

**MJAL2:1JANUARY2010****ISSN 0974-8741****Distributions of different parts of speech in different parts of a text and in different texts****by Hemlata Pande and H. S. Dhama****Distributions of different parts of speech in different parts of a text and  
in different texts****Hemlata Pande and H. S. Dhama (Email: drhsdhami@yahoo.com)****Dept. of Mathematics,  
K.U. S. S. J. Campus, Almora,  
Uttarakhand (INDIA)****Abstract**

The present work is an attempt in the direction of predicting sequences of building blocks of communication, the parts of speech, in any text. Five novels and five short stories of English language have been tagged under four different treatments and then statistical distributions of the parts of speech have been obtained. The goodness of fit of the distributions has been testified by  $\chi^2$  -test while the validation of the four different treatments has been done by the application of random block design and completely randomized design of experiments, under the assumption that the mean of distribution scores for applied treatments is same for different parts and sub-parts of the POS. The mean scores of distributions of different POS have been also compared by random block design.



### **Introduction**

POS tagging is the process of marking up the words based on its definition as well as its context. When either statistical or rule based methods are applied to this arena, automatic POS tagging comes under the purview of natural language processing. POS tagging can be used in partial parsing, which refers to various levels of details of syntactic analysis. Tagging and partial parsing, taken together, can be used for improving the performance of information retrieval. Brill (1992) has described a rule based tagger which has many advantages over stochastic taggers, like vast reduction in stored information and perspicuity of a small set of meaningful rules as opposed to the large table of statistics needed for stochastic taggers.

The problem of identifying the correct parse—the parse that humans perceive—among the possible parses is a central application of stochastic grammars in computational linguistics. POS tagging has been applied, in the recognition of table of contents by Belaid et al (2000), in genre classification by Y. Kim and S. Ross (2006) and in sentiment classification by Pang and Liltian (2002). A novel statistical model for speech recognition and POS tagging has been generated by Yuan and Chen (2006). On the basis of acquired knowledge from the book of Balakrishnan and Nevzorov (2004) and Krishnamoorthy (2006), we propose to fit theoretical distributions to the sequence of parts of speech in a text as well as in different texts in order to know how closely the observed sequences of parts of speech approximate to theoretical statistical distributions.

Guilpin & Guilpin (2005) have used Poisson distribution (by assuming that if a random variable follows Poisson law then the events follow uniform law and applying Kolmogorov criterion) as distribution of part of speech in a text and have validated them by their application to the study of two Greek texts. Stevenson and Paolo Merlo (2000)



**MJAL2:1JANUARY2010**

**ISSN 0974-8741**

**Distributions of different parts of speech in different parts of a text and in different texts**

**by Hemlata Pande and H. S. Dhani**

have classified verbs and lexical semantic classes based on indicators of verb alternations and in a later work; they (2001) have presented a report on supervised learning experiments to automatically classify three major types of English verbs. We have used Binomial, Poisson, 2-Poisson, Katz (1996) K-mixture as models to find the distribution

of various POS. The last three distributions have been suggested as models for information retrieval in the book of Manning (1999). The validation work has been accomplished by the applications of  $\chi^2$  test and Design of experiments.

### **Modus operandi**

In order to determine distributions of parts of speech we have tagged ten corpora by Go-tagger<sup>1</sup>, using the rule files of Brill (1992), under following cases-

1. The novel 'A Few Quite Days' has been taken as a combination of 139 documents each of 500 word tokens, where the first document corresponds to first 500 words, II to subsequent 500 words and so on.
2. The novel 'Bird Flu' has been taken as a combination of 62 documents each of 500 word tokens.
3. We have considered the novel 'A Few Quite Days' as a combination of 69 documents, of 1000 word tokens each.
4. In this case different chapters and pages of 5 novels and 5 short stories (links have been mentioned in appendix) have been selected in the form of different documents as depicted in following table –

---

<sup>1</sup> <http://uluru.lang.osaka-u.ac.jp/~k-goto/download/download.cgi?name=GoTagger07.zip>



Table 1-Documents selected for treatment IV

Text	Documents	Number of documents
Fire Storm	Acknowledgement, Epilogue & 72 Chapters	74
Bird Flu	Prologue & 28 Chapters	29
My name's Jack	19 Episodes	21
A half Life Of one	19 Chapters	29
Few Quite Days	27 Pages	27
Short Stories		5
Total		177 documents
Minimum document length =154word tokens, Maximum length of document =6669 word tokens, Average length of documents = 1701 word tokens		

In order to find the distribution of POS corresponding to above-mentioned four cases, we have applied four treatments. In first three treatments,  $x$  has been taken as the frequency of occurrence of a POS in a document and  $f(x)$  as the number of documents in which it occur  $x$  times, then we can infer about the occurrences of different POS for our corpora under study. One such result has been demonstrated with the help of graph in figure 1.

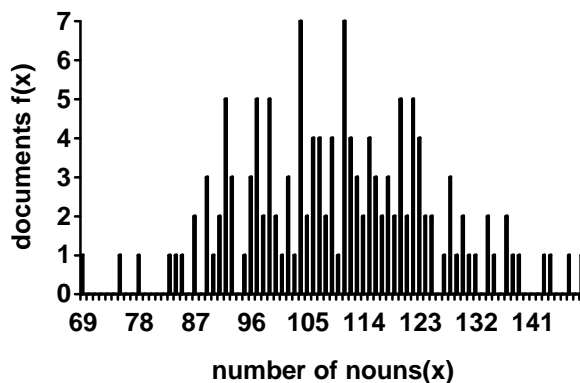


Figure 1. Number of documents in accordance with number of nouns.

It has been observed that range of  $x$  is large for almost all POS, for example in case of noun it is 80. It is better if we take class intervals by considering the mid points and frequencies of class intervals. This explanation results in classification of 139 documents in the form of  $p$  class intervals of equal width by the formula  $\log_2 N+1 = p$  as elucidated in the following figure 2.

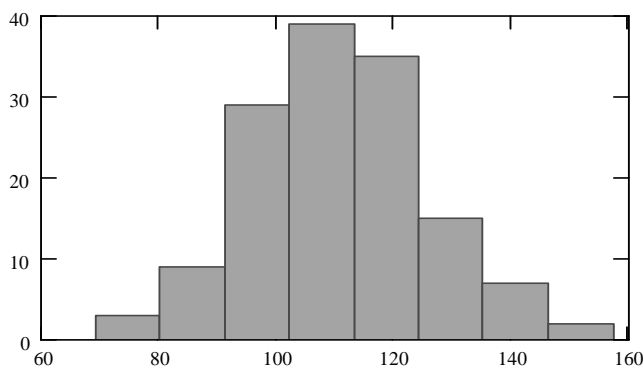


Figure 2. Histograms for number of nouns in A Few Quite Days.



We have also observed that the number of occurrence of a POS in a document increases with the increase of length of the document. Therefore in treatment IV, we have tested the distribution of 'part of speech average word token length of documents' by taking  $x = \frac{P}{n} A$ , and have classified the data in the form of class intervals for  $x$ , where  $P$  is number of the POS in a text,  $n$  length of the text in word tokens and  $A$  has been obtained equal to 1701, which is the average length of documents.

### Distributions of POS tags in texts

If  $a_i$ ,  $i = 1, 2, \dots, p$  (8 here) be the mid point of  $i^{\text{th}}$  class interval and  $h$  be the common width of classes then we can take a new variable  $y = (a_i - a_1)/h$ , where  $i = 1, 2, \dots, p$  and if corresponding frequency of class is denoted by  $F(y)$  then in order to find the distribution of  $F(y)$ , we have tried following four distributions -

#### 1. Binomial

$$F(k) = N {}^a C_k p^k q^{(a-k)}, a \text{ and } p \text{ are parameters and } q = 1 - p. k = 0, 1, 2, \dots, a \quad \dots (1)$$

#### 2. Poisson

$$F(k) = N e^{-a} \frac{a^k}{\text{fact}(k)} \quad a \text{ is parameter. } k = 0, 1, \dots \quad \dots (2)$$

#### 3. 2-Poisson

$$F(k) = N \left[ \alpha e^{-a} \frac{a^k}{\text{fact}(k)} + (1 - \alpha) e^{-b} \frac{b^k}{\text{fact}(k)} \right], \alpha, a, b \text{ are parameters} \quad \dots (3)$$

The parameters for Binomial, Poisson, and 2-Poisson have been determined by the 'method of moments' for the frequencies of respective distributions. In case of Katz K mixture we have made the use of the following formula



## 4. Katz's K mixture

$$F(k) = N \left[ (1 - \alpha)\delta + \frac{\alpha}{\beta + 1} \left( \frac{\beta}{\beta + 1} \right)^k \right] \quad \left. \begin{array}{l} \text{where } \delta = 1 \text{ if } k = 0 \text{ otherwise } \delta = 0 \text{ and} \\ \alpha \text{ and } \beta \text{ are parameters and are given as} \end{array} \right\} \dots(4)$$

$$\beta = \frac{N(\lambda - 1) + F(0)}{N - F(0)}, \quad \alpha = \frac{\lambda}{\beta} \quad \left. \begin{array}{l} \lambda \text{ is observed mean and } F(0) \text{ observed frequency} \\ \text{corresponding to } k = 0 \end{array} \right\}$$

The goodness of fit of these four distributions has been tested by  $\chi^2$  test (with  $n-k-1$  degree of freedom, where  $n$  is total number of classes with at least frequency 5 and  $k$  is total number of parameters determined, ) at the 5% level of significance. In case when 5% level of significance is not applicable we have used 1% level of significance and if the number of total classes are less than or equal to  $(k + 1)$ , distributions with theoretical probability and empirical probabilities of different classes differ only by small amount less than 0.05 have been used (in case of proper noun plural total classes with frequency  $>5$  are 3 and for K-Mixture  $k=2$ ). Figures 3, 4, 5 and 6 represent the appropriate fitted distributions for the four respective treatments I II, III and IV.



Abbreviating the four distributions as B for Binomial, P for Poisson, 2P for 2-Poisson and K for K-Mixture, we have -

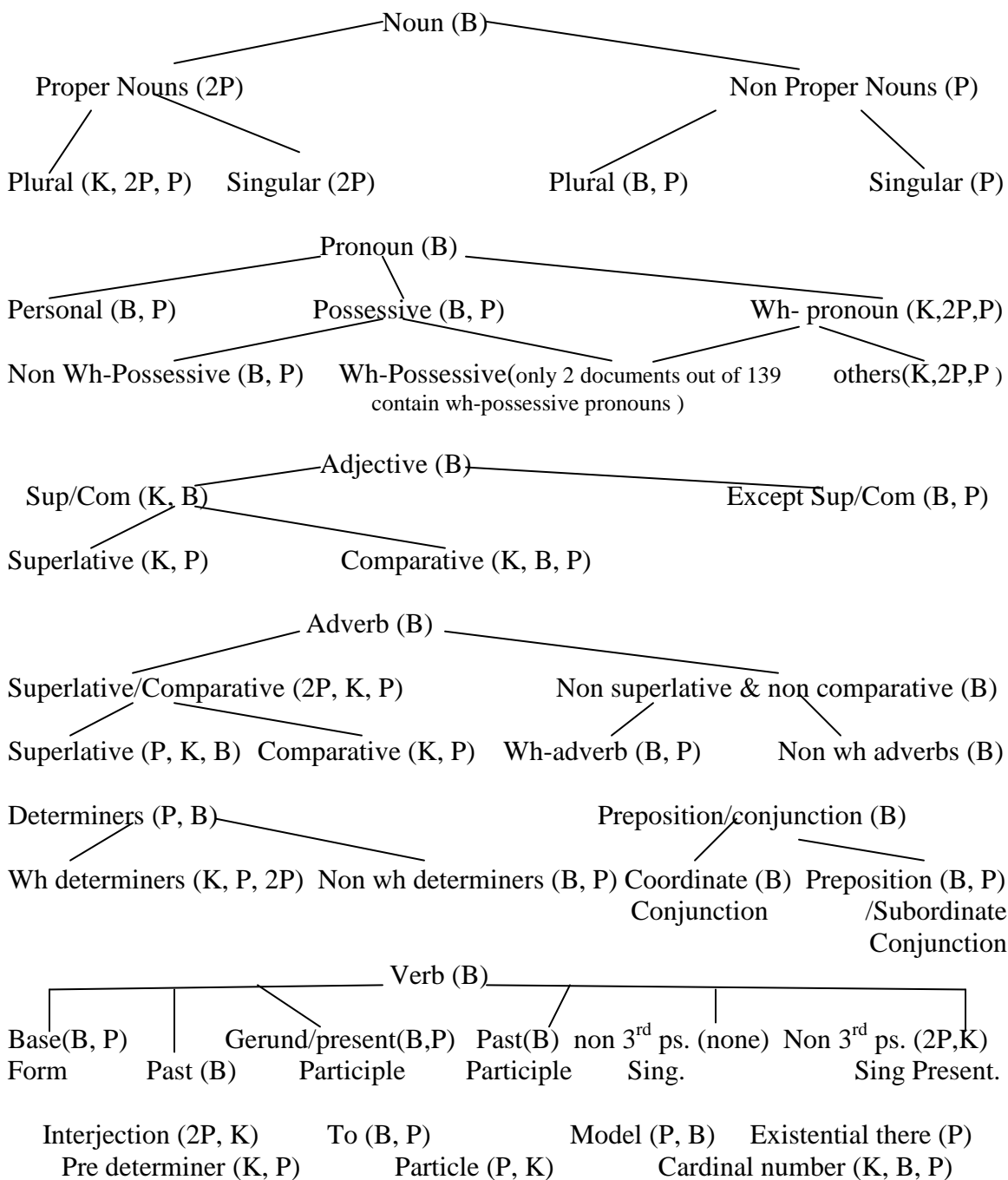




Figure 3. Distributions obtained for different parts of speech under treatment I

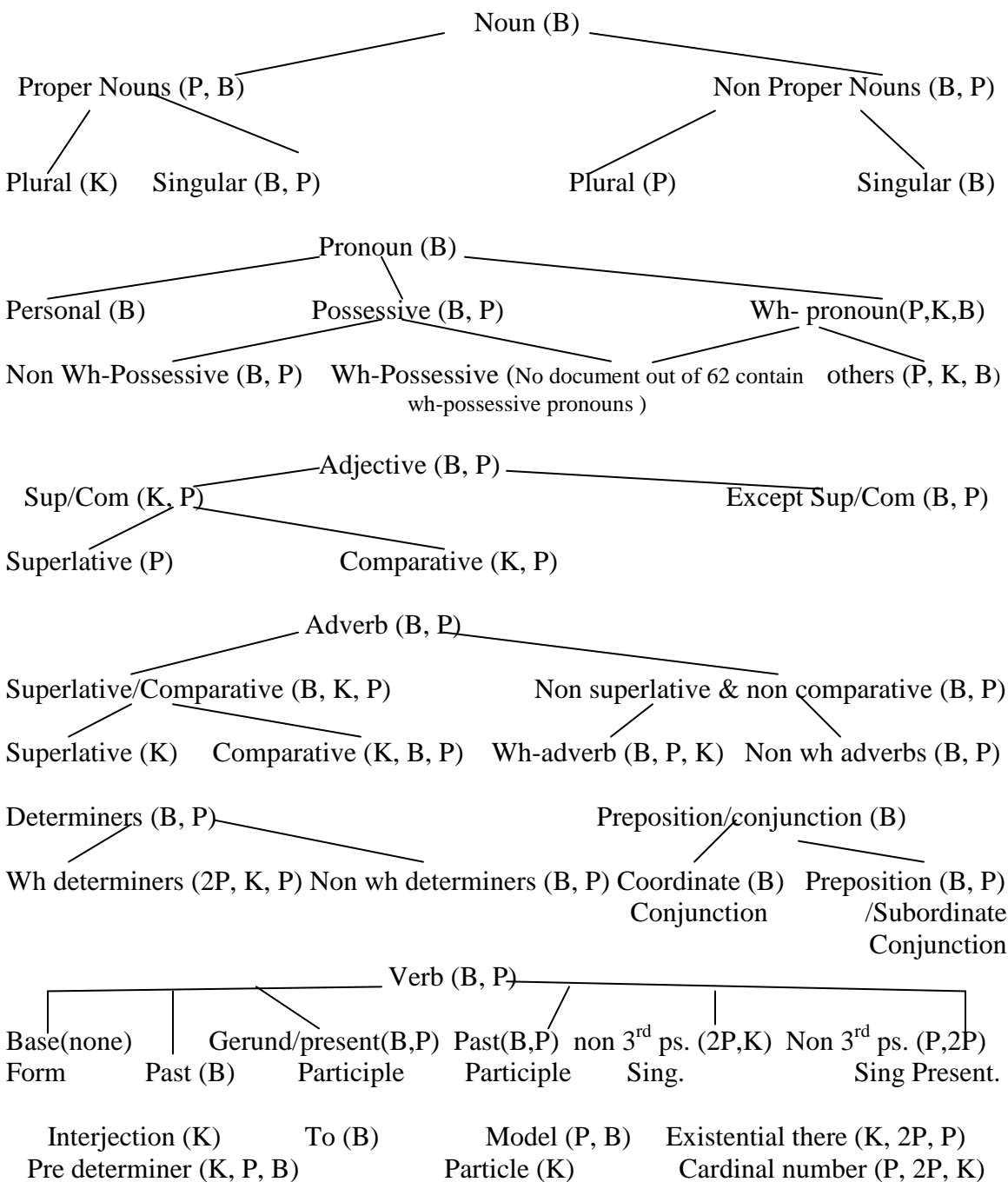




Figure 4. Distributions obtained for different parts of speech under treatment II

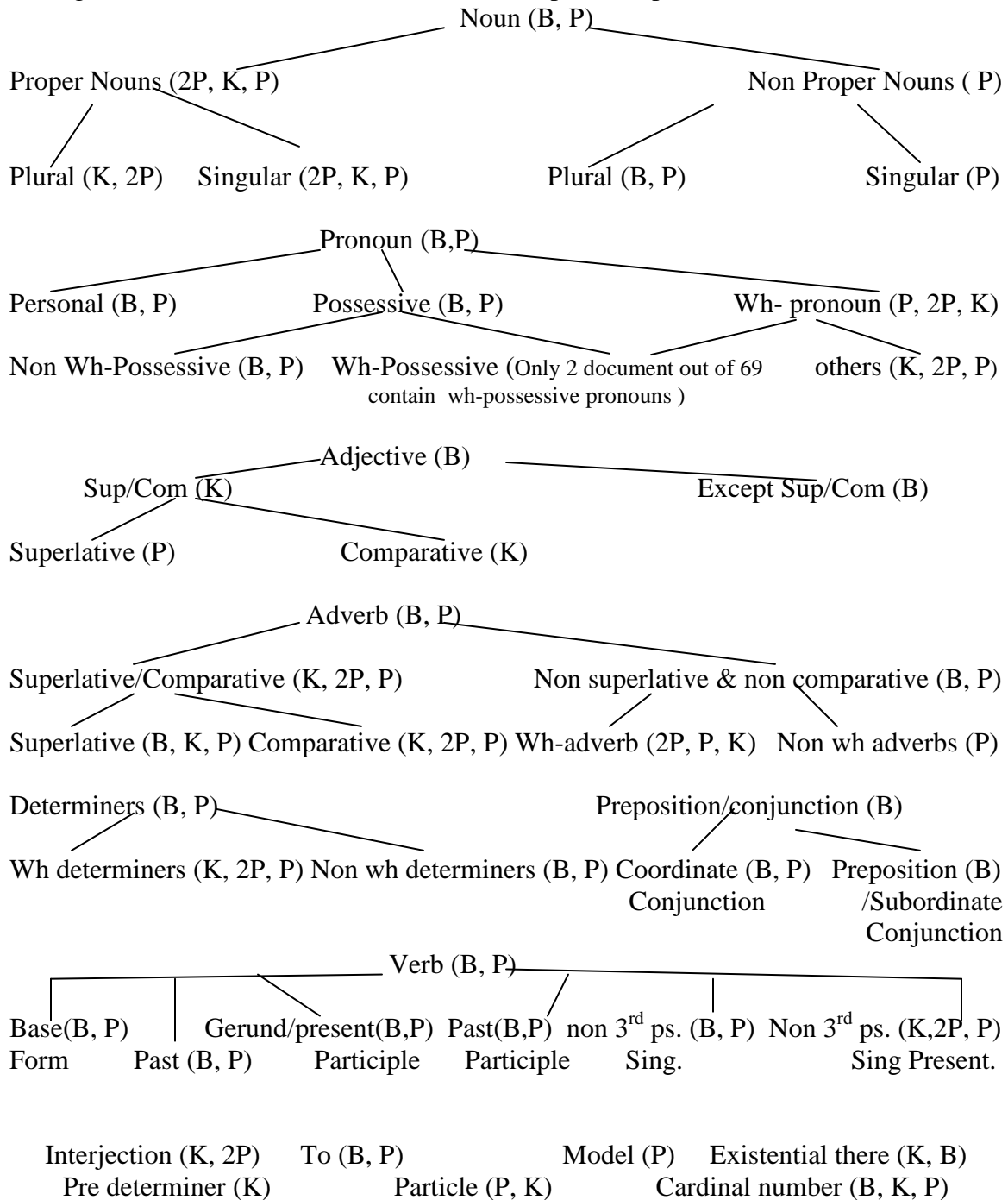




Figure 5. Distributions obtained for different parts of speech under treatment III

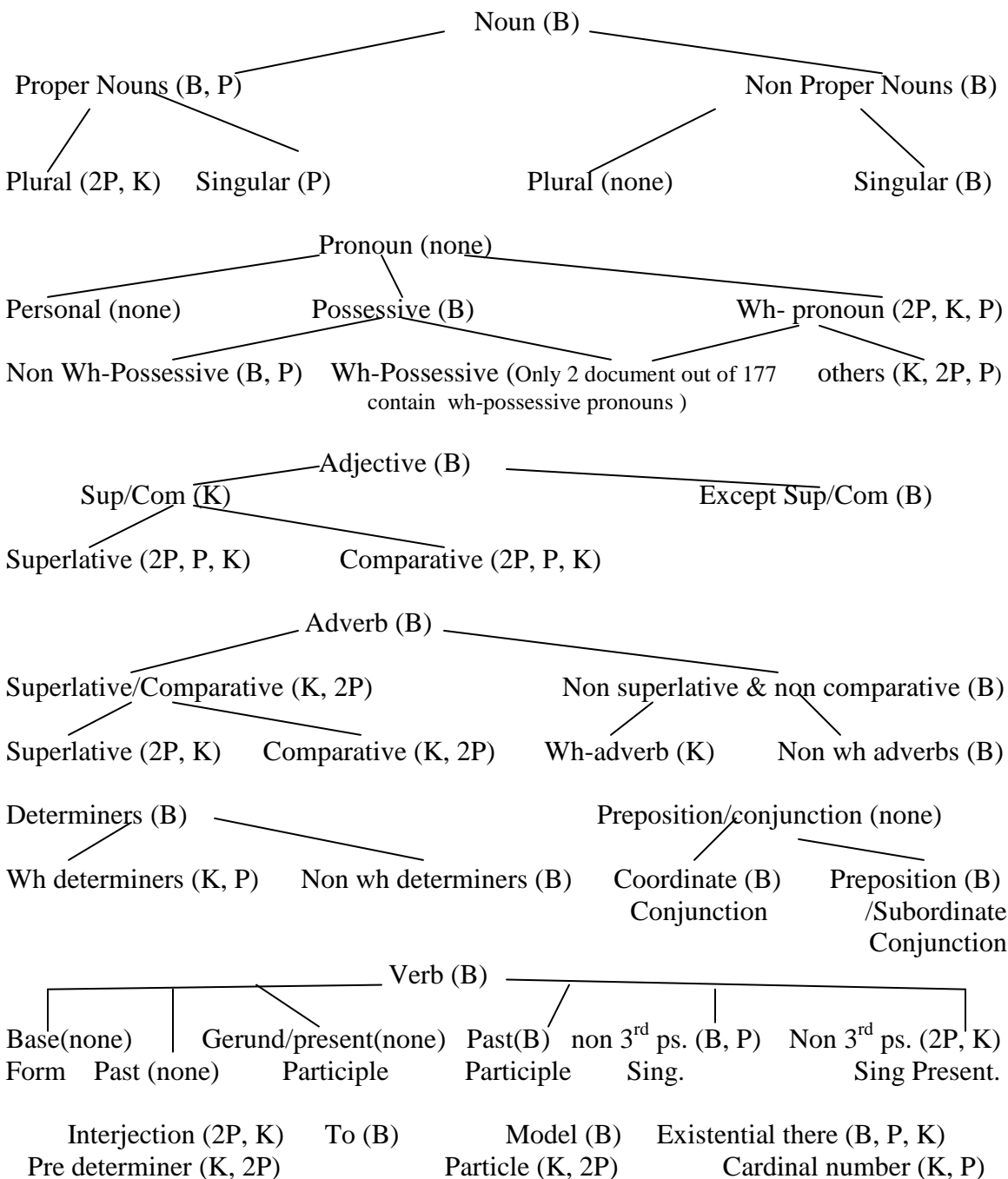


Figure 6. Distributions obtained for different parts of speech under treatment IV



The cumulative counts for the four significant distributions are -

$$f(B)=105, f(P)=105, f(2P)=37, f(K)=66.$$

The inverses of these frequencies were obtained and each value was divided by lowest of these so as to obtain the scores of each distribution as under-

$$r(B)=1, r(P)=1, r(2P)=2.8378, r(K)=1.5909.$$

Design of experiment has been utilized to test the hypothesis that the four different treatments shall not affect the merit of the procedure.

The total scores of fitted distributions for each POS and for its subparts have been calculated. As in case of noun, following two tables are generated

Table 2 and Table 3, for completely randomized design by taking two hypotheses:

$H_0$  : The means of scores for different treatments are equal.

$H_1$  : At least two means are not equal.

Table 2. Total scores of distribution for different parts of noun.

	Noun	Proper Nouns	Other Nouns	Proper Plural	Proper Singular	Non proper Nouns Plural	Non proper Nouns singular	Total xi	$\chi^2$
Treatment 1	1	2.8378	1	5.4288	2.8378	2	1	16.1044	259.3517
Treatment 2	1	2	2	1.5909	2	1	1	10.5909	112.1672
Treatment 3	2	5.4288	1	4.4288	5.4288	2	1	21.2864	453.1108
Treatment 4	1	2	1	4.4288	1	0	1	10.4288	108.7599
Total								58.4105	933.3896



$$\text{Correction factor } C.F. = \frac{x_{..}^2}{N} = 121.8495$$

$$\text{Total Sum of Squares } T.S.S. = \sum_i \sum_j x_{ij}^2 - C.F. = 64.43181$$

$$\text{Sum of squares between methods } S.S.B. = \sum x_i^2 - C.F. = 11.49185$$

$$\text{Sum of squares within methods } S.S.W. = T.S.S. - S.S.B. = 52.93996$$

Table 3. Analysis Of Variance Table.

Source of variation	Sum of Squares	Degrees of Freedom	Mean squares	F	Tabulated F.05(3,24)
Between methods	11.49185	3	3.830616	1.736586	3.01
Within methods	52.93996	24	2.205832		
Total	64.43181	27			

As the estimated value of F is less than the tabulated value at 5% level with 3 and 24 degrees of freedom therefore  $H_0$  is not rejected or means of distributions applied to different categories of noun are significantly not different for the four treatments. Similar process was also applied for all other POS- pronoun, adjective, adverb, determiner, conjunction/preposition and verb for all these and their subcomponents also the means are not differ significantly.

For relative comparison of the distributions of different POS in different treatments we have applied Randomized Block Design for which analysis tables are shown as Table 4 and Table 5.



Table 4. Scores for different parts of speech and different treatments.

Treatments \ POS ↓	POS						To	Existential There	Pre Determiner	Particle	Model	Determiner Preposition/ Conjunction	Cardinal Number	
	Noun	Pronoun	Verb	Adjective	Adverb	Interjection								
1	1	1	1	1	1	4.4287	2	1	2.5909	2.5909	2	2	1	3.5909
2	1	1	2	2	2	1.5909	1	5.4287	3.5909	1.5909	2	2	1	5.4287
3	2	2	2	1	2	4.4287	2	2.5909	1.5909	2.5909	1	2	1	3.5909
4	1	0	1	1	1	4.4287	1	3.5909	4.4287	4.4287	1	1	0	5.4287

C.F. = 244.1519

Total Sum of Squares= 105.3495

Sum of Squares Blocks= 1.089628

Sum of Squares (varieties) =67.8751

Taking  $H_{01}: \mu_1 = \mu_2$  .....Taking  $H_{02}: \mu_{.1} = \mu_{.2}$  .....

Table 5. Analysis Of Variance.

Source Of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F	Tabulated F
Blocks	1.089628	3	0.363209	0.389315	2.848
Varieties	67.8751	13	5.221161	5.596437	1.992
Error	36.38481	39	0.932944		
Total	105.3495324	55			



Since the calculated value of F treatments is less than tabulated value at 5% with 26 and 2 degree of freedom  $H_{01}$  is accepted thus the block means are not significantly different at 5% level but calculated value of F for varieties is much higher than tabulated value at 5%, so we reject  $H_{02}$ , that is, variety means are highly significant.

For the pairs of part of speech for which means are significantly different

$$total\ x_{.i} - total\ x_{.j} > 4t \sqrt{\frac{2s_e^2}{r}} \dots\dots\dots (5)$$

When we Substitute the tabulated (2-tail) value of t for 39 degrees of freedom and 5% level of significance, in the above equation (5), we infer that parts of speech according to difference in there mean scores can be divided in four groups.

I.	Noun, Pronoun, Adjective, Adverbs, Verb, Determiners, Preposition/conjunction, Model, To
II.	Cardinal numbers, Interjection, Existential there
III.	Pre Determiners
IV.	Particles

Cardinal numbers and Interjections and Existential there (part of speech taken in Go-tagger), form pairs with all parts of speech in group I to have significant difference in their mean score of fitted distributions. The mean scores of Pre Determiners are significantly different as compared with all members of Group I, except determiners and cardinal numbers. Similarly mean scores of Particles are significantly different with scores of Noun, Pronoun, Adjective, Prepositions and Cardinal Numbers.



**MJAL2:1JANUARY2010**

**ISSN 0974-8741**

**Distributions of different parts of speech in different parts of a text and in different texts**

**by Hemlata Pande and H. S. Dhani**

### **Conclusion**

By using design of experiments for the comparison of distributions of parts of speech in (a) different parts of a text (b) different parts of another text and (c) in different texts (by considering parts of speech per token) we conclude that -

- Noun, pronoun, adjective, adverb, determiner, conjunction/preposition and verbs and their sub parts follow same types of distributions for the four treatments.
- If we consider only fourteen parts of speech- noun, pronoun, adjective, adverb, verb, to, determiner, preposition/conjunction, model then for all of them Binomial or Poisson distributions are more appropriate but in case of cardinal numbers, particle, interjection , pre determiners and existential there, distributions shall be 2P or K mixture.
- The distributions for the four treatments are of similar types.



**MJAL2:1JANUARY2010**

**ISSN 0974-8741**

**Distributions of different parts of speech in different parts of a text and in different texts**

**by Hemlata Pande and H. S. Dhani**

### **Acknowledgement**

The authors are grateful to the Council of Scientific & Industrial Research (CSIR), New Delhi, for providing financial assistance to carry out the research work on this interesting field of applied Mathematics in the form of a senior research fellowship to the first author.



**MJAL2:1JANUARY2010**

**ISSN 0974-8741**

**Distributions of different parts of speech in different parts of a text and in different texts**

**by Hemlata Pande and H. S. Dhani**

### References

- Balakrishnan , N., Nevzorov, V. B.(2004). *A Primer on Statistical Distributions*. Wiley, IEEE Press.
- Belaid, A., Valverde, N. (2000). Part-of-speech tagging for table of contents recognition. In *Proceedings of Fifteenth International Conference on Pattern Recognition*.(451-454). Barcelona, Spain.
- Brill, Eric (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third ACL Applied Natural Language Processing* (pp. 152-155) Trento, Italy.
- Guilpin, P., Guilpin, C. (2005). Linguistic and statistical analysis of the frequency of a particular word at different times (diachrony) or in different styles (synchrony). *Journal of Quantitative Linguistics*, 12(2-3), 138-150.
- Katz, S. M.(1996). Distribution of content words and phrases in text and language processing. *Natural language Engineering*.2, 15-59.Cambridge University Press.
- Kim, Y. & Ross, S.(2006).Genre Classification in Automated Ingest and Appraisal Metadata. J. Gonzalo et al. (Eds.): *ECDL 2006, LNCS 4172* (pp. 63–74),Springer-Verlag Berlin Heidelberg.
- Krishnamoorthy, K.(2006). *Handbook of Statistical Distributions and Applications*. CRC Press & HALL/ CRC. Taylors & Francis Group.
- Manning, Christopher D. and Hinrich Schutze(1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.
- Pang, B., Lee, L., Vaithvanathan, S.(2002). Thumbs up?:Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*(pp.79-86). Morristown, NJ, USA.
- Stevenson, S. and Merlo, P.(2000). Automatic lexical acquisition based on statistical distributions. *Proceedings of the First Conference on Computational Linguistics* (pp.815-821).



**MJAL2:1JANUARY2010**

**ISSN 0974-8741**

**Distributions of different parts of speech in different parts of a text and in different texts**

**by Hemlata Pande and H. S. Dhani**

Merlo, P and Stevenson, S.(2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3),373-408.

Yuan, L., Chen, Z. (2006).A novel statistical model for speech recognition and POS tagging. *IEEE International Conference on Video and Signal Based Surveillance AVSS'06*.

### Appendix

A Few Quite days Paul Azul	<a href="http://chtrust.com/novel">http://chtrust.com/novel</a>
A Half Life Of One Bill Liversidge	<a href="http://halflifeone.blogspot.com/">http://halflifeone.blogspot.com/</a>
Bird Flu Dawn Meier	<a href="http://www.blueunicornpublishing.com/">www.blueunicornpublishing.com/</a>
Blue Seaweed	<a href="http://www.foxglove.co.uk/shorts/bluegrass.html">www.foxglove.co.uk/shorts/bluegrass.html</a>
The statements	<a href="http://www.foxglove.co.uk/Sherlock/famous.html">www.foxglove.co.uk/Sherlock/famous.html</a>
The Bloody Sock and Other Tales	<a href="http://www.foxglove.co.uk/tadeusz/bloody-sock.html">www.foxglove.co.uk/tadeusz/bloody-sock.html</a>
The Dissatisfied Voter	<a href="http://www.foxglove.co.uk/Sherlock/voter.html">www.foxglove.co.uk/Sherlock/voter.html</a>
Firestorm 2034 Jonathan Hayward	<a href="http://jonathanscorner.com/writings/firestorm/firestorm.html">http://jonathanscorner.com/writings/firestorm/firestorm.html</a>
My name's Jack J. Charles Cripps	<a href="http://www.trivigo.com/MY%20NAME'S%20JACK[1].html">www.trivigo.com/MY%20NAME'S%20JACK[1].html</a>
The Determined Existentialist	From: <a href="http://www.foxglove.co.uk/shorts/index.html">www.foxglove.co.uk/shorts/index.html</a>