Annals of **F**uzzy **M**athematics and **I**nformatics
Volume 6, No. 1, (July 2013), pp. 141–149
ISSN: 2093–9310 (print version)
ISSN: 2287–6235 (electronic version)
http://www.afmi.or.kr

Application of fuzzy approach for the characterization of graphemes of English

Hemlata Pande

Received 12 June 2012; Revised 30 July 2012; Accepted 14 October 2012

Abstract. Present paper is an attempt in the direction to specify the properties of English language regarding the occurrence of graphemes of the language alphabet in various texts with the application of fuzzy membership function.

2010 AMS Classification: 97M80, 91F20

Keywords: Grapheme, English, Text, Membership.

Corresponding Author: Hemlata Pande (hlpande@rediffmail.com)

1. Introduction

The discipline of scientific study of natural languages by utilization of mathematical techniques and mathematical tools is known as Mathematical Linguistics. Different methods have been applied for the purpose, for instance, with reference to the methods in algebraic linguistics the works of Polkowska [19], Béchet et al [2], Lambek ([10, 11]), Casadio and Lambek (eds.) [4] and Lambek [12] can be cited and for the applications of various methods and models in quantitative or statistical linguistics, the books by Köhler et al (Eds.)[8], Köhler and Rieger (Eds.)[9] and Mohanty and Köhler (Eds.)[15] can be refered to mention only a few. An introduction to the statistical analysis of language can be found in the books of Manning and Schütze [13] and Baayen [1]. Applications of various mathematical and especially statistical methods to natural language processing have been extremely successful as these approaches often utilize large text corpora to build approximate generalized models of linguistic events. Construction of these models depends on the actual examples of the events supplied by the text corpora, without the accumulation of major linguistic knowledge. The large availability of text and speech corpora has cooperated crucially in the success of these approaches.

In typography, the elementary unit for processing of written language is grapheme and in most cases grapheme corresponds to letter. The graphemes of English language alphabet are twenty six letters of the alphabet. The frequencies of letters in text have often been studied for the exploitation in cryptography, in modern international morse code techniques and in data-compression techniques. Linotype machines are also based on letter frequencies of English language texts. The reference for these applications can be had from Nation Master- Encyclopedia ("Letter frequencies"51). For the analysis of natural languages on the basis of graphemes, we can cite the works of Good [6], for identifying the equation for ranked distribution of grapheme frequencies; Solso and King [21] for investigating the frequency and versatility of letters in English language; Bell and Witten [3] for presenting letter statistic for brown corpus; Grzybek and Kelih [7] for developing a theoretical model for grapheme frequencies of Slavic alphabets; Eftekhari [5] for concerning the fractal geometrical approach to letters of English language; Pande and Dhami([16, 17]) for the analyses of graphemes of English and letters of Hindi language respectively and of Martindale et al [14] for presenting the equations regarding the letter frequencies of different languages. Sanderson [20] has mentioned the fact that distribution of letters in a text can assist in the language determination process.

Regardless of the freedom to express one's views, a considerable amount of text of natural language follows some regularities concerning to the pattern of occurrence of different components and these regularities can be identified and can be expressed in the methodical form. In the present paper, I shall use the fuzzy membership function as a tool to represent the regularities of different English language texts for the pattern of occurrence of graphemes. My attempt is to show that a set of fuzzy membership functions can be defined which is helpful to identify the pattern of graphemes of English in different texts.

2. Methodology

For the present work, I have selected nine texts (mentioned from S. No. 1 to S. No. 9 in the appendix A) and have determined the normalized frequency of occurrence corresponding to each grapheme in these texts. Pande and Dhami [16] have defined the fuzzy set 'set of ranks close to equal percent' for graphemes of English language alphabet with the help of the fuzzy membership function $\chi(x)$. By applying a similar approach, I have defined three fuzzy membership functions $\chi_1(x)$, $\chi_2(x)$,$\chi_3(x)$ for the three sets: graphemes of higher proportion, graphemes of middle range proportion and graphemes of lower proportion respectively in the form of the following equations (2.1), (2.2) and (2.3):

$$\chi_1(x) = \begin{cases} 1 & \text{if } x \geq 0.073 \\ (x - 0.048)/0.025 & \text{if } 0.048 < x < 0.073 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

$$\chi_2(x) = \begin{cases} 0 & \text{if } x \geq 0.073 \text{ or } x \leq 0.005 \\ (0.073 - x)/0.025 & \text{if } 0.048 \leq x < 0.073 \\ (x - 0.005)/0.025 & \text{if } 0.005 < x \leq 0.03 \\ 1 & \text{if } 0.03 < x < 0.048 \end{cases} \tag{2.2}$$

$$\chi_3(x) = \begin{cases} 1 & \text{if } x \leq 0.005 \\ (0.03 - x)/0.025 & \text{if } 0.005 < x < 0.03 \\ 0 & \text{otherwise} \end{cases}, \tag{2.3}$$

where x represents the normalized frequencies of the graphemes and the value of x, for English language texts, is generally in the range from 0 to 0.14 (or 0.135 approximately) for each element in the set $\{\mathrm{a}, \mathrm{b}, \mathrm{c} \ldots \mathrm{z}\}$. These three sets of graphemes of higher proportion, middle range proportion and lower proportion have been denoted by A, B and C respectively.

Here the membership function $\chi_2(x)$ corresponding to the set B is same as the membership function $\chi(x)$ defined by Pande and Dhami [16] related to the 'set of ranks close to equal percent' except for the facts that, in the present case it is defined corresponding to graphemes instead of ranks and in this instance the variable x has been taken as the proportion of graphemes while % proportion has been used in Pande and Dhami [16]. This membership function was determined by us by assumption that if the occurrence of each grapheme is equally probable then each grapheme should form 3.85 % part of a text in terms of total number of graphemes in text. This membership function has been selected in Pande and Dhami [16], by the study of proportion of graphemes in various texts, in such a manner that for the elements of the rings G_1, G_2 and G_5 (defined in Pande and Dhami [16]) its value is zero for different texts while in case of field G_3 the value is in range from 0 and 0.97 and for field G_4 between 0.02 and 1.

For the purpose to define the remaining two functions $\chi_1(x)$ and $\chi_3(x)$ corresponding to the sets of graphemes of higher and lower proportion, I in the present study have assumed that if a grapheme has such a proportion in a text that is not close to the fraction of equal proportion then it is either highly occurred in text or its rate of occurrence is lower in it. In the case of the function $\chi_3(x)$, I have chosen the membership function such that the graphemes of the ring G_5 (mentioned in Pande and Dhami [16]) always have membership value 1 regarding to the function. If the limit of x is slightly varied for example to put at 0.0055 or 0.0045 in equation (2.3) in place of 0.005 then it will not very much effect the procedure, in that case also members of G_5 will have full membership regarding to the function and the same process can be applied and characterization of the graphemes (whatever be in that case) can be defined and can be verified. Virtually the goal of this paper is to determine a sort of fuzzy membership functions and to apply them for graphemes' classification. It is not claimed that these are the only functions for which the graphemes follow the hypotheses a) to e), defined later in this section, but it is tried to show that 'for English language, this set of functions (defined in equations (2.1), (2.2) and (2.3)) is such that it characterizes graphemes according to the hypotheses'. Similarly the function $\chi_1(x)$ selected in such a manner, that graphemes of the rings G_1 and G_2 (mentioned in Pande and Dhami [16]) always have membership value 1 regarding to the function. For the text 'Bird Flu' these paradigms of fuzzy sets of graphemes are as mentioned in equations (2.4), (2.5) and (2.6):

$$A = \{(\mathrm{a}, 1), (\mathrm{e}, 1), (\mathrm{h}, 0.4459), (\mathrm{i}, 0.788791), (\mathrm{n}, 0.84121), (\mathrm{o}, 1), (\mathrm{r}, 0.462659), (\mathrm{s}, 0.354421), (\mathrm{t}, 1)\} \tag{2.4}$$

(2.5)
$$B = \{(\text{b}, 0.384543), (\text{c}, 0.698491), (\text{d}, 1), (\text{f}, 0.541801), (\text{g}, 0.659673), (\text{h}, 0.554058), (\text{i}, 0.211209), (\text{k}, 0.293023), (\text{l}, 1), (\text{m}, 0.851215), (\text{n}, 0.15879), (\text{p}, 0.459064), (\text{r}, 0.537341), (\text{s}, 0.645579), (\text{u}, 0.945569), (\text{v}, 0.126698), (\text{w}, 0.981554), (\text{y}, 0.60357)\}$$

(2.6)
$$C = \{(\text{b}, 0.615457), (\text{c}, 0.301509), (\text{f}, 0.458199), (\text{g}, 0.340327), (\text{j}, 1), (\text{k}, 0.706977), (\text{m}, 0.148785), (\text{p}, 0.540936), (\text{q}, 1), (\text{u}, 0.054431), (\text{v}, 0.873302), (\text{w}, 0.018446), (\text{x}, 1), (\text{y}, 0.39643), (\text{z}, 1)\},$$

where the graphemes with zero membership value have not been included in the sets. Graphs of the membership values of the graphemes when the graphemes are arranged in descending order of their frequencies of occurrence have been shown in the FIGURE 1:

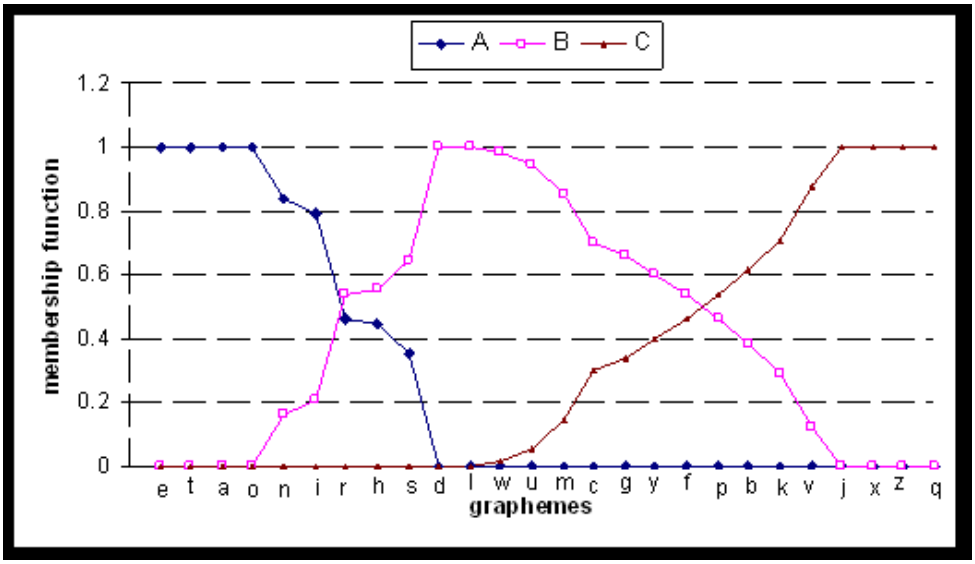

FIGURE 1. Membership values of graphemes in the three sets A, B and C.

After determination of the membership values for the three sets, I have determined the sets $A \cap B$,$B \cap C$, $A \cup B$ and $B \cup C$ (by standard union and standard intersection method of fuzzy sets) for each of the nine texts and subsequently to check the general pattern of occurrence of each grapheme in various English language texts, I

have determined the average values of the membership function for each grapheme corresponding to the seven sets : A, B, C, $A \cap B$,$B \cap C$, $A \cup B$ and $B \cup C$. In order to determine the characteristics of different graphemes, graphemes having the average membership values greater than 0.5 for any set A, B or C and less than 0.35 for remaining two sets have been checked. On the basis of the average values of the membership functions, I have characterized each grapheme as higher proportion, middle range proportion or lower proportion. 21 graphemes have been classified according to the hypotheses:

a). The graphemes 'a', 'e', 'i', 'n', 'o', 't' belong to the set A with the value of membership function $\chi_1(x)$ greater than 0.5.
b). The graphemes 'c', 'd', 'g', 'l', 'm', 'r', 'u', 'w', 'y' belong to the set B with the value of membership function $\chi_2(x)$ greater than 0.5.
c). Graphemes 'j', 'k', 'q', 'v', 'x', 'z' belong to the set C with value of membership function $\chi_3(x)$ greater than 0.5.

The remaining graphemes, b, p, f, h and s are such that these belong to $B \cup C$ or $A \cup B$ with the average value of the membership functions at most 0.637 and belong to $B \cap C$ or $A \cap B$ with the average membership value at least 0.36. For these graphemes, I have used the following hypotheses:

d). Graphemes 'b', 'f'and 'p' belong to the set B or to C with membership value greater than 0.5.
e). Graphemes 'h' and 's' belong to the set A or to B with membership value greater than 0.5.

Thus, on the basis of this arrangement it can be stated that there is a clear classification of graphemes of English language: out of total 26 graphemes, 21 graphemes belong to any of the sets A, B or C (discussed above) with a membership value greater than 0.5 and the remaining graphemes 'b', 'f', 'p', 'h' and 's' are such that the first three graphemes out of these belong to the set B or to C and last two graphemes to the set A or to B with membership values greater than 0.5.
The graphemes a to z not only occur in English but some other languages also use these in their writing system, however their rates of presence in different languages' texts are not same, as mentioned by Sanderson [20]. The technique, discussed in this paper, can be used to determine the similarities and dissimilarities of such kinds of diverse languages for the proportion of graphemes $\{\mathrm{a}, \mathrm{b}, \mathrm{c} \ldots \mathrm{z}\}$. The formation of the functions $\chi_1(x)$ and $\chi_3(x)$ is such that the hypotheses a) and c) not only affirm the graphemes belonging to the sets A and C but these will also confirm the not belonging of these graphemes to C and A respectively with a membership value greater than 0.5. If such occurrence is happened in a text then it shows a great dissimilarity from the pattern of graphemes in English language texts and in such a situation the chances of the text to be of language other than English are higher.

3. Validation of the classification

After classification of the graphemes in different sets, for validation, I have checked their nature in different texts. First the nature of the graphemes have been checked in the above mentioned nine texts (S.No. 1 to S. No. 9 in the appendix A) which

contain 47,877 to 3,84,945 graphemes of English alphabet in all. For all the nine texts, all graphemes occur in the texts according to the above defined hypotheses except for the grapheme 'r' in the texts 'Average Jones' and 'Little Eve Edgarton'. For example, in the case of the text 'Average Jones' the values of fuzzy membership functions which are greater than 0.5 have been shown in the table below (TABLE 1). The table dipicts that for this text all graphemes belong to the sets A, B or C

TABLE 1. Membership values greater than 0.5 corresponding to the three fuzzy functions

Grapheme	$\chi_1(x)$	$\chi_2(x)$	$\chi_3(x)$	Hypotheses52	Grapheme	$\chi_1(x)$	$\chi_2(x)$	$\chi_3(x)$	Hypotheses52
a	1			T	n	0.809637			T
b			0.613877	T	o	1			T
c		0.790296		T	p		0.564945		T
d		1		T	q			1	T
e	1			T	r	0.536973			F
f		0.69111		T	s		0.503562		T
g		0.629331		T	t	1			T
h		0.597795		T	u		0.971855		T
i	0.81485			T	v			0.744735	T
j			1	T	w		0.660351		T
k			0.831278	T	x			1	T
l		1		T	y		0.631807		T
m		0.824836		T	z			1	T

according to the hypotheses except one grapheme 'r'. So the accuracy rate of the classification is 96.15%.

I have applied the same approach to ten texts containing 5,972 to 3, 25, 710 graphemes in all, which have been used by Pande and Dhami [16] for the analysis, mentioned in S. No. 10 of the appendix A, and to the a corpora formed by the compilation of different 32 texts (texts used by Pande and Dhami [[18]] for study of word length frequencies) having total 26, 64, 572 graphemes. The classification of all the graphemes has found correct in these cases also except for grapheme 'g' in the text 'The Dissatisfied Voter' and for grapheme 'y' in the text 'Slide of Passions' (mentioned in S. No. 10 of appendix A). Thus the characterization hypotheses determined in the paper can be taken as general hypotheses for the classification of graphemes of English in different sets.

4. Discussions and further work

In this paper, I have characterized the graphemes of English language alphabet for their occurrence in different texts on the basis of their membership in fuzzy sets

of higher proportion, middle range proportion and lower proportion by utilizing the fuzzy membership functions specified in the equations (2.1), (2.2) and (2.3). This kind of tactic can be applied to different types of linguistic components and for various languages. This approach is an innovative approach which depends solely on the membership values of the graphemes. Previously for the analyses of graphemes, the rank frequency approach has been selected by researchers.

References

[1] R. H. Baayen, Analyzing Linguistic Data: A Practical Introduction to Statistics Using R, Cambridge University Press, Cambridge, 2008.

[2] D. Béchet, A. Dikovsky, A. Foret and E. Garel, Optional and Iterated Types for Pregroup Grammars, In C. Martin-Vide, F. Otto and H. Fernau (Eds.), Proc. of the 2d Intern. Conf. on Language and Automata Theory and Applications (LATA 2008). Tarragona, Spain, March 13-19, 2008. LNCS 5196, pp. 88–100, Springer.

[3] T. C. Bell and I. H. Witten, Source models for natural language, Retrieved February 1, 2009, from http://hdl.handle.net/1880/46172, 1988.

[4] C. Casadio and J. Lambek (Eds.), Computational Algebraic Approaches to Natural Language, Polimetrica International Scientific Publisher, Monza/Italy, 2008.

[5] A. Eftekhari , Fractal geometry of texts: An initial application to the works of Shakespeare, Journal of Quantitative Linguistics 13(2-3) (2006) 177–193.

[6] I. J. Good , Statistics of language, In A. R. Meetham and R. A. Hudson (Eds), Encyclopedia of Linguistics, Information and Control (pp. 567–581). Oxford: Pergamon, 1969.

[7] P. Grzybek and E. Kelih, Towards a general model of grapheme frequencies in Slavic languages, In R. Garabík (Ed.), Computer Treatment of Slavic and East European Languages, pp. 73–87, Bratislava: Veda, 2005.

[8] R. Köhler, G. Altmann and R. G. Piotrowski (Eds.), Quantitative Linguistik / Quantitative Linguistics, Mouton de Gruyter, 2005.

[9] R. Köhler and B. B. Rieger (Eds.), Contributions to Quantitative Linguistics, Springer, 1993.

[10] J. Lambek, A computational algebraic approach to English grammar, Syntax 7(2) (2004) 128–147.

[11] J. Lambek, From word to sentence: a pregroup analysis of the object pronoun who(m), J. Log. Lang. Inf. 16(3) (2007) 303–323.

[12] J. Lambek, From Word to Sentence: A Computational Algebraic Approach to Grammar, Polimetrica, 2008.

[13] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, 1999.

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

[15] P. Mohanty and R. Köhler (Eds.), Readings in Quantitative Linguistics, New Delhi: Indian Institute of Language Studies, 2008.

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

[17] H. Pande and H. S. Dhami, Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language, SKASE Journal of Theoretical Linguistics 7(2) (2010) 19–38.

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

[19] M. S. Polkowska, The theory of configurations in algebraic linguistics, Int. J. Comput. Math. 14(3 and 4) (1983) 239–257.

[20] R. Sanderson, COMP527: Data Mining, http://cgi.csc.liv.ac.uk/~azaroth/courses/current/comp527/lectures/comp527-28.pdf. (2008) .
[21] R. L. Solso and J. F. King, Frequency and versatility of letters in the English language, Behavior Research Methods and Instrumentation. 8 (1976) 283–286.

5. Appendix A

Different texts used for study

S. No.	Text	Source/Author
1	Bird Flu	www.blueunicornpublishing.com/
2	My names Jack	www.trivigo.com/MY%20NAME'S%20JACK[1].html
3	Both Sides of Moon	http://www.thenightwriter.co.uk/page42.html
4	Jack and Jill	By 'L. M. Alcott'
5	Firestorm 2034	http://jonathanscorner.com /writings/firestorm/firestorm. html
6	The $30,000 Bequest	By 'Mark Twain'
7	Average Jones	By 'Samuel H. Adams'
8	Little Eve Edgarton	By 'Eleanor H. Abbott'
9	All's For the Best	By 'T S Arthur'
10	(a) A Few Quite days (b)Susan (c)A Half Life Of One (d)Hamlet (e)The Dissatisfied Voter (f)Blue Seaweed (g)The statements (h)The Bloody Sock and Other Tales (i)A Cord of seven Strand (j)Slide of Passions	These texts have also been used by Pande and Dhami [16]

Hemlata Pande (hlpande@rediffmail.com)
Department of Mathematics, Kumaun University, S. S. J. Campus Almora, Almora-263601, Uttarakhand, INDIA

1Letter Frequencies. Nation Master- Encyclopedia. Retrieved 17 Oct. 2008 http://www.nationmaster.com/encyclopedia/Letter-frequencies

2 T=True, F=False