Humanities Computing in Italy

Antonio Zampolli

1. Historical Introduction

It is unnecessary to give more than a brief introduction here since the origins and the development of humanities computing activities and of computational linguistics in Italy have already been described in previous articles (cf. A. Zampolli, 1967; U. Bortolini, C. Tagliavini, A. Zampolli, 1971). The present article is designed to cover development between 1968 and 1973, the years following A. Duro's article "Humanities Computing Activities in Italy."[1] From the beginning three themes were apparent. The first of these, which could be called lexicological and lexicographical, begins with the work of Father R. Busa, who is recognized as a pioneer in this field not only in Italy but in general (cf. R. Wisbey, 1965). His first experiments date back to 1949, and he announced and published his first results in 1951 (R. Busa, 1951). In 1953, he founded the first independent center devoted entirely to the analysis of literary and linguistic data: the Center for the Automation of Literary Analysis at Gallarate. This continued to function until 1966, when the processing of the *Index Thomisticus* was transferred to the Centro Nazionale Universitario di Calcolo Elettronico (National University Computing Center) at Pisa (cf. R. Busa, A. Zampolli, 1968). This Computing Center, founded at Pisa in 1965 between I.B.M. Italia and the University, celebrated its inauguration by publishing the electronically produced *Indexes* and *Concordances* of Dante Alighieri's *Divine Comedy* (cf. C. Tagliavini, 1965). Meanwhile in 1964, A. Duro and A. Zampolli had begun their first experiments to create a large Lexical Archive of the Italian Language on behalf of the Accademia della Crusca in Florence. From this *A Treasury of the Italian Language From the Origins* was derived, followed by *A Historical Dictionary of the Italian Language.*[2] In 1966 also the Accademia della Crusca entrusted the processing of its texts to C.N.U.C.E. The Academy's example and authority and the success of the operation managed to crush the scepticism and distrust which the Italian universities, despite Father Busa's accomplishments, had long nourished with regard to computers, and which had until then limited the development and diffusion of methods for mechanical analysis of data in the humanities.[3] The cited article by A. Duro deals with this point, listing all the projects undertaken in 1967 and noting that C.N.U.C.E. "gathers and coordinates virtually all the research conducted in Italy by use of computers."

[1] *Computers and the Humanities* 3, i (Sept. 1968): 49-72.

[2] Cf. A. Duro (1973).

[3] Cf., for example, "Le applicazioni dei calcolatori elettronici alle Scienze Morali e alla letteratura," *Almanacco Letterario Bompiani,* 1962; the results of an investigation are published here with the significant title: "The Two Cultures." See also the report of the Twenty-First National Philosophy Conference which took place in Pisa in 1967, with the title *Man and the Machine.*

Antonio Zampolli, a professor of linguistics at the University of Pisa, is coordinator of linguistics and literary activities at C.N.U.C.E.

The second theme can be labeled linguistic statistics. Here too Italian scholars, after several precedents in the previous century,[4] ignored the application of statistical methods, which had had a notable success in America and Europe, thanks above all to the demands for educational instruments by linguistics teachers, and to the structuralist basis created for European linguistic statistics. The theoretical and methodological interest with which the first Italian studies in this field were conducted in about 1960 shows that this was a result not of conservatism on the part of Italian linguists, but rather of a deliberate delay in expectation of more mature and suitable methods (cf. for example, the works of L. Heilmann and G. Rosiello). It was immediately realized that electronic examination permitted the verification, on an adequate informative basis, of the hypotheses and "laws" on quantitative structures in linguistic data, the naïveté and superficiality of which had roused the justified scepticism of many scholars. A. Zampolli's thesis (1960), written under the guidance of C. Tagliavini, was the first case of the complete statistical study of a text at the phonological, morphological, lexical, syntagmatic, and syntactic levels. At C.N.U.C.E. all aspects of this theme find natural support from the quantitative data supplied by the concentration of text processing.

The third theme, which began with research for mechanical translation, on behalf of the Centro di Cibernetica e di Attività Linguistiche di Milano, led by S. Ceccato, suffered from the general world-wide crisis in mechanical translation. It should be remembered that the famous lectures by J. Bar-Hillel which spelled out the reasons for the crisis, were given in Venice at the 1961 NATO Advanced Institute for Machine Translation. A similar fate overtook the information retrieval projects, in which the use of refined linguistic instruments was proposed, and which were carried out by groups working closely with the mechanical translation team. It should be emphasized, however, that these Italian experiments helped to formulate, and often to anticipate, judgments on the real difficulties of mechanical translation (cf. G. Lepschy, 1966, Appendix). Recently some research into the field of mathematical models of language and phonemic, syntactic, and semantic analysis of a text has been started at C.N.U.C.E. In a way, at the level of basic research and with a different methodological premise, this occupies the place left vacant by the third theme.

2. The C.N.U.C.E. Linguistics Branch

It is still true today, as A. Duro observed in 1968, that almost all the studies in progress in Italy in the field of computational linguistics and humanities computing activities are co-ordinated, in their informative and computational aspect, by the Linguistics Branch of C.N.U.C.E. at Pisa, and the associated processing is carried out on the center's computers.[5] By its very nature, it is designed to carry out original research. At the same time it makes available to humanists not only the electronic systems and the technical staff required to carry out the operations, but also the necessary scientific advice, not to mention the analysis and editing of new programs when the utility programs already available in the branch are inadequate.[6] This, however, has become less

[4] One of the earliest studies is the small volume by Niccolò Tommaseo (1843). In 1880 F. Mariotti published in Florence a short work entitled *Dante and Language Statistics,* which contains the word and the grammatical category frequencies of the *Divine Comedy.* In Italy, as in other countries, the first statistical studies of phonetics are due not to linguists but to specialists in stenography (see G. Aliprandi, 1940) and in audiometry (E. Bocca, A. Pellegrini, 1950). The great statistician and economist M. Boldrini published some articles on stylometrics in 1948.

[5] The work on the *Index Thomisticus* took place in Pisa between 1966 and 1969 and after a brief interval at Boulder, Colorado, it is now being continued at Venice, in close collaboration with C.N.U.C.E.

[6] The C.N.U.C.E. Linguistics Branch is directed by A. Zampolli and divided into 4 sections: the section that assists the users (8 people), the software section (8 people), the research section (16 people), and the section for conferences and teaching activities (4 people). The branch uses the I.B.M. 370/158 system full time and, for interactive works, the I.B.M. 360/67.

and less necessary, at least as far as the lexical and statistical examination of texts is concerned, because scientific collaboration with some eighty university and research institutes concerned with linguistic science and humanities, and almost seven years of experience, have allowed us to create a series of procedures, which are applicable to texts in any language and which are concerned with the following main aspects:

Lexical Examination: the production of lists of lemmas and forms in the various traditional arrangements (alphabetic, frequency, inverse spelling, and so forth), of line and rhyme indexes, of various types of concordances, of context cards, etc.[7] As will be explained later, we are working on a new procedure which links a kind of *information retrieval,* interacting with a bank of linguistic data, to the printing and publishing of concordance indexes.

Phonetic Examination: algorithms for the semi-automatic phonetic transcription of texts, division of syllables, recognition of clusters and phonetic patterns, etc.

Consultations of Automatic Lexicons: memorization and consultation of large-scale automatic lexicons, with particular attention paid to the problem of optimization.

Statistical Elaborations: the counting of various linguistic units, calculation of various indexes (dispersion, usage, entropy, correlations, significance tests, etc.).

Philological Elaboration: dealing with variants, publishing critical editions.

This is not the place to discuss the programs in detail (on this point see Zampolli, forthcoming) and I will simply outline the main principles according to which we worked.

The readers of *Computers and the Humanities* are too well acquainted with the problem of standardization in the recording of corpora and of texts and basic elaboration programs for me to have to comment on it. B. Quemada gave a clear picture of the problem at the *colloque* at Besançon in 1961, and subsequently it has been raised at nearly every congress on the subject, up to the recent *Colloque de Terminologie et Traduction*[8] at Montreal in 1972, and in numerous articles on the subject, in particular, that of M. Kay.[9] The problem is posed usually for at least three good reasons, each dependent on the other. From the financial point of view, the cost of the recording of the texts in the phases of pre-editing, punching (or equivalents), and corrections is higher than the cost of processing, at least for those institutes with access to university computing centers. From the scientific point of view, the uniformity of criteria for recording and analysis allows research, particularly statistical and stylometric research, to take place on large corpora containing individual texts or groups of texts worked on by different researchers. From the point of view of future developments in research, it seems necessary to create great data archives, *banques de mots,* formed dynamically and in such a way that the various researchers can, on the one hand, question the bank for their own purposes, and on the the other, deposit in the bank the results of their analyses, so as to proceed gradually and coherently towards the ideal aim of the maximal indexation of the corpora.[10]

[7] The context cards are normal mechanographic cards that refer each to an individual occurrence in the text. On the card is printed the lemma, the reference, the date, and 900-character context. Some of this information is also punched to allow any subsequent rearrangements in the archive.

[8] See the *Actes du deuxième colloque international de linguistique et de traduction,* 1973.

[9] M. Kay, "Standards for Encoding Data in a Natural Language," *Computers and the Humanities* 1, v (May 1967):170-77.

[10] In May 1973 a conference on the theme of "L'indexation Maximale de Corpus" was held at the Centre de Philologie Romane at the University of Strasbourg, in which some thirty scholars discussed two main points. The first concerned the choice of the most convenient indexing levels, given the present state of linguistics, to carry out the indexation of units and linguistic facts, present in a corpus at various linguistic levels, in a 'neutral fashion' with regard to the various schools, so that the results are universally usable. There are different possibilities. To prepare different types of indications for each theory, to limit oneself to a minimal indexation, that is to say the accurate recording of the text. The second point concerned the execution of the analysis by means of electronic data processing, and in particular the application to linguistic corpora of the most modern instruments of computational linguistics (parser, cognitive network, etc.).

The various Italian institutes concerned with the electronic elaboration of texts have achieved this unification of methods and procedures. The coding system in use at the C.N.U.C.E. Linguistics Branch is the only one used, and our electronic library contains over 80 million words in over 30 different languages. The very existence of this library is a good argument to persuade anyone undertaking new research in Italy to adopt the C.N.U.C.E. standard. This ensures that the text will be recorded without omitting any useful information, and spares the scholar the boring technical problems of coding and processing. We can say that for any text in any language, it takes only a few dozen minutes to adjust by control cards the procedures, including the codification, the required type of contextualization (there are many possible definitions of context: variable or fixed length, fixed or variable position of contextualized word), and the rules of alphabetic arrangement, etc. Naturally, the continual evolution of systems technology demands, fortunately, a continual revision of procedures. In any case, the existence of an adequately formalized description permits the adoption of the systems available in "real time."

3. Work in Progress

Although the projects in progress are too numerous to be mentioned in detail, it is useful to indicate the main spheres of activity.[11]

Organizing Linguistic Data

Lexical Archives for the Editing of Large Historical Dictionaries of a Language. Archives for the *Treasury of the Italian Language from the Origins* and for the *Historical Dictionary of the Italian Language,* both carried out for the Accademia della Crusca of Florence, contain about 30 million words of running text, of which approximately 40 percent will eventually be represented in the archives. For the thirteenth and fourteenth centuries, in particular, literally all texts and documents have been recorded in the computer. In the last five years the input has been almost completed, and now our human and computational resources are devoted principally to the lemmatization of texts. The lexicological phases have all reached a noteworthy level of formalization, and altogether they form a procedure that acts as a model for the elaboration of all other Italian projects. These experiments, like the other lexicographic undertakings which sprang up in Italy in this period, have their roots in the rich European lexicographic tradition, and in particular the Italian tradition: the *Vocabolario della Crusca,* which appeared in 1612, was the first of the great European historical dictionaries.

On the other hand they are also part of the modern tendency to create large archives and inventories, determined largely by the arrival of computers as well as by the development of descriptive linguistics. The idea of creating lexicological archives is slowly making headway in each country. New methods of inventory-making have profoundly changed the conditions and perspectives of lexicology, and above all the positive principle of cumulative works has come to the fore. The experiments taken up a few years ago on a very large scale as national undertakings at Besançon, Florence, Nancy, Leiden, etc., have been transformed, from one point of view, into routine undertakings in which large teams of lexicographers work. But it seems clear that a new experimental era is about to begin: the quantity of material collected–tens of millions of references–means that the

[11] The following paragraphs should be considered not as a classification but merely as a convenient division to simplify the exposition. The classifications so far proposed are, in any case, not fully satisfactory (cf. S. Lamb, 1961 and 1965, J. Gardin, 1965, P. Garvin, 1962, C. A. Montgomery, 1969, D. G. Hays, 1969). The three main themes mentioned in the first part of this article can be placed as follows: the first group in *Organizing Linguistic Data* and *Text Corpus Editing,* the second in *Linguistic Statistics,* and the third in *Mathematical Models.*

computer must be used, not merely to assemble archives, but also to organize and analyze them, with an interactive procedure of selection and collection. The identification and retrieval of linguistic units and data will be carried out, or at least facilitated, by automatic procedures of linguistic analysis that would create a preliminary filter between the editor and the archive. Objectively this aim does not appear to have a manageable immediate realization, but undoubtedly it is already attracting the attention of many disciplines adjacent to lexicology and lexicography, whose archives are proposed as basic catalyzing instruments for linguistic and literary research in general. In the same field I note the *Project for a Historical Dictionary of the Romanian Language in the Sixteenth Century* being directed with a collaboration between Pisa and Bucharest. F. Dimetrescu (1973) or the *Lexicon of Cuneiform Hittite* by Prof. Meriggi (1973) of Pavia are examples of lexical archives of languages with non-Latin alphabets.

Lexicographic Documentation Designed to Edit Historical Dictionaries Related to Particular Disciplines. The planned *Historical Dictionary of the Italian Language* (cf. Ciampi, 1973) represents projects characterized by the necessity of making a large selection because the lexicon to be displayed in the analyzed works and placed in the archives is clearly a specialized one; that is to say, the percentual relationship between the occurrences in input and the occurrences that become *headings* in the archives is very low. So these seem interesting economic experiments in relation to the present demands of technological dictionaries, which interest researchers from many scientific and technical fields.

Banks of Dialect Data. The archive of the *Italian Linguistic Atlas* includes some 9 million words in phonetic transcription gathered during investigations in 1000 Italian localities and based on a questionnaire containing 7000 items. Some of the analyses of this questionnaire have already been published (*Atlante Linguistico Italiano*, 1971, 1973) and the answers are now being converted to machine-readable form. The adoption of electronic data processing methods has meant the reversal of the traditional order of operations in which the material is plotted on the maps before the indexes and analyses are made.[12] The present plan includes the publication of the atlas after the publication of regional indexes, local dictionaries, maps regarding particular linguistic phenomena, etc.

We are already considering a project to extend the archives by new investigations structured on a dynamic bank of dialect knowledge. The renewed interest in dialectology in Italy is motivated in part by the need to avoid losing our regional cultural heritage, and in part by the privileged position of dialect and regional studies in socio-linguistic research which, at the moment, is very much felt in the context of the social and geographic mobility of groups and individuals.

Italian Machine Dictionary. A group of fifteen young scholars from the Linguistics Branch of C.N.U.C.E. is creating an Italian Machine Dictionary (I.M.D.), its nucleus to consist of some 150,000 lemmas obtained by the union, in the set theory sense, of the nomenclature of the major Italian dictionaries. This nucleus will then be constantly enriched by new terms emerging from the examination of those texts which are gradually enriching the C.N.U.C.E. electronic library. The following information, in a very simple formalism, has already been recorded for every lemma: its place in determined sectors of the lexicon (regional, archaic, technical, etc.), the homographic or other links[13] with

[12] See the example given by C. R. Wood at the I.C.C.L. 1969 at Stockholm.

[13] Experience of analysis carried out on some 2,000 Italian texts shows that the researcher's divergences on lemmatization criteria can be reduced to a few possible alternatives for certain words, for example to distinguish the substantive and adjectival use, the adverbial and prepositional use, etc. The I.M.D. is organized in such a way that the lexicon is divided into classes of words for which the same alternatives are valid. It only takes a few control cards to choose the preferred alternative for each class.

other lemmas, a preliminary grammatical classification,[14] some phonetic comments,[15] and a series of inflection codes. A generative kind of algorithm has produced from this information all the possible forms according to the Italian system. Thus the I.M.D. consists of some 150,000 lemmas and two million corresponding forms, recorded in machine-readable form according to both current spelling practice and phonetic transcription.

We are now adding further information to each lemma: etymology, division into roots and affixes (which will permit us to create an inventory of morphemes alongside the inventory of lemmas), and the distinction between polysemies. This latter is obviously a very engrossing task, given the present state of the art in semantic theory. Until a theory is produced which can be applied to something other than the usual lexical micro-subset, our approach is essentially pragmatic. We apply to the various accessions a codified classification, of the same type as R. Hallig's and W. von Wartburg's *Begriffsystem,* as well as a few classical lexical features, of the animate-inanimate type, etc. The central part of the work, however, consists of formulating a brief meaning definition formalized only in its structure, not in the contained definers.[16] These definitions try also to express, as far as possible, the links between the various accessions. We expect to be able in the future to make the definitions more coherent, economical, and compact, applying to them algorithms now under study.

The aims of the I.M.D. can be summed up in three groups:

Semi-Automatic Lemmatization of Italian Texts

As a result of consultation of the I.M.D., the forms from the texts are linked directly with their lemmas if they are univocal; if they are homographic they are marked out during lemmatization, accompanied by the contexts in which they appear and by all the lemmas to which they belong. We have two algorithms for consulting the I.M.D.: one searches the text for the token in the forms contained in the I.M.D., the other breaks the token into segments and, according to morphological rules and tables of endings, seeks the root of the words in the list of roots in the I.M.D. The former algorithm is the most convenient when dealing with a corpus of several hundred thousand words, the latter for smaller numbers. It could be argued that there are algorithms in existence which could eliminate the consultation of a machine dictionary by means of morphological rules, probabilistic rules, and context-sensitive rules.[17] The applications to which the I.M.D. is put do not seem to permit the use of such algorithms for various kinds of reasons: for example, the need for a high exactitude threshold, and the need to link every token in the text not only to the lemma and the grammatical category but also to additional information contained in the I.M.D.

Other objections to the use of a machine dictionary in the development of large text archives are based mainly on the problem that it represents a personal point of view and a crystallization of a linguistic system which would be imposed on the texts. The results would include the risk of treating as univocal those words which should be

[14] For now, this is of the traditional type: substantive, adjective, etc. In all, we have 64 different classes. Later we will apply the categories which relate to formal grammar of Italian. In particular we have begun by expressing the different types of verbal construction. See for example the works of M. Gross (1969) and of A. K. Zholkovsky and I. A. Melchuk (1970).

[15] The algorithm for the automatic phonematic-phonemic transcription of Italian texts requires that some information which cannot be determined algorithmically be previously inserted: position of accent, open *e* and *o*, etc.

[16] Definitions are intended to express two kinds of linking, word-object and word-word. The former is performed by a plain description (complex definition) which can be translated in linguistic features. The latter consists both in one word (synonym: word-to-word linkage) and in syntactic transformation (derivate-to-primary linkage).

[17] See, for example, S. Hellberg (1972) and S. Klein *et al.* (1963).

considered as homographs and vice versa. It can be shown (A. Zampolli, 1973, pp. 118-61) that these difficulties can be almost totally solved by an appropriate structuralization of the links between the lemmas in the machine dictionary, providing the possibility of choosing, by means of a few control cards, between different levels of analysis explicitly formulated in the machine dictionary's definition of linguistic units.

On the other hand, the advantages of using the machine dictionary are obvious. From the economic point of view the human time saved varies, particularly if for the solution of homographs interactive systems (terminal, video, etc.) are used, between 70 percent and 90 percent. From the scientific point of view the machine dictionary favors the comparability and even the normalization of the lemmatization of the text, which is necessary in the light of the present multiplication of text examinations.

Text Corpus Editing

Lexicons, indexes of concordances, for an author's complete works

Among the many examples, the numerous ones that relate to the complete works of philosophers and thinkers continue the activities of Busa. The lexical index is considered not only the key which permits an easy search for those passages where ideas are defined or explained, but also as an instrument of interpretation. Apart from the *Index Thomisticus* already mentioned, the first of the planned 60 volumes which should begin publication within the year, I cite only the archive of the complete words of the Italian nineteenth-century philosopher, A. Rosmini, which was completed this year.

Examination of single works

The list of works which appears regularly in *Computers and the Humanities,* in the section on "Verbal Materials in Machine-Readable Form," represents for the most part stylistic studies, often carried out as degree theses, on a lexical and statistical basis, using concordances and indexes as favorite tools. It may be interesting to quote some concordance experiences that differ from the traditional forms.

The concordances that we call *contrastive* have been shown to be very useful in the comparison of an original text with its satellite texts (for example translations into other languages or later editions). Every form of the original text has as its context not only the normal context but also the corresponding contextual units in the satellite texts. Equally every form of every satellite text has as its context its own normal context with the corresponding contexts in the other satellite texts and in the basic text. This method has permitted useful critical considerations on the mechanism of translation of literary text: e.g., the work on the various Italian translation of some passages from Goethe's *Faust.*

The concordances we call thematic have been requested for the analysis of the structure of poetic works, above all, those based on oral tradition. In these concordances the entry consists of the code of a theme or a subtheme, and the context is the whole corresponding passage. Obviously all the possible lists of correlations between lexicon and theme are also produced. The most advanced work in this sector is the structural analysis of the lyrical components of Pindar and Bacchylides.

Dissimilar applications which have in common the fact that they deal with texts in a philological way include the *Tabulae Iguvinae* (A. Prosdocimi, 1970), the *Götische Bibel,* and collected Latin drama. D'Arco Silvio Avalle's project (1973), on the poetic language of Italy before the end of the thirteenth century, is particularly interesting because it represents a complete cycle of operations which, in interaction with the computer, transcribes original manuscripts and prints by photocomposition of the text and its various analyses (indexes, concordances, etc.). This project should become a model of its kind, since the text is printed and the indexes and concordances are produced with only

one transcription of the text. There is, therefore, only one correction, with the additional advantage that the interaction with the computer allows that correction to be more accurate than is possible through traditional proofreading. It has now been proved by the experiences of several researchers that the indexes and concordances are very efficient instruments for the detection of the editor's mistakes, oversights, and inconsistencies in the text. The indexes and concordances produced every day at C.N.U.C.E. bring to light mistakes and errors in critical editions rightly held to be safe and excellent. It is therefore a great advantage to be able to publish an edition of a manuscript after having been able to analyze it with the computer.

Another typical problem is the treatment of variants. A few attempts to automate the *collatio* and the *recensio codicum* phases have produced positive results in the study of the relative algorithms, but none of these appears to be destined for practical application (cf. G. P. Zarri, 1973). In the effort of some Italian philologists at a methodological definition of the treatment of variants in texts for which indexes, concordances, and statistical analyses are to be published, the most varied methods have been tried. These range from the simplest solution that just applies a symbol to mark those words with variants, to the most complicated solutions in which all the information in the critical apparatus is introduced during analysis so that the variants and the links between them are thus presented in the concordances and indexes. The most advanced project is the examination of the three editions of L. Ariosto's *Orlando Furioso* by C. Segre, the president of the International Semiotic Society.

Historical Research

A degree thesis at Turin University will be based on research into medieval Italian documents. At the Institute of Modern History, Paleography and Diplomatics of the University of Pisa, a fairly large corpus of medieval documents is being worked on. In both these projects the pre-editing phase is of great importance. It includes the compilation of an informative card for every document, the division of the text for the diplomatic studies, the insertion of commentaries into the text, and the explanation of the links between the various commentaries and, in turn, between them and the text (L. Fossier, 1973, P. Scalfati, 1970).

Documentation

The information retrieval studies carried on by industry and the state-owned corporations, although originally numerous, applying methods of automatic meaning extraction, with instruments for linguistic, syntactic, and semantic analysis, have now been virtually abandoned. Instead, the whole text is being used and the research carried out with the help of thesauri (R. Borruso et al., 1969) or machine dictionaries. Two very important projects dealing with legal texts are the Istituto per la Documentazione Giuridica del C.N.R. (C. Ciampi, 1973) and the project under the patronage of the Chamber of Deputies on all the Italian Acts of Parliament from 1848 to the present day (Camera, 1972). The latter project, which requires the analysis of some hundred million words, will use the I.M.D. A statistical-linguistic approach to information retrieval which has many supporters is employed at Euratom in Ispra, near Milan.

Linguistic Statistics

The electronic library of C.N.U.C.E. offers the researcher in linguistic statistics a collection of easily usable corpora of about 80 million words plus the benfits of uniform recording and analyzing methods. Since almost all the researchers who are making lexical or stylistic investigations of these corpora use statistical methods and techniques, it is impossible to list them all. In general, these studies into the statistical composition of

contemporary Italian have the dual aim of providing a knowledge of the frequencies of units and "rules" in the Italian linguistic system at all levels (phonological, morphological, lexical, syntactic, etc.) and at the same time establishing models of the quantitative characteristics of the language which would be more satisfactory than the models produced in the period following the Second World War (e.g., by P. Guiraud and G. Herdan), which the advances in electronic analysis of texts have rendered obsolete.

Statistics on the Lexicon

The I.M.D., which contains morphemes and words in phonological transcription, can be considered as representative of the Italian lexical system, despite the theoretical and practical difficulties of this task.[18] I still consider valid the suggestion of N. Trubetzkoy (Trubetzkoy, 1939, ch. 7) which explicitly and energetically affirmed that statistics were complementary in the dictionary and in the texts. Although this suggestion was put forward by the Prague school and accepted by many European linguists, such as A. Martinet (1960) and C. Tagliavini (1968) it has never, except for very small samples, been put into practice in any language.[19] Only a machine dictionary permits such researches, just as it permits the creation of an inventory of all the minimal pairs which, despite all the justified doubts on the subject,[20] remains the basis of the calculation of the functional yield of a phonematic system. At the level of composition and derivation of words, satisfactory statistical studies exist for other languages,[21] and are now being applied to Italian.

Statistics on Texts

Phonemic Statistics

The corpus at C.N.U.C.E. has been used for statistics on a phonemic, lexical, and syntactic level. At the first, a problem is created by the many different regional systems, and by the absence of any universally accepted phonemic inventory for any regional system. In the speech of educated Florentines, with which standard Italian tends to be identified, appear certain phonetic entities, e.g., geminated consonants, which resist the analytic techniques of the phonological schools. An inventory of phonetic units has been proposed, such that if their frequencies are known, it is possible by simple arithmetical operations, to obtain the frequencies of the phonemes which comprise the various proposed inventories (A. Zampolli, 1968). The data obtained, from a corpus of five million phonemes, with regard to phonemes, syllables, word structures, consonant and vowel groups, etc., are about to be published. These data refute the widely held hypothesis that the frequency of the phonemes is independent of the kind of text under consideration.

Lexical Statistics

Because Italian was the only language that had no frequency dictionary, the *Lexicon of the Frequencies of the Contemporary Italian Language* by U. Bortolini, C. Tagliavini, and A. Zampolli filled an obvious lack. This work, which closely followed the

[18] For example, whether the lexicon of a language should be considered a closed or open list, whether one can talk of the lexicon of an ideal speaker or whether one should consider the lexicon of limited social groups, etc. For these problems and the related bibliography see A. Zampolli (1973c).

[19] For example, see N. Saramandu (1966).

[20] The most lucid exposition of the theoretical and practical difficulties connected with functional load measurement is probably that of J. Lyons (1968, 2.4).

[21] For example, see E. Sapir (1921), J. H. Greenberg (1960), P. Guiraud (1954), and J. Dubois (1966).

example of A. Juilland's collection (1964, 1965, 1971), can be placed in the sphere of research intended to discover statistical models. These replace previously proposed models which can definitely be traced back to the system of chance extraction, and found their definitive shape in the parallel that C. Herdan established between *langue* and *parole*, on the one hand, and the statistical relationship sample and universe on the other. That one cannot talk of the probability of a word in a linguistic system is now proved also for the high frequency words and the so-called empty or asemantic words. The *Lexicon of Frequencies* showed, for instance, that the frequency of articles in a corpus of 500,000 words taken from the cinema, theater, novels, newspapers, and textbooks varies between 5000 and 15,000. Equivalent variations are shown by the frequencies of prepositions, adverbs, conjunctions, etc., whether taken individually or as categories. The *Lexicon* showed surprising regularities, however. Some thirty parameters were analyzed (frequency of grammatical categories, average length of syllables, distribution of frequencies in decreasing order, etc.). The frequencies of these parameters in the subsections of the corpus, considered in the order cinema, theater, novels, newspapers, and textbooks, vary in only two ways. They either rise regularly from one to two, to three, to four, to five, or decrease in the same order; there is never a different sequence.

It therefore seems justified to search for a model which is based, not on a language as one universe, but on several synchronic and diachronic levels of a language, and which at the same time divides the words into classes with different statistical characteristics (theme words, 'mots disponibles', etc.). Such a model, depite the undeniable difficulties of a stylometric conception which presupposes a linguistic norm against which stylistic exceptions can be measured (cf. R. Dyer, 1973), is the only one that seems to provide a chance of well-based advances in linguistic statistics, both as the study of quantitative macromodels and as an instrument of stylistic research.

Syntactic Social Linguistic Studies

T. de Mauro, G. Policarpi, G. Proverbio, and A. Zampolli have begun a frequency study of traditional syntactic types. The choice does not derive, as some might think from the consideration of the negative theoretical position of the first followers of Chomsky towards linguistic statistics, a standpoint which today has been at least partially overcome,[22] but from the inadequacy of all the generative-transformational grammars for the description of the chosen corpus, which includes, not only literary texts but texts from popular Italian. The aim is to bring into the light of scientific research important linguistic areas which today are still in the shadow. If there are numerous phonetic and phonological studies of the Italian diasystem, there are still few studies of the morphological and lexical differences, and studies on syntax are almost nonexistent. This extension is linked with the predominant currents in studies in Italian socio-linguistics, where the most important problem is the cultural integration of emigrants (F. Alberoni, G. Bablioni, 1965) and above all of their school-age children who are disorientated when they come into contact with different social and linguistic realities. The statistics cover the frequency of various types of sentences and phrases, and relationships and the sequences of the phrases in the sentence. Not only the quantitative data is interesting but also the new syntactic types which, of necessity, have to be introduced into the traditional inventory.

Psychology and Psychiatry

These fields of research are an example of an application of statistical linguistics to extralinguistic disciplines. Psychiatrists and psychologists from the universities of Pisa

[22] For example, see C. W. Hayes (1966 and 1968), R. Ohmann (1964), I. Rosengren (1971).

and Milan have shown that statistical differences between the various categories of mental illness, particularly in the use frequencies of grammatical categories, could be used in the diagnosis of, and a more profound understanding of, mental sickness. Studies carried out on conversations of schizophrenics, maniacs, and mentally retarded children have shown interesting links with distribution of frequencies found in the subsets in the *Lexicon of Frequencies*. Research now continues at the syntactic and contextual level (Castrogiovanni et al., 1967, Castrogiovanni, 1973).

Mathematical Models

Discovery Procedures

Inventory of Popular Italian Songs. For some years the Italian Committee for Demologic Studies, directed by A. M. Cirese, has been developing a systematic documentation for the available information on Italian folklore. Published and unpublished collections of popular songs from all the Italian regions have been put in machine-readable form. This corpus of 300,000 lines has been submitted first of all to the now-standard processing: concordances and indexes of words and of lines. But the methods used have unusual characteristics (A. M. Cirese, 1973). A remarkable number of groups of fixed and crystallized words (formulae or models which, passing from one text to another, represent the basic elements and units of the lexicon) constitute popular poems. A critical edition of these texts would not follow the principles which would be valid for the texts of a single, known author. Rather than seek to reconstruct the version closest to the original by eliminating the alterations introduced by tradition, folklorists adopt the sum of the variants as their "text." One of the main goals of the research, therefore, is to identify formulae, and to discover the syntactic rules which link them within certain metric schemata. We are asked whether discovery procedures exist for formulae whose structure remains constant, while the structural elements can vary. Fixed groups of words or ordered and even discontinuous successions of known elements can be found by means of the frequencies of such binary, tertiary, *n*-ary groups. The same can be said of structures which vary in a known context. The discovery procedure is more complicated from the algorithmic point of view when the identification of the formula requires the recognition of a particular logical and syntactic structure whose nodes can be occupied by particular classes of meanings, rather than by explicitly enumerable keywords. The oral and multi-dialectical nature of the texts enormously complicates the problem of parsing and of thesauri. For the moment very simple procedures are being studied.

The normal kind of rhyming dictionary, which divides the lines of the text in classes of equivalence on the basis of the relationship "having the same ending" (that is to say, the same string of letters starting from the final tonic vowel included), is insufficient in relation to the actual conditions of these texts, for it does not recognize the need to understand the metric mechanisms. The learned Italian poetry uses the so-called *perfect rhyme* (completely identical endings) and two kinds of imperfect rhyme: *assonance* (endings with identical vowel sounds and dissimilar consonants, e.g. *are* : *ate*) and unstressed *consonance* (endings with different tonic vowel but otherwise identical). In popular poetry not only these three types are found, but also lines which share a different relationship between the respective endings (e.g., *arrivàre : amáro,* and *andáre : avéte*). Our aim is to deduce algorithmically the isophonic types and levels of popular poems, without applying predetermined metric schemes. One point of departure is a system that associates a triple set with each line ending. Each triple of the set expresses the relationship of that ending with all the endings that precede it in the same unit of composition. The first character of the triple expresses the identity or the difference between the tonic vowels, the second between the consonants, the third between unstressed vowels. Naturally the need to work on the phonemic transcription and not on the conventional

spelling presents a difficult problem. The I.M.D. opens the way to a basically automatic solution.

Word Formation

The formation of words by means of affixes has always interested linguists from various schools. Recently it has become part of the projects of linguistic planning. For example, a committee for the terminology of all the Romance languages has been formed with the aim of creating the world-wide norms for the development of scientific terminology. Naturally a knowledge of the rules and of the potentialities of word-formation by means of affixes is essential for such a purpose. Finally even the generative schools have felt the need to introduce into their grammar not only lists of morphemes, but also the rules for their composition. The report "Morphology in a Generative Grammar," presented by M. Halle to the 11th International Congress of Linguists in Bologna, 1972, focuses on this topic (M. Halle, 1973). He presented a series of facts to show that, apart from the list of morphemes and the rules of word formation, which together define the *potential* words of the language, a needed filter, when applied to the potential words, produces the set of *actual* words. Although this model has already been criticized, it strikes us that, whether it is accepted or replaced by other models, the following information has to be gathered: a list of morphemes of a language and their division into classes of equivalence with regard to the possibility of combination. The procedure for acquiring this information, we think, should break the lexicons of the I.M.D. into morphemes, establish the classes and rules of possible combination, produce the potential lexicon automatically, and compare with the I.M.D.

Grammar Tester and Automatic Syntactic Analysis

For the moment we have abandoned the idea of creating a new program. We have been allowed to implement on our own electronic systems such extant programs as J. Friedman's transformational grammar tester (1969) and the ATEF and CETA of B. Vouquois's team in Grenoble. This decision was taken mainly because of the state of grammatical studies of Italian, which are far from allowing us to obtain a formal grammar adequate for the texts, even though research was multiplied in the last five years (Società di Linguistica Italiana, 1969 and 1971, Centro per lo Studio..., 1972, Accademia della Crusca, 1971).

The computational treatment of grammar seems to us useful in the attempt to coordinate our grammatical research into a coherent pattern. The parsing algorithms necessary to make the processing required by the projects listed in the above (e.g., to distinguish the homographs brought to light in the I.M.D.), unfortunately, are rendered utopian by the synchronic and diachronic variety of the texts. But it is certainly possible to bring into use some parsers *ad hoc* to achieve a notable saving of time and money with such problems as high-frequency grammatical homographs. *The Problem of Semantic Representation in a Computerized Textual Grammar,* the doctoral thesis of Hans-Dieter Rauschner, at Scuola Normale Superiore of Pisa, is based on a synthesis of generative semantics and recent models of German *Textlinguistik,* so that the notion of *sentence* will no longer be considered. His hypothesis will be verified in a man-machine dialogue system for understanding and generating natural language, the object of the communication process being limited to a defined sector of French verb morphology, but the computer will have a complete knowledge of this sector. Thus, in any dialogue concerning morphology, the grammar will be able not only to judge the grammatical correctness of an input, but also to understand it, relating it by identification procedures to the right particulars of the system's universe. The final aim of the research is to prove that the object of communication can be enlarged. This will be tried by giving the system a

complete knowledge of the grammar on which the system itself is based, thus identifying language and metalanguage.

History of Art and Archeology

The first studies in the mechanization of cataloging and the organization of museums are beginning which are obviously necessary to handle the extraordinary richness of the Italian artistic heritage, and to reduce the continual dangers to which it is exposed. The Institute of Archeology of the University of Pisa, under the direction of P. Arias (1973) is continuing the development of D. Beazley's catalog of Greek ceramics.

Music and Musicology

Attempts to study musicians' styles are rare although the extreme formalization of musical texts would seem to present fewer difficulties for automatic analysis than literary stylistics. So-called computer music has, however, rapidly gained ground. The Musicological Laboratory of the Florence Conservatory, directed by Maestro P. Grossi, has introduced programs at C.N.U.C.E. to execute music. The programs allow the creation of a fairly high number of original rhythms and intervals impossible for other instruments. Also the computer accepts a series of instructions which permit the immediate modification of the tempo and frequencies of any musical text previously executed or simply memorized. A language has been created for the representation of musical texts, many concerts have been given, and some records have been published.

Teaching Activities

A course in mathematical and computational linguistics became official at the University of Pisa's Faculty of Letters and Philosophy in 1970. A course was introduced in 1973 at the School of Postgraduate Linguistic Studies at the same university.[23] In the last two years in Italian universities some twenty theses on the uses of computers for the stylistic analysis of literary texts have discussed, e.g., the poetic works of Hart Crane, W. C. Williams, Ronsard, Saba, Montale, Ungaretti, five Italian translations of Goethe's *Faust,* the *Divine Comedy,* and some Catalan poems.

Every year in the field of C.N.U.C.E.'s teaching activities two- or three-week courses are given to introduce humanists to the use of computers. Similar courses are held for computer music, and starting this year, a course also in the history of art. Beyond these specific courses, programming instruction is customarily given, suitable to researchers not particularly familiar with mathematics. The assistants at C.N.U.C.E.'s Linguistic Branch have given courses for humanists and linguists at other Italian universities such as Rome, Turin, and Milan. Since 1970 a biennial summer school is held at C.N.U.C.E. on linguistic and literary electronic data processing, inspired originally among others by J. Raben and R. Dyer. The school alternates introductory and specialized classes; in 1970 it concentrated above all on stylistics and stylometrics.[24] In 1972 there were two parallel sessions: the formal grammars of natural languages and their automatic treatment and applications of informantics to lexicology and lexicography.[25] In 1974 it will probably

[23] See A. Zampolli (1969).

[24] Methods of describing the syntax of natural languages which can also be treated automatically: Courses: M. Gross, distributional and transformational methods, G. Faucunnier, formal grammars and automatoms, M. Kay, application of formal grammars to the automatic syntactical analysis of natural languages, automatic treatment of grammatical rules, J. Lyons, Semantics.

[25] The possibilities and perspectives of the application of computers to the needs of lexicographical analyses and realization of various types of dictionaries: Courses: B. Quemada, The problems and methods of contemporary lexicology and lexicography aided by computers, Ch. Muller, elements of lexical and lexicometrical statistics, A. Zampolli, the technology of computers and terminals, their applications to lexicological and lexicographical aims.

be devoted to formalization in semantics and its computational reflexes.[26]

Some Conclusions

The positive, and in a sense exceptional, development of research in Italy without doubt derives, not so much from the rapid proliferation of projects and experiments which has overcome the standstill of 1955-1965, as from the creation of C.N.U.C.E.'s electronic library, and the standardization of text-recordings and the main analyzing programs. These unquestionable advantages enjoyed by Italian researchers are freely available to others who wish access to all the texts prepared in Italy as one corpus, gathered together in one place, and treated according to homogeneous criteria.

At recent European meetings the hope has been expressed of creating similar centers or at least similar corpora, one for each language. The links between researchers from different countries would certainly be facilitated if a network of national centers could be created for the exchange of material, for the exchange of information, for co-ordination of experiments and so on.

For the future, therefore, our work seems to show solid guarantees, in particular for the creation of an archive of linguistic and literary data. This archive is rapidly growing to fill the needs of the entire range of Italian linguistics studies, both in their diachronic and in their stylistic, regional, and sociolinguistic aspects. This situation, although reasonable, makes particularly worrisome and urgent to Italian researchers a problem which has been emphasized, more or less explicitly, in various places.[27] The development of applications along what can now be considered the classic lines of the 50's and 60's has reached saturation point. If current methodology continues to be followed, there are very few concrete prospects of development. For example, while the computers are becoming more and more rapid, and the programs more and more sophisticated, the lexicographers cannot benefit in proportion because present methodology already produces far more data than an editing team of reasonable size can analyze when working with present procedures. If we accept the limits of today's procedures we will continue to produce concordance or lexical archives without adequate classification or analysis. Such an operation is postponed to a later editing stage which creates enormous difficulties due to the quantity of documentation and materials. It seems necessary to render automatic at least part of lexical analysis. The diffusion of results, by publishing of the concordances and indexes, seems difficult, and perhaps not very useful. This form of diffusion should be replaced by a linguistic data bank, which could be consulted in an interactive fashion through a network of terminals. (For a detailed discussion, see A. Zampolli 1973a.)

The task which awaits Italy's researchers, and perhaps not only them, seems to be

[26] The following conferences have taken place in Italy:

Man and the Machine, XXI Congresso Nazionale di Filosofia, Pisa, 1967, see Proceedings, *L'Uomo e la Machina* (1967).

Electronic automation and its scientific, technical and social implications. L'Automazione Elettronica e le sue Implicazioni scientifiche, techniche, e social, Accademia dei Lincei, Rome; 1967, see Atti, 1968.

Séminaire International sur le Dictionnaire Latin de Machine, C.N.U.C.E., Pisa, 1968, see Actes (1968).

Colloque International sur l'Elaboration Electronique en Lexicologie et en Lexicographie, C.N.U.C.E., Pisa, 1970, see the Proceedings in A. Zampolli, 1973B.

1er Table ronde international des directeurs d'entreprises lexicographiques, Accademia della Crusca, Florence, 1971, to be published.

The techniques of classification in linguistics, Accademia dei Lincei, Rome 1972 (to be published).

International Conference on Computational Linguistics, International Committee on Computational Linguists, C.N.U.C.E., Pisa, 1973 (to be published).

[27] See, for example, A. J. Aitken (1971 and 1973), Bailey (1972 and 1973), P. Lehman (1972), Venezky (1972).

to put to common use in their activities some instruments already studied in depth by other computational disciplines. Here I am thinking in particular of the machine dictionaries, the thesauri, interactive man-machine techniques, data base management, automatic or semiautomatic parsing systems, even at a not particularly sophisticated level.[28] Because of their variety and their size, the corpora being constituted for linguistic and humanist study, present not merely particular problems for computerization, but also, and above all, a variety of linguistic facts for the formal descriptions of which present knowledge and linguistic methodology are far from being adequate.

[28] See, for example, D. Ross (1973), L. Milic (1973).

References

Accademia Della Crusca. *Il Canzoniere di F. Petrarca. Indici e Concordanze,* Firenze, 1971.

———. *Indice dei testi sottoposti a spoglio lessicale,* Firenze, 1972.

———. *Studi di Grammatica Italiana,* Firenze, 1971.

———. "Actes du deuxième colloque international de linguistique et de traduction (Montréal 4-7 octobre 1972)," in *META* 18, 1-2 (1973).

Aitken, A. J. *Quelques problèmes de la lexicographie historique: comment l'ordinateur peut-il donner de l'aide?* in press in A. Zampolli, 1974.

———. *Historical Dictionaries and the Computer,* in R. A. Wisbey, ed., 1971.

Aitken, A. J., R. W. Bailey, N. Hamilton-Smith, eds. *The Computer and Literary Studies,* Edinburgh, 1973.

Aliprandi, G. "Frequenze Dattilografiche," in *Bollettino dell'Accademia Italiana di Stenografia di Padova,* 1940, p. 273.

———. *Almanacco Letterario Bompiani 1962,* Milano, 1962.

Arias, P. E. "Mise en fiche des vases grecs; problèmes et discussion," in A. Zampolli (1973B), pp. 243-47.

Atlante Linguistico Italiano. *Questionaria, I Testo,* Torino, 1971.

———. *Questionario, II Indici,* Torino, 1973.

———. "Actes de Seminaire International sur le Dictionnaire Latin de Machine," in *Calcolo,* Vol. 5, Suppl. n. 2 (1968).

———. *International Conference on Lexicography in English,* New York, 1972.

Bar-Hillel, Y. *Language and Information–Selected Essays in Their Theory and Application,* Jerusalem, 1964.

Bartoletti Colombo, A. M. *Per un vocabolario delle Costituzioni di Giustiniano,* Firenze, 1973.

Bocca, E., A. Pellegrini. "Studio statistico sulla composizione fonetica della lingua italiana e sua applicazione pratica all'audiometria con la parola," in *Archivio Italiano di Otologia, Rinologia e Laringologia* 56 (1950) suppl. n. 5, pp. 116-141.

Borruso, R., A. Falcone, E. Caporta, V. Novelli. *Sistema di ricerca elettronica della giurisprudenze,* Roma, 1969.

Bortolini, U., C. Tagliavini, A. Zampolli. *Lessico di frequenza della lingua italiana contemporanea,* IBM Italia, 1971.

Busa, R., S. J. *Sancti Thomae Aquinates Hymnorum Ritualium Varia Specimina Concordantiarum,* Milano, 1951.

Busa, R., A. Zampolli. "Centre pour l'Automation de l'Analyse Linguistique (C.A.A.L.) Gallarate," in *Les machines dans la linguistique,* Prague, 1968.

Castrogiovanni, P., A. Telara, "Primi risultati di un'analisi statistica morfologica e lessicale delle risposte al test di Rorschach nella prospettiva di uno studio dei rapporti tra psicopatologia e linguaggio," in A. Zampolli (1973B) pp. 307-23.

Castrogiovanni, P., G. Maffei, P. J. Pasquinucci, N. Lijtmaer, S. A. Cerri, G. Torrigiani, A. Zampolli. "Analisi linguistica delle risposte al test di Rorschach di schizofrenici e neurotici e dei rispettivi familiari," in *Neuropsichiatria* 34, iv (1968) pp. 811-37.

Centro per lo Studio dell'Insegnamento all'Estero dell'Italiano, *Scritti e ricerche di grammatica italiana,* Trieste, 1972.

Ciampi, C. "Les Projets de recherche automatique des informations juridiques," in A. Zampolli (1973B), pp. 249-68.

Cirese, A. M. "Inventaires et répertoires lexicaux, formulaires et métriques des chants populaires italiens," in A. Zampolli (1973B), pp. 209-32.

D'Arco Silvic Avalle. "Projet pour une liste des concordances de la langue poétique en Italie avant la fin du XIII Siècle," in A. Zampolli (1973B), pp. 19-27.

De Mauro, T., G. Policarpi. "Ricerche sulla strattura del periodo italiano," in Società Italiana di Linguistica, *L'Insegnamento dell'Italiano la Italia e all Estero,* Roma, 1971, pp. 683-93.

Dimitrescu, F. "Projet d'un dictionnaire de la langue roumaine du XVI siècle," in A. Zampolli (1973B), pp. 41-48.

Dubois, J. *Etude sur la dérivation suffixale en français moderne et contemporaire,* Paris, 1962.

_____. "Utilisation des statistiques lexicographiques pour l'étude structurale du lexique," in *Statistique et Analyse Linguistique (Colloque de Strasbourg, 1961),* Paris, 1966, pp. 95-98.

Duro, A., A. Zampolli, "Analisi lessicali mediante elaboratori elettronici," in *Atti di Convegno sul tema: L'Automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali* (Accademia Nazionale dei Lincei, Roma, 1967), Roma 1968, pp. 121-39.

Duro, A. "Elaborations électroniques de textes effectuées par l'Accademia della Crusca," in A. Zampolli (1973B), pp. 53-75.

Dyer, R. "The Measurement of Individual Style," in A. Zampolli (1973B), pp. 325-48.

Faedo, C. *Concordanze e Frequenze dell'opera poetica di H. Crane.* Unpublished thesis, Pisa, 1971.

Friedman, J. "Applications of a Computer System for Transformational Grammar," in *ICCL,* 1969.

Froger, D. J. *La critique des textes et son automatisation,* Paris, 1968.

Gardin, J. "A Typology of Computer Uses in Anthropology," in D. Hymes, ed., *The Uses of Computers in Anthropology,* 1965, pp. 103-17.

Garvin, P. L. "Computer participation in linguistic research," *Language,* 38, iv (1962), pp. 385-89.

Giraud, P. *Les caractères statistiques du vocabulaire,* Paris, 1954.

Grassi, G. "Perspectives de l'emploi de l'elaborateur electronique en géographie linguistique et en dialectologie," in A. Zampolli (1973B), pp. 233-39.

Greenberg, J. *Essays in Linguistics,* Chicago, 1957.

Gross, M. *Lexique des constructions completives,* Paris, 1969.

Guardi, T. "Le lexique du théatre latine," in A. Zampolli (1973B), pp. 101-102.

Hayes, C. W. "A Transformational-Generative Approach to Style: Samuel Johnson and Edward Gibbon," in *Language and Style,* I (1968).

Hays, D. G. "Applied Computational Linguistics," in G. E. Perren, J. E. M. Trim, eds., *Applications of Linguistics,* Cambridge, 1971, pp. 65-84.

Hallig, R., W. von Wartburg, *Begriffssystem als Grundlage für die Lessicographie,* Berlin, 1952.

Hellberg, S. "Computerized Lemmatization without the Use of a Dictionary: A Case Study from Swedish Lexicology," in *Computers and the Humanities* 6, iv (1972), pp. 209-12.

Heilmann, L. "Considerazioni statistico-matematiche e contennio semantico," in *Quaderni dell' Istituto di Glottologia* (Università di Bologna) 7 (1962-1963), pp. 34-45.

Juilland, A., D. Brodin, C. Davidovitch. *Frequency Dictionary of French Words,* The Hague, 1971.

Juilland, A., E. Chang-Rodriguez. *Frequency Dictionary of Spanish Words,* The Hague, 1964.

Juilland, A., P. M. H. Edwards, I. Juilland. *Frequency Dictionary of Rumanian Words,* The Hague, 1965.

Kay, M., K. Spark Jones, "Automated Language Processing," in E. Cuadro, pp. 141-66.

Kay, M. "Standards for Encoding Data in a Natural Language," in *Computers and the Humanities* 4, v (1967), pp. 170-77.

Klein, S., R. F. Simmons. "A Computational Approach to Grammatical Coding of English Words," *Journal of the ACM* 10 (July 1963), pp. 334-47.

Lamb, S. M. *Linguistic Data Processing.*

_____. "The Digital Computer as an Aid in Linguistics," in *Language* 37 (1961), pp. 382-412.

Lehmann, W. P. "On the Design of a Central Archive for Lexicography in English," preprint for the *International Conference on Lexicography in English,* New York, 1972.

Lyons, J. *Introduction to Theoretical Linguistics,* Cambridge, 1968.

Losano, M. G. "Lexicographie computationelle et information juridique," in A. Zampolli (1973B), pp. 299-303.

L' Uomo e la Macchina, Atti del XXI Congresso Nazionale di Filosofia (Pisa, 22-25 Aprile 1967), Torino, 1967.

Meriggi, P. "Un lexique de l'Hittite cuneiforme," in A. Zampolli (1973B), pp. 111-13.

Milic, L. "Autocoding in Computational Stylistic," in B. Kachru, H. F. W.Stahlke,*Current Trends in Stylistics,* also in A. Zampolli (1974).

Montgomery, C. A., "The 1969 International Conference on Computational Linguistics: A Progress Report," in *Computers and the Humanities* 4, iii (1970), pp. 193-198.

Ohmann, R. "Generative Grammars and the Concept of Literary Style," in *Word* 20 (1964), pp. 423-439.

Parisi, D. "Un modèle componentiel du signifié dans l'étude du lexique et de la syntaxe," in A. Zampolli (1973B).

Penazzo, D. *Analisi comparativa di cinque traduzioni italiane del Faust di Goethe,* unpublished thesis, Parma, 1972.

Prosdocimi, A. L. "Kutef persnima ařepes arves: analisi interna e problemi redazionali nelle invole iguvine," in *Mille. I Dibattiti del Circolo Linguistico Fiorentino, 1475-1970,* Firenze 1970, pp. 185-207.

Quemada, B. "Bilan des applications de l'informatique aux études lexicologiques," in *Actes du deuxième colloque international de linguistique et de traduction* (Montréal 4-7 octobre 1972), in *META* 18, i-ii (1973), pp. 87-102.

Raben, J. "Computer Research in the Study of Literature," in A. Zampolli (1973B), pp. 391-455.

Raben, J., R. L. Widmann, "Information Systems Applications in the Humanities," in E. Cuadros, ed., pp. 439-69.

Rosengren, J. "The quantitative concept of a language and its relation to the structure of frequency dictionaries," in *Etudes de linguistique appliquée* (nouvelle série) 1 (1971), pp. 103-26.

Ross, D. "Beyond the Concordance: algorithms for descriptions of English clauses and phrases," in A. J. Aitken *et al.* eds., 1973.

Sapir, E. *Language,* New York, 1921.

Saramandu, N. "Sur le rendement fonctionnel des types de structures phonématiques en roumain," in *Cahiers de linguistique théorique et appliquée,* III (1966), pp. 147-160.

Scaliati, S. P. P. "Notizie e studi a proposito della edizione delle pergamene pisane" (Secoli VIII-XII) in *Archivi e Culaura,* 4, i-ii (1970), pp. 181-95.

Società di Linguistica Italiana. *La grammatica, La lessicologia,* Roma, 1969.

———. *La sintassi,* 1969.

———. *L'insegnamento dell'italiano in Italia e all'estero,* 1971.

Stolz, W. S., P. H. Tannenbaum, F. V. Carstensen, "A Stochastic Approach to the Grammatical Coding of English," in *CACM,* 8, 6 (1965), pp. 399-405.

Tagliavini, C. *Concordanze della Divina Commedia,* Pisa, 1965 (Ci si riferisce all'introduzione dell'edizione IBM Italia).

———."Applicazioni dei calcolatori elettronici all'analisi e alla statistica linguistica," in *Atti del Convegno sul tema: l'automazione elettronica e ie sue implicazioni scientifiche, tecniche e sociali* (Accademia Nazionale dei Lincei, Roma 1967), Roma (1968), pp. 111-118.

Trubetzkoy, N. S. "Grundzüge der Phonologie," in *Travaux du Cercle Linguistique de Prague,* VII (1939).

Venezky R.L."Computer Applications in Lexicography," preprint for the *International Conference of Lexicography in English,* New York, 1972.

Von Stechow, A. *Syntactic Analysis of Italian,* in A. Zampolli (1973B).

Winograd, T. *Understanding natural language,* Edinburgh, 1972.

Wisbey, R. A. (ed.) *The Computer in literary and linguistic research,* Oxford, 1971.

———. "Computers and Lexicography," in D. Hymes, ed., *The Uses of Computers in Anthropology,* La Haye, 1965.

Wood, C. R. "Dialectology by Computer," in *ICCL,* 1969.

Woods, W. A., R. M. Kaplan. *The Lunar Sciences Natural Language Information System,* BBN, Cambridge (Mass).

Zampolli, A. *Studi di statistica linguistica eseguiti con impianti IBM* (Tesi di Laurea Dattiloscrieta), Padova, 1960.

———. "Notà tecnica," in *Raccoltà Barbi di Canti Popolari Italiani Esperimento di Elaborazione elettronica F 1/RB,* Pisa, 1967, pp. II-XI.

———. "Intervento sul tema 'Il dizionario italiano di macchina,'" in *Calcolo,* v, suppl. n. 2 (1968A), pp. 109-126.

———. Recherche statistique sur la composition phonologique de la langue italienne éxecutés avec un système IBM, in *Les Machines dans la linguistique,* Praga, 1968B.

———. "L'elaboratore elettronico negli studi linguistici," *Rivista IBM* 2 (1968C).

———. Appunti per l'intervento alle Giornate di Studio sul tema 'La preparazione del Personale per la elaborazione automatica dei dati,'" in *Italia* (AICA), Roma, 1969A.

———. Due conversazioni sul panorama attuale della linguistica computazionale, Pisa, 1969B.

———. "Notà Tecnica," in A. Bartoletti Colombo, *La Costituzione della Repubblica Italiana del 1947. Testo, Concordanze, Indici.* Firenze, 1971.

———. "Cronaca. Notizie, spunti e appunti," in *Archivio Glottologico* LV (1970), fasc. 1-2, pp. 272-279.

_____. "L'automatisation de la recherche lexicologique: état actuel et tendences nouvelles," in *Actes du deuxième colloque international de linguistique et de traduction* (Montréal 4-7 octobre 1972), in *META* vol. 18, N. 1-2 (1973A), pp. 103-136.

_____. "La Linguistica Matematica e i Calcolatori," *Proceedings of the First International School and of the International Conference (Pisa, 1970)*, Firenze, 1973B.

_____. *La Section Linguistique du C.N.U.C.E.*, Firenze, 1973B.

_____. *Mathematical and Computational Linguistics II. Proceedings of the 1973 International Conference on Computational Linguistics. Pisa 1973*, Firenze 1974 (in print).

_____. *Mathematical and Computational Linguistics, Proceedings of the Second International Summer School. Pisa 1972.* Firenze 1974 (in press).

Zarri, G. P. *Algorithms, Stemmata Codicum and the Theories of Dom H. Quentin,* in A. J. Aitken et al. (eds.) 1973.

Zilletti, U. "The Works to Prepare and Draw up a Vocabulary of the Justinian Legislation," in A. Zampolli (1973B), pp. 201-200.

Recent Publications

GEORGE E. HEIDORN, *Natural Language Inputs to a Simulation Programming System* (NPS-55HD72101A). Monterey, CA: Naval Postgraduate School, 1972.

A. J. SZANSER, *Automatic Error Correction in Natural Texts - Supplement* (NPL Report COM 63). Teddington, Middlesex, UK: National Physical Laboratory, Division of Computer Science, 1973.

JURIMETRICS JOURNAL 13, ii (Winter 1972), is almost totally devoted to detailed descriptions of courses with such titles as "International Response to Science and Technology" (Harvard and MIT), "Computers and the Law" (Cornell), "Law, Logic and Computers" (Connecticut), and "Law, Cognition and the Computer" (Stanford). Published by the American Bar Association Standing Comittee on Law and Technology, this journal is published at 1155 East 60th Street, Chicago, IL 60637 ($10 for one year, $3.00 for single issue).

SYSTEM, A NEWSLETTER FOR EDUCATIONAL TECHNOLOGY AND LANGUAGE LEARNING SYSTEMS, edited by Norman F. Davies and John R. Allen. The first issue contains an editorial, a select bibliography, a calendar of conferences, and the following articles; "Language Laboratory Methods in Old English" (O. D. Macrae-Gibson); "The Use of a Desk Computer as a Language Teaching Aid" (Nevile Shrimpton); "Stand und Entwicklung des technologischer Fremdsprachunterricts" (Heinrich Schrandt); "Current Trends in the Use of Language Learning Laboratories in Sweden (l)" (Hans Jalling). Free subscriptions are available from Professor Davies at the University of Linköping, Sweden.

Recent publications of the National Language Research Institute, 3-9-14 Nisigaoka, Kita-ku, Tokyo (in Japanese): *Studies in the Vocabulary of Modern Newspapers,* vol. 4 (1973); *Basic Study of the Relation between Social Structure and Language (3)* (1973); *The Development of Syntactic Structures in Children's Speech from 3 to 6* (1973); *Studies in Computational Linguistics,* vol. 5 (1973); the latter contains an article by Shiro Hayashi, "Computer-based Linguistic Study of Literary Text" (in English).

INFORMATION RETRIEVAL, a 132-page publication, the third in a series of handbooks on managing information retrieval, provides guidance in the use of information retrieval systems. $1.25 per copy from the Superintendent of Documents, Government Printing Office (GPO Catalog No. GS4.6:IN3/2), Washington, DC 20402.

Continued on page 372.