American Journal of Computational Linguistics *Microfiche 28*

SUMMARIES OF CONTRIBUTIONS

TO THE

FIRST NATIONAL CONFERENCE

ON THE APPLICATION OF MATHEMATICAL MODELS

AND COMPUTERS IN LINGUISTICS

Part II Russian Language Contributions

VARNA, BULGARIA MAY 3 - 9, 1975

A Ljudskanov
Institute of Mathematics and Mechanics
Chairman of the Organizing Committee

Copyright © 1975
Association for Computational Linguistics

Editorial Note

Summaries originally prepared in Roman-alphabet languages were reproduced on AJCL Microfiche 20. The present fiche contains translations of 46 summaries originally submitted in Russian.

AJCL thanks the organizers for permission to reprint the summaries and David Ackerman who made the translations

--D G H.

TABLE OF CONTENTS

GENERAL

PHONETICS - PHONOLOGY

LEXICOGRAPHY - LEXICOLOGY

GRAMMAR

SEMANTICS - DISCOURSE

LINGUISTICS

COMPUTATION

DOCUMENTATION

THE LOGICAL FOUNDATIONS OF MEASUREMENT OF SEMANTIC INFORMATION

NIKOLAJ STANULOV

A method is described for determining the relative quantity of semantic information (QSI) in a written natural-language text (the QSI method). The construction of the method is based on the ideas of logical semantics, mathematical logic, and the so-called thesaurus approach. The logical bases are given with the help of axioms of semantic information and their corollaries Information, and hence semantic information, is described in set-theoretic fashion. The natural language enters the role of object language; proceeding from it a formal language is constructed whcih is conditionally termed an algebraic information (AI) language; this is the fragmentary metalanguage defined by the correspondence of the object language. The translation of texts from this language into the AI language (i.e. the compilation of AI expre-sions) allows the determination of the quantity of information in object-language texts. The quantity of semantic information is expressed by the element number of semantic space of the type of algebraic field of real (and possibly of other) numbers. The QSI method includes as a frequent case the solving of verbal problems in mathematics (not excluding the measurement among the numbers of the QSI) Several means for the analysis of texts are discussed, as are the rules for compiling AI expressions

The proposed method is suitable for application to any texts, particularly in informatics--in information processing (indexing and the like) of bibliographic materials.

ON DISCRETE (NON-SHANNON) CALCULATION OF ENTROPY

IVAN DOBREV

R. V L. Hartley (1928) introduced the following mathematical expression of entropy H. H= log n, where n is the number of equiprobable experimental outcomes This formula can be understood only as a postulate and by no means as a formal, logical conclusion from more elementary proposals, since there is an error in Hartley's chain of deductions. From the mathematical point of view, the conclusion of C. Shannon (1942) that H = log n is fully rigorous, but it is the consequence of very strong preliminary assumptions (axioms) on entropy as a function of the number of equiprobable outcomes n According to Shannon, the function H = f(n) must be a continuous, additive and monoton function. How-ver continuity and additivity are not necessary conditions. The identification of one of the n equiprobable outcomes of the experiment by the method of binary questions is a discrete process. A simple experiment with n = kl equiprobable outcomes and a sequence of two simple experiments with k and 1 equiprobable outcomes are two different things.

Assuming that entropy is the measure of the average quantity of binary questions necessary for the identification of one of the equiprobable outcomes of the experiment we set $\phi(1) = 0$ and the recurrence relation $\phi(n) = 1 + \phi(n/2)$ for all even values and $\phi(n) = 1 + \frac{n-1}{2n}\phi(\frac{n-1}{2}) + \frac{n+1}{2n}\phi(\frac{n+1}{2})$ for all odd values. It is immediately verified that $\phi(n)$ $\log_2 n$ and that the difference does not exceed 0.09 Using the definition of the function $\phi(n)$ and the approximation $\phi(n) \simeq \log_2 n$ with an absolute error 0 09, one may calculate the entropy of the experiment with independent but not equiprobable results. $H \sim -\sum_i p_i \log_2 p_i$.

AUTOMATION OF THE QUANTITATIVE ANALYSIS OF TEXTS

R. K. KIRKOVA

Under examination are some means of automating the process of quantitative analysis of texts through the utilization of computers.

Algorithms are cited for the creation of an integral system of machine processing of textual information, where the basic goal is the acceleration of the computer's work and the significant abbreviation of the volume of the text subject to processing by the linguist, by means of the compilation and utilization of frequency dictionaries.

Also cited are some results received in the conduct of experiments in the Computing Center of the Institute of Mathematics and Mechanics of the Bulgarian Academy of Sciences on material of Bulgarian artistic and scientific texts

ON THE QUANTITATIVE FOUNDATION OF LINGUISTIC CLASSIFICATIONS

MIROSLAV YANAKIEV

In the process of working out a frequency dictionary of contemporary Bulgarian, it has become clear that the classification procedures, while at least having a place in Bulgarian linguistic investigations, are implicitly founded on the information, intuitively accumulated in the knowledge of linguists, on the statistical characteristics of distribution of the language facts studied in texts. Unfortunately the linguist does not by any means always succeed in studying texts in such quantity that the automatisms accumulated in his knowledge of statistical information on the phenomena being studied might create a sufficiently reliable fulcrum for the execution of classification procedures, which are "thinner" than those already long known in linguistic tradition.

In this we can see the profound reason for the cul-de-sac into which contemporary linguistics has wandered. Inasmuch as the application of computers for the extraction of information on the statistical characteristics of the distribution of language facts makes work possible with enormous files of text, computer linguistics is the exit from this impasse.

AN ALGORITHM FOR THE RECOGNITION OF ISOLATED WORDS BY THE ENVELOPE OF THE SIGNAL

YU. MARINOV, S. TSOCHEV, M. KHARDALOV, AND B. ZHECHEV

This work advances the results of computer modeling of the recognition, of a limited application, of words by their envelopes To this end the first ten numbers, pronounced by different speakers, are used. The indications by which standards are formed are invariant with respect to compression-type linear transformation and homothety.

If $\{A_i\}$ is the set of extreme values of a given realization and $\{t_i\}$ are the moments at which the envelope reaches them, then the speech signal is represented as two ordered sequences of discrete values of the indicated characteristics:

$$\eta_i = \frac{A_i - A_{i+1}}{A_{i+1} - A_{i+2}}, \quad \tau_i = \frac{t_{i+1} - t_i}{t_{i+2} - t_{i+1}} \tag{1}$$

The averaging of the characteristics formed as n-dimensional vectors permits the formation of standards for comparison. The metrics in the waveform space are determined in the following manner:

$$d(x,y) = \sum_{i=1}^{N} K_i (x_i - y_i)^2 \tag{2}$$

Here $\{K_i\}$ are weighted coefficients determined by study and $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$ are vectors of characteristics. The program realization of the algroithm, carried out on the M220-M computer in Algol, has testified to the algorithm's good functioning. The simplicity of sign-formation renders it attractive for a number of applications, especially when work on the real-time scale is necessary.

SOME PROBABILITY PARAMETERS OF A STRUCTURAL-SEMANTIC DESCRIPTION OF THE BULGARIAN LEXICAL SYSTEM

D. A. RAJNOVA

The problem of creating such a grouping of the dictionary content of a language in which the separate word would neighbor those words with which it relates semantically and functionally, and in directo proximity with the word from which it is formed, has long awaited solution.

Practically, this means the consideration of the word as boundary set of subsystems. We have carried out such an examination with the help of punched cards with edge-notching and machine information processing. We have termed the working method structural semantics. In the process of this development, the necessity has arisen for new terms. The term "conceptual bundle", which we have accepted conditionally, embraces words with one and the same internal linguistic motivation khe characteristics of the bundle are defined as the external valency of a certain nuclear pireme and as the sum of the frequencies of the separate lexical units entering into its composition, determined on the basis of their greatest separate weight of their functioning in syntactic constructions.

The method we have worked out for the structural-semantic description of real lexical systems has permitted the determination of a number of probability parameters for the Bulgarian lexical system.

The examination of some of these is the object of this work.

THE QUANTITATIVE AND SEMANTIC CHARACTERISTICS OF COMMUNICATIONS VERBS IN JOURNALISTIC TEXTS

GERGANA D. MIKHAJLOVA

The object of research is the class of communications verbs (Vc). The class Vc is delimited with the help of four filters. I. a Vc is copied from dictionaries or other linguistic handbooks; II. a Vc is isolated from a text in a logical-intuitive manner; III. the Vcs isolated by filters I and II are verified in a combination matrix; IV. the presence of the semantic factor "input f communication" is sought out, with the help of componential analysis; for each verb having passed through filters I, II, and III.

The investigation of the class Vc is carried out on a selection from journalistic texts (on one theme or another), 10,000 word-forms in volume. The quantitative distribution of Vcs in the given selection is observed, taking into account X, B, V. On the basis of these data, conclusions are drawn (with a certain degree of reliability) on the appearance of Vcs in texts which are homogeneous in style and theme.

The lexical and grammatical combinability of Vcs with the communications agent (N) is studied. The communications agents are unified by the semantic factor of "inputs of communication" Those cases are followed up in which the distributive relations (lexical, grammatical) for various Vcs coincide, and account is taken of the influence of the distribution on the semantic proximity of the verbs of the class Vc. The intensity of combination between N and Vc is determined.

The "input mode of communication" (S) is examined. Basic constructions are distinguished with the help of the Vcs. The characteristic construction(s) for each verb of the class Vc is determined. For verbs of communication input semantic nearness is found by the same method. The value of the Vc-S correlation coefficient is given.

A FREQUENCY DICTIONARY OF CONTEMPORARY BULGARIAN JOURNALISTIC LANGUAGE

RADOSLAV MUTAFCHIEV

The aim of the frequency dictionary of contemporary Bulgarian journalistic language is to reflect the statistical character of Bulgarian journalistic style over the past ten years (1966-1975)

The dictionary is compiled on the basis of a textual file of 500,000 grammatical forms. It embraces all journalistic genres and is compiled by methods accepted abroad.

Each separate lexeme is represented with the following statistical indices. absolute frequency, relative frequency, arithmetic mean, mean-square deviation, and the coefficient U. The entire selection is divided into 25 samples of 20,000 word-forms each. Under each lexeme is given its allolexemes with their absolute frequencies and their statistical indices.

The dictionary will contain the following divisions an alphabetic list of lexemes and their allolexemes, a ranked list of lexemes with their absolute frequencies, a diagram of the frequencies of all grammatical categories encountered in the selection.

The dictionary is intended for research in the field of linguistics. It is being compiled by the Group in Structural, Quantitative, and Applied Linguistics of the Bulgarian Language Institute of the Bulgarian Academy of Sciences. The mathematical guarantee and also the compilation of algorithms and programs is conducted by Docent D. Dobrev and R. Kirkova of the Institute of Mathematics and Mechanics.

THE ALGORITHM AND PROGRAMS FOR THE COMPILATION OF A FREQUENCY DICTIONARY OF WORD-COMBINATIONS

NIKOLA RAJKOVSKI

The quantitative investigation of word combinations in texts is of interest to the theory and practice of quantitative linguistics.

The algorithms developed make possible the compilation of frequency dictionaries and the investigation of various quantitative characteristics of word combinations of two and/or three word forms in a text.

The algorithm consists of the following parts

1. An algorithm for the construction of a list of word combinations.

2 An algorithm for the compilation of a permutational dictionary of word combinations

3. An algorithm for processing the list of word combinations, obtaining the quantitative characteristics, and printing out the alphabetic and frequency dictionaries of word combinations

The algorithm is programmed in Fortran and has been tested on the ISL 1900 computer.

The programs find application in the compilation of thesauri the quantitative analysis of texts, and other linguistic investigations.

ON THE "STRENGTH" OF THE INTERMORPHEMIC BOND (ON THE MATERIAL OF MEDIEVAL TEXTS)

M. YANAKIEV, K. NAJDENOV, M. KOTAROV AND N. V. KOTOVA

It is accepted as a working hypothesis that the linguist appraise the "strength" of the intermorphemic bond (the "strength" of a junction of morphemes) as a certain function of two conditional probabilities--the probability of the presence of the first morpheme before the second (it being known that the second already exists) and the probability of the presence of the second morpheme after the first (it being known that the first already exists)

The explanatory power of the accepted hypothesis is clarified. It turned out that the behavior of linguists in the process of morphemic segmentation of a communication is much easier to explain if one proceeds from the hypothesis being verified than if one proceeds from the previously tacitly accepted hypothesis on the evaluation of the "strength" of the intermorphemic bond for the appearance in the report of a second morpheme after the presence of the first in the report became known.

A SYSTEM FOR THE AUTOMATIC SEGMENTATION OF BULGARIAN WORD FORMS INTO MORPHEMES

KHRISTINA BRAJKOVA

The subject of discussion is the work which serves as the basis for the graduating thesis prepared in the Laboratory of Mathematical Linguistics of the OKTU sector of the Institute of Mathematics and Mechanics of the Bulgarian Academy of Sciences under the direction of Doctor of Philological Sciences A Ljudskanov. Its goal is the realization of automatic segmentation according to the morphological principle of prefixes and "prefix like" elements at the beginning of the word form and the establishment of the leftmost limit of the root of the word form "Prefix like" elements are the series of graphemes which may form the initial portion of complex words.

This work forms part of a system for automatic segmentation of Bulgarian word forms into morphemes.

In the work presented one can differentiate information and logic portions The information portion consists of a list of prefixes and prefix-like elements and of lists of "exceptions" for each prefix, the logic portion includes rules for working with the above and indicates the succession of the computer's work.

In the present work the information portion is limited to the letter "z" of the Bulgarian alphabet, but this in no way disturbs the universality of the proposed algorithm

The programming of the algorithm is being carried out in the PL/1 language on the Ye.S. 1040 computer

A MODEL OF SOUND LINKAGE AND MODIFICATION IN BULGARIAN SUFFIXES

STEFAN KHRISTOV

The linkages established by means of a mathematical model of Bulgarian phonetics are fully explicated in the particular structure of the language, in the given instance in suffixes.

The mathematical model reveals a lnaguage to be a system of structures and mechanisms, the function of which obeys the laws of thought. Such regularities in the internal logic are observed in Bulgarian in mutually conditioned changes, within the framework of each complex pair of alternating consonants and vowels.

These regularities can serve as the criterion in creating a model of human speech activity, which can be utilized in the creation of a scientifically based theory, both for stenography and for the creation of a machine for recording of speech.

The mathematical model of Bulgarian phonetics for the complex pairs of alternating cònsonants and vowels can be examined dialectically, i.e. in its unity and contradiction

In such a case it can be bifurcated into two consonants or into two vowels with contrastive structures, which are located between each other in a certain congeneric linkage.

ON THE FUNCTIONING OF PERSONAL PRONOUNS IN RUSSIAN AND BULGARIAN

G. V. YERMOLENKO
U. S. S. R.

We have analyzed a Russian text and its Bulgarian translation (Young World Fiction Writers, a Collection of Stories and Tales, a Selection from Iv. Ruzh, "Folk Culture", Sofia, 1963) The length of the original is 44,000 word-uses, of the translation, 50,000 word uses. Such a ratio is a consequence of the analyticity of the Bulgarian language in the area of nouns

Total quantity of personal pronouns		Singular					Plural		
		1.p	2.p	3rd Person masc	fem	ntr	1.p	2 p	3.p
Russian	3100	782	333	1287	263	3	140	62	230
Bulgarian	2622	497	263	1173	283	26	116	36	228

The frequencies of the 3rd person, singular, neuter pronouns and 2nd person plural pronouns are not reliable. The difference in numbers is an indication that the function of indicating person in the formation of verb-forms is weakened in Bulgarian personal pronouns. If one addresses one's attention to the instances of the pleonastic use of pronouns (nego go, nas ni, etc) and also to the reprise of nouns (Az Mayakovski go uvazhavam, etc then the divergence in frequencies appears even more significant A chi-square test shows that the divergence between frequencies of first and second person singular pronouns in the nominative case is significant for the two languages, and in the oblique cases not significant. The personal endings of the verb show the person of the subject. This is why the difference in frequencies is reached at the expense of the nominative case The third person singular pronoun functions in Bulgarian a little differently than in Russian, in the oblique cases, "toj" and

"tya" are used more frequently than their Russian counterparts. The frequency of the dative analytic is too insignificant; we encountered only 29 forms. This provides grounds to consider that the synthetic, and not analytic, type of declension is still more characteristic of Bulgarian pronouns.

A QUANTITATIVE MODEL OF RUSSIAN WORD FORMATION (ACCORDING TO DATA FROM A RUSSIAN CLUSTER-FREQUENCY DICTIONARY)

L. N. ZASORINA
U. S. S. R.

The materials of the Russian Frequency Dictionary, created by coworkers at the Leningrad and Gor kovskoe State Universities, have served as the basis for a Cluster-Frequency Dictionary (compilers. L. M. Akul enko, E. S. Andreeva, L. N. Zasorina, M. I. Privalova and E. V. Tisenko). It was initiated from a glossary of 39,268 units received in a selection of 1,056,382 word uses compiled from four groups of texts--magazines and newspapers, scientific publications, prose, and drama

The original glossary was subjected to a preliminary ordering Nonnormative variants of words (of the type rodnyj, shkap, botinka) were reducted to normative ones, and participles were reduced to infinitive forms. As a result, the compression of the glossary to 38,000 lexemes was achieved.

The distribution of the vocabulary into clusters was conducted by way of unifying all derivatives having motivation on the same root (an identical root and a general semantic component). A word with a nonproductive base is accepted as the index of the cluster, e.g. GOVOR (say), ZHIT' (live), ZNAT' (know) A cluster is considered a group of words consisting of not less than three words.

In the process of clustering, a class of nonclustering words was isolated It embraced 3000 units, among which stands out a small group of auxiliary words (prepositions, conjunctions, and interjections) and a group of nouns and adjectives, both proper and ethnonymically productive (Kievan, Londonite, Fraunhoferian, etc.). Borrowing constitutes the major part of this class (akr, acre; analoj, lectern; ar (as in hectare), etc) and etymologically isolated lexemes (borshch, proso, millet, rataj, peasant-plowman, etc.).

The distribution by clusters of the basic file has proven uneven. In all, 2500 different clusters have been isolated (this result merits attention in comparison with the familiar data T. F. Efremova and A. I. Kuznetsova received 4500 roots, and D Worth et al. 10,593 clusters). By our calculations, the 200 highest-frequency clusters embrace 62% of all different words. At the same time the 105 highest-frequency clusters ($x > 1000$) includes 30% of the glossary and covers 27% of the entire selection. The second group of 100 clusters (95) forms 32% of the words and covers 12% of all word uses. All remaining lexemes (about 16,000) are distributed among the lesser-numbered clusters.

For a quantitative model of word formation there exist two root-word indices. The word-forming capacity of the root word is determined by the quantity of products of a given root in the dictionary--the ratio of the number of products to the number of all different entries. By this ratio the roots IMET' (possess) VODIT' (lead), LEZHAT' (lie), DAT' (give), ROD (gen-), DELO (matter, affair, business), ZHIT' (live), BIT' (be), and others are found in the leading portion of the distribution.

The text-forming capacity is characterized by the mean frequency of the derivatives in the cluster--the ratio of the total frequency of the cluster to the number of its derivatives A specific property of pronoun roots is displayed, compare ETO (ETOT, POETOMU), this (pronoun), this (demonstrative adjective) for this reason, 13,334/3 = 4,444, MOJ (POMOEMU, MY), my, in my opinion, we, 7976/3 = 2,658, CHTO (what, which), 18636/35 = 407 etc. A comparatively large text-forming capacity is also found among the numbers ONE and TWO and among nearby and semisignificant roots BIG, OTHER, OBLIGED/MUST. Among the significant words located in the leading portion of the distribution by mean frequencies are the roots SAY (84), WANT (71.7), SPEAK (49 6), and NEW (49 5)

A full explication of the fundamental pool of Russian roots in a reliable fashion in their probability relationship to the data file is revealing a new approach to the study of dictionary distribution types, which are important for various practical applications.

A MODEL OF THE GENERATION OF SUBSTANTIVE FORMS IN THE CONTEMPORARY ARMENIAN LANGUAGE

R. L. URUTJAN

A paradigmatic system for each morphological meaning in the contemporary Armenian language is formed from a certain initial word-form (the pure stem) by way of addition or exclusion of certain phoneme complexes. We note that the formation of forms with the help of alternation is easily carried out by means of negation and the addition of a phoneme complex

If we represent the process of form-formation analytically, then we obtain the following expression

$$A(((+\bar{a}) + \bar{b}) + c) + d,$$

where A is the pure stem, $a = a \vee -a$ (a is addition, -a is exclusion) is basically the category of number, $\bar{b}$ $b \vee -b$ in even order is the category of case and class, $c = c \vee -c$ is the category of case, and finally $\bar{d} = d \vee -d$ is the article.

The work examines the questions of constructing a certain generator which automatically generates substantive forms in the contemporary Armenian language.

A method has been worked out for recording the initial word forms, and a classification of the frequency model of generation has been made. Inputing an initial word form, we obtain a system of paradigms of the given word form as output

The given model may be applied in systems for the automatic processing of texts.

Considering, also, that one morphological meaning can be generated by various mechanisms, the author considers it advisable to supply the contemporary Armenian dictionary with generation models, making it possible to obtain a system of paradigms for each initial word

SEMANTIC STRUCTURES AND SYNTACTIC REPRESENTATION

V. D. KLIMONOV
GERMAN DEMOCRATIC REPUBLIC

A linguistic model is being examined in which cognitive structures (the semic representation) (1) are transferred by means of a system of rules into complex semantic ones (the sememic representation) (2) and the latter into syntactic ones (the lexemic representation) (3). Between levels (from the highest on down) single- and multi-structured correspondences are present which are determined by the resources of the (given) language system and regulated by communicative-pragmatic factors

Structures with predicates of motion (selected illustrations are in abbreviated notation). Notation z^1/z^2, a moving body, v/w, initial/final point, x, motive force, r(es) o(bjects, ab(lativus), al(lativus), ac(tor), 1/2/3, number of the level A.1. non (esse (z^1, v)) & fieri (esse (z^1, w)), 2. $movere_1$ (z^1, v, w), 3. The ball (ro) flew out from the courtyard (ab) onto the street (al). B.1. causare (pertinere ad (x, z^2), non (esse (z^2, v)) & fieri (esse (z^2, w))), 2. $movere_2$ (x, z^2, v, w) 3. The boy (ac + ab) thrust the shovel (al + ro) into the earth (al). C.1. A 1 & B 1 with z^1 in place of x, 2. $movere_3$ (z^1, z^2, v, w); 3. The bus (ro + ac) carries passengers (al + ro) ... / The passengers (al + ro) ride on the bus (ro + ac)..

PROGRAM REALIZATION OF AUTOMATIC ANALYSIS OF THE SEMANTIC CONTENT OF TEXTS

D. M. DOBREV AND R. K. KIRKOVA

On the basis of algorithms for the automatic analysis of the semantic content of texts developed by A. Ljudskanov and D Ero, corresponding mathematical program realizations have been created. Applying these programs, a series of experiments has been performed on the material of Bulgarian scientific texts These experiments indicated the advisability of changing the algorithms by introducing pseudolinguistic particles in the corresponding dictionaries.

Certain experimental results are also cited.

THE STRUCTURE AND DISTRIBUTION OF HERMIT WORDS (SLOVO-ODINOCHKA) IN BULGARIAN AND RUSSIAN

A. I. KUZNETSOVA AND D. A. RAJNOVA
U. S. S. R. BULGARIA

In a synchronic structural-semantic description of the Bulgarian and Russian lexical systems, a group of hermit words was isolated.

Hermit words are distinguished not only with respect to structural semantics, but also by a special quantitative character It was this character which provided the basis for their special investigation

The study of hermit words in language allows the specification of the semantic-system concept and its utilization toward the aims of practical grammar and machine processing of information.

The comparison of the hermit words in Bulgarian and Russian has shown a recurrent pattern which affirms the necessity of coordinating frequency-research methods for lexical systems with the principles of actual derivation.

AN ALGORITHM TO DISCLOSE THE SET OF POTENTIAL ANTECEDENTS OF PRONOUNS IN BULGARIAN AND RUSSIAN SCIENTIFIC TEXTS

MARIA KULEVA

The Laboratory of Mathematical Linguistics of the Institute of Mathematics and Mechanics of the Bulgarian Academy of Sciences and the Sector for Automatic Processing of Natural Languages of the Zentralinstitut für Sprachwissenschaft of the German Academy of Sciences are conducting joint work on the creation of a model and algorithms for the automatic identification of actual antecedents (Aa) of some classes of some classes of pronouns (P_7, P_5, and P_p in the singular) in Bulgarian, Russian, and German scientific texts. The problem of identifying the Aas is examined in this work in regard to their selection from the set of potential antecedents (Ap). It is evident that, for the realization of this selection, it is necessary to automatically establish beforehand the set Ap and the working unit (WU) within the framework of which the machine will amass the information necessary to carry out the indicated selection. This is the subject of the author's graduation thesis, which is being carried out in the Sector of the IMM under the direction of Dr. A Ljudskanov, the basic positions of which are expounded in this report

The logical content of the system is this any WU includes the two sentences to the left of the sentence in which the current P is located, consequently, in order to establish the limits of the WU, the machine must be in a position to establish sentence boundaries, and for this it must "distinguish" segmenting punctuation marks from nonsegmenting ones (this task is carried out in Block 1, which has been worked out with the use of dialog in the prospect of including in the future an automatic reading mechanism) Then the classes of P which are relevant to the system are automatically identified, and each of these is

assigned "its own" WU (this is carried out by Blocks 2 and 3) Finally, on the basis of the formal grammatical relations between the Ps and their possible antecedents on the level of superficial structure, the set of Aps in each WU is automatically disclosed (this task is fulfilled by Block 4).

The system is realized on the machine Ye.S. 10/40, and the programs are compiled in Fortran. The proposed algorithm acts upon both Bulgarian and Russian texts.

ON THE AUTOMATIC IDENTIFICATION OF PRONOUN ANTECEDENTS IN RUSSIAN, BULGARIAN AND GERMAN

G. KLIMONOVA AND S. KARAG'OZOVA
GERMAN DEMOCRATIC REPUBLIC BULGARIA

The report discusses some results of the joint work of two collectives the Laboratory of Mathematical Linguistics of the Mathematics Institute of the Bulgarian Academy of Sciences and the Sector of Automatic Language-data Processing of the Zentral institut für Sprachwissenschaft of the (East) German Academy of Sciences.

A series of criteria is proposed for the automatic identification of the referential mechanism linking certain pronouns and their antecedents in Russian, Bulgarian, and German

1. The morphological criterion selects out those word forms in the working text which may be formally antecedents (i.e by agreement with the given pronoun in gender and number)

2. Only those syntactic criteria are cited which are common to all three languages (disregarding the criteria resting on the peculiar features of the syntactic structure of any one of these languages) the condition of branching (the pronoun and antecedent cannot be located on one branch of the tree of dependency), the condition of copredication, exclusion by analogy, coordinated predicates, the criterion of restriction of the search for the antecedent with the help of paired language elements, etc

3 Finally, attention is addressed to the following semantic criteria for exposing mechanisms of pronominal reference the selection criterion, which embraces solely the semantic subclasses of nouns in the presence of verbs, the criterion of lexical combination/noncombination (the entry of nouns into semantic relations of the type "X has (a) Y" "X and its Y"), and the criterion of restricting the search for an antecedent with the help of antonyms.

A textual example is given for each adduced criterion

THE MODELING OF NONINDIVIDUAL LANGUAGE AGENTS

SLAVCHO PETKOV

In the contemporary stage of modeling of language phenomena the "individualistic" approach is predominant, first of all the activity of the individual is modeled--the translation, analysis, and synthesis of speech, the analysis and generation of the text, the coding, indexing, classification, and search for documents, speech patholog , compilation of dictionaries, and so on.

The significance of such an approach is beyond doubt, but together with this, apparently equally timely, is the approach in which the object of modeling becomes the activity of nonindividual language agents. Well-known experiments in the given direction suggest a clarification of a series of theoretical and methodological problems.

The author distinguishes four types of nonindividual language agents: binary, group, collective, and complex. In such an approach the field of mathematical and machine modeling is expanded, and the transition from complex to super-complex problems is effected. Inasmuch as these systems are grounded on linguistic principles, modeling of a special type is held in view (it could be termed linguonic), in which cybernetic and bionic principles can be viewed as partial instances.

The report describes the experimentally tested algorithm for machine processing of data from linguistic experiments which prove the qualitative distinction between individual and nonindividual agents in the construction of a text.

THE APPLICATION OF THE FUZZY SET MECHANISM TO THE STUDY OF ARTIFICIAL LANGUAGES

P. BYRNJEV AND V. DIMITROV

Any artificial language, particularly an algorithmic one, can be viewed as a permissible language (the set of all permissible language elements), as an individual language (the set of language elements used by the discrete user), and as a collective language (the set of language elements used by a certain group of users).

Individual and collective languages may be defined as the fuzzy sets indicated in the set of permissible language elements

The degree of affinity of a language element to one or another fuzzy set, which corresponds to a certain individual or collective language, is proportional to the frequency with which this element is used by the individual and collective users.

The new approach which is proposed for the study of the properties of individual and collective languages is based on the theory of fuzzy sets

The report examines several theoretical methods of formulating the proposed approach. These methods are illustrated by the concrete practical results of the investigation of certain algorithmic languages.

A QUANTITATIVE INVESTIGATION OF THE ACTUAL UTILIZATION OF ALGORITHMIC LANGUAGES

P. BYRNJEV

It is characteristic of artificial languages that their creation precedes their utilization, whereas natural languages form and develop in the process of their use and under the influence of numerous factors, among which the foremost are the users of the language.

Algorithmic languages and, in particular, programming languages, constitute an important class of artificial languages The languages seen below are primarily of this kind

The simplicity, harmony, and completeness of artificial languages are their essential advantage when they are applied to special purposes. On the other hand in work with the artificial languages it is impossible to exhaustively foresee which language elements will be required. For this reason, differences arise between the requisites of the user and the resources of the language.

A sharp boundary between artificial and natural languages does not exist. On the one hand artificial languages also fill out and develop. On the other hand artificial languages are used, so to speak, in a "natural" manner. People select certain permissible language elements. From this point of view, a given language can be viewed as a permissible language (the set of all permissible elements) as an individual language (the set of elements used by the discrete user), and as a collective language (the set of elements used by a certain group of people) Inasmuch as individual and collective languages (and in the case of natural languages, permissible ones) are not strictly determined, it is convenient to employ the mechanism of fuzzy sets.

The report is founded on the necessity or studying the actual utilization of algorithmic languages, and results are

presented of a statistical analysis of the utilization of the language elements of some widespread algorithmic languages.

In this light several conclusions are drawn relative to the interaction between algorithmic languages and their users.

The report brings out considerations which reinforce the assertion that it is necessary to distinguish artificial languages from, and not liken them to, natural languages

Some of the questions examined are controversial in nature.

BASES FOR CONTEXT-FREE LANGUAGES

A. A. RADENSKI

This work examines the task of finding bases for context-free languages. The location of bases has theoretical and applied significance. At the time of translation of the programs of an algorithmic language, the translator in the first place 'edits' the program, exchanging some of the basic symbols of the language for metalinguistic variables which are defined by these symbols. In Algol 60 such variables could be 'identifier', 'unsigned integer', 'integer', and so forth The aggregate of these variables and of the remaining basic symbols is termed the basis. The use of the basis heightens the rapidity of action of the translator in constructing the syntactic tree and hence leads to the improvement of some other characteristics of the translator.

The concept of the basis of a language is formally defined in the work. The graphs G_i, $i = 1, 2, 3$ are defined, linked with the language's grammar It is shown that the location of the bases of a certain special class reduces to the location of the maximally empty subgraphs of the graph G_1 A method is given for the location of all bases without repetitions. A modification of this method permits the exposure of only those bases answering certain conditions. These methods are realized in the form of a program for the Minsk 32 computer. Some results of the investigation, using this program, of the syntax of Algol 60 and Fortran, are cited.

ON THE LINGUISTIC PROBLEMS OF AUTOMATIC INFORMATION PROCESSING

EVA BENESHOVA
PRAGUE

The basic question in linguistic research directed toward the creation of an information language in information-retrieval type systems is the establishment of an approach serving to apportion the concise material content of language communication As distinguished from machine translation, the question at hand is selection of the language material which must be subjected to processing. If the final record must be not only a coordination of the terms of the given text, but presumably also a record of the syntactic relationships between terms, the corresponding information language must contain a certain measure of grammatical analysis. It is assumed that the expression of the communication's structural relations can proceed from the semantic predicative construction (from the so-called word relations) With the help of a dictionary of word relations and of grammatical analysis, the action (with its sphere of valency) is defined in the role of the center of material information. Actions can be grouped in broader classes having a general semantic character in common, the essential quality of such classification is that verbs belonging to one semantic class have the same type of actants. The information arrived at permits, besides purely documentary use of the information, the solution of other, more complex, problems (automatic abstracting, factographic informati retrieval, and the like)

AN ALGORITHM FOR THE ACCELERATED CALCULATION OF SET-ALGEBRA EXPRESSIONS

KOSTADIN GRUEV

Under examination are the set-algebra calculations

$$\mathcal{G} = \{\mathcal{P}(\theta), \cup, \cap, \setminus\}$$

where $\mathcal{P}(\theta)$ is the family of subsets of the set θ, and $\cup$, $\cap$ and $\setminus$ are symbols of the set operations of union, intersection, and relative complementation. It is assumed that expressions in this algebra are written in infix form with arbitrary depth of term recursion. In term notation an arbitrary number of bracket pairs may be used. Object variables designate the identifiers of the elements of $\mathcal{P}(\theta)$.

It is shownt that the majority of tasks solved in information retrieval systems (IRS) are reducible to the solution of similar expressions. So, for example, if θ is viewed as the object of an IRS, then the retrieval of object(s) from θ possessing a certain combination of signs is reducible to the calculation of the expressions from $\mathcal{G}$ corresponding to this combination. By force of the large dimensionality of set-expression files, conventional calculation methods are ineffective.

The proposed algorithm stands out, by way of comparison with the essential archetypal algorithms, because of its better convergence and obviation of the necessity for the storage of intermediate results. This is achieved at the expense of the reorganization of the initial expressions and of the organization on this basis of a parallel-sequence calculation routine The algorithm has been verified in practice The program module, worked out by the author for the ISL series 1900 calculating system, constitutes about 700 commands.

A MATHEMATICAL MODEL OF THE DISTRIBUTION OF INFORMATION IN RETRIEVAL SYSTEMS

M. DOMBROWSKI
POLAND

Many different methods exist for information retrieval These methods were created with the aim of minimizing the computer's retrieval time and volume of storage. Some of these also allow for the property of multiple inquiries The report will present a general model for the distribution of information in retrieval systems. It will be shown that frequent isntances of this model represent the distribution of information in such retrieval methods as complete scanning, inversion chaining, and the methods of Ghosh-Abraham, Lune, Shaw, and Wong-Chang.

ON AN APPROACH TO THE CREATION OF A DESCRIPTOR LANGUAGE WITH ELEMENTS OF GRAMMAR

N. STANCHEV

This is a proposal for a specific approach in the creation of a vocabulary of a descriptor-type information language (IL) which facilitates the isolation of synonyms, homonyms, etc It entails the entering of each concept on a separate card, together with all the words with which it forms word combinations A principle for creating descriptor-language grammar is being proposed. With this aim, representatives of each group of descriptors (class of concepts) are selected, from which matrices of the form "ACTION-OBJECT", "MEANS-ACTION-OBJECT", and "CIRCUMSTANCE-ACTION-OBJECT" are formed. The matrices enter into the dictionary pool of the IL in the capacity of lexical units. With the help of these matrices, all possible logical considerations are stated which are adequate to the consideration of the concepts.

An algorithm is suggested for indexing the content of documents and inquiries. A search mode for documents (SMD) and inquiries (SMI) is created as a chain of elements, the mutual relations of which are set with the help of matrices. The matrices are the basic link of the chain, to which are joined the usual descriptors elucidating it.

A method is proposed for avoiding parasitic combinations by utilizing a method of "semantic overlap" in the creation of the SMD, i.e. the SMD is created from the successively arranged elements, the neighbors of which encompass one identical concept.

The IA vocabulary is grouped into the facets "MEANS' "OBJECTS", "CIRCUMSTANCES", and so forth

The advantages of the approach outlined are as follows The IL, in its ultimate stage, approaches natural language, the subjectivity in the indexing of documents and inquiries is

reduced; parasitic combinations between lexical units are prevented, conditions are created for expansion and contraction of search scope without loss of information, the process of creating the IL is facilitated.

The approach described is realized in one of the branches of transportation.

THE SEARCH PROCESS IN AN AIRS

SONJA KH'NGIKJAN

Described here are the language and logic of the AIRS which is being developed in the Academy of Social Sciences and Social Government. The criterion of semantic correspondence which aligns the output in four echelons is being investigated.

The dependence of the search process on the information language is revealed The necessity for reverse communication with the user is shown, as well as its influence on the search.

Reverse communication with the user is examined from the point of view of two procedures--a presearch procedure (compilation of search instructions up to the execution of the search process) and a postsearch procedure (the compilation of final search instructions after computer processing results have been received). The stages and forms of these two procedures in the AIRS and the possibility of future transition to dialog with the computer are indicated.

A PROGRAM PACKAGE FOR THE REALIZATION OF ADIS A MIXED-TYPE IRS

VASIL DRAGULEV AND VASIL METODIEV

The subject of the report is a program package for realization of an automatic documentary information system, ADIS. The package was developed in the computing center of the Academy of Social Sciences and Social Government The ISI 1904 computer is utilized. The work was conducted in 1974.

In the report an attempt is made to create a classification diagram of IRSs, and the place of ADIS in this diagram is indicated. The basic functional structures of the IRS are examined and functions exceeding the resources of the package are indicated

The second part of the report discusses the fundamental demands and principles which accompany the work. A generalized flowchart of ADIS is proposed The fundamental demands and technological solutions in the package's program realization are examined particularly the questions

- the creation and actualization of the data base
- the independence in structure of the data base from the basic programs of the package
- the language for the giving of retrieval instructions
- the realization of direct and sequential retrieval

In the end several practical results are discussed of the utilization of the program package and the trends for future development

AUTOMATIC INDEXING AND CLASSIFICATION IN INFORMATION SYSTEMS

V G. ZAJTSEV
U. S. S. R.

The report describes a model for automatic indexing and classification of documents and for the inquiries in information systems that operate in dialog mode. The model is built as a set of modules that realize the following functions

automatic translation of the content of documents (synopses) into a descriptor-type information-retrieval language with grammar
automatic classification of documents (synopses)
automatic reception of a field of ordered structures of standard thematic inquiries
automatic translation of the content of inquiries into a descriptor-type IR language with grammar
automatic classification of inquiries into a field of ordered structures of standard thematic inquiries
retrieval of relevant documents (in dialog mode)

The realization of the model in its original form was carried out on an ASVT M-2000 computer on a file of texts (about 2000 synopses) along the theme "Automatic Control Systems and Calculating Techniques" and produced a positive result. An analysis of the results received from the operation of the model has permitted further expansion of its functions The classification module that has been worked out is based not on the a priori classification of the acquired knowledge, but on the descriptor description of documents and inquiries which makes possible the multi-level reception with the input of the next portion of documents and inquiries. Besides that, the use of grammatical means in indexing (indices of connection) significantly improves the information language's index of 'accuracy

A THESAURUS FROM THE POSITION OF A SYSTEMS APPROACH

V. PEJCHEVA AND S. GABROVSKA

The systems approach, as a universal method, can be successfully used in the consideration of a thesaurus as systems of categories and concepts with fixed semantic linkages between them. From the point of view of the systems approach, a thesaurus possesses the following system-forming characteristics Content--the elements of a thesaurus, i.e. its dictionary composition, including the idea of a certain field of knowledge or practical activity; Relational composition--the links unifying its components into a whole. The thesaurus's relations of interdependence are considered gender-aspectual (gender-aspect and aspect-gender) and the portion of associative linkages, and the relations of determination are considered the associative linkages of the reason-consequence type. Pertaining to descriptors which lack genetic or associative linkages among themselves, it can be said that they are in the relationship of constellation, since they are combined within the framework of the system, The function of the thesaurus--its ability to serve as the major component part in a system of higher order, an information-retrieval system, The management of the thesaurus--the process of its improvement by means of changes in its content, structure, and relations.

In constructing the thesaurus, one must heed the demands of the system's methodology, for integrity, compatibility of components (especially urgent in the development of polythematic thesauri), horizontal and vertical decomposition of the system, minimization of the interaction of different subsystems, etc

A SET OF ALGORITHMS AND PROGRAMS FOR AUTOMATED THESAURUS CONSTRUCTION IN INFORMATION RETRIEVAL SYSTEMS

ANGEL MARCHEV

The report examines a sequence for the formation of a thesaurus in information-retrieval systems. This task consumes time and labor in large quantity, and the possibilities of its facilitation by means of partially automated work on thesaurus formation are being examined. Toward this aim information files have been defined, and a set of algorithms and corresponding programs has been worked out allowing the automation of the most time-consuming operations in the creation of the thesaurus (the creation of a dictionary of word-forms and word-combinations, alphabetic and frequency ordering, verification and correction of the dictionary, evaluation of the semantic bonds between descriptors, verification of the classification routine, the statistical profile of the dictionary composition, etc.)

The programs are written in FORTRAN IV and intended for the ISL 1904A computer.

The programs which have been worked out are universally applicable to the formation of thesauri and various types of dictionaries. In the given instance, they are utilized for the creation of an economic information thesaurus in compliance with the proposed sequence of its construction

AUTOMATED INFORMATION-RETRIEVAL SYSTEMS WITH BILINGUAL INPUT

A. LJUDSKANOV

Bilingual (multilingual) input into automated information systems (AIRS) makes possible the processing of documents in different languages without the creation of many systems or even the preliminary translation from a foreign language into the national one. The problem of bilingual (multilingual) input to an AIRS presents itself in connection with automated indexing of documents and inquiries in different natural languages The realization of bilingual input is founded on certain informatic media and mathematical program realization

I. The informatic media 1. The standard forms of information cards (IC) of documents and of information blanks (IB) of questions are coded with a special code (0 and 1) for the guarantee of automatic selection of the file (Bulgarian and Russian) given in the corresponding natural language 2 The processing of the factological portions of the ICs and IBs is effectuated on the basis of code catalogs. 3 The resume (annotations) and questions are processed with the help of a bilingual (multilingual) descriptor dictionary in which the pairs, triplets, etc. of different-language descriptors are linked by a general code to the metalanguage of the system

II. The mathematical program realization includes programs for forming code catalogs and a machine descriptor dictionary, a program for automated indexing, and a program for the search and delivery of responses The first experiments have shown that only the indexing program requires some changes in connection with the discrimination of the constants of the different natural languages. The block organization of the program makes possible with the help of a key the activation of that block which is necessary for the processing of the text in the corresponding language.

III. The automated information-retrieval system of the Institute of Radioelectronics is in the experimental phase with the goal of including Russian input The collective is also carrying on preliminary investigations toward the realization of tertiary English input into the system

INFORMATIONAL-LINGUISTIC INTERACTIONS IN THE SYSTEM "INFORMATION-USER INFORMATION-RETRIEVAL SYSTEM"

NIKOLA RAJKOVSKI

In the process of information service between the information user and the information-retrieval system (IRS), certain informational-linguistic interactions (ILI) are established which flow within the framework of the following language systems

1. The information-retrieval language YaS1 in contact with the information files and the description of the semantic content of inquiries and documents.

2 The user language YaS2 in contact with the system (computer and under control of the process of retrieving and processing information.

3 The programming language YaS3 according to the formulation and transcription of processes in the solution of information tasks in the computer.

4. The natural language YaS4 in user-contact with the information in the system.

The following aspects of the ILIs are examined

1 The semantic aspect--the semantic content of the ILI is formulated in terms of YaS1

2 The program aspect--the presence of program media in carrying out the commands of communication and control by processes in the computer.

3 The pragmatic aspect--the satisfaction of the information needs which are the goal and the sources of interaction.

4. The organizational-technical aspect--the ILI is carried out in a given structure with the help of certain technical means

THE NEW INFORMATION LANGUAGE "POLITIKA"

KHRISTO TS. GEORGIEV

Stage I of language compilation. As a result of a statistical inspection of Bulgarian foreign-policy texts, 50,000 word forms in length, frequency and reverse-frequency dictionaries have been compiled The 10,000 word forms entered cover the foreign-policy texts to 90%. Both dictionaries have permitted the transition to the study of lexical and morphological semantics of the foreign-policy (FP) vocabulary on the paradigmatic plane

Stage II of language compilation Based on the paradigmatic comparison of the lexical meanings of natural-language words (synonymy, gender, aspect, and so forth, the FP thesaurus has been compiled. 235 semantic groups (SG) have been singled out comprising 800 different words. AN SG contains two or more words having one and the same elementary meaning (EM). The artificial designation of an SG, i.e the EM, is the root of the artificial-language word Furthermore paradigmatic relations (gender and so on) have been established between roots (ESs). The gender roots of a given root are called prefixes of the artificial word Besides prefixes, two numerical suffixes are appended to the root; the first indicates the form of connection of the concrete natural-language word within the framework of the SG, the second indicates the grammatical role of the given word (adjective, gender, number, etc)

In this fashion if one natural-language word enters into several SG's, it then has several ESs

The information language Politika is a language of indirect documentation, descriptor type (having two layers of vocabulary), with grammar, branched and, to the extent of synthesis, synthetic

On the basis of the description which has been carried out on the FP vocabulary, algorithms have been drawn up for the auto-

matic translation of text from natural language to the artificial language Politika (including algorithms for the removal of polysemanticity in natural-language words and algorithms for the automatic recognition of the semantic form on the level of indexing, annotation, and synopsis.

THE ANALYSIS AND RETRIEVAL OF JURIDICAL INFORMATION

M. KUTA
POLAND

The report gives the nature of juridical texts with respect to their structure and language Starting from these characteristics, three approaches are described for the solution of the problems of analysis and retrieval in the field of law,

- a method based on the indexing of documents
- a full-text method
- a heuristic approach to the question of research on legal texts.

Then examples are presented of information-retrieval systems based on these methods. In the end an appraisal of the adequacy and effectiveness of the above methods will be made

AN AUTOMATED INFORMATION SYSTEM TO SERVE SCIENTIFIC CONGRESSES

VALJA PETROVA

This report is based on the graduation thesis which is being prepared in the Institute of Mathematics and Mechanics under the direction of coworker Avram Eskenazi.

The subject of work is an information system for processing data related to the conduct of scientific congresses, conferences symposia, etc.

The majority of the programs are written in the RPG language Besides this, the system's sorting programs are used

Several of the basic blocks have been worked out

For input of data extracted from requests entered and for creation of a basic file containing information relating to each participant;

For file renovation by new requests and for correction of errors which slipped through in the original listing,

For the output of amassed information in edited, expanded, and concise format,

For sorting records by various indices and output of the information so established in a certain format (sorting by governments and names, sections and names, the indicated category of participants and names, etc),

For output of reports to participants and postal addresses

The system has been applied for the First National Conference on the Use of Mathematical Models and Computers in Linguistics

A LANGUAGE DESCRIBING THE STRUCTURE AND FUNCTIONING OF AN ORGANIZATIONAL-ADMINISTRATIVE TYPE CONTROL SYSTEM

G. PETROVA AND ST. DZHEDZHEV

The processes of obtaining information are always accompanied by the processing of information of varying content volume, and form of presentation. For the information management of these processes under a certain control system (CS), it is necessary to construct a suitable information system (IS) allowing the correct packaging and processing of information (including the actually existing control system), which is received at different times from different sources and carriers. The best conditions in this regard are provided by factographic information systems.

Some questions on the construction of an information language in the planning of a factographic type information system are the theme of this report. The presence of a language which ensures the capabilities enumerated above creates the conditions for the proper flow of the automated CSs real-system analysis and synthesis processes.

In their report the authors found the reflection chiefly of the problem of selecting the basic semantic unit and the means of epresenting the relationships between objects in the real environment--objects about which information is to be stored in the system's memory (particularly in the case where the objects are separate information systems)

A SYSTEM FOR PROCESSING INFORMATION FROM QUESTIONNAIRES AND FORMS

PET'R STANCHEV

The system includes a language for processing questionnaire information and a translator for this language. A unique feature of this language is the division of input information and its processing with the aim of logical and formal control of data and obtaining an accurate data bank.

The realized translator utilizes tree-like search methods and the notation of the language's operators and signs via direct access by initial letter, where single-letter signs are found directly.

An operator exists in the language for processing textual responses with a standardized word order These responses are evaluated with the help of a fuzzy set (Zade) mechanism For this purpose an algorithm exists for locating various parts of the expression, locating them in the translator's dictionary, and evaluating them. A program has also been created for the compilation of a frequency dictionary of open terms depending on their type.

The translator is realized in the Fortran and Assembler languages for the Minsk 32 computer

THE LINGUISTICS OF A TEXT AND MACHINE TRANSLATION

R. G. PIOTROVSKIJ
U. S. S. R.

I Our knowledge of the nature of language and of the mechanisms of text formation, the technical resources of calculating techniques, and the contemporary state of programming do not permit us to hope for the successful execution of a full translation of a foreign text in the ensuing decades The existence of the antinomy between synchronism and diachronism, of a "soft" language system and its "coarse" deductive description places in doubt the very possibility, in principle, of global machine translation.

II. At present it is expedient to realize lexical (automatic dictionaries of word-forms and locutions) and lexico-grammatical translation on the computer on the basis of the linguistics of the text.

III The linguistics of a text is the inductive-deductive theory which proceeds from the assumption that the generation of a text is the interaction of the language's productive system, which limits this system to a norm and also to that situation, which is described in the text Each text consists of units of expansion dictated by the language's system and norm and of semantically dominant units prompted by the situation.

IV The linguistics of a text uses two systematic approaches the investigation of the text with the help of probability-statistical and information methods and the deductive description of the productive system, which must be carried out with the help of the theory of fuzzy sets and algorithms.

V. The first approach permits the tracing of the information distribution in the text, the disclosure of the probability-informational nature of text formation, and the separation of the dominant units of the text from the units of expansion.

A SYSTEM OF MACHINE TRANSLATION UNDER DEVELOPMENT IN THE MATHEMATICAL LINGUISTICS LABORATORY OF THE INSTITUTE OF MATHEMATICS AND MECHANICS

S. KARAGEZOVA, A. LJUDSKANOV, E. PASKALEVA, AND T. SHAMRAJ

I. 1. Problem, strategy, and transformation of the more "profound" analyses into more "superficial" ones.
2 The language material and preliminary investigations
3. Analysis as the selection of signified items for assigned signifiers, translation.

II. 1. The lexemic dictionary, intention, compilation methods, type of units (premorphemic analysis), structure--the zones A, B, and C, the 16 information columns and their content, the volume of memory.
2. The morphological dictionary. Analogy with the lexemic.

III. The logic of the system preliminary automatic processing, search, analysis routines, translation routines, synthesis routines.

IV. 1. Preliminary processing, automatic segmentation of the text into sentences.
2. Dictionary search, registration of "zero endings", alternation of the stem, complex words, unfound elements.
3. The routines "Lexical homonymy" "Lexical polysemanticity", "Stable combinations", and "Morphological polysemanticity". Linguistic stems, "contrastive grammar", algorithmic solution, outputs.

V Machine realization and work on the "oriented" syntax and semantics. Dialog and paths of development for future work

These operations make possible the automatic separation of the theme and rhematics of the text, which is in turn the condition for the realization of machine indexing and abstracting of the foreign or native-language text, and also for correct, lexico-grammatical machine translation.

VI. The theory of fuzzy sets and algorithms is more helpful than contemporary discrete-mathematic methods in describing a language's system and norm and also in formalization of text-formation methods. This theory, which is based on probability logic, plays apparently the deciding role in the creation of semantic machine translation which works on the computer.

AN EXPERIMENT ON THE STATISTICAL ANALYSIS OF INTERLINGUAL IDIOMATICITY

M. M. KOPYLENKO
U. S. S. R.

In translated Old Bulgarian (OB) inscriptions from the 9th and 10th centuries, 94 verbs form combinations of the type "verb + abstract noun" (V+Nabstr). Of the uses, 74.5% fit V+Nabstr, and are formed by four verbs: TVORITI (create), IMITI (possess), DATI (give); and PRIDTI (come). Their equivalent combinations in the Ancient Greek (AG) originals are combinations with a more frequently used verb of the given meaning, a less used verb with the given meaning, and verbs of another meaning, univerbs and those different from the V+Nabstr construction. The four types of equivalents correspond to the four degrees of interlingual idiomaticity. A large portion of the uses fits into the pattern of type 1 equivalents. The average number of equivalent AG combinations formed with a dominant verb is 70% and the mean square deviation is 10.3%, the coefficient of variation does not exceed 15%. The remaining three types of equivalents do not stand out by way of regularity--the coefficient of variation exceeds 40% in all cases. All four types of AG equivalents for the OB combinations formed by the remaining 90 verbs vary over a very broad interval. The quantitative indices of interlingual idiomaticity presented by other authors agree with our data.

The number of hypo-idiomatic correspondences, in the first place, differs from one language to another, and, in the second place, noticeably falls in recent time by comparison with antiquity. However, general patterns are nevertheless observed the large number of hypo-idiomatic equivalents to combinations formed by high-frequency verbs and the negligible variability of this number. These patterns must be subject to interpretation on the material of various combination types and various languages.

AN EXPERIMENT ON THE USE OF LANGUAGE MODELING IN PSYCHOLINGUISTIC EXPERIMENTATION

A. P. VASILJEVICH
U. S. S. R.

In research on spoken language activity a certain role is assumed by questions concerning the generation by a human of connected text in a situation wherein the sense is preassigned. In the majority of cases the required meaning can be expressed in several manners. However it is evident that the selection of a specific text (e.g., of a sentence) transmitting the assigned meaning is not fortuitous. Hence, in our experiment, various ways suggested themselves to the group of fifty informants to express in Russian the meaning of the phrase, "The Master (M) defeated the Grand Master (G.M.)" The informants indicated an obvious preference for certain sentences among the possible periphrases (e.g., "The M. won a victory over the G.M.", "The G.M. lost to the M.", and so forth). For an explanation of the preferences received, it was natural to turn to a linguistic description of the many synonymic phrases with the sense of "The G.M. defeated the G.M.". We find such a description in the model "Meaning ↔ Text" by I. A. Mel'chuk. Within the framework of this model one may, in particular, impart a quantitative sense to the concept of the "syntactic complexity" of a phrase and grade all periphrases by this index. The results of our experiment, analyzed from this point of view, provide grounds for the conclusion that, in the generation of sentences conveying an assigned meaning, preference is shown for "simple" phra$es.

THE OBJECT AND METHODS OF STYLOSTATISTICS

M. N. KOZHINA
U. S. S. R.

1. In connection with the contemporary scientific-technical revolution, the integration of the sciences and the mathematics of the methods of the more frequently-encountered sciences are, along with other traits, characteristic of the development of science, as is well known. Linguistics, too, is caught up in this mathematization, including stylistics. However the application of quantitative methods in linguostylistics has yet to receive its own sufficient theoretical groundwork, the specifics for applying these methods here are inadequate, as are the questions of interpretation of the stylostatistical indices received. Hence, not infrequently, annoying errors crop up both in the organization of collected material and in the treatment of the data received; at times, false conclusions are also drawn

2. One of the important bases for the elimination of the defects outlined in item 1, along with strict observance of the principles of mathematical-statistical method, is the precise definition of the object and specific character of stylostatistical method.

3. The object of stylostatistical research is the statistical structure of the expression (text) in one or another sphere of intercourse, corresponding to the concrete form of social consciousness and activity (scientific, artistic, legal, and so forth) which represents the specific organization of the language units, which create a special functional style of speech. The category of the functional style of speech is quantitative-qualitative. By the latter component is meant the semantic aspect possessing, in context, stylostatistical meaningfulness.

The method of stylostatistics proves qualitative-quantitative (statistical) and hence comparative, since the specific character

of separate functional styles (variety of speech) is most effectively revealed in their comparison. The utilization of this method must be based on the theory of functional stylistics (the concept of style, its functions, the extralinguistic bases, specificity, and speech organization of style, and interstylar and intrastylar differentiation in speech.

One of the important conditions of applying this method must be to take into account semantics and the functional meaningfulness of the language units in the given context (and in the functional style), inasmuch as frequently the very same quantitative data with respect to the same language units without taking into account their semantics (in the broad sense of the word, including grammatical semantics) can belong to different speech varieties, but have at the same time different stylistic significance. (Compare, for example, the close quantitative index for the use of verbs in the present tense in scientific language and dramaturgy.)

5. Turning to statistical methods for the investigation of different types of speech, including scientific speech, it is necessary to clearly define the principles for the selection of material and accurate stylistic borders of the latter (meanwhile in the selection of texts for analysis which actually represent the given functional style, and represent it in sufficient variety for objective conclusions to be made, they are encountered in investigations of omission.

The contemporary questions for stylostatistics are thus the selection and organization of material, the volume of subsampling (in the investigation of different levels of speech), and the problem of interpreting quantitative data, to which a significant part of the report is devoted

STYLISTICS AND THE MOST FREQUENT WORDS OF A LANGUAGE

ANNA VASIL'EVNA ORLENKO

The most frequent words in the Bulgarian language, from the point of view of their statistical distribution in texts, are not stylistically neutral

The material from a very large selection of Bulgarian texts has shown that, for various language styles, the most frequent words, which form the beginning of a frequency-ranked list, are disposed in different ways in this list, but in a fully derined way for texts of a given language style.

It is proved, for example, thán in journalistic texts, the most frequent word is always the preposition NA (on, onto)· In Bulgarian art-literature texts, on the other hand, NA never appears more frequently than the conjunction I (and).

A SYSTEM FOR AUTOMATIC TEST GENERATION

AVRAM YESKENAZI

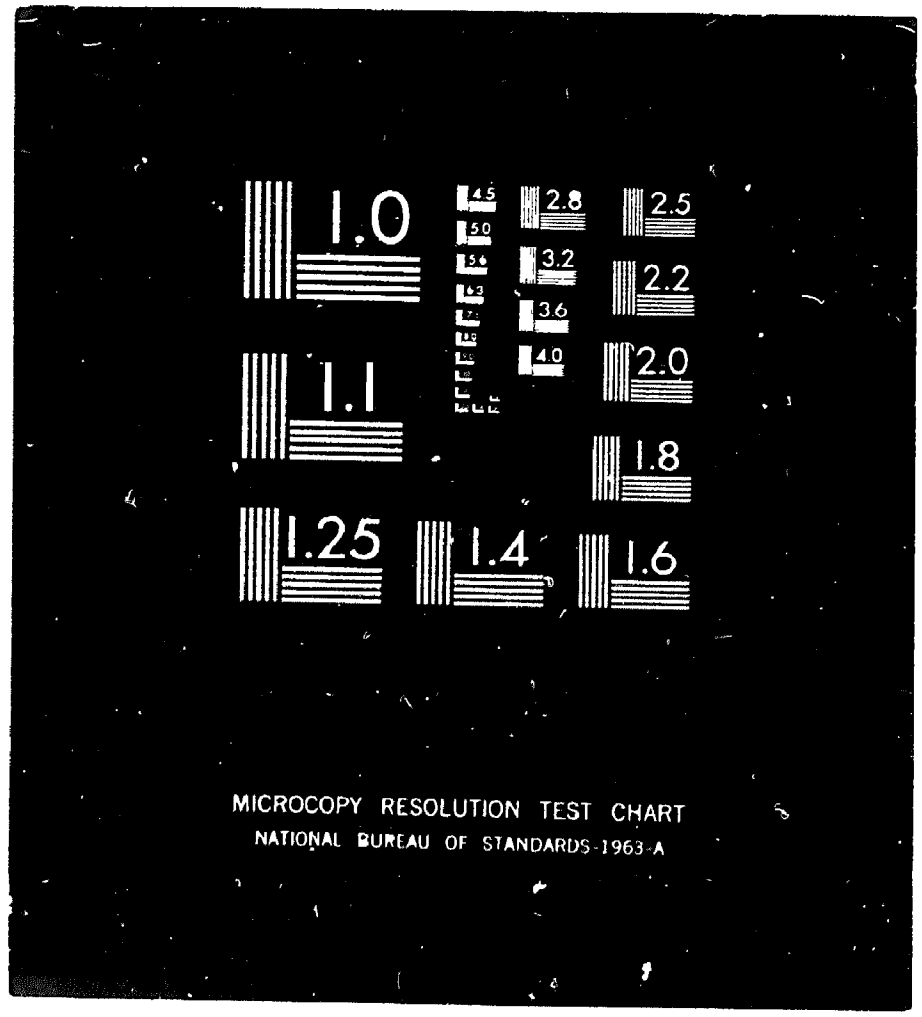
A dictionary of concepts and their characteristics (main idea, membership, functions, coefficient of difficulty, and other traits) is preassigned With the help of this dictionary and user-supplied parameters--the number of questions, the number of answers to each question, the sum of the coefficients of difficulty, etc.--a test is generated. Algorithms have been created which generate questions and corresponding answers of a certain syntactic structure. Besides this, certain means of solving linguistic problems have been developed The diversity of tests is achieved at the expense of using an appropriate form of random-number generator

The first variant of the system has been worked out in the algorithmic language PL/1 on the Ye S. 1040 computer The experiment was conducted on a concept dictionary on the theme "The operating system of the DOS/ES Work on this system is continuing.

NAME INDEX

The transliterations used here have not been verified by comparison with published lists or indexes. The editor will be glad to receive corrections which will be used in cumulative indexes of AJCL.

END


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A