

Contents

	Page
1. Introduction	1
1.1. Historical sketch	1
1.2. Language as a mass phenomenon – ‘Quantity Survey’ of language	2
1.3. Chance as a factor of linguistic expression and language structure	3
1.4. Structuralism and statistical linguistics	4
1.5. Language as choice and chance	6
1.6. DE SAUSSURE’S ‘Principe Linéaire’ and geometrical duality	7
1.7. Literary statistics, a new branch of applied statistics	8
1.8. Plan of the book	11
I. Language as Chance I – Statistical Linguistics	14
2. Stability of Linguistic Distributions	14
2.1. A fundamental law of communication	15
2.2. Frequency distributions of linguistics – Experimental data	16
2.3. The statistical interpretation of DE SAUSSURE’S ‘langue-parole’ dichotomy	27
2.4. Comparison by rule-of-thumb methods	29
2.5. Comparison by methods of statistical inference	34
2.5.1. Standard error test	35
2.5.2. Chi-square test	36
2.6. Interpretation of test results	41
2.7. Simple and complex distributions	42
2.8. A practical criterion of stability of linguistic distributions	43
3. Explanation of Stability of Linguistic Distributions	44
3.1. Overlap between texts in vocabulary and frequency of occurrence	44
3.2. The relation between grammar and lexicon	48
3.3. The ‘grammar load’ of a language – Methods of assessment	49
3.4. Grammar as a factor of the stability of linguistic distributions . .	55
3.5. The mutually limiting action of grammar and lexicon components	58
3.6. Doubts about the stability of the phonemic (alphabetic) distribution	58
4. Application of the Theory of Stability of Alphabetic Distributions to a Problem of Language Mixture	60
4.1. Problems in connection with language mixture	60
4.2. The alphabetic distribution of nouns	61
4.3. The multiplicative law of the noun-initial distribution	65
4.4. Comparison of the LR component of English with Mediaeval Latin	66
Bibliography	68
II. Language as Choice I – Stylostatistics	70
5. Style as a Statistical Concept	70
5.1. Quantitative features of style	70
5.2. Using statistics for determining the chronological order of texts .	71

	Page
5.3. Richness of vocabulary	72
5.4. How text length in English is accounted for by vocabulary	73
5.5. The general relation between vocabulary and text length	75
5.6. Vocabulary ratios	77
5.7. Special and total vocabulary – Romance vocabulary in CHAUCER'S 'Canterbury Tales'	78
5.8. Generalisation of the quantitative law of language mixture	83
5.8.1. Explanation of the quantitative law of language mixture	85
5.9. Unsuitable mathematical models in language statistics, and their consequences	87
5.9.1. The ZIPP law as an unsuitable model	88
5.9.2. The MANDELROT Canonical Law – Shortcomings from the theoretical and practical angles	89
5.9.3. The so-called 'Law of Least Effort' in language	90
6. Word Count Mathematics	91
6.1. Central values and values of dispersion	92
6.2. The frequency distribution of vocabulary	94
6.3. Sampling methods for word counts	95
6.4. Illustration – The Russian word count	96
6.5. A statistical paradox and its explanation	100
6.6. A new statistical parameter – The 'Characteristic'	101
6.7. Style as a statistical concept	102
6.8. YULE'S experiment	105
6.9. v_m as a measure of the 'langue-parole' duality	106
6.10. Characteristic and Entropy	112
6.11. Summary	113
6.12. Words and concepts – Professional codes	115
6.12.1. Size vs. content of concepts	120
6.13. Stability of the distribution of grammar forms – Recurrence of particular grammar forms as stabilising factor	121
6.13.1. The Russian grammar-form count	121
6.13.2. Discussion	123
6.14. The chance distribution of grammar forms	127
6.15. The sound and symbol duality (Chinese)	130
6.15.1. The Chinese dictionary – Radical and Phonetic	131
6.15.2. The duality principle of a Chinese dictionary	132
6.15.3. Distribution of characters according to stroke number of phonetic	133
6.15.4. Distribution of sub-classes to radicals according to the number of ideograms per sub-class	143
6.15.5. Taxonomic structure of the Chinese dictionary – Chance as a factor of Chinese lexicography	143
7. Style Relationships – Bi-Variate Stylostatistics	147
7.1. Joint word occurrence in different authors	147
7.1.1. A statistical study of political vocabulary	148
7.1.2. Sampling methods	148
7.1.3. The distribution of political vocabulary	150
7.2. Correlation of authors through vocabulary	151
7.3. Vocabulary overlaps between authors – Significance tests	155
7.4. Correlation between authors through frequency of use of words	157
7.5. Interpretation of correlation between authors	160
7.6. Correlation and disputed authorship	161

	Page
8. A Guide to Stylo-statistical Investigations	163
8.1. Preparing the punched cards (or tape) for processing linguistic information	163
8.1.1. The word as the elementary unit of running texts	164
8.1.2. The word as elementary lexical unit	165
8.1.3. Conclusions	168
8.2. Word categories to be included, and the size of sample	169
8.2.1. Type of word categories to be included in the word count	169
8.2.2. Size of sample	170
8.3. The fallacy of determining style by differences in frequency of a few grammar ('function') words	172
Bibliography	174
III. Language as Chance II - Optimal Systems of Language Structure	176
III.(A) Combinatorics on the Phonemic (Alphabetic) Level	176
9. The Combinatorial Structure of Words	176
9.1. Linguistics as a branch of semiology	176
9.2. Combinatorial structure of composite alphabetic code symbols	178
9.3. A de-coding experiment	182
9.4. Comparison of alphabetic and phonemic codes	185
9.5. Discussion - Conformity vs. discrepancy of alphabetic and phonemic codes	189
9.6. Consonant combinations in Czech and German	190
9.7. Non-random sequences of phonemes	191
9.8. The patterning of Semitic verbal roots subjected to Combinatory Analysis	194
10. Optimality of the Word-Length Distribution	198
10.1. Redundancy of coding in natural languages	198
10.2. Lognormality of the word-length distribution	201
10.3. Lognormality and Optimality	204
11. Combinatorics applied to Problems of Classical Poetry	206
11.1. The sequence of dactyls and spondees in the Latin hexameter	206
11.2. Sentence length and caesurae in the early Greek hexameter	210
III.(B) Combinatorics on the Lexicon Level	214
12. Random Partitioning of Vocabulary - Vocabulary Connectivity	214
12.1. The deterministic view of the use of words and some facts against it	214
12.2. Chance, the ever-present alternative	215
12.3. Fitting the Random Partitioning Function to the results of empirical vocabulary connectivity	218
13. The Generalised Random Partitioning Function and Stylostatistics	219
A. The Pauline Epistles	219
13.1. Derivation of formula for the generalised Random Partitioning Function	220
13.2. Application to the Pauline Epistles	223
13.3. The mathematical definition of uniformity of style	231
13.4. Totals of vocabulary, observed and calculated, per Epistle	234
13.5. Graphical representation of the Random Partitioning Function	235

	Page
B. The New Testament	238
13.6. Application of the Random Partitioning Function to the New Testament in Greek	238
13.7. Comparison of results with current Bible exegesis	247
13.8. Graphical presentation and vocabulary totals per part	248
14. The "New Statistics" on the Vocabulary Level	249
14.1. Quadratic vs. linear fluctuations	249
14.2. Quantum statistics of language	252
14.2.1. How the need for the "New Statistics" arose in Physics	253
14.2.2. The Norm of Vocabulary Connectivity as corresponding to Black Body radiation	255
Bibliography	258
III.(C) Information Theory	259
15. Principles of Information Theory	259
15.1. Relation between information theory and statistical linguistics	259
15.2. The binary code – The Entropy	260
15.3. The linguistic interpretation of entropy and redundancy	264
15.4. Efficiency of a code	267
15.5. Derivation of the entropy from the multinomial law	270
15.6. An inequality relation between the entropy and the repeat rate (and its sample statistic K)	271
15.7. Efficiency of coding – The law of optimal redundancy	274
15.7.1. The condition for optimal coding	274
15.7.2. Binary coding as optimum strategy of enquiry	275
16. Information-Theoretical Analysis as a Tool of Linguistic Research	282
16.1. Language as an efficient code	282
16.2. The statistical study of word-length	284
16.3. Pitman's Shorthand as an efficient code	289
16.4. Stability of word-length distributions	291
16.5. The mechanism of the linguistic development towards monosyllabism in the light of information theory	294
16.6. Entropy and Entropy	298
16.7. Word-length in terms of the number of phonemes (letters)	299
16.8. Relation between syllable and letter number per word	301
16.9. Different interpretation of the entropy according to the linguistic unit	302
17. Language Translations as Bi-Variate Distributions of Coding Symbols	304
17.1. Bi-variate information theory	305
17.2. The criterion of quantitative relationships between original and translation	308
17.3. The experiment – Bi-variate syllable counts	309
17.4. Stability of bi-variate syllable counts	311
17.5. Interpretation of the stability of bi-variate distributions of word-length	318
17.6. The conditioned entropy on the lexicon level	319
17.6.1. Word counts in their relation to vocabulary, word association and grammar	324
Bibliography	326

	Page
IV. Language as Choice II – Linguistic Duality	328
18. The Four-fold Root of Linguistic Duality	328
18.1. Boolean law of duality	328
18.2. Duality and probability	329
18.3. The principle of duality in higher mathematics	330
18.3.1. The principle of geometrical duality in language – Interchangeability of Type and Token in linguistics statements	332
18.4. The Type-Token duality – Combinatorics of sentence formation	335
18.4.1. Combinatorics and the Alphabet-Square	335
18.4.2. Discussion	338
18.4.3. The diachronic aspect of planned combination	339
19. Duality as Correcting Factor – Inadequacy of Truly Semiotic Codes	341
19.1. DE SAUSSURE's 'signifiant-signifié' relation and linguistic duality	341
19.2. The restless universe of language	344
19.3. Stunted development of languages through lack of duality	345
20. Duality and Language Translation	348
20.1. Variability of translational equivalence	348
20.2. Relation between word-length and meaning	350
20.3. The translation matrix of meaning	350
20.4. Duality of meaning as an obstacle to machine translation	351
20.5. The concept of comparative stylistics	355
20.5.1. Description of G. Barth's work	355
20.5.2. Statistical results	356
20.5.3. Graphical analysis (sequential sampling method)	360
20.6. The qualitative aspect of style	365
Bibliography	369
V. Statistics for the Language Seminary	371
V.(A) Statistics of Language in the Mass	371
21. Descriptive Statistics	371
21.1. Statistical distributions and elementary statistical constants	371
21.2. Empirical facts about statistical constants	377
21.3. Arithmetic mean and standard deviation of composite statistical masses	378
21.4. The Gaussian or Normal Law	381
21.4.1. Form and statistical constants of the normal distribution	381
22. Statistical Inference – The Binomial Case	384
22.1. Mathematical tools for the combinatorial technique	385
22.2.1. The argument from text to sample	389
22.2.2. The argument from sample to text	391
22.2.3. The argument from one text sample to another	394
22.3. Statistical Inference for Great Collectives	396
22.3.1. Inference from a very great statistical collective (Bernoullian Problem)	397
22.3.2. Inference from sample to very great statistical collective (Bayes' Problem)	399
22.3.3. The chance distribution of rare events – The law of small numbers	400

23. Statistical Inference in the Case of Multiple Classification of Events	401
23.1. Inference from total to sample	402
23.2. Inference from sample to total	403
23.3. Inference from one sample to another	403
23.4. Inference when dealing with great statistical masses	404
23.5. Testing two distributions for compatibility – The X-square test	405
23.6. Analysis of the internal structure of a statistical mass – Lexis' L	408
24. Theory of Correlation	410
24.1. Functional relation vs. statistical relationship	410
24.2. The line of regression	412
24.3. Fallacies of interpretation	414
24.4. The correlation coefficient	417
24.5. Significance of the correlation coefficient	419
24.6. Bernoullian correlation – The coefficient of contingency	421
V. (B) Statistics of Language in the Line	422
25. The Dimension of Time in Language Statistics	422
25.1. Statistics in the "Region of Lost Dimensions"	422
25.2. Statistics of language in the line	422
25.3. Sampling on the lexicon level	423
25.4. Random partitioning	425
25.5. A mathematical model of language mixture	426
26. Linguistic Duality and 'Parity'	430
26.1. Language statistics and statistical physics	430
26.2. The problem of conservation of parity in fundamental physics	431
26.3. Laterality of the speech function in the brain and linguistic duality	435
Bibliography	437
Appendix – A Survey of Past and Present-day Statistical Linguistics	438
Author Index	446
Subject Index	448