PERSPECTIVE PAPER: QUANTITATIVE LINGUISTICS*

Wolf Moskovich
The Hebrew University of Jerusalem

Introduction

Qualitative analysis of language is based on the study of the opposition *presence-absence* of a certain linguistic phenomenon in the structure of language without taking into consideration the frequency of the phenomenon. Quantitative analysis, on the other hand, is a way of describing a linguistic system based on an estimate of the relative frequencies of the phenomena under investigation. Quantitative linguistics (statistical linguistics, linguistical statistics) is a part of modern linguistics "using statistical methods for investigation of acts of speech and the system of language" (Akhmanova, 1966, p. 219).

While analyzing the application of quantitative linguistics to information science, one has to take into consideration all the three apices of the triangle: mathematical statistics, linguistics, and information science. Representatives of all three disciplines are involved in research in the field, and their attitudes towards, and judgments of quantitative linguistics are often influenced by their professional background. In the following discussion, we shall come across conflicting approaches to quantitative linguistics and its application to information science, and our judgment will always be on the side of quantitative linguistics as a sovereign linguistic field of study.

Though it is commonly accepted that quantitative linguistics is a part of mathematical linguistics (or computational linguistics), every now and then there are attempts to exclude it from these disciplines. This opposition usually comes from certain mathematicians who consider mathematical linguistics to be a mathematical and not a linguistic discipline. They reduce it to a study of deductive *linguistic calculi* or *linguistic algorithms* (e.g., Gladkij and Mel'čuk, 1970). They argue that statistical techniques are common for all sciences and therefore their application to language analysis does not belong to linguistics.

There are other mathematicians who think that mathematical linguistics belongs both to linguistics and to mathematics, being "a science that discovers and studies mathematical structures really existing in linguistic objects" (Šrejder, 1975, p. 7). But the majority of experts agree that quantitative linguistics wholly belongs to mathematical linguistics. "When we study the non-grammatical side of language, almost all of the problems themselves are of a quantitative nature," writes Kiefer (1964). "No doubt, statistical linguistics represents another approximation to language structure, complementary to the algebraic one. Statistical methods and theories are thereupon indispensable in mathematical linguistics" (p. 26).

---

*This paper is presented on behalf of KVAL Institute for Information Science.

Wolf Moskovich

In the book by Sparck Jones and Kay (1973), the terms *quantitative linguistics, linguistical statistics,* and *statistical linguistics* do not appear; the whole camp of linguistics as opposed to the camp of information science is referred to as *linguistics, theoretical linguistics,* or *computational linguistics.* There is no chapter or section of the book entitled *quantitative linguistics.* Such a situation seems paradoxical since the authors claim that the most substantial contribution linguistics made to information retrieval is in the use of vocabulary statistics. One cannot evade the impression that Sparck Jones and Kay presented the material in terms of direct application of statistical techniques to document analysis without making it explicit whether the field belongs to linguistics or not. The authors try in the first chapter of the book to set some boundaries between linguistics and information science being aware that "much of what must be explained in order to make information science amenable to computer techniques belongs to what linguists have undertaken to explain for quite independent reasons" (p.3).

However, when it comes to the application of statistical techniques, described in Chapter 6, *Semantics,* on more than 40 pages of the book, only one and a half pages (pp. 171-172) are allotted to linguistics proper (the section *Automatic classification in linguistics*). Another section of the book, *Statistical syntax* (pp. 110-111), is not larger -- less than a full page. The whole plan of the book and the discussion of the relevant material makes it clear that here the reader has to do with terminological inconsistency on the part of the authors rather than with conscious distortion of the picture.

It is a serious drawback of the book that it adequately describes only work done in the USA, Britain and France. Research performed outside these countries is mentioned only in passing (if mentioned at all). The corresponding publications in other countries were not analyzed, and as a rule, references are made only to their English abstracts taken from *Abstract Journal: Informatics.* Consequently, the conclusions of the book are based only on a partial survey of work done in the field. The book, furthermore, gives little attention to theoretical and linguistically-oriented research in the field of our concern. It is as though no valid research independent of information science applications has been done in quantitative linguistics.

In order to give an account of the contribution of quantitative linguistics to information science, we need to evaluate current work in the field from the point of view of quantitative linguistics. The questions which are to be examined are the following: To what extent does application of statistical techniques to text analysis belong to linguistics and to what extent to information science? What is in fact the current contribution of quantitative linguistics to information science and what may be this contribution in the future? What are the main lines of research in quantitative linguistics, what are the scientific results achieved, and what is the relevance of these results to information science?

We shall try to answer these questions, not by relying mainly on the publications reviewed in the Sparck Jones and Kay book but by taking into consideration a wider range of literature belonging to the field of quantitative linguistics with due attention to original work produced both in western countries and in the Soviet Union.

The Interrelation of Quantitative Linguistics and Information Science

Information science deals with storage, retrieval and transmission of information. As a theoretical discipline it studies the rules and laws according to which semantic information is created and transformed and here it shares its interests with theoretical linguistics. As an

applied field it uses various methods of handling information, both developed within its limits for its specific needs and borrowed from adjoining sciences. The more sophisticated and complex the techniques of adjoining sciences are, the more difficult is their application. Quantitative characteristics of messages are studied by methods of mathematical statistics known to every system analyst, and it seems only natural to use them for information retrieval purposes.

Starting from H. P. Luhn's first successful experiments in automatic statistical indexing and abstracting in the late 1950s, statistical techniques were taken up by a number of researchers in information science. Most of the systems created are based on ad hoc techniques, and although the results obtained are as a rule satisfactory and encouraging, usually no qualitative analysis of the material subjected to counting is given.

> *It is noteworthy that many of the contributors to the 1964 Washington Symposium on Statistical Association Methods for Mechanized Documentation, which can be regarded as the culmination of the period of enthusiasm for the new approaches to the central problem of document description by statistical techniques, are no longer engaged in research in this area... The need to test bright ideas, when combined with a growing awareness of the complexity of a retrieval system, and higher standards of experimentation, brought research workers up against the prospect of long hard labor with a very uncertain outcome* (Sparck Jones and Kay, 1973, p. 12)

Apparently there is a certain limit beyond which the *brute force* statistical approach cannot penetrate, and additional techniques and criteria of a qualitative nature are needed.

The differences between the use of statistical techniques in information science and quantitative linguistics are twofold: (1) Quantitative information science usually relies on ad hoc techniques, whereas qualitative linguistics tries to develop a general theory or model of linguistic behaviour, and interprets its findings in the light of this theory (or model). (2) Quantitative information science usually is satisfied with a pragmatic result and does not study the qualitative characteristics of counted objects; quantitative linguistics compares quantitative characteristics of counted linguistic units with their qualitative characteristics and tries to find correspondences between them.

Although statistical techniques implemented in information science and linguistics may be the same, the aims of the two disciplines are different. These differences should be a main factor in defining the scope of quantitative linguistics and its potential contribution to information science. The aim of quantitative linguistics is to find and describe the laws governing the statistical organization of texts, and to discover the structure of language through quantitative analysis of the behaviour of linguistic units in texts. The aim of information science is to apply statistical techniques for document analysis, storage and retrieval in order to build workable information systems.

It is clear that information science can benefit from using the experience of quantitative linguistics, and that statistical regularities of language behaviour discovered and described in quantitative linguistics may serve as a basis for application in information science. It goes without saying that some refinements of the statistical procedures created within information science for its specific needs, as well as some facts relating to the statistical characteristics of text units discovered within information science may be useful for research in quantitative linguistics.

Wolf Moskovich

The State of the Art of Quantitative Linguistics

According to the classification by Karlgren (1975a), the quantitative models so far used within linguistics are divided into three main groups, as they try to provide

1. Quantitative arguments for qualitative issues

2. Quantitative descriptions of language phenomena

3. Quantitative explanations of language phenomena.

As an example of the model of the first group, a statistical procedure for determining the sequence of various themes in a text is given (Karlgren, 1975b). A computation based on a statistical model, according to which the probability for recurrence of a word is characteristically greater when the theme is the same, suggests tentative demarcations between text sections. But the hypothesis put forward -- that there is a borderline at that and that point in the text -- is not a quantitative statement. The model is used to produce and corroborate a qualitative hypothesis.

Examples of the models of the second group are the following: word frequency distributions, quantitative descriptions of sentence length, syntactic complexity, semantic uniformity of texts, semantic distances between linguistic units. Such quantitative descriptions are aids for determining qualitative issues, but they have also an intrinsic interest.

Models of the third group explain the language mechanism in quantitative terms. For example: the relation between frequency and length of words is explained in terms of a mathematical model to minimize communicative effort.

Most of the research in quantitative linguistics potentially useful for information science belongs to the second group of models in this classification. Nowadays, frequency dictionaries and concordances to texts in various languages are being produced on a mass scale in line with the trend of recent years. The use of computers for their compilation facilitates the work, and makes it possible to compute a larger number of quantitative characteristics of linguistic units. For example, the statistical data given in the frequency dictionary of present-day American English by Kučera and Francis (1967) include the frequency of words, the number of genres and the number of samples in which they have occurred, word-frequency distributions within various subsets of the corpus, information about type-token ratios in the corpus and various subsets, and word length and sentence-length statistics.

The advance of computers, as well as inner forces of development within linguistics brought two major tendencies into modern quantitative linguistics: (1) an increased attention to the analysis of subsets of natural language; and (2) a shift from statistics of isolated words to statistics of word combinations. Both tendencies have major implications for the further cooperation between quantitative linguistics and information science. Statistical analysis of subsets of natural language provides quantitative data and special frequency vocabularies of narrow scientific and technological fields which can be immediately used in information retrieval. The former orientation of quantitative linguistics towards the analysis of texts of general and literary character was of much less (if any) interest for information retrieval. For the statistical study of word combinations,

the traditional statistical apparatus used in counting isolated words was not suitable. A new apparatus had to be used, and new results were then achieved with its help. It marked a real revolution in quantitative linguistics which thus can study statistical interactions of words in text and not just their frequency. It brought new perspectives into information science because such statistical techniques and data provide access to associative term structures which are of particular significance for information retrieval purposes.

A notion of *sublanguage* was put forward and a multitude of frequency dictionaries of words and word combinations was produced in recent years (at least for the USSR, where more than 100 such dictionaries were compiled in the past 10 years, a complete list is presented in Moskovich, 1967-1974). For each sublanguage the following statistical materials can be, and in some instances partly are, compiled (N is the number of typical subcorpuses for a sublanguage):

1. N rank frequency dictionaries of word forms.
2. N alphabetical frequency dictionaries of word forms.
3. General rank frequency dictionary of word forms (indicating for each word form and for each of its representations with a definite marknote: total frequency, total ratio frequency, the number of subcorpuses, frequencies -- absolute and ratio -- in these subcorpuses).
4. General frequency dictionary of word forms.
5. N frequency-distribution tables.
6. General frequency-distribution tables for word forms.

7-12. Analogous inventories for lexemes.

13. N tables for sentence-length distribution.
14. General table for sentence-length distribution.
15. N frequency dictionaries of binary word combinations.
16. General frequency dictionary of binary word-combinations.
17. Alphabetical dictionaries of lexical-syntagmatic word distributions based on semantic roles of the word in the sentence (this dictionary consists of two parts: distributions of the government words and distributions of the subordinate words).
18. N alphabetical frequency dictionaries of complete lexical-semantic complexes with a given nucleus (both binary and more than binary).
19. General alphabetical frequency dictionary of lexical-semantic complexes.
20. N frequency-distribution tables for lexical-semantic complexes.
21. General table of complex-frequency distribution.
22. N frequency lists of semantic syntagmatic relations.
23. General frequency list of semantic syntagmatic relations.
24. Dictionary indicating the semantic-relational productivity of lexemes (measured by the number of different binary combinations which the given word enters). (Gorodeckij, 1972)

Computer compilation of most of these materials presupposes the use of pre-machine text marking, and means a transition from collection of mechanical contexts (concordances) to more complex statistical descriptions of linguistic distribution based on real syntactic and semantic relations. Three main types of markings may be used: (1) dividing lines between morphemes; (2) grammatical or semantic indicators of word forms; (3) indicators of syntactic roles and relations.

On the basis of such preliminary markings (syntactic class, subclass, governing word, the type of syntactic connection between words), an *automatic grammar* for texts of American patent claims was built (Moskovich, 1966).

In one of a number of similar studies, researchers tried to clarify the following question: Do the semantics of the components and the type of syntactic links between them condition the morphological structure of noun combinations in modern Ukrainian? For the configuration *N Nominative case and N Genitive case*, when the governing component designates various appliances and is characterized by suffixes *-ač* or *-tor*, the following syntactic-semantic links between components of noun combinations are predicted: (1) *to have as an object of action* -- 70 per cent; (2) *to be a part* -- 20 per cent; (3) *to be an instrument* -- 10 per cent (Skorohod'ko, 1964). The use of pre-machine text marking broadens the possibilities for quantitative analysis and links it to other linguistic techniques.

A completely different approach based on statistical discovery procedures of the kind usually applied in decipherment of messages was developed in recent years. The approach is conceived as an extension and perfection of classical techniques of distributive linguistic analysis aimed at delimitation and classification of text units. "Distributive-statistical analysis may be defined as analysis of text consisting of algorithmic procedures with a wide use of statistics and based only on information about distribution in text of objectively delimited text elements" (Ivanova and Šajkevič, 1970, p.79).

Distributive-statistical analysis can be applied to the delimitation and classification of units on various levels of language. On the basis of conditional probabilities of letters in text and in different positions in words, individual morphemes as well as classes of morphemes can be delimited. Several algorithms for distributive-statistical morphemic analysis were successfully tested (Ivanova and Šajkevič, 1970; Andreeva, 1969).

More spectacular results were achieved in the application of distributive-statistical techniques on the lexical-semantic level of language. The usual procedure is to compare the theoretically expected and actual cooccurrence of terms in texts. The strength of connections among terms is determined by the value of deviation of their actual occurrence from the theoretical one. In such a way, semantic networks are built.

It can be shown that the development and application of statistical association techniques took place in linguistics and information science in a parallel fashion (Moskovich, 1972a). Sparck Jones and Kay describe the history of research in this field as an event within the confines of information science. After reviewing relevant publications in a way that implies that they belong to information science, they draw the following conclusion: "The impact of all this on the main stream of linguistics is not discernible, and the impact on linguistics of any kind has been slight. However, a few experiments have been performed in which text based statistical associations have been exploited to obtain semantic characterizations of words" (1973, p. 171). While this judgement may be supported by the relative number of publications on the subject in information science and linguistics, it

hardly does justice to quantitative linguistics.

The most prominent publication on statistical association techniques in information science is that of the proceedings of the symposium held in Washington in 1964 (Stevens et al., 1965). According to this book and other sources, statistical association techniques were first applied to the study of texts by psychologists. In a paper published in 1942, an American psychologist, A. L. Baldwin (1942), used the values of the cooccurrence of words in the letters written by a female patient as indicators of corresponding connections among ideas in her mind.

The first attempts to use these techniques for information retrieval were made in the late 1950s (Needham, 1961; Doyle, 1959; Giuliano and Jones, 1962). In linguistics, there was a parallel development. In the late 1950s, several statistical procedures for measuring syntagmatic and paradigmatic proximity of words were suggested and tested by Andreev (1959, 1961). At the same time, Šajkevič (1961, 1963) conducted his experiments on the discovery of semantic fields of language on the basis of statistical analysis of cooccurrence of words in text. Cooccurrence of 1078 adjectives in a text sample of 2 million words of English poetry was analyzed. The interval in which cooccurrence was registered was a line of poetic text (about 5-6 words). As a result of the analysis, paradigmatic groups of words were discovered. These included synonyms (*dumb - mute; weak - faint - feeble; vile - base - mean; foul - filthy - loathsome; gentle - mild*), antonyms (*mortal - immortal; new - old; great - small; dead - alive; strong - weak*), and groups of words with similar meaning (*green - fresh - new; lofty - divine - heavenly - sacred; heroic - brave - honourable - noble*). All these links of words create a complex associative net.

In later years, the ideas put forward by Andreev (1965, 1967, 1969) were mainly developed in the direction of distributive-statistical analysis of morphology and those by Šajkevič in the direction of lexical-semantic studies and their application to information science needs (Pritsker, 1964; Moskovich, 1965, 1969; Ivanova, 1967, 1969; Ivanova and Moskovich, 1968; Šajkevič, 1970b). Particular attention in this latter set of papers was paid to the detection of statistical associative links among words in various intervals of text. The results obtained on large samples of text suggest that the most interesting linguistic results may be obtained in the minimal and medium intervals of text (from three words to a paragraph). As a rule, in the minimal interval, syntagmatic relations are detected, whereas in the medium - both syntagmatic and paradigmatic ones are found. First-generation and second-generation profiles of words were built, and it was shown that second-generation profiles reflect paradigmatic links with more accuracy. The same distributive-statistical procedure was applied to translations of the same text into different languages, and important typological features of semantics of these languages were discovered. Text-oriented specialized thesauri for various technical fields were created with a view to their subsequent application within information retrieval systems.

One of the most serious projects on the use of distributive-statistical techniques both for linguistic and information retrieval purposes was that of the Cambridge Language Research Unit. In the experiments of this group, cooccurrence of words in the maximal text interval (i.e., the whole text) was counted and the possibility was shown of detecting groups of words similar to those found in Roget's *Thesaurus of English Words and Phrases* (Sparck Jones, 1962, 1964).

Harper (1965, 1966) counted the cooccurrence of nouns in the interval of one sentence in a sample of 120 thousands words of Russian texts on physics. Cooccurrence of every noun with its governing and subordinate words was registered. Semantic distance between any two nouns was defined as the ratio of the quantity of governing and subordinate words

common for both of them, to the product of the frequencies of these nouns. Semantic groups of nouns were discovered which showed a good approximation to intuitively felt groups of semantically related words. Similar experiments with Russian verbs are described by Apresjan (1966,b). In some recent studies, cooccurrence of nouns with the same verbs as subjects (or objects) was used as a measure by which nouns and verbs were grouped into classes. Results obtained showed good agreement with a classification of nouns and verbs into semantic classes derived manually (Hirschman et al., 1975; Sager, 1975).

In the paper of Lewis et al. (1967), successful results of experiments on classification of synonyms and antonyms are reported. The authors used ten variations of formulae for measuring semantic distances. Three of these formulae were selected as giving the best approximation to an ideal classification of synonyms and antonyms.

Several attempts were made to build semantic networks on the basis of data on syntagmatic distances between words in texts (Ratceva, 1965, 1966a,b). Along the same lines, research on lexical collocations in text was performed. It seems that a level of linguistic organization distinct from both syntax and semantics exists, since statistically significant cooccurrences of words were found within a span of three words (Bulaševa, 1969) and of four words (Zueva, 1970; Jones and Sinclair, 1972) on either side of the study words. A number of such collocations display strong semantic cohesion.

A study of the notion of synonymy applicable both in linguistics and information retrieval was undertaken by Brodda and Karlgren (1969).

Only a part of linguistic research in statistical association analysis relied on ideas and methods developed by information scientists. An overview of the papers discussed above shows that almost all the research done in the USSR in the 1960s was initially independent of events in information science. This is explained by the fact that in those years advanced computers were relatively scarce in that part of the world, and statistical associative techniques could not have been used successfully on a large scale. In the USSR, there was almost no independent research on statistical association techniques in information science in the 1960s. All the ideas and experience came from quantitative linguistics. Actual machine experimentation for information retrieval purposes was started recently (Borodin and Kozokina, 1971). In the West, the situation was apparently different; pragmatic applications came first. But there is good evidence that at least part of that research was purely linguistic and that another part had from the start both information retrieval and linguistic aims.

Distributive-statistical text analysis as an objective method for detecting lexico-semantic links of words has great significance for modern computational linguistics, which relies heavily on semantic networks and dictionaries. In a special linguistic study (Moskovich, 1971, 1972b), the heuristic potential of three methods of detecting lexico-semantic links of words was tested: (1) distributive-statistical text analysis; (2) psycholinguistic word association analysis; (3) analysis of word definitions. Comparison of the results obtained by application of these methods brought the author to the conclusion that the most detailed and exact associative profiles of words are obtained with the help of the distributive-statistical method. Such profiles reflect the character of texts chosen for analysis.

A specific part of information science dealing with automatic synthesis of words (which are to serve as trademarks) and automatic retrieval of words-trademarks benefited from research in quantitative phonology. Corresponding information systems for processing trademarks were often devised by linguists and based on preliminary research of the laws

of phoneme combinatorics (Brodda and Karlgren, 1964; Moskovich, 1968). As was shown in these studies, combinatorial properties of phonemes depend upon their acoustical distinctive features, and phonemic proximity of trademarks was measured on the basis of the distinctive features of the phonemes composing the trademarks (Moskovich, 1968; Muljačič, 1967).

Research in statistical extraction techniques for automatic abstracting and indexing was comparatively independent of preliminary linguistic work, although some of the systems were designed by professional linguists. The method of automatic statistical abstracting of Agraev et al. (1963) is based on the evaluation of the *semantic weight* of a sentence which is regarded as a *coefficient of connection* of the analyzed sentence with the whole text of the document. The *coefficient of connection* is calculated according to a special formula, where the frequency of occurrence of words of the analyzed sentence in the whole text, the length of the text and the length of the sentence are the variables. Another method of statistical abstracting takes into consideration only the frequency of terms in the text and the number of words (terms and non-terms) in it (Purto, 1961).

Methods of statistical indexing which are usually based either on relative frequencies of words in documents or on the comparison of the frequency of words in a document with their frequency in the whole collection of documents are similar to those used in linguostatistics for defining themes of a text (Guiraud, 1960). Iker (1974, 1975) devised and tested a computer system, SELECT, which isolates the major themes of a text by selecting those words of the text which are most correlated with all other words and the set of most frequent words (excluding function words) and using them as input to a factor analysis.

For various applications, it is essential to know laws governing the distribution of words in text. Most of the relevant research in this area was done within quantitative linguistics not by linguists, but by expert mathematicians. The law of rank distribution of words known under many names as the law of Zipf, Bradford, Lotka, Estoup, Mandelbrot, or Wyllys has been discussed in many publications. Some of authors consider this law neither a linguistic nor a mathematical one (Herdan, 1960, 1964); others point to its correspondence to facts of different languages (Sambor, 1969; Orlov, 1970; Arapov et al. (1975a,b). Studies of distributions of linguistic units in texts showed that the character of their distributions is not constant and varies according to the interval of text in which distribution is measured (Šajkevič, 1970a). However, for practical reasons it is accepted that words of medium and low frequency are distributed according to Poisson's law. Measures for evaluating the strength of connections in distributive-statistical association nets are sometimes based on this assumption (Šajkevič, 1963).

Frequencies of linguistic units in text reflect their roles and properties in the structure of language. A notion of productivity of linguistic units was suggested in (Moskovich, 1969). The productivity of a linguistic unit is defined as its importance and role in the structure of language. The productivity of words can be measured by the quantity of its meanings, quantity of its derivatives and quantity of set expressions with this word as one of the elements. Frequency of words strongly correlates with their productivity. The position of an element within linguistic structure is reflected in its frequency (Moskovich, 1969). This rule applies not only to the lexico-semantic level of language, but to the morphological one as well. Productivity of an affix is defined as the quantity of words-derivatives with this affix in a language. Recent investigations showed that there is a striking correlation between frequency of affixes and their productivity. Among the 50 most frequent and 50 most productive suffixes in Russian, 36 are common to both lists; among the 30 most frequent and 30 most productive prefixes, 28 are common (Kuznecova and Lavrenova,

1975).

The idea of comparing word frequencies with their semantic properties is not new. There were suggestions for creating a new type of semantic dictionary of the type of Roget's Thesaurus with frequencies of individual words and semantic groups indicated. This idea was put into practice in quantitative studies of poetic vocabulary (Abramova, 1974). Further studies of this kind may provide a new insight into the structure of language.

Evaluating the state of the art of quantitative linguistics, one notices that although considerable efforts were made in the last 20 years to study the quantitative side of linguistic events and valuable results were achieved, there remain more questions than answers. Research in quantitative linguistics consumes too much time for mechanical work. This work is like "a gigantic wall that a researcher built between himself and the text; it serves as excuse of the fact that he did not think of the text itself" (Guiraud, 1963, p. 45). Even if the actual counting is not done manually but with the help of computers, linguostatistical data received as a computer output are too voluminous to be evaluated easily. That is why special attention should be paid to the quality of the statistical hypotheses suggested for testing. Statistical data reflect various tendencies and constraints in language structure, often of a conflicting character, and it is not always easy (or even possible) to disentangle them. A negative attitude to quantitative linguistics on the part of some linguists is a result of identifying the whole discipline with boring statistical computations. The best side of quantitative linguistic research, the one bringing new linguistic insights into the structure of language, remains unknown, or is considered unimportant by the critics. Modern trends in theoretical linguistics are directed more towards linguistic competence rather than performance.

Discussing conflicting approaches to language of theory-oriented linguists and data-oriented researchers in automated language processing, Montgomery writes:

> *If one takes a negative point of view, these dichotomies represent irreconcilable differences in the basic conception of language; more positively, they may be regarded as complementary perspectives on the nature of language. The initial issue is thus one of determining which view is correct. Should the positive view be adopted, there is a more fundamental question as to the potential for unifying the two approaches to provide a balanced attack on problems of natural language analysis and description... It is reasonable to consider the two approaches complementary, since the specific weaknesses of the data-oriented position are offset by corresponding strengths in theoretical orientation, and conversely* (1969, p. 12).

In our opinion, the inductive study of language structure, although it is now out of fashion, will always be indispensable both for theoretical linguistics and for its applications. Quantitative linguistics not only provides an explanation for a whole dimension of language structure, but is of immediate use for various applications of linguistics.

Contribution of Quantitative Linguistics to Information Science (Retrospect and Prospect): Conclusion

On the preceding pages, we returned several times to the assertion that although the aims of research in quantitative linguistics and information science are different, these disciplines treat the same material with the same or similar methods, and can only benefit by mutual cooperation. After the revealing studies of Salton and his associates on the SMART project

(Salton, 1971) and the appearance of the book by Sparck Jones and Kay (1973), no one can deny the fact that the application of vocabulary statistics to document retrieval brings more immediate results than the use of other more sophisticated linguistic techniques. However, although the pragmatic result is there, the work of statistical techniques has not received an adequate linguistic explanation within that or another linguistic theory, and it is not always clear if statistical operations slide over the surface structure of text or touch deep structure layers of language as well. There is no doubt that without touching the deep structure layers of language, quantitative techniques could not produce meaningful results; but few attempts were made to explain the inner mechanics of quantitative language analysis. Providing the explanation of how and why statistical methods produce the results they do, may be a major contribution of quantitative linguistics to information science. Such an explanation may give a solid theoretical basis for the creation of new, more effective information retrieval systems.

Quantitative linguistics contributes to information science at least in three domains:

1. It helps to create a lexicographic basis for information systems.
2. It helps to solve the problems of automatic indexing, abstracting and comparison of documents.
3. It helps to elucidate complex quantitative laws of text organisation.

In recent years, there has been a noticeable tendency towards mutual understanding and cooperation between quantitative linguists and informatior scientists. A number of information retrieval projects were devised by quantitative linguists, and special quantitative studies of texts and compilation of frequency dictionaries of words and word combinations were conducted prior to actual building and testing of an information retrieval system (e.g., the Laboratory of Applied Linguistics of the Moscow State University devised a large information system for processing data on military field operations built entirely on quantitative linguistic criteria (Kolgushkin, 1970); all the linguistic components of this system are based on previously compiled frequency dictionaries of words, word pairs and n-tuples). Authors of frequency dictionaries often declare the aim of their work to be information science applications.

Of particular importance both for linguistics and information science is research in the field of distributive-statistical techniques of text analysis. Associative nets of words built by distributive-statistical techniques have some peculiar characteristics which make them particularly suitable for information retrieval purposes: they include only the terms that actually appeared in documents; they are created by an algorithmic procedure which can be repeated by the computer on texts of any length and any subject (that makes the procedure extremely attractive for new, rapidly developing areas of science for which no dictionaries or thesauri are available); they reflect specific statistical links of words which cannot be detected by any other known linguistic method. In the course of testing and implementing the distributive-statistical method of thesauri construction on large text samples, some major problems of automatic information retrieval are being solved as a by-product (e.g., elementary statistical parameters, such as average frequency and dispersion, are used as criteria for discriminating terms from nonterms; measures of cooccurrence of words lead to lists of set word combinations which are to be included as independent lexical units into dictionaries of automatic information retrieval systems). Though much was done in this area, much more has to be done since a whole gamut of theoretical problems has not yet been solved, such as criteria for delimitation of parts of an association net into semantic fields; criteria for distinguishing substantial links of terms from unsubstantial ones; criteria

for distinguishing the effects of different associative term structures on retrieval performance from those of other information retrieval system components; etc.

One can foresee a wider application of machine-built thesauri based on distributive-statistical text analysis in the years to come. Initial hopes that the use of such thesauri might lead to fully automatic information retrieval systems seem to have given way to more realistic expectations of a wider use of *automatic thesauri* in man-machine information retrieval systems of interactive searching. As more knowledge is gained on the linguistic nature of these thesauri, of their properties and the influence of these properties on retrieval performance, distributive-statistical thesauri will be more and more subjected to human postediting and correction with some semantic links being removed and others, more useful for retrieval purposes, given priority or introduced. Distributive-statistical thesauri may prove to be extremely useful for artificial intelligence systems.

With the advent of user-oriented natural language information retrieval systems, a tendency to incorporate statistical and distributive-statistical algorithms into more general algorithms for *total* automatic text analysis (including morphological, syntactic and semantic analysis) may be expected. The role of the linguostatistical component of such systems will be extremely important for the automatic formation of thesauri, text compression and weighting of keywords, sentences and larger parts of text.

Contrary to the opinion prevailing among research workers in automatic data processing that precoordinate indexing systems will become obsolete in the near future, a wide use of such systems of document classification as Universal Decimal Classification, International Patent Classification and possibly other document classifications of new types is to be expected. These events will call forth more attention to work in automatic classification indexing based on linguostatistical criteria. There are various possibilities of organizing classification work in terms of man-machine interaction, e.g., with the computer providing a tentative list of classification tags for a document to be subsequently used by an indexer.

Quantitative linguistics has still to answer seemingly easy, but actually extremely complicated questions: What is a semantic unit for counting? What are the criteria for delimiting words in running text? In what instances is a word combination to be considered a semantic unit equivalent to a word in counting?

In future uses of linguostatistical techniques for information retrieval, more attention will be paid to morphological analysis reduction of words to stems, conflation of singular and plural forms of the same word, etc. Such simple devices considerably improve the performance of information retrieval systems.

Quantitative linguistics will have its impact on future research on text units larger than a sentence. Work on automatic abstracting is dependent upon progress in this area. It remains to investigate which units may be chosen as representatives of the contents of a document and by what criteria they may be selected.

In conclusion, it remains to add that justified criticism or reevaluation of the contribution of linguistics to its various applications, be it the ALPAC report (1966) or the review book by Sparck Jones and Kay, only furthers the progress of linguistics. It helps us to identify our weak points, and to develop strategies for coping with the current situation in applied fields. It would be wrong to draw exclusively negative, pessimistic conclusions from the existence of this criticism. Most of the problems of automatic text processing, including information retrieval, are of a linguistic nature, and will be solved in the course of time by

proper, more advanced linguistic techniques. In future developments, the techniques of quantitative linguistics will be interlocked more closely with other linguistic techniques and without any doubt will be extremely useful for various applied fields. The specific character of quantitative linguistics as a linguistic discipline comes from the fact that it is the properties of linguistic units, and not just their frequencies that stand at the focus of its attention. The linguist begins quantitative linguistic research with an initial phase of establishing a unit of counting and formulating a hypothesis. But his linguistic work proper starts only when the numbers have been counted and he has to interpret them.

It is this author's contention therefore, that quantitative linguistics is indeed a major area within linguistics, and is not, as some authorities contend, merely a branch of applied mathematics beyond the boundaries of linguistics. Moreover, it is, we feel, an area which has already contributed much, and through bold and innovative research, has much more to add to the progress of information science.

References*

ALPAC. *Language and Machines: Computers in Translation and Linguistics.* Publication 1416. National Academy of Sciences, Washington, D.C., 1966.

Abramova, N. I. *Poetičeskaja Leksika Francuzskogo Jazyka (na Materiale Francuzskoj Poezii XIX Veka) Poetic Vocabulary of the French Language* (on the Material of French Poetry of the XIX Century). 1 MGPIIJA. Moscow, 1974.

Agraev, V. A., Borodin, V. V., and Glebskij, Ju. V. "O Nekotoryx Metodax Avtomatičeskogo Referirovanija" (On Certain Methods of Automatic Reviewing). In: *Ucenye Zapiski Gor'kovskogo Gosudarstvennogo Universiteta im. N.I. Lobačevskogo, 66. Serija Filologija, Prikladnaja Lingvistika i Metodika.* Gor'kij, 1963.

Akhmanova, O. S. "Slovar' Lingvističeskix Terminov" (Dictionary of Linguistic Terms). Moscow, *Sovetskaja Enciklopedija,* 1966.

Andreev, N. D. "Modelirovanie Jazyka na Baze Ego Statističeskoj i Teoretiko-Množestvennoj Struktury" (Modelling Language on the Basis of Its Statistical and Set-Theoretical Structure) In: *Tezisy Soveščanija po Matematičeskoj Lingvistike.* Leningrad State University, Leningrad, 1959. Pp. 16-22.

Andreev, N. D. "Vozmožnyj Put' Modelirovanija Semantiki Jazyka" (A Possible Way for Modelling Semantics of Language). *Doklady na Konferencii po Obrabotke Informacii, Mašinnomu Perevodu i Avtomatičeskomu Čteniju Teksta,* Issue 10. Moscow, VINITI, 1961.

Andreev, N. D., ed. *Statistiko-Kombinatornoe Modelirovanie Jazykov (Statistical-Combinatorial Modelling of Languages).* Moscow - Leningrad, Nauka, 1965.

Andreev, N. D. *Statistiko-Kombinatornye Metody v Teoretičeskom i Prikladnom Jazykoznanii.* Leningrad, Nauka, 1967.

---

*Titles in parenthesis are English translations of the titles of publications in various Slavic languages.

Andreeva, L. D. *Statistiko-Kombinatornye Tipy Slovoizmenenija i Razrjady Slov v Russkoj Morfologii* (*Statistical-Combinatorial Types of Word Inflexion and Categories of Words in Russian Morphology*). Leningrad, Nauka, 1969.

Apresjan, Ju. D. "Algoritmy Postroenija Klassov po Matritse rasstojanij" (Algorithms for Building Classes on the Matrix of Distances). *Mašinnyi Perevod i Prikladnaja Lingvistika*, 1966, 9, 3-18. **(a)**

Apresjan, Ju. D. *Idei i Metody Strukturnoj Lingvistiki* (*Ideas and Methods of Structural Linguistics*). Moscow, Prosvescenie, 1966. **(b)**

Arapov, M. V., Efimova, E. N., and Šrejder, Ju.A. "Rangovye Raspredelenija v Tekste i Jazyke" (Rank Distributions in Text and Language). *Naučno-Texničeskaja Informacija, Series* 2, 1975, 2, 3-7. **(a)**

Arapov, M. V., Efimova, E. N., and Šrejder, Ju. A. "O Smysle Rangovyx Raspredelenij" (On the Substance of Rank Distributions). *Naučno-Texničeskaja Informacija*, Series 2, 1975, 1, 9-20. **(b)**

Baldwin, A. L. "Personal Structure Analysis: A Statistical Method for Investigating the Single Personality." *Journal of Abnormal and Social Psychology*, 1942, 37, 163-183.

Borodin, V. V., and Kozokina, S. M. "Postroenie Grafa Sovmestnoj Vstrečaemosti na EVM" (The Building of a Graph of Cooccurrence on a Computer). In Moskovich, W., ed., *Voprosy Lingostatistiki i Avtomatizacii Lingvističeskix Rabot*, Volume 5. Moscow, Patent, 1971. Pp. 59-67.

Brodda, B., and Karlgren, K. "Relative Positions of Elements in Linguistic Strings." *Statistical Methods in Linguistics*, 1964, 3, 49-101.

Brodda, B., and Karlgren, H. "Synonyms and Synonyms of Synonyms." *Statistical Methods in Linguistics*, 1969, 5, 3-17.

Bulaševa, N. S. "Statistika Trexslovnyx Sočetanij s Opornym Častotnym Slovom iz Tekstov po Russkoj Radioelektronike" (Statistics of Three Word Combinations with a High Frequency Main Word from Russian Texts on Radioelectronics). In Piotrovskij, R.G., ed., *Statistika Teksta*, Volume 1. Minsk, 1969. Pp. 376-382.

Doyle, L. B. *Library Science in the Computer Age.* Report SP-141. System Development Corporation, Santa Monica, California, 1959.

Giuliano, V. E., and Jones, P. E. *Linear Associative Information Retrieval.* Report CACL-2. Arthur D. Little, Inc., Cambridge, Massachusetts, 1962.

Gladkij, A. V., and Mel'čuk, I. A. *Elementy Matematiceskoj Lingvistiki* (*Elements of Mathematical Linguistics*). Moscow, Nauka, 1970.

Gorodeckij, B. Ju. "Review of H. Kučera and W.N. Francis. Computational Analysis of Present-Day American English." *Linguistics*, 1972, 84.

Guiraud, P. *Problèmes et Méthodes de la Statistique Linguistique.* P.U.F., Paris, 1960.

Guiraud, P. "La Mécanique de l'Analyse Quantitative en Linguistique." In: *Études de Linguistique Appliquée,* Volume 2. Paris, 1963.

Harper, K. E. *Measurement of Similarity Between Nouns.* Memorandum RM-4532PR. The Rand Corporation, Santa Monica, California, 1965.

Harper, K. E. *Some Combinatorial Properties of Russian Nouns.* Memorandum AD-638924. The Rand Corporation, Santa Monica, California, 1966.

Herdan, G. *Type-Token Mathematics.* The Hague, Mouton, 1960.

Herdan, G. "Quantitative Linguistics or Generative Grammar?" *Linguistics,* 1964, 4, 56-65.

Hirschmann, L., Grishman, R., and Sager, N. "Grammatically-Based Automatic Word Class Formation." *Information Processing and Management,* 1975, 11, 39-57.

Iker, H. P. "SELECT: a Computer Program to Identify Associationally Rich Words for Content Analysis. I. Statistical Results." *Computers and the Humanities,* 1974, 8, 4, 213-319.

Iker, H. P. "SELECT: A Computer Program to Identify Associationally Rich Words for Content Analysis. II. Substantive Results." *Computers and the Humanities,* 1975, 9, 3-12.

Ivanova, N. S. "Ustanovlenie Smyslovyx Svjazej Meždu Slovami na Osnove Statističeskoj Metodiki." In Moskovich, W., ed., *Voprosy Lingvostatistiki i Avtomatizacii Lingvostatističeskix Rabot,* Volume 1. Moscow, Patent, 1967. Pp. 52-61.

Ivanova, N. S. "K Voprosu ob Avtomatieskom Postroenii Tezaurusa" (On Automatic Thesaurus Construction). *Naučno-Texničeskaja Informacija,* Series 2, 1969, 6, 17-20.

Ivanova, N. S., and Moskovich, W. Automatic Compiling of Thesauri on the Basis of Statistical Data." In: *Information Retrieval Among Examining Patent Offices.* VII Annual Meeting, BIRPI, Geneva, 1968.

Ivanova, N. S., and Šajkevič, A. Ja. "Distributivno-Statističeskoje Opisanije Amerikanskix Patentnyx Tekstov" (Distributive-Statistical Description of American Patent Texts). In Moskovich, W., ed., *Voprosy Lingvostatistiki i Avtomatizacii Lingvističeskix Rabot,* Volume 4. Moscow, Patent, 1970. Pp. 77-221.

Jones, S., and Sinclair, J. McH. "English Lexical Collocations." *Cahiers de Lexicologie,* 1972, 23(2), 15-61.

Karlgren, H. "Quantitative Models - of What?" *Statistical Methods in Linguistics,* 1975, 25-31. (a)

Karlgren, H. "Text Connexivity and Word Frequency Distribution." In Ringbom, H., ed., *Style and Text.* Studies Presented to Nils Erik Enqvist. Stockholm, Skriptor, 1975. Pp. 1-14. (b)

Kiefer, F. "Some Aspects of Mathematical Models in Linguistics." *Statistical Methods in Linguistics,* 1964, 3, 8-26.

Kolgushkin, A. N. *Lingvistika v Voennom Dele (Razrabotka i Ispol'zovanie Častotnyx Slovarej Voennoj Leksiki) (Linguistics in Military Science (Compilation and Use Of Frequency Dictionaries of Military Words)).* Moscow, Voenizdat, 1970.

Kučera, H., and Francis, W. N. *Computational Analysis of Present-Day American English.* Providence, Rhode Island, Brown University Press, 1967.

Kuznecova, A. I., and Lavrenova, O. A. "O Suščestvovanii Korreljacii Mezdu Produktivnostju i Upotrebitel'nostju Affiksov v Russkom Jazyke" (On the Existence of Correlation Between Productivity and Frequency of Affixes in Russian). In Zvegincev, V.A., ed., *Issledovanija po Strukturnoj i Prikladnoj Lingvistike.* Moscow State University, Moscow, 1975. Pp.83-99.

Lewis, P. A. W., Baxendale, P. B., and Bennett, J. L. "Discrimination of the Synonymy/Antonymy Relationship Between Words." *Journal of the ACM,* 1967, 14, 20-44.

Montgomery, C. A. "Linguistics and Automated Language Processing." *COLING.* Preprint No. 1. Stockholm, 1969.

Moskovich, W. "Opyt Kvantitativnoj Tipologii Semantičeskogo Polja" (An Experiment in Quantitative Typology of a Semantic Field). *Voprosy Jazykoznanija,* 1965, 4, 80-91.

Moskovich, W. "Avtomatizacija Nekotoryx Aspektov Lingvističeskoj Raboty" (Automatization of Some Aspects of Linguistic Work). *Voprosy Jazykoznanija,* 1966, 1, 102-111.

Moskovich, W., ed. *Voprosy Lingvostatistiki i Avtomatizacii Lingvostatističeskix Rabot (Problems of Linguistic Statistics and Automatization of Linguostatistic Work).* Vols. 1-6. Moscow, Patent, 1967-1974.

Moskovich, W. "Typological Classification of Information Retrieval Languages and Transcriptions." In: *Information Retrieval among Examing Patent Offices.* VII Annual Meeting. BIRPI, Geneva, 1968. Pp. 248-269.

Moskovich, W. *Statistika i Semantika (Statistics and Semantics).* Moscow, Nauka, 1969.

Moskovich, W. *Informacionnye Jazyki (Information Languages).* Moscow, Nauka, 1971.

Moskovich, W. "Distributivno-Statističeskij Metod Postroenija Tezaurusov: Sovremennoje Sostojanije i Perspektivy" (Distributive-Statistical Method of Thesaurus Construction: The State of the Art and Perspectives). *Naučno-Texničeskaja Informacija,* Series 2, 1972, 3, 12-21; 4, 15-24. (a)

Moskovich, W. "Metody Obnaruženija Leksiko-Semanticeskix Svjazej Slov" (Methods of Detection of Lexico-Semantic Links of Words). *Inostrannye Jazyki v Skole*, 1972, 6, 10-22. (b)

Muljačič, Z. "La Combinabilité des Phonemes sur l'Axe Syntagmatic dépend-elle de leurs Traits Distinctifs?" In: *Phonologie der Gegenwart.* Wien, 1967.

Needham, R. M. *Research on Information Retrieval, Classification and Grouping*, 1957-1961. Ph.D. Thesis, Cambridge University, 1961.

Orlov, Ju.K. "O Statističeskoj Strukture Soobščenij, Optimal'nyx dija Čeloveceskogo Vosprijatija" (On the Statistical Structure of Messages Optimal for Human Perception). *Naučno-Texničeskaja Informacija*, Series 2, 1970, 8, 11-16.

Pritsker, A. J. "Distributivno-Statističeskij Analiz Semantičeskogo Polja" (Distributive-Statistical Analysis of a Semantic Field). In: *Problemy Formalizacii Semantiki Jazyka. Tezisy Dokladov*, Moscow, 1964. Pp. 130-132.

Purto, V. A. "Ob Avtomatičeskom Referirovanii na Osnove Statističeskogo Analiza Teksta" (On Automatic Abstracting on the Basis of Statistical Text Analysis). In: *Doklady na Konferencii po Obrabotke Informacii, Masinnomu Perevodu i Avtomatičeskomu Čteniju Teksta.* Moscow, VINITI, 1961.

Ratceva, I. I. "Algoritmizacija Issledovanija Smyslovych Svjazej" (Algorimization of Investigation of Semantic Links). *Naucno-Texniceskaja Informacija*, 1965, 8, 35-42.

Ratceva, I. I. "Eksperimenty po Avtomaticeskomu Vyboru Smyslovych Kategorij na Dvujazyčnyx Tekstax" (Experiments on Automatic Choice of Semantic Categories in Bilingual Texts). In: *Vsesojuznaja Konferencija po Informacionno-Poiskovym Sistemam i Avtomatizirovannoj Obrabotke Naučno-Texničeskoj Informacii*, 3-ja, Volume 2. Moscow, VINITI, 1966. P. 249. (a)

Ratceva, I. I. "Problema Vybora Znacenija Slova i Smyslovye Rasstojanija" (The Problem of Selection of Word Meaning and Semantic Distances). *Naučno-Texničeskaja Informacija*, 1966, 5, 36-47. (b)

Sager, N. "Computerized Discovery of Semantic Word Classes." In Grishman, R., ed., *Directions in Artificial Intelligence: National Language Processing.* Courant Institute of Mathematical Sciences, New York University, New York, 1975. Pp. 27-48.

Šajkevič, A. Ja. "Raspredelenie Slov v Tekste i Vydelenie Semantičeskix Polej Jazyka" (Distribution of Words in Text and Discovery of Semantic Fields of Language). In: *Tezisy Dokladov Mežvuzovskoj Konferencii po Primeneniju Strukturnyx i Statističeskix Metodov Issledovanija Slovarnogo Sostava Jazyka.* Moscow, 1961.

Šajkevič, A. Ja. "Raspredelenie Slov v Tekste i Vydelenie Semantičeskix Polej Jazyka" (Distribution of Words in Text and Discovery of Semantic Fields of Language). In: *Inostrannye Jazyki v Vysšej Škole*, Volume 2. Moscow, 1963. Pp. 14-26.

Šajkevič, A. Ja. "Interval Teksta i Xarakter Statističeskix Raspredelenij Jazykovyx Edinic" (Interval of Text and Character of Statistical Distribution of Linguistic Units). In

Moskovich, W., ed., *Voprosy Lingvostatistiki i Avtomatizacii Lingvisticeskix Rabot*, Volume 3. Moscow, Patent, 1970. Pp. 15-22. (a)

Šajkevič, A. Ja. "Korreljacionnyj Analiz v Lingvostatistike i Ponjatie Intervala Teksta" (Correlational Analysis in Linguostatistics and the Notion of Text Interval). In Moskovich, W., ed., *Voprosy Lingvostatistiki i Avtomatizacii Lingvostatističeskix Rabot*, Volume 2. Moscow, Patent, 1970. Pp. 254-274. (b)

Salton, G., ed. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, New Jersey, Prentice-Hall, 1971.

Sambor, J. *Badania Statystyczne nad Slownictwem (Statistical Investigation of Vocabulary).* Warszawa, 1969.

Skorohod'ko, E. F. "Pro Zastosuvannja Elektronnyx Cyfrovyx Masyn u Lingvistyčnyx Doslidženjax" (On the Use of Computers in Linguistic Research). In: *Ukrains'ka Respublikans'ka Naukova Konferencija z Pytan' Metodologii Movoznavstva.* Kiev, 1964. Pp. 49-50.

Sparck Jones, K. "Mechanized Semantic Classification." Paper 25. *1961 International Conference on Machine Translation of Languages and Applied Language Analysis.* National Physical Laboratory Symposium No. 13. Teddington, England, 1962.

Sparck Jones, K. *Synonymy and Semantic Classification.* Ph.D. Thesis, University of Cambridge, 1964.

Sparck Jones, K., and Kay, M. *Linguistics and Information Science.* New York, Academic Press, 1973.

Šrejder, Yu. A. "O Statuse Matematičeskoj Lingvistiki" (On the Status of Mathematical Linguistics). In: *Voprosy Informacionnoj Teorii i Praktiki.* Sbornik No. 27. Moscow, VINITI, 1975. Pp. 7-18.

Stevens, M. E., Heilprin, L., and Giuliano, V. E., eds. *Statistical Association Methods for Mechanized Documentation.* Miscellaneous Publication 269. Washington, D.C., National Bureau of Standards, 1965.

Zueva, T. R. "Statističeskaja Xarakteristika Četyrexslovnyx Sočetanij Russkogo Jazyka po Elektronike" (Statistical Characteristics of Four Word Combinations in Russian Texts on Electronics). In: *Pervaja Naučnaja Konferencija Čimkentskogo Pedinstituta.* Cimkent, 1970. Pp. 48-53.