Journal of Mathematical Modelling and Application
2013, Vol. 1, No. 8, 84-90
ISSN: 2178-2423

Mathematical modeling of the dependence of pattern of occurrence of graphemes in different texts with reference to the occurrence in various words in intra text perspective

Hemlata Pande
Dept. of Mathematics,
Kumaun University, S. S. J. Campus Almora
Uttarakhand, INDIA
hlpande@rediffmail.com

Abstract
Natural language texts have regularities for the pattern of various components of language and these regularities can be put in theoretical framework to identify several characteristic features of language and to determine the relation between different components by gathering the database for the components in different texts. In the present paper, the pattern of occurrence of different graphemes of English language has been investigated in intra text perspective. Graphemes have been studied for their incidence in different texts and in words of different length in these texts. A correlation has been determined between the two patterns and on the basis of the correlation; pattern of occurrence in any text has been represented in terms of their pattern of occurrence in different words of text in the form of linear equations. Results have been compared for the empirical and theoretical frequencies for the compiled text formed as the composition of all the considered texts.

Key words: Text, frequency, grapheme, model.

1.Introduction

Linguistics is the scientific study of natural languages and the applications of mathematical tools to study the language known as mathematical linguistics. In mathematical linguistics we work with the aim to elucidate and to illustrate the multitude of linguistic observations circling around us in the form of conversation, writing and other media. Hiemstra (2000) has mentioned that "Mathematical models are used in many scientific areas with the objective to understand and reason about some behaviour or phenomenon in the real world". In mathematical linguistics the process of generating mathematical models for the pattern of various components is quite familiar as in this context we can cite the works of Naranan and Balasubrahmanyan (1998) to investigate models of power law relations in linguistics and information science; of Trillo(2001) for presenting the mathematical model for the analysis of variation in discourse; of Kracht (2003) for discussing different mathematical models of language. A compendium of use of different mathematical models as Zipf law, Zipf Mandelbrot law, Poisson distribution, 2-Poisson distribution, Katz's K-mixture can be seen in the book of Manning and Schütze (1999).

A grapheme is the fundamental unit for processing in written language and in most cases it corresponds to the letter of alphabet of language. In the area of mathematical models for the graphemes or letters of different languages we can cite the works of Good (1969), for formulation of an equation for ranked distribution of grapheme and phoneme frequencies; Solso and King (1976) for examining frequency and versatility of letters in English language; Bell and Witten (1988) for presenting letter statistic for brown corpus and of Martindale et al (1996) for comparison of the rank frequency distributions of graphemes and phonemes; Grzybek and Kelih (2005) for expressing the relation of rank and frequency for grapheme frequencies of Slavic alphabets in the form of negative hypergeometric distribution etc., to mention only a few. Eftekhari (2006) has applied fractal geometric approach to texts by using frequencies of letters. Sanderson (2007) discovered the fact that distribution of letters in a text can potentially help in the process of language determination. Pande and Dhami (2009) have given a relation for the distribution of 26 graphemes of English language in a text in the form $F_r = a(r^2 + k_r r + c)^{-b}$, where F_r is the frequency corresponding to rank r; a, b ,c are

parameters and k_r takes different values in different layers of rank r, in 2010 we (Pande and Dhami, 2010) have presented the Yule's distribution as the model for frequencies of occurrence of different letters in a text and in the initial position of words of a text for Hindi language. All these studies of graphemes and letters have been considered in the inter text perspective i.e. for the pattern of different letters in any text. In this paper, I am making an attempt for their determination in intra text viewpoint i.e. how each letter occurs in different texts.

2.Method

We have initiated our work by taking 32 different texts available at different links given in the Appendix A (in parenthesis, the total number of non-numerical words in each text has been mentioned). The frequencies of occurrence of 26 graphemes (letter 'a' to letter 'z' of English language alphabet) in these 32 texts have been calculated. Here the occurrence of a letter as uppercase letter as well as lowercase letter has been considered. I have also calculated the frequencies of occurrence these graphemes in the words of lengths two, words of length three, length four and so on for each text separately. The determined database for a text "Adrift in New York: Tom and Florence Braving the World" written by H. Alger and containing 52,875 non-numerical words has been shown in the Appendix B. For each grapheme, we have determined the correlation between the frequencies of each grapheme in the whole text and in words of different lengths. Words of length one have not been considered as in each text almost all the one length words are 'A' and 'I'. Similarly words of length greater than 13 have not been considered due to their lower rate of occurrence in any text. As for the grapheme 'a', the determined values of correlation coefficient between the occurrence of grapheme 'a' in whole texts and in words of different lengths n, n=2, 3, 4.....13, have been listed in the Table 1.

Table 1. Correlation for occurrence of grapheme 'a' between its occurrence in whole text and words of length n

Length of word (n)	Correlation coefficient	Length of word (n)	Correlation coefficient
2	0.982939	8	0.96892
3	0.977497	9	0.953861
4	0.994032	10	0.931831
5	0.98755	11	0.923345
6	0.984889	12	0.829446
7	0.931939	13	0.752456

I have selected for each letter the length of the word for which the occurrence of the grapheme is highly correlated with its occurrence in whole text. In the case when the difference between correlation coefficient for words of length n and the highest correlation coefficient is less than 0.0055, the words of length n have also been taken into account. For example in the case of the grapheme 'a' the highest correlation coefficient is between the occurrence of 'a' in the whole texts and its occurrence in words of length 4 and there is no other word length for which the correlation coefficient differs with the highest coefficient 0.994032 by a number less than 0.0055 so for the occurrence of 'a' the length '4' has been taken into account. Similarly the case of each grapheme has been investigated separately. After selecting the lengths for each grapheme, an attempt has been made for expressing the frequency of occurrence of a grapheme in whole text in terms of its frequency of occurrence in the words of specific selected lengths. In the case when for a grapheme more than one length has been selected, total frequency of occurrence of grapheme in all the selected words has been taken into account.

Let the frequencies of occurrence of graphemes a, b, c.... have been represented by f_a, f_b, f_c and so on and f_{na}, f_{nb}, f_{nc} .... respectively represent the frequencies of occurrence of graphemes a, b, c.... in n length words, where n varies from 2 to 13, then the obtained results for the 26 graphemes have been represented in the Table 2. In this table the given values of the determination coefficient (R^2) have been determined by the formula given in equation (1) below:

Erro! (1)

Where the data set has *n* observed values y_i, i = 1, 2 … n each of which has an associated modeled value f_i and $\overline{y}$ is the mean of the observed data. SS_{tot} and SS_{err} represent the total sum of squares and the sum of squared error (or the residual sum of squares).

Erro! and **Erro!** (2)

Table 2. Dependence of frequencies of various graphemes in different texts with their frequencies in the words of selected lengths

Grapheme	Relation between the occurrence of grapheme in whole text and in words of selected length n	Determination coefficient for the relation
a	$f_a = 5.0367(f_{4a}) + 255.87$	0.9881
b	$f_b = 6.1803(f_{5b}) + 127.56$	0.9641
c	$f_c = 1.717(f_{5c} + f_{6c} + f_{7c} + f_{8c}) + 32.87$	0.9956
d	$f_d = 3.3383(f_{4d} + f_{5d}) + 191.49$	0.9905
e	$f_e = 3.6738(f_{5e} + f_{6e}) + 202.27$	0.9917
f	$f_f = 1.9425(f_{2f} + f_{4f} + f_{7f}) - 1.0651$	0.9937
g	$f_g = 2.3969(f_{5g} + f_{7g}) + 45.278$	0.9948
h	$f_h = 1.6907(f_{3h} + f_{4h}) + 90.636$	0.9967
i	$f_i = 2.0248(f_{2i} + f_{4i} + f_{7i} + f_{8i}) + 229.75$	0.9931
j	$f_j = 22.334(f_{7j}) - 71.364$	0.8378
k	$f_k = 1.4743(f_{4k} + f_{5k}) + 39.509$	0.9914
l	$f_l = 2.8409(f_{5l} + f_{6l}) - 4.2049$	0.9944
m	$f_m = 3.5252(f_{4m}) + 102.95$	0.9795
n	$f_n = 2.7038(f_{4n} + f_{6n} + f_{7n}) + 101.89$	0.9954
o	$f_o = 1.7532(f_{2o} + f_{4o} + f_{6o} + f_{7o}) + 55.193$	0.9958
p	$f_p = 7.3993(f_{8p}) + 30.817$	0.9888
q	$f_q = 2.4791(f_{5q}) + 20.349$	0.845
r	$f_r = 3.2636(f_{4r} + f_{6r}) + 73.969$	0.9926
s	$f_s = 2.8096(f_{4s} + f_{7s} + f_{8s}) + 112.63$	0.9972
t	$f_t = 1.837(f_{3t} + f_{4t} + f_{5t}) + 210.37$	0.9956
u	$f_u = 2.3169(f_{5a} + f_{6a} + f_{7a}) + 65.218$	0.9966
v	$f_v = 6.3813(f_{6v} + f_{9v}) + 21.385$	0.9813
w	$f_w = 1.9364(f_{4w} + f_{5w}) + 52.162$	0.9891
x	$f_x = 2.3868(f_{5x} + f_{7x} + f_{9x}) + 8.0974$	0.9788
y	$f_y = 3.0676(f_{4y} + f_{6y}) + 127.44$	0.956
z	$f_z = 4.8365(f_{7z}) + 7.433$	0.9109

The values of the determination coefficients are greater than 0.955 for all the graphemes(and these are greater than 0.988 for 18 graphemes out of 26 graphemes) except in the case of 'j', 'q' and 'z'. The graphemes j, q, x, z, are lower-proportion graphemes of English texts, as mentioned in Pande and Dhami(2009) and therefore their occurrence is not very regular in different texts. On the basis of the values of the determination coefficient it can be assumed that the frequencies of occurrence of graphemes in different texts of English language depend on their occurrence in certain words of specific lengths for which the equations have been mentioned in Table 2.

For example Figure 1 depicts the frequencies of occurrence of grapheme 'a' in the words of selected length 4 and corresponding frequencies of 'a' in the whole text the frequencies determined by the linear relation $f_a = 5.0367(f_{4a}) + 255.87$ along the straight line shown in the figure. The line closely represents the trend of increase of frequencies in text with increase of frequencies in selected length 4.

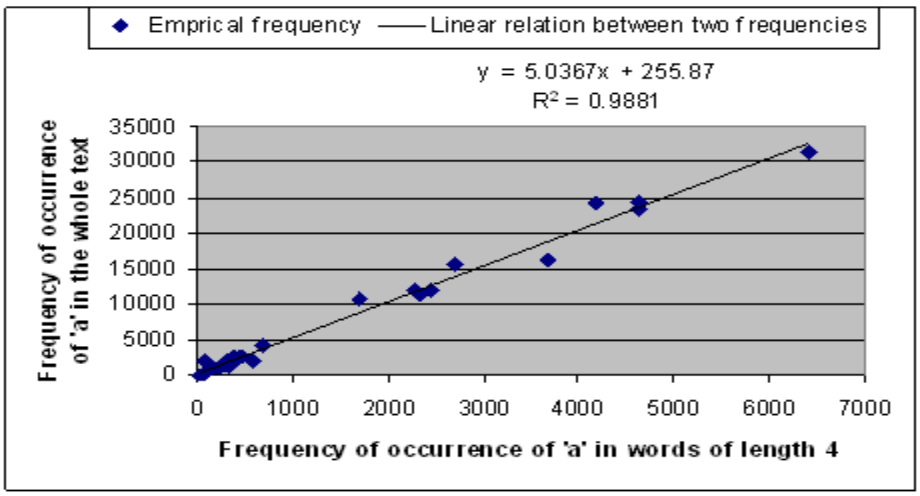

Figure 1. x-y plot of frequency of occurrence of 'a' in different texts and in words of length 4 in these texts and corresponding fitted linear relation

3.Results

After investigating the relation for each grapheme, for validation, the frequencies of each grapheme has been determined in the single text formed by the composition of all the considered texts containing 6, 28, 650 non numerical word tokens in all. The calculated values of the frequencies by applying the relations given in Table 1 and corresponding actual frequencies have been shown in the Figure2.

Figure 2 shows that the calculated values of the frequencies are almost the same as that of empirical ones, and the value of the discrepancy coefficient for the empirical and theoretical frequencies for all graphemes in the compiled text is 0.0012 which is much less than the value 0.01 required for very good fit.

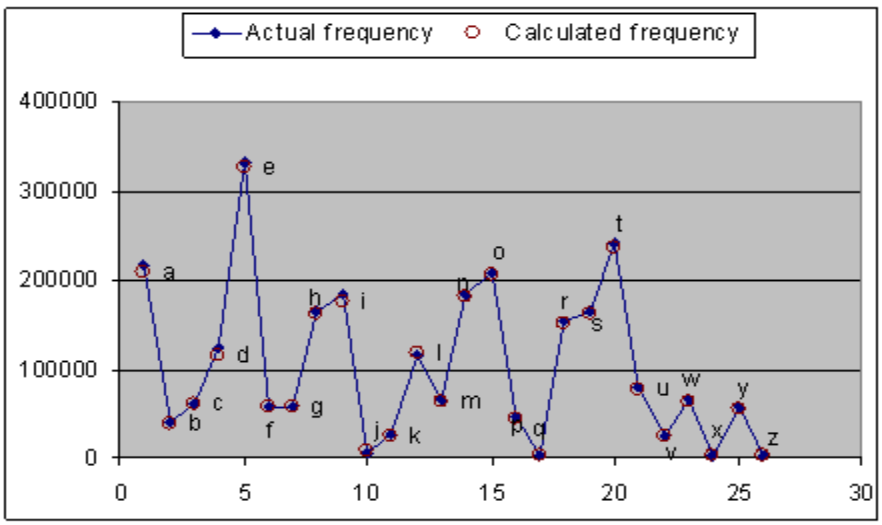

Figure 2. Calculated and actual frequencies of the graphemes in the composed text

4.Conclusion

On the basis of the above discussion, it has been concluded that the occurrence of various graphemes in different texts depends linearly on their occurrence in some specific words or in other way it can be said that the pattern of occurrence in the texts can be determined in terms of their pattern of occurrence in words of various lengths by linear relations. These equations define the occurrence of a grapheme in intra text viewpoint, and represent the characteristic equations of English language which shows that how the occurrence of various graphemes is given if any considerable amount of English language text has been selected.

References

Bell, T. C. and Witten, I. H. (1988). Source models for natural language, *Available at* <http://hdl.handle.net/1880/46172>.

Eftekhari, A. (2006). Fractal Geometry of Texts: An initial Application to the works of Shakespeare, *Journal of Quantitative Linguistics, Vol.13, No. 2-3*, 177-193.

Good, I. J. (1969) Statistics of language, *In A. R. Meetham and R. A. Hudson(Eds.)Encyclopaedia of linguistics , information and control*, Oxford: Pergamon, 567-581.

Grzybek, P. and Kelih, E. (2005). Towards a General Model of Grapheme Frequencies in Slavic Languages, *In: R. Garabík (Ed.), Computer Treatment of Slavic and East European Languages Bratislava: Veda* , 73-87.

Hiemstra, D. (2000). Using language models for information retrieval, *CTIT Ph.D. Thesis Series No. 01-32, Centre for Telematics and Information Technology Netherlands*. Available at: http://wwwhome.cs.utwente.nl/~hiemstra/papers/thesis.pdf

Kracht, M. (2003). *The Mathematics of Language*, Berlin: Mouton de Gruyter.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press.

Martindale, C., Gusein-Zade, S. M., Mekenzie, D. and Borodovsky, M. Y. (1996). Comparison of equations describing the ranked frequency distributions of graphemes and phonemes, *Journal of Quantitative Linguistics, Vol. 3, Issue 2*, 106-112.

Naranan, S. and Balasubrahmanyan, V. K. (1998) Models of Power Law Relations in Linguistics and Information Science, *Journal of Quantitative Linguistics, Vol. 5, Issue 1-2*, 35-61.

Pande, H. and Dhami, H. S. (2009). Generation of a model for grapheme frequencies and its refinement and validation by group theoretic aspects. *Journal of Quantitative Linguistics, Vol. 16, Issue 4*, 307-326.

Pande, H. and Dhami, H. S. (2010). Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language. *SKASE Journal of Theoretical Linguistics, Vol. 7, No. 2*, 19-38.

Pande H. and Dhami, H. S. (2012). Model generation for word length frequencies in texts with the application of Zipf's order approach. *Journal of Quantitative Linguistics, Vol. 19, Issue 4*, 249-261.

Sanderson, R. (2007). COMP527: Data Mining. *Available at* <*www.csc.liv.ac.uk/~azaroth/courses/current/comp527/lectures/comp527-28.pdf*>

Solso, R. L. and King, J. F.(1976). Frequency and versatility of letters in the English language, *Bahavior research methods and instrumentation, 8*, 283-286.

Trillo , J. R. (2001). A Mathematical Model for the Analysis of Variation in Discourse. *Journal of Linguistics, Vol. 37, No. 3*, 527-550

Appendix A

Analyzed texts; the quantities within parentheses are number of non numerical word tokens in the texts (These texts were also used in our previous study (Pande and Dhami, 2012).

Texts	**Author**
Firestorm 2034 (60,447)	J. Hayward
Bird Flu (33,781)	D. Meier
Both Sides of the Moon (66,872)	M. G. Kimber
The Statements of the famous (4,746), The determined Existentialist (4,809), The case of the dissatisfied voter (5,139)	A. Starling
The Bloody Sock and Other Tales (1422)	T. Deregowski
My Name's Jack(36,880)	J. C. Cripps
The House Of Silvery Voices(5,625), A Patroness Of Art(8,075), Plooie Of Our Square(7,570), Triumph(6,222), Average Jones(69,531)	S. H. Adams
Jack and Jill(94,202)	L. M. Alcott
Adrift in New York: Tom and Florence Braving the World(52,875)	H. Alger
"Party Cries" In Ireland(363), The $30,000 Bequest(11,051)	M. Twain
The Begging- Letter Writer(3,236), Our School(3,013)	C. Dickens
The White Linen Nurse(47,691), Little Eve Edgarton(30,781), The Blinded Lady(6,045)	E. H. Abbott
Christmas Revived(1,949)	A. W. Abbot
The Birthday of the Infanta(7,451)	O. Wilde
I Will!(3,056), For The Fun Of It (4,237), Haven't The Change(1,113), Children--A Family Scene(4,452), Coals of Fire(4,255), Thou Art The Man!(1,420), Brandy As A Preventive(5,319), All's For the Best(35,022)	T. S. Arthur

Appendix B

Frequencies of occurrence of graphemes in the text "Adrift in New York: Tom and Florence Braving the World" and in words of different lengths

Grapheme	Occurrence in whole text	Occurrence in words of length 2	Occurrence in words of length 3	Occurrence in words of length 4	Occurrence in words of length 5	Occurrence in words of length 6	Occurrence in words of length 7	Occurrence in words of length 8	Occurrence in words of length 9	Occurrence in words of length 10	Occurrence in words of length 11	Occurrence in words of length 12	Occurrence in words of length 12
a	16276	1017	3468	3677	1862	1475	1107	867	644	414	198	169	78
b	3178	506	664	343	351	503	293	294	96	57	39	25	6
c	5009	0	198	690	672	998	522	818	467	274	114	156	81
d	9841	348	1905	1846	1203	1869	918	790	447	279	116	73	39
e	26420	2016	4040	4998	3161	3874	2592	2752	1518	755	329	232	121
f	4411	1120	536	646	285	489	346	568	211	114	34	46	12
g	4180	95	217	457	746	927	801	402	302	113	61	32	19
h	12207	1049	4250	3092	1593	751	775	368	157	88	23	44	9
i	15395	2286	1444	2779	1337	1618	1331	981	759	443	269	232	104
j	280	0	6	160	18	34	24	9	17	11	1	0	0
k	2300	0	56	1032	569	374	129	77	22	28	8	2	3
l	9432	0	586	2641	1524	1627	1076	993	428	231	137	139	34
m	5906	1288	974	1501	560	440	343	249	280	139	55	33	34
n	14393	1302	2356	2384	1556	1829	1371	1554	922	500	272	200	116
o	17758	3711	3448	3294	1703	2086	1106	1156	571	322	149	155	36
p	3166	146	66	474	442	448	502	414	249	207	105	76	24
q	194	0	1	0	73	4	45	25	8	7	7	22	2
r	12158	257	1529	1839	1764	2350	1391	1560	731	375	161	111	65
s	13049	1172	2591	2212	1534	1620	1348	952	766	404	160	189	80
t	17690	2658	3387	4192	1848	1913	1231	855	727	415	200	165	69
u	7247	162	2136	833	1294	967	697	420	293	211	90	105	32
v	2003	0	44	836	306	224	245	124	122	42	27	20	11
w	5098	93	1247	2054	726	489	231	160	61	17	11	5	1
x	312	1	14	15	18	56	53	23	65	35	20	9	3
y	5827	517	2334	939	639	442	301	232	184	92	63	56	21
z	144	0	0	6	21	7	24	53	15	7	2	8	1