Romanian Reports in Physics, Vol. 60, No. 3, P. 501–509, 2008

Dedicated to Prof. Ioan-Iovitz Popescu's 75th Anniversary

A GENERALIZATION OF THE GEOMETRIC DISTRIBUTION AND ITS APPLICATION IN QUANTITATIVE LINGUISTICS

JÁN MAČUTEK

Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia, e-mail: jmacutek@yahoo.com

(Received February 1, 2008)

Abstract. A new discrete distribution which is a generalization of the right truncated geometric distribution is presented. Its basic properties are studied. The distribution is applied to modelling rank frequencies of graphemes.

Key words: discrete probability distributions, geometric distribution, quantitative linguistics.

1. MOTIVATION

Modelling rank frequencies of graphemes (*i.e.*, models of grapheme frequencies ordered from the most frequent to the least frequent) has quite a long history in quantitative linguistics. One can find a short overview of previously suggested models in [3]. However, almost all those models are either language specific or they fit data from short texts only. The only one which is generally valid is the negative hypergeometric distribution (cf. [8, pp. 465–468]). But linguists still face problems in this field. An important drawback of the negative hypergeometric distribution is that, as it is shown *e.g.* in [6], it is derived from binary urn scheme considerations, *i.e.*, balls of two colours are drawn from an urn. The binarity – a situation which is unrealistic in grapheme frequencies modelling – makes it difficult (if possible at all) to interpret parameters, even if some regularities in parameters behaviour were reported in [1]. Therefore we offer an alternative model, a new discrete distribution which is not a result of a binary urn scheme. As far as we know, the distribution is presented for the first time, at least it is not included in the dictionary [8] containing 750 discrete distributions.

2. DEFINITION AND BASIC PROPERTIES

Consider the distribution

$$P_x = cp^{x-1}\left(1+\frac{a}{n-x+1}\right), \quad x = 1, 2, \ldots, n,$$

with the parameters $p \geq 0$ and $a \geq -1$, c being a normalization factor. Obviously, the right truncated geometric distribution (cf. [8, pp. 572–574]) is a special case of the distribution for $a = 0$.

Define $\Phi_n(z,s,v) = \sum_{j=0}^{n} \frac{z^j}{(v+j)^s}$, *i.e.*, $\Phi_n(z,s,v)$ is the incomplete Lerch function. We can express the normalization factor

$$c^{-1} = \sum_{j=1}^{n} p^{j-1} + a\sum_{j=1}^{n} \frac{p^{j-1}}{n-j+1} = \frac{1-p^n}{1-p} + ap^{n-1}\left(1 + \frac{p^{-1}}{2} + \frac{p^{-2}}{3} + \ldots + \frac{p^{-(n-1)}}{n}\right) =$$

$$= \frac{1-p^n}{1-p} + ap^{n-1}\Phi_{n-1}\left(p^{-1},1,1\right)$$

and the probability generating function

$$G(t) = \sum_{j=1}^{n} P_j t^j = ct\sum_{j=1}^{n} (pt)^{j-1} + cat\sum_{j=1}^{n} \frac{(pt)^{j-1}}{n-j+1} =$$

$$= ct\sum_{j=1}^{n} (pt)^{j-1} + cap^{n-1}t^n\left(1 + \frac{(pt)^{-1}}{2} + \frac{(pt)^{-2}}{3} + \ldots + \frac{(pt)^{-(n-1)}}{n}\right) =$$

$$= ct\left[\frac{1-(pt)^n}{1-pt} + a(pt)^{n-1}\,\Phi_{n-1}\left((pt)^{-1},1,1\right)\right].$$

Denote $\lceil x \rceil$ the smallest integer which is greater or equal x and define any sum with the lower limit greater than the upper one to be equal 0. The (cumulative) distribution function is

$$F(x) = P(X < x) = \sum_{j=1}^{\lceil x-1 \rceil} P_j = c\sum_{j=1}^{\lceil x-1 \rceil} p^{j-1} + ca\sum_{j=1}^{\lceil x-1 \rceil} \frac{p^{j-1}}{n-j+1} =$$

$$= c\frac{1-p^{\lceil x-1 \rceil}}{1-p} +$$

$$+cap^{n-1}\left[1 + \frac{p^{-1}}{2} + \frac{p^{-2}}{3} + \ldots + \frac{p^{-(n+1)}}{n} - \left(1 + \frac{p^{-1}}{2} + \frac{p^{-2}}{3} + \ldots + \frac{p^{-(n-\lceil x-1 \rceil-1)}}{n-\lceil x-1 \rceil}\right)\right] =$$

$$= c\frac{1-p^{\lceil x-1 \rceil}}{1-p} + cap^{n-1}\left[\Phi_{n-1}\left(p^{-1},1,1\right) - \Phi_{n-\lceil x-1 \rceil-1}\left(p^{-1},1,1\right)\right].$$

Next, we obtain the mean of the distribution

$$\mu = \sum_{j=1}^{n} jP_j = c\sum_{j=1}^{n} jp^{j-1} + ca\sum_{j=1}^{n} p^{j-1}\left(\frac{n+1}{n-j+1}-1\right) =$$
$$= c\left[\frac{1-(n+1)p^n}{1-p} + \frac{p(1-p^n)}{(1-p)^2}\right] +$$
$$+ca(n+1)p^{n-1}\left(1+\frac{p^{-1}}{2}+\frac{p^{-2}}{3}+\ldots+\frac{p^{-(n-1)}}{n}\right) - ca\frac{1-p^n}{1-p} =$$
$$= \frac{c}{1-p}\left[(1-p^n)\left(\frac{1}{1-p}-a\right) - np^n\right] + ca(n+1)p^{n-1}\Phi_{n-1}\left(p^{-1},1,1\right).$$

To obtain the variance we first derive

$$E\left(X^2\right) = \sum_{j=1}^{n} j^2 P_j = c\sum_{j=1}^{n} j^2 p^{j-1} + ca\sum_{j=1}^{n}\frac{j^2 p^{j-1}}{n-j+1}.$$

It holds

$$c\sum_{j=1}^{n} j^2 p^{j-1} = \frac{c}{(1-p)^3}\left[1+p-(n+1)^2 p^n + \left(2n^2+2n-1\right)p^{n+1} - n^2 p^{n+2}\right]$$

and

$$ca\sum_{j=1}^{n}\frac{j^2 p^{j-1}}{n-j+1} = ca\sum_{j=1}^{n} jp^{j-1}\left(\frac{n+1}{n-j+1}-1\right) =$$
$$= ca(n+1)\sum_{j=1}^{n} p^{j-1}\left(\frac{n+1}{n-j+1}-1\right) - ca\sum_{j=1}^{n} jp^{j-1} =$$
$$= ca(n+1)^2 p^{n-1}\Phi_{n-1}\left(p^{-1},1,1\right) - ca(n+1)\frac{1-p^n}{1-p} - \frac{ca}{(1-p)^2}\left[1-(n+1)p^n + np^{n+1}\right],$$

hence (using the well known identity $D(X) = E(X^2) - \mu^2$) we have

$$D(X) = \frac{c}{(1-p)^3}\left[1+p-(n+1)^2 p^n + \left(2n^2+2n-1\right)p^{n+1} - n^2 p^{n+2}\right] +$$
$$+ca(n+1)^2 p^{n-1}\Phi_{n-1}\left(p^{-1},1,1\right) - ca(n+1)\frac{1-p^n}{1-p} -$$
$$-\frac{ac}{(1-p)^2}\left[1-(n+1)p^n + np^{n+1}\right] - \mu^2.$$

Another distribution characteristic which is often used in quantitative linguistics as a measure of diversity is the repeat rate (known also as the Herfindahl index)

$$rr = \sum_{j=1}^{n} P_j^2 = c^2 \sum_{j=1}^{n} p^{2j-2} \left(1 + \frac{a}{n-j+1}\right)^2 =$$

$$= c^2 \sum_{j=1}^{n} p^{2j-2} + 2ac^2 \sum_{j=1}^{n} \frac{p^{2j-2}}{n-j+1} + (ac)^2 \sum_{j=1}^{n} \frac{p^{2j-2}}{(n-j+1)^2} =$$

$$= c^2 \frac{1-p^{2n}}{1-p^2} + 2ac^2 p^{2n-2} \left(1 + \frac{p^{-2}}{2} + \frac{\left(p^{-2}\right)^2}{3} + \ldots \frac{\left(p^{-2}\right)^{n-1}}{n}\right) +$$

$$+ (ac)^2 p^{2n-2} \left(1 + \frac{p^{-2}}{2^2} + \frac{\left(p^{-2}\right)^2}{3^2} + \ldots + \frac{\left(p^{-2}\right)^{n-1}}{n^2}\right) =$$

$$= c^2 \left[\frac{1-p^{2n}}{1-p^2} + 2ap^{2n-2}\Phi_{n-1}\left(p^{-2}, 1, 1\right) + a^2 p^{2n-2}\Phi_{n-1}\left(p^{-2}, 2, 1\right)\right].$$

Almost all discrete models in linguistics satisfy the general equation

$$\frac{P_x}{P_{x-1}} = 1 + a_0 + \sum_{i=1}^{k_1} \frac{a_{1i}}{\left(x - b_{1i}\right)^{c1}} + \sum_{i=1}^{k_2} \frac{a_{2i}}{\left(x - b_{2i}\right)^{c2}} + \ldots,$$

see [9] for its derivation. For the distribution introduced in this paper we have

$$\frac{P_x}{P_{x-1}} = p + \frac{-\frac{ap}{a+1}}{x-(n+1)} + \frac{\frac{ap}{a+1}}{x-(n+a+2)}$$

if $a \neq -1$, and

$$\frac{P_x}{P_{x-1}} = p + \frac{-p}{(x-n-1)^2}$$

if $a = -1$. Hence the distribution is a special case of the general model and it can be used within existing linguistic theories.

3. APPLICATION

The considered distribution will be applied to modelling rank frequencies of graphemes in four Slavic languages in this section. Observed rank frequencies in Russian, Slovak, Slovene and Ukrainian were published in [3–5] and [2], respectively. As the value of the Pearson χ^2 statistics increases approximately linearly with the sample size N (which is often hundreds of thousands or even more in linguistics), we use the discrepancy coefficient $C = \frac{\chi^2}{N}$ as the goodness of fit criterion. The value $C < 0.02$ (set empirically) indicates a good fit.

The parameter n is the inventory size (*i.e.*, the number of graphemes in a language), the other two parameters are estimated by the minimum χ^2 method (minimalization procedures in the statistical software *R* are used for computations). In the procedures, the initial value of the parameter p is the estimated value of the right truncated geometric distribution parameter obtained by the Altmann-Fitter (a software for fitting and estimating parameters of 200 discrete distributions). Finally, the initial value of the parameter a is determined as follows. It holds

$$\frac{P_{n-1}}{P_{n-2}} = \frac{3}{2} p \frac{a+2}{a+3}$$

and from the quotient $\frac{P_n}{P_{n-1}}$ we have

$$p = \frac{P_n}{2P_{n-1}} \frac{a+2}{a+1} .$$

We obtain the quadratic equation

$$a^2 \left(\frac{4P_{n-1}}{P_{n-2}} - \frac{3P_n}{P_{n-1}} \right) + a \left(\frac{16P_{n-1}}{P_{n-2}} - \frac{12P_n}{P_{n-1}} \right) + 12 \left(\frac{P_{n-1}}{P_{n-2}} - \frac{P_n}{P_{n-1}} \right) = 0 .$$

The initial value for the estimation of the parameter a is a solution (the one which is greater or equal –1) of a similar equation, where probabilities P_{n-2}, P_{n-1}, P_n are replaced with observed frequencies f_{n-2}, f_{n-1}, f_n.

Results are presented in the following four tables. The fit is satisfactory for Russian, Slovene and Ukrainian. For Slovak the discrepancy coefficient value slightly exceeds 0.02, but the sample size is relatively small (less than one half of other three sample sizes) which may be a reason for a worse fit. A new investigation for Slovak (and some more languages) will have to be done when new data are available. But tentatively we can accept the considered distribution as an alternative model for rank frequencies of graphemes.

Table 1

Grapheme frequencies in Russian

i	f(i)	NP(i)	i	f(i)	NP(i)	i	f(i)	NP(i)
1	982048	856864.36	**12**	272216	293857.56	**23**	101994	96871.16
2	763584	777822.41	**13**	262459	266360.80	**24**	93156	86940.07
3	701891	706021.88	**14**	248196	241371.58	**25**	77999	77768.03

(continues)

Table 1 (continued)

i	f(i)	NP(i)	i	f(i)	NP(i)	i	f(i)	NP(i)
4	593949	640799.23	**15**	222221	218657.40	**26**	69870	69225.61
5	563581	581551.54	**16**	195629	198006.45	**27**	54464	61157.94
6	532783	527730.98	**17**	181684	179225.62	**28**	30584	53351.25
7	456610	478839.75	**18**	163449	162138.49	**29**	24421	45450.26
8	423657	434425.48	**19**	156929	146583.53	**30**	22314	36711.69
9	403285	394077.08	**20**	151944	132412.29	**31**	9578	25019.94
10	353818	357420.91	**21**	146832	119487.47	**32**	2257	0.00
11	295548	324117.33	**22**	138459	107680.92			

$a = -1$ $N = 8697949$
$n = 32$ $\chi^2 = 79827.61$
$p = 0.9075$ $C = 0.0092$

Table 2

Grapheme frequencies in Slovak

i	f(i)	NP(i)	i	f(i)	NP(i)	i	f(i)	NP(i)
1	14194	13783.47	**17**	2676	2934.81	**33**	402	611.19
2	13772	12518.85	**18**	2660	2671.85	**34**	346	552.10
3	12701	11370.01	**19**	2408	2424.82	**35**	297	498.21
4	9285	10326.34	**20**	2262	2200.41	**36**	270	448.98
5	8323	9378.22	**21**	1954	1996.55	**37**	253	403.91
6	7099	8516.92	**22**	1825	1811.37	**38**	172	362.50
7	6562	7734.49	**23**	1685	1643.13	**39**	131	324.26
8	6534	7023.71	**24**	1611	1490.29	**40**	124	288.64
9	6164	6378.02	**25**	1593	1351.43	**41**	47	255.00
10	6091	5791.48	**26**	1465	1225.26	**42**	27	222.45
11	5731	5258.65	**27**	1422	1110.61	**43**	10	189.51
12	5659	4774.64	**28**	1395	1006.41	**44**	3	153.07
13	5103	4334.96	**29**	1294	911.71	**45**	2	104.32
14	4121	3935.56	**30**	1073	825.60	**46**	0	0.00
15	3845	3572.75	**31**	947	747.29			
16	3135	3243.18	**32**	719	676.05			

$a = -1$ $N = 147392$
$n = 46$ $\chi^2 = 3444.94$
$p = 0.9087$ $C = 0.0234$

Table 3

Grapheme frequencies in Slovene

i	f(i)	NP(i)	i	f(i)	NP(i)	i	f(i)	NP(i)
1	32036	34217.99	**10**	14043	13399.61	**19**	5055	4936.69
2	31891	30856.43	**11**	13034	12055.39	**20**	4608	4378.14
3	31122	27821.44	**12**	10517	10839.96	**21**	2606	3821.70
4	27150	25081.21	**13**	10514	9740.32	**22**	2554	3272.29
5	22905	22607.00	**14**	10216	8744.59	**23**	2463	2685.65
6	16088	20372.82	**15**	9568	7841.86	**24**	1675	1938.54
7	16084	18355.22	**16**	7446	7021.97	**25**	497	431.61
8	15221	16532.97	**17**	6413	6275.29			
9	14668	14866.89	**18**	5361	5592.43			

$a = -0.8594$ $N = 313735$
$n = 25$ $\chi^2 = 3498.97$
$p = 0.9031$ $C = 0.0112$

Table 4

Grapheme frequencies in Ukrainian

i	f(i)	NP(i)	i	f(i)	NP(i)	i	f(i)	NP(i)
1	37267	33319.42	12	13697	13488.71	23	5074	5283.15
2	32774	30703.82	13	12959	12413.77	24	4625	4824.45
3	25080	28291.74	14	12949	11421.83	25	3876	4395.13
4	24639	26067.30	15	12398	10506.32	26	3843	3990.73
5	21053	24015.89	16	10584	9661.12	27	3565	3605.92
6	20941	22124.00	17	8944	8880.58	28	2857	3233.71
7	20075	20379.18	18	8877	8159.45	29	2790	2863.46
8	19171	18769.95	19	7487	7492.82	30	2484	2476.18
9	16296	17285.70	20	6888	6876.12	31	2407	2030.25
10	16240	15916.65	21	6406	6305.00	32	506	1404.53
11	13936	14653.77	22	5850	5775.35	33	78	0.00

$a = -1$ $N = 386616$
$n = 33$ $\chi^2 = 2995.54$
$p = 0.9224$ $C = 0.0077$

4. FIRST CLASS MODIFICATION

Probability distributions must be sometimes modified in linguistic modelling. In some cases a variation of parameters itself cannot capture all factors (authors, genre, etc.) which have influences on data. Modifications of one or more classes are investigated in [10]. We limit ourselves to the first class modification here.

According to [10] define

$$Q_1 = 1 - \alpha(1 - P_1),$$
$$Q_x = \alpha P_x, \qquad x = 2, 3, \dots, n,$$

where $0 < \alpha < (1 - P_1)^{-1}$ and $\{P_x\}$ is the distribution defined in Section 2. Obviously $\{Q_x\}$ is a probability distribution.

We apply the modified distribution to modelling rank frequencies in Tamil. Our Tamil corpus consists of seven texts which were taken from [7]. As in the mentioned paper only grapheme relative frequencies rounded to three decimal places together with the sample sizes can be found, our reconstruction of grapheme frequencies may not be exact, however, possible tiny differences cannot be significant as far as goodness of fit is considered.

The proportion f_1 (Table 5) is apparently too high for the original (*i.e.*, non-modified) distribution to yield a good fit. We obtain the estimations $a = -1$, $n = 30$ and $p = 0.9067$, resulting in $\chi^2 = 2373.33$ and $C = 0.0300$.

Table 5

Grapheme frequencies in Tamil

i	f(i)	NP(i)	i	f(i)	NP(i)	i	f(i)	NP(i)
1	11462	12117.09	**11**	2902	2792.75	**21**	1087	993.54
2	6050	6853.24	**12**	2807	2525.17	**22**	970	889.72
3	5844	6205.90	**13**	2599	2282.51	**23**	899	794.11
4	5423	5619.17	**14**	2440	2062.39	**24**	755	705.32
5	5083	5087.37	**15**	2214	1862.66	**25**	674	621.75
6	4157	4605.34	**16**	2019	1681.37	**26**	623	541.19
7	3729	4168.41	**17**	1763	1516.72	**27**	420	460.03
8	3526	3772.35	**18**	1669	1367.07	**28**	382	370.76
9	3331	3416.32	**19**	1510	1230.92	**29**	201	252.13
10	319	3087.84	**20**	1329	1106.85	**30**	0	0.00
$a = -1$ $n = 30$ $p = 0.9176$ $\alpha = 0.9424$			$N = 78987$ $\chi^2 = 350.21$ $C = 0.0044$					

Parameters of the modified distribution are again estimated by the minimum χ^2 method. The initial values of the parameters a and p (denoted p_{in} in the next equation) are taken from fitting the original distribution, n is, as above, the inventory size. The initial value of the parameter α is obtained from the definition of Q_1, *i.e.*,

$$\alpha = \frac{1 - \frac{f_1}{N}}{1 - p_{in}}.$$

As can be seen in Table 5, the fit is substantially improved by the first class modification.

Acknowledgement. Supported by the grant VEGA 1/3016/06.

REFERENCES

1. P. Grzybek, *On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies*, Glottometrics, **15**, 82–91, 2007.
2. P. Grzybek, E. Kelih, *Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph*, G. Altmann, V. Levickij, V. Perebyinis, Eds., *Problems of Quantitative Linguistics,* Ruta, Chernivtsi, 2005, 159–179.
3. P. Grzybek, E. Kelih, G. Altmann, *Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: Modele der Häufigkeitsverteilung*, Anzeiger für Slavische Philologie, XXXII, 25–54, 2004.
4. P. Grzybek, E. Kelih, G. Altmann, *Graphemhäufigkeiten im Slovakischen. Teil II: Mit Digraphen*, R. Kozmová, Ed., *Sprache and Sprachen im mitteleuropäische Raum,* Filozofická fakulta UCM, Trnava, 2006, 641–664.
5. P. Grzybek, E. Kelih, E., Stadlober, *Graphemhäufigkeiten des Slovenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik*, Anzeiger für Slavische Philologie, XXXIV, 41–74, 2006.
6. N. L. Johnson, S. Kotz, *Urn Models and Their Applications: An Approach to Modern Discrete Probability Theory,* Wiley, New York, 1977.
7. G. Siromoney, *Entropy of Tamil prose*, Information and Control, **6**, 297–300, 1963.
8. G. Wimmer, G. Altmann, *Thesaurus of discrete uniariate probability distributions,* Stamm, Essen, 1999.
9. G. Wimmer, G. Altmann, *Unified derivation of some linguistic laws*, R. Köhler, G. Altmann, R. G. Piotrowski, Eds., *Quantitative Linguistics. An International Handbook,* de Gruyter, Berlin, 2005, 791–807.
10. G. Wimmer, V. Witkovský, G. Altmann, *Modification of probability distributions applied to word length research*, Journal of Quantitative Linguistics, **6**, 257–268, 1999.