064	0.473
PE17	680	0.302
LotR	234	0.104
S	214	0.095
WJ	185	0.082
UT	89	0.04
VT42	84	0.037
VT45	74	0.033
PM	71	0.032
Letters	67	0.03
SD	60	0.027
RGEO	51	0.023
VT46	51	0.023
VT48	49	0.022
VT47	39	0.017
VT50	32	0.014
WR	31	0.014
VT44	29	0.013
MR	25	0.011
LB	20	0.009
RC	20	0.009
PE19	17	0.008
TC	17	0.008
VT41	14	0.006
TI	11	0.005
LR	10	0.004
PE18	8	0.004
RS	7	0.003
PE13	7	0.003
TAI	4	0.002
PE11	4	0.002
VT39	1	0.0
sum	3269	
unique entries total	2251	

Of course, a good amount of words is attested in various sources, so that the added count is higher than the actual entry count. The Venn diagram in figure 3 shows how words are shared across the two top sources (*The Etymologies* and *Parma Eldalamberon 17*) and the rest.

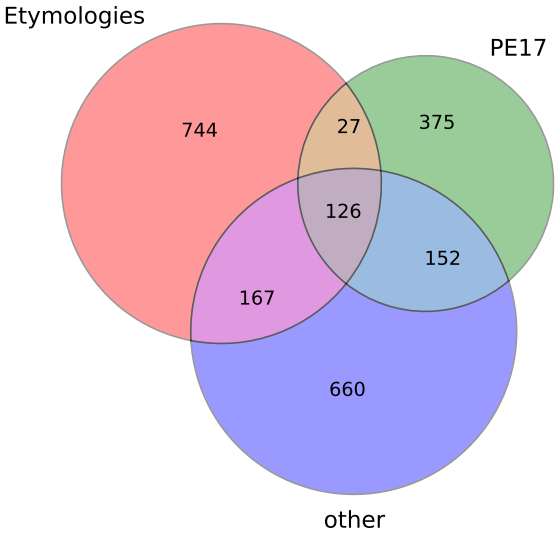

Figure 3: Sindarin vocabulary sources

References

[1] G. K. Zipf, *Relative frequency as a determinant of phonetic change*, Harvard studies in classical philology, Vol. 40 (1929), pp. 1-95

[2] С. М. Гусейн-Заде, О распределении букв русского языка по частоте встречаемости, Пробл. передачи информ. 24:4 (1988), 102–107

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

[4] W. Li, P. Miramontes, G. Cocho, *Fitting ranked linguistic data with two-parameter functions*, Entropy 2010, 12, 1743-1764

[5] J. Higgins, *RP phonemes in the Advanced Learner's Dictionary*, `http://myweb.tiscali.co.uk/wordscape/wordlist/phonfreq.html`

[6] C. E. Shannon, *A mathematical theory of communication*, The Bell system technical journal, 27, 379-423, 623-656 (1948)

[7] C. E. Shannon, *Prediction and entropy of printed English*, The Bell system technical journal, 30(1), 50-64 (1950)

[8] *Hiswelóke Sindarin dictionary* statistical charts `http://www.jrrvf.com/hisweloke/sindar/online/sindar/charts-sd-en.html`