REPORTS AND COMMUNICATIONS

A Giant Alphabetic Index to a Medium UDC Edition

In 1975 an account was given in this journal of a method for preparing alphabetic indexes of UDC editions by computer. It was written by Professor Elvia de Andrade Oliveira, unfortunately recently deceased, and described a system called ELVCDU3 for making automatic indexes (I.C. 2 (1975) 2, p. 90–93). The system was described by the aid of examples from two English UDC editions, classes 50/53 of the Full Edition (1943), and the Abridged Edition (1961). The computer had been programmed for rephrasing complicated sentences, and each final sentence was rotated so that the significant words were successively brought forward as key-words. All words were considered as significant that were not found in a list of non-significant words included in the program. In many cases it was found necessary to supplement existing texts with words from superior headings to make them intelligible when appearing alone.

Professor Oliveira mentions that her system has been used in processing two Portuguese UDC editions, Class 32, Politics, of the full edition, and a Medium Edition. The present article will be devoted to a study of the latter edition, which has since been published (1), and especially to its 'automatic' alphabetic index.

In the middle of the 1960-ies a working group within FID/CCC devoted considerable time to examining in detail a German proposal for selecting about 33 000 UDC numbers for a Medium UDC Edition. Agreement was rapidly reached, and in 1967 the German *DK-Handausgabe* and the French *Edition Moyenne* were published. They were followed by a Russian edition in 1969 and a Japanese in 1974. They have all indexes, three of them alphabetic, of course, and the Japanese one syllabic. Various parts of the medium edition have furthermore been published in several languages.

The Portuguese edition now before us consists of three volumes, all published in 1976 as the Brazilian contribution to the Dewey centenary jubilee. From a UDC point of view it might perhaps be said that it appeared at a somewhat unfortunate moment. The editors have not been able to include changes in the system approved and published later than June 1972. After that time the revision work has been unusually active, largely in view of the preparations for an entirely new medium edition to be published simultaneously in several languages, the *International Medium Edition*. The English and German versions will appear any day now, and the French will follow very soon. This edition will contain all changes, partly quite extensive, up to June 1977, and the selection of items from the full edition will be rather different. Much as the appearance of a medium edition in an Iberian language is welcomed, one must regret that it will probably rather soon be felt to be somewhat out of date.

The systematic part of the Portuguese edition follows DK-Handausgabe very closely. It even contains a revised version of the detailed introduction written by Dr. Karl Fill for the German edition. It is hardly of any real interest outside the Portuguese-speaking wolrd, but on the other hand the alphabetic index must be said to be of considerable general interest. One is of course immediately struck by its formidable size. While the systematic part is an ordinary volume of 456 pages (DK-Handausgabe has 440 p.), the alphabetic index consists of two thick volumes of altogether 1443 pages, i.e. about 3.5 times as many. Against 35 759 UDC numbers in the systematic volume there are no less than 154386 sentences in the index, or 4.3 times as many. This might be compared with the 43 000 terms in the alphabetic part of DK-Handausgabe, which, due to very space-saving printing in three columns, only require a volume of261 pages.

In view of the above, the Portuguese index can well be called enormous. It is by no means the first index made by computer. In fact, hardly any indexes to day are made without some help of a computer, but generally they have a size that one could call normal, i.e. somewhat like the index of DK-Handausgabe just mentioned. But there is one computer-made index that could well be compared with the Portuguese as far as size is concerned. In 1975 an index to the Dutch UDC edition was published. This edition is abridged with only about 16 000 UDC numbers in the systematic tables, but the index prepared by K. J. Dreese is a volume of no less than 730 pages or slightly more than half the size of the Portuguese index. The method used by Dreese is rather different from that of Oliveira. The significant words are not brought out by rotation but picked out, and they appear as key-words in the index followed by numbers in UDC order, and with all texts where a key-word appears given in full. These texts are generally easier to read than the permuted Portuguese sentences.

However, it is the Portuguese index with which are primarily concerned here, and let us study the successive permutation of a sentence. I have chosen the UDC number 547.3 with the following text according to the full English edition (1972): 'Acyclic unsaturated hydrocarbons and their monosubstituted derivatives.' The text in the Portuguese edition runs as follows: *Hidrocarbonetos acíclicos não saturados e seus compostos monovalentes.* In the index we find this sentence in permuted form after each of the five significative words (given here in the order of the words in the original sentence). (For technical reasons the one-line sentences are here carried over to a second line.)

Hidrocarbonetos 547.534, 547.535, 547.536, 549.88, 661.715
  acíclicos não saturados e seus compostos monovalentes 547.3
acíclicos
  não saturados e seus compostos monovalentes, Hidrocarbo (= 547.3
saturados
  e seus compostos monovalentes, Hidrocarbonetas acíclicos (= 547.3
compostos -033.2, 668.539
  monovalentes, Hidrocarbonetos acíclicos não saturados e s (= 547.3
monovalentes
  Hidrocarbonetos acíclicos não saturados e seus compostos 547.3

This example shows us several different things. The text after a key-word must not exceed one line. In this case, where the sentence is fairly long, the end of the text has been cut off in three cases (indicated by the sign (=), with particularly unfortunate result in the third sentence, where the omission of the word *não* completely changes the meaning (from 'unsaturated' to 'saturated'). Unfortunately we find in the original index a very disturbing misprint (corrected here). The very important word *aclclicos* is written *acídios*, which makes the sentence meaningless. Curiously enough we find after the neighbouring number 547.4 another misprint of the same word, *acldicos*, just as meaningless. Both these words appear as key-words in the index in addition to the correct word. These misprints have been carried over from the systematic tables. In connection with these sentences it should be pointed out that the diacritic signs shown here do not appear in the index, which is reproduced by offset from the computer listing. The text consists of capitals only and is in some places rahter badly reproduced.

Two of the five key-words in the above example are followed by UDC numbers. This simply indicates that under these numbers the key-word in question stands as independent word after the proper manipulation by the computer. Look at the first number after *compostos* in the tables. We find

-033.2 *Estuque, cimentos e compostos*

The corresponding English text is 'Hardsetting plasters, cements and compositions'. These three words are separated by the computer and appear as separate sentences of one word. The other number after *compostos* 668. 539, refers to rather obscure types of essential oil used as cosmetic. It would be a great mistake to believe that the UDC numbers following the key-words are specially important.

If we study the list of sentences after the key-word *compostos* we find to our surprise that their number is as high as 483. Of these texts 34 are incomplete because they are too long. This is not always serious, but there definitely are cases where the most important part of the sentence has been removed. It is certainly not easy to find your way in a list of alphabetically arranged texts of almost four full pages. The fact that the lists are arranged in two columns on lying A4 format rather increases the difficulty of handling the heavy volumes. Of course you don't have to look at the longest list. After *hidrocarbonetos* there are 67 lines and after *saturados* 86. *Monovalentes* is still better, only 17 lines. *Aciclicos* should be our best choice with its five lines, but unfortunately we won't find what we want on account of the misprint mentioned.

Even if we find a short list which can be rapidly surveyed, the texts are often difficult to read in their permuted form, and may, as mentioned above, be seriously mutilated. To find the way among very specialized UDC numbers is possible only for those who don't need to use an alphabetic index. The ordinary users definitely need more help, e.g. of the type given in the index of DK-Handausgabe as shown by the following example:

Lithium (Anorg. Chemie) 546.34
Lithium (Metallhüttenkunde) 669.884
Lithiumverbindungen (Chem. Erzeugnisse) 661.834
Lithiumverbindungen (Lagerstätten) 553.663

Long lists of permuted texts do in fact dominate the index to an alarming extent. After *máquina* and *máquinas* (machine, machines) we find lists of altogether 1174 texts, occupying 11 full pages. The closely related key-words *materia(s)* and *material, materiais* have lists with 978 texts and fill nine pages. *Sistema(s), industria(s)* and *fabricaçao* (1229 texts) fill another 12 pages. These are just a few examples.

I am convinced that key-words followed by very long lists will practically never be used. Who will look under *fabricaçao* if he wants information on wine making? He finds everything under *vinho.* Almost 50 texts, but he very rapidly finds *vinho, fabricaçao de.* Another example. Under *maquinas* there are 40 texts starting with *(maquinas) de costura.* All these texts refer to UDC numbers which have six figures in common, 687.053, 'Sewing machines and accessories'. One of these is 'sewing machines for straw hats. This latter sentence is of course also found after 'straw' and 'hats' *(palha* and *chapeus).* In this case I think the most obvious reaction is that these details are quite unnecessary. We only need a reference to the number for sewing machines. Further details are easily found in the systematic tables and need not encumber the index. Incidentally we find in the index to DK-Handausgabe a reference to sewing machines but also two subdivisions:

Nähmaschinen (Bekleidungsindustrie) 687.053
Nähmaschinenantriebe (Bekleidungsindustrie) 687.053.5
Nähmaschinennadeln (Bekleidungsindustrie) 687.053.2

This is definitively enough. On account of the permution the 40 texts of the Portuguese index form no less than 135 lines after various key-words.

The system for simplifying complicated sentences seems on the whole to have functioned quite well. However, for one type of sentence the result is not satisfactory. I will show one example. We find in the table the following sentence:

612.799 *Crescimento e fisiologia das unhas e pelas*

This means 'Growth and physiology of nails and hair'. In the index the sentence has been converted into two shorter sentences, which of course also appear in converted form:

Crescimento de unhas	Growth of nails
Fisiologia de pelos	Physiology of hair

But you won't find 'growth of hair', nor 'physiology of nails'. This type of sentence is not very common, but a rapid survey has given me a few examples, e.g.

631.248 Dung pits and hills for manure and refuse.

They all show the same defect as the first one. Another similar sentence shows the danger of relying too much on the computer:

624.014.26 Iron and steel structures, screwed and pegged.

Here we find in the index 'iron structures, screwed' and 'steel structures, pegged', while the expression 'iron and steel' is certainly used collectively in this sentence.

However, much more serious is the fact, that in a very large number of cases the texts have not been supplemented, and you are left with sentences or very often single words that are unintelligible or difficult to interprete. We have discussed above the word *compostos*, and I should like to return to the second UDC number after this key-word in the index and look at the first part of the text in the systematic table:

668.539 *Outros compostos*

This means 'other compounds'. The word 'other' can not very well serve as a key-word, and we stand there with the isolated term 'compounds'. What compounds? We get this information easily enough by looking up the table, which tells us that the text of 668.53, the superior heading, is 'principal constituents of essential oils'. There are many types, in 668.531/538 we find hydrocarbons, alcohols, aldehydes, ketones, etc. All stand isolated as key-words, with no reference to the fact that they are constituents of essential oils when the UDC number begins with 668.53. *Compostos* is more difficult, we must look at the remainder of the text after 668.539. After the text above stands a colon followed by the following words: *combinaçoes de nitrogenio e enxofre.* The full text in English is as follows:

668.539 Other compounds: nitrogen and sulphur compounds.

The words after the colon have found their way to the index, unfortunately in completely mutilated form. We find the following texts:

Combinaçoes de nitrogenio com enxofre
Nitrogenio com enxofre, combinaçoes de
Enxofre, combinaçoes de nitrogenio com

This can only mean compounds of nitrogen *with* sulphur, which is something quite different from the original text. You might get away from the isolated *compostos* by formulating two texts somewhat as follows:

Nitrogen compounds as principal constituents of essential oils
Sulphur compounds as principal constituents of essential oils

I will add a few examples of isolated key-words from other parts of the UDC system:

Escolas	355.714	*Velocidade*	531.112	*Legendas*	78.089.1
Prisões	355.715	*Aceleração*	531.113	*Metro*	781.62

We have to the left schools and prisons in military establichments, in the middle velocity and acceleration in cinematics, and to the right text and metric in music.

In an article a few years ago (DK Mitt. 16 (1972) 2, p. 7–10) I criticised some indexes to existing UDC editions. As the article was written in German, I took most examples from German editions, among them DK-Handausgabe. This criticism dealt with details, and need not concern us here. Important is that these indexes were all prepared with the users' interest in view. I think we can safely say that users of the indexes of (say) DK-Handausgabe, various abridged editions, and the numerous parts of the full English edition published in recent years will reasonably well lead to the right place in the corresponding systematic table. That must be the necessary aim of an alphabetic index. I am sorry to have to say that in my opinion the Portuguese index is not up to the mark when seen from the point of view of the user. Many details criticised here can of course be improved, but unfortunately the permution of a very considerable number of sentences has given us a far too bulky index with many far too general key-words followed by practically unsurveyable lists of texts.

Even if I have found it necessary to criticise the alphabetic index rather severely, I am full of admiration for the progranuning work behind the index. The fact that the full Portuguese edition can now be transferred to tape and that printouts can be made of any type of edition is of course very valuable. I don't see why it shouldn't be possible to utilize the existing material also to prepare an alphabetic index that is really serviceable. Not with permuted texts and not entirely automatic. It requires considerable brainwork to make the proper choice of key-words.

A final word also about the Dutch index. The key-words are not obtained automatically in that system but are picked out, which means a good deal of brainwork at a stage when the Portuguese edition is supposed to flow entirely automatically. The texts are not processed but kept in their original form and are thus necessarily often very long. They are sometimes repeated many times as every text must stand after each key-word it contains. I have found one text of six lines in the Dutch index that occurs no less than 15 times. In my view neither of these two indexes means the last word in computer-made indexes. But a study of both may produce useful ideas with regard to the formulation of texts in the systematic tables. They are undoubtedly often far too clumsy at the present time. On the other hand one can hardly accept Dreese's argument for the very general keywords, namely that they may be useful for an analytic study of the UDC. An alphabetic index is not the right place for such studies if it means introduction of key-words that make the use of the index more difficult. The users' interests must come first.

Einar Öhman

(1) Classificação Decimal Universal. Edição media em Língua Portuguesa. Vol. 1 Tabela, XXVII + 456 p.; Vol. 2/3 Indice, IV + 1443 p. (FID 541). Rio de Janeiro 1976.

The Society of Indexers International Conference 1978

Over 100 indexers from 16 countries gathered for the first international conference, July 14–16, 1978, at Digby Stuart College, London. Tribute was paid to the President, G. Norman Knight, whose book – 'Indexing, the Art of' – was announced for publication in the near future by the publisher George Allen & Unwin. The winner of the Library Association Wheatley Medal for an outstanding index was named as the Index to Archaeologia Cambrensis, 1901–1960. This was compiled by Mr. T. Rowland Powel and the award was accepted, in his unavoidable absence, by Miss Dorothy Moore.

In an international gathering such as this, it was appropriate that one paper, delivered by Dr. Hans H. Wellisch of the University of Maryland, should be on *Multilingual and Multiscript Indexes.* Even more appropriate was that the series of texts chosen – Herbals – should include the first published index (in the late 15th) and also the first published use of a "see reference" in Fuchs' Herbal of 1543. It is interesting that in those days such importance was given to the index that there is an example where the index is prominently named on the front page even before the copyright notice, and several examples of the index preceeding the text. In some of these cases the index proved to be so heavily used that it was totally destroyed and only the text remains now.

Also on the subject of foreign language indexes Dr. James D. Pearson, Professor of Bibliography at the

School of Oriental and African Studies, gave a valuable paper on *Languages of Asia, with Special Reference to the Islamic World.* The indexing of these languages, even confining the problem to that of personal names, is very complex. The Burmese have no family names. Thai people are addressed by their first name and the Cairo telephone directory is also arranged in this way. The 15 official Hindu languages in India are rendered a greater problem by haphazard word division, with inconsistencies in the names when written in English. According to the Guinness Book of Records the shortest name is U (Uncle or Mr.) E.

Publishing has, of economic necessity, become an industry. Even University presses find it ever more difficult to produce what was once thought of as a quality book. So said Mr. Archie Turnbull, Secretary of Edinburgh University Press, in a paper entitled *The Economics of Academic Publishing.* An index could now only be included if it would add to the sale of the book in question. The publisher is not there for the public good, but for obtaining the material of which books are made and then selling them. He must know what the public need and find emergent disciplines. Thus, uncertainty and indecision are basic ingredients of a publisher's state of mind.

Indexes are the antithesis of lateral thinking, according to Dr. Edward de Bono in his paper *Lateral Thinking and Indexing.* This is because an index directs to a particular location for a particular piece of information – lateral thinking implies that free thought and association should be applied. A further interesting point made is that it is easier to learn backwards, for example when the solution is known before the problem is analysed.

Mr. Michael Marland, Head of Woodberry Down Comprehensive School, London, in his paper, *Reading to Learn and Using the Index*, stressed that the distinction between 'Work' books and 'Reading' books, should not be made. Greater emphasis must be made on reading at all stages of education as a process of continual development. Pupils should be taught to read textbooks as well as works of literature and this includes the teaching of how to use an index. Education, as is indexing, is the reorganisation of knowledge.

Dr. William S. Heckscher presented his paper on *Unconventional Indexes* in a highly unconventional manner, basing his approach on an analysis of the woodcuts and texts of Albrecht Dürer. His thesis was that the modern tendency towards rigid standardisation may produce efficient indexes (it very often doesn't) but it destroys their power to illuminate the works to which they refer.

The Conference was a very happy occasion which realised a new interest in indexing, particularly in representatives from overseas. It was generally felt that the atmosphere is now ripe for preparations to begin for another, even bigger, conference on a world wide scale.

The proceedings and most of the papers will be published in the October issue of the Indexer and the remainder in the April 1979 issue.

Enquiries to: Mr. J. Ainsworth Gordon, 28, Johns Avenue, London NW4 4EN. (Press Release)

The Committee on Conceptual and Terminological Analysis (COCTA)

The Committee on Conceptual and Terminological Analysis (COCTA) is a Research Committee of the International Political Science Association (IPSA) and the International Sociological Association (ISA), and is also a Standing Committee of the International Social Science Council (ISSC). It promotes the clarification of key concepts used in all the social sciences and seeks to accelerate the processes whereby terminological agreements can be reached that will facilitate the easy identification of these concepts and the formulation, communication and retrieval of scientific statements in which they are used. To achieve these goals COCTA prepares and distributes relevant guidelines and a supporting glossary of metaconcepts, serving as a link between UNESCO's INTERCONCEPT project and interested working groups of social scientists.

1. *Working through* IPSA *and* ISA, COCTA sponsors panels, roundtables and workshops at which theoretical and concept papers are presented.

a. *Theoretical papers* deal with the general problems and approaches to conceptual and terminological analysis, methodologies, metaconcepts and strategic questions. One outcome of these activities has been the publication of THE TOWER OF BABEL: ON THE DEFINITION AND ANALYSIS OF CONCEPTS IN THE SOCIAL SCIENCES, by G. Sartori, F. Riggs, and H. Teune (Pittsburgh, Pa., 1975). Currently COCTA is engaged in the analysis and clarification of "metaconcepts" required for the study of concepts and terms used in the social sciences. A preliminary glossary of such metaconcepts is contained in the *Guidelines for Concept Analysis*, by G. Sartori, presented at the ISA Congress, Uppsala, Sweden, August 1978. COCTA studies on metaconcepts involve the utilization and reconciliation of basic perspectives, methods, and concepts drawn from the fields of logic, philosophy, classification research, information science and the language sciences (including psycholoinguistics, sociolinguistics, language planning, lexicology and terminology).

b. *A pilot project* designed to illustrate and clarify the methodology involved in carrying out concept analyses led to the preparation of the *Guidelines* (mentioned above) and the presentation at various conferences, over a period of years, of papers on selected terms. A number of such papers, presented at the ISA Congress in Uppsala will be combined in a symposium volume edited by G. Sartori.

2. *Working through the* ISSC, COCTA offers the results of its methodological studies and pilot projects as a basis for the establishment of conceptual and terminological projects by research committees, not only of ISA and IPSA, but also of any interested member associations of the ISSC. In order to offer concrete help to interested groups of scholars in any subject field – and representing any theoretical or philosophical orientation– COCTA is preparing, under an ISSC contract with UNESCO, a set of guidelines that will eventually be published for general utilization. A workshop on this subject has been held at Skokloster, Sweden, on August 12–13, 1978.

3. COCTA *cooperates with* UNESCO's INTERCON-

CEPT *program*, which is designed to provide, in cooperation with the social science community, a computerized terminology bank that can be used by working groups of social scientists, as well as others interested in the retrieval and study of social science materials. An advisory committee, chaired by F. Riggs, has been established by UNESCO to provide overall guidance for this new program. Its first meeting was held in Hattenheim, Federal Republic of Germany, from May 3 to June 1, 1978. The fundamental goals and methods of INTERCONCEPT were laid down by a Meeting of Experts held at UNESCO house in Paris, May 1977. The Guidelines to be prepared by COCTA/ISSC are explicitly designed to facilitate the effective utilization by scholarly working groups of INTERCONCEPT facilities.

4. *Cooperation with other agencies.* From its inception, COCTA has cooperated with various associations, universities, and research centers, on the basis of mutual interests. Panel sessions have been organized at conferences of the International Studies Association and the American Political Science Association. An experimental terminology bank was started at the University of Pittsburgh's Information Utilization Laboratory, with the support of C. Beck, under the editorship of G. Graham. Joint panel sessions have been held with the research committees on Political Sociology (IPSA and ISA) and Sociolinguistics (ISA), and the Ad Hoc Group on Alienation (ISA). The U.S. Political Science Information System (Pittsburgh) has agreed to add definitions to the abstracts of articles covered by its service, using for this purpose definitions of key terms included in the articles indexed.

5. *Organization.* COCTA is an unfunded voluntary effort. Membership is open to any interested scholar. There are now about 100 members in some 23 countries.

An executive committee gives intellectual leadership. Its members are: G. Graham, A. Judge, J.-E. Lane, P. Ludz, J. Pool, F. Riggs (secretary); G. Sartori (chairman), G. Sjoblom, and H. Teune.

News of COCTA activities is carried in the newsletters of IPSA and ISA. The CISS Bulletin (now Comparative Research) published by the Comparative Interdisciplinary Studies Section (CISS) of the International Studies Association also regularly includes news of COCTA activites. From time to time the secretary distributes memoranda to members on current plans and activities, and the ISSC also gives information about COCTA. A series of working papers – now terminated – was also distributed to COCTA members through the good offices of the International Studies Association. Anyone interested in learning more about COCTA or becoming a member should communicate with: Professor F. W. Riggs, Political Science Department, University of Hawaii, Honolulu, Hawaii, 96822, USA.

F. W. Riggs

7th International Conference on Computational Linguistics (COLING 78)

The "International Conference on Computational Linguistics", a permanent institution (COLING), takes place every two years. The meetings, sponsored and organized by an independent scientific committee under the chairmanship of Professor Vauquois, Grenoble, have the purpose of bringing together, at regular intervals, people working in the field of computational linguistics – rerearchers, university teachers and practicians.

This year's conference, held at the University of Bergen, Norway, from August 14–19, 1978 was attended by more than 250 participants from 25 countries. The wide range of topics dealt with all current aspects of linguistic analysis by computer and offered a comprehensive review of the present state of the art and of new trends. A rough breakdown of papers, panel discussions and demonstrations (over 80 in all) shows that the emphasis was on the following subject fields: artificial intelligence and text theory, automatic translation including computer-aided translation, automatic syntactic and morphological analysis.

Artificial Intelligence

The field of artificial intelligence (AI) figures prominently among research efforts, as proven by the fact that as many as 20 papers were devoted to the linguistic aspects of this subject matter. Whereas ten years ago, sentences like "Mary seeks a unicorn. Mary seeks him" were first analyzed morphologically and syntactically and later semantically, this is now done by projecting linguistic utterances onto models of intensional logic. The analysis of vague natural language elements (fuzzy semantics) is carried out by mapping the elements onto the "world knowledge" of a certain domain under certain linguistic constraints.

Two AI projects in which the computer has to "understand" natural language should be mentioned. One of them, developed at the University of Michigan, analyzes short stories and finds the point with linguistic means (not keywords). The means are stylistic and grammatical elements, such as repetition, negative statement, unusual items or expressions (pointing to unusual events), time frames, events and the key event, direct and indirect speech. The main characteristics of the story are evaluated, weighted and scored and the point (happening) is inferred by the computer in a concise manner.

The second project is concerned with the automated understanding and summarizing of newspaper articles. FRUMP (Fast Reading Understanding and Memory Program) of the Yale University AI Group consists of a stored Parser and Understander which analyze articles and extract information with the aid of a "sketchy script" (data structure skeleton fot typical events or situations). The computer picks the matching script from a variety of scripts and produces a summary of a few lines. Required CPU time: approximately 4 seconds on average. About 200 sketchy scripts have been prepared so far for the most important daily newspaper events, such as holdups, raids, highjackings, plane crashes, natural disasters, etc. For example, the computer must be able to analyze and extract the following facts from a lengthy report on an earthquake: location, strenghth on the Richter scale, duration, number of people killed, number of people injured, amount of damage, measures taken. Newspaper agencies and PR departments of industrial firms have shown interest in this project. What makes the task difficult is the fact that individual readers consider different things to be important and that events are seen by the reader against a certain background knowledge which gives the information its proper perspective and varies from reader to reader (elections in the USA are seen differently from those in a South American country, as are hostages taken in the Netherlands (by the Moluccans) or during a "regular" bank holdup, etc.) The analysis, structuring and storing of this "world knowledge" in such a way as to enable the computer to make inferences from it, is one of the main difficulties. Research concentrates on finding new approaches to this problem, other than the methods used by automatic abstracting.

Question-Answering Systems

Question-answering systems, a subfield of artificial intelligence, aim at building models of spoken language (dialogues) in a given situation and testing them on the computer. Participants in such a dialogue are, e.g., a master craftsman and an apprentice who is ignorant of the task (performance of a certain activity, such as repairing an appliance) and the instruments (tools). By posing questions (using linguistic tools) the aim is to find a common store of knowledge which enables the apprentice to find the proper tool among a variety of "misleading" tools and to perform the work intended. Experiments of this kind serve to build closed world models (data bases) that are static, have a one-way flow of information, assume perfect knowledge, and penalize uncertainty and error. The ultimate aim is to arrive at models that are open.

Machine Translation

An interesting hybrid form between automatic text understanding and machine translation is another AI project which is based on the assumption that a text can be translated only if it is fully understood. This is also true of the computer. SAM (Script Applier Mechanism, Yale University) "translates" in two phases:

Phase 1 consists of understanding the source-language text and constructing a language-free representation for it;

Phase 2 comprises forming a language-free summary or paraphrase from the memory representation and expressing it in the target language.

In other words, SAM transfers information from the source language to the target language rather than translating linguistic elements. This project proves the feasibility of storing and inferring "world knowledge" to a limited, but nontrivial extent.

Machine translation systems which try to integrate this aspect of artificial intelligence are referred to as knowledge-based MT. At the moment, however, they have the character of theoretical models such as SALAT (System for Automatic Language Analysis and Translation), a special research project of the University of Heidelberg.

One of the numerous – more practically oriented – projects of machine translation, presented at the Congress, is TAUM-Aviation (Traduction Automatique de l'Université dé Montréal) which seems to be of special interest due to its source language, which is English, and its goal: translation of technical manuals for a certain type of aircraft. Now that the predecessor, TAUM-Météo, has been in operation for some time, translating, with considerable success, weather bulletins (which are more or less stereotype texts), the University of Montréal has been assigned the task of developing a system for the automatic translation of the documentation for the Aurora CP-140 aircraft. The required special vocabulary (some 50,000 dictionary entries) is being supplied by the translators of the Bureau for Translation of the Department of the Secretary of State, Ottawa. TAUM-Aviation contains a pre-editing phase, in which homogeneous text passages are marked (e.g. maintenance instructions in the imperative). Syntactic-semantic analysis is particularly difficult with technical texts in English since it has to cope with: ellipses, delimitation of string compounds, disambiguation of identical noun, verb and adjective forms, etc. Example: "Cap all open lines and cover pump shaft" or "In case of leakage reservoir bellows should be discarded". This English/French system calls for a thorough post-editing phase and is to be operational by the end of 1979.

The FR-II (French/Russian) machine translation system presented by Professor O. Kulagina of the Institute for Applied Mathematics of the Academy of Science of the USSR, follows a familiar approach: morphological analysis, dictionary lookup, syntactic analysis, transfer, synthesis. The syntactic analysis uses so-called linguistic filters which exclude certain grammatical patterns and have to be passed through again and again until the matching sentence diagram (tree) has been found. FR-II translates sentences comprising a maximum of 45 units (= words or groups of words). The speaker gave the following information as to the effectiveness of the USSR system: 84 % of the sentences are translated correctly or adequately, although it is difficult to define correctness or adequacy.

Statistical evaluations of the translations have produced the following results:

	good	average	poor
comprehensible	63.6 %	24.9 %	11.5 %
grammatically correct	53.5 %	32.7 %	13.8 %
adequate	66.7 %	22.5 %	10.8 %

It could not be ascertained whether or not FR-II belongs to the third generation of MT which is to be extended by the new dimension of "world knowledge" as understood by artificial intelligence. Up to a few years ago, this dimension used to be referred to as the "pragmatic level" of translation.

The advent of the microprocessor and its extension to almost every sector of everyday life have caused industry and the universities in the USA to consider the feasibility of processing natural language on microcomputers. One such project is "Micro-Based English" (University of North Carolina) which is a program package consisting of three parts: McSNOBOL (an adaptation of SNOBOL to microcomputers) LEXASH (LEXical hASHer) which provides access to parts of a large associative lexicon in the master computer, and PARSYN (PARse SYNtactic categories) for syntactic analyses. Micro-Based English could prove valuable for linguistic research on a smaller scale and should enable the researcher to use the microcomputer as an intelligent terminal from his desk.

As far as continuous speech recognition is concerned, it was reported that 220 standard-type sentences, spoken by three speakers (selected on the basis of their pronunciation), were recognized with a 93 % accuracy (Sperry Univac). Research efforts of this kind aim at replacing keyboard input by the spoken word.

Computer-Aided Translation and Terminology

The panel discussions covered the following topics: computer linguistics as seen through its own eyes, programming languages for linguistic phenomena and computer-aided translation (CAT) which, in turn, would not be feasible without intensive terminology efforts. The following is a summary of the opinions voiced on machine-aided translation:

In the years to come the need for high-quality translations will continue to rise, rather than diminish. One reason for this is: the number of publications is increasing by 12.5 % annually, i.e. it is doubling every eight years. Translation is becoming more and more complex and is recognized, even by theoretical linguistics, as an independent and important discipline of applied linguistics. At present, computer-aided translation offers the best chances of successful and effective utilization of existing capabilities, particularly if combined with text processing systems and possibly with partial machine translation. The principle applied by practicians (e.g. those working for Siemens, the European Communities and the Bundessprachenamt) is as follows: What is needed is not just intelligent machines, but intelligent and well-trained users (linguists) who can correct every system-induced error via on-line terminals. Access to data bases is also very important. The conflation of existing dictionaries, glossaries, terminologies etc. to form large universal data banks, as suggested by some participants, does not seem to be practicable since the masses of terminological data could hardly be harmonized by terminologists. It is preferable to have a small, consistent data base in a well-defined subject field rather than a huge collection of ad-hoc terminologies from a variety of sources in random disciplines.

Standardization of terminology on a global scale is also a task too large to handle, although this demand – made by the USA – is quite justified. Great efforts in this direction have been undertaken by ISO and international professional associations (IEC, FID, CCITT, ECMA etc.) for some years, but their work is a mere drop in the ocean, owing to the rapid increase in technical vocabulary in all languages.

Furthermore, it should be borne in mind that it is the technician and engineer – experts in the field of technology – who have undertaken the essential task of national and international standardization of terms. The linguist (or terminologist) is a newcomer to this field. In our view, it was therefore inappropriate to discuss questions of standardization at a conference for computational and theoretical linguistics.

Among the profusion of impressions and optimistic opinions on AI research, there was one statement which – though somewhat exaggerated – was not contradicted by those present. Professor Karlgren from Stockholm said that, for all its undisputed success, computational linguistics has not yet reached the level of the linguistic ability required of a school beginner.

H. J. Schuck, E. Hoffmann
(Bundessprachenamt)

IN MEMORIAM

Rasmus Mölgaard Hansen

Our colleague Rasmus Mölgaard Hansen, a well-known Danish scientist and scholar in the field of classification research, passed away on 18th July 1978 at the age of 74.

For some 25 years of his life RMH played an important part in international cooperation on developments in the field of documentation in general and classification in particular.

During this period RMH impressed himself on his many colleagues all over the world as a brilliant expert. To those privileged to collaborate with him he became a good and highly-valued acquaintance, to many a one even a close friend.

RMH rendered a significant contribution to the organization of international teamwork and showed himself able to transform into reality what he had projected.

From 1952 to 1962, as secretary of the International Building Classification Committee (IBCC), he was the editor of six important reports and a co-editor of the abridged UDC tables for the subject field "Building", the so-called ABC tables, published in twelve languages and unique in their field.

From 1962 to 1974, as chairman of FID/CR, he edited the important series of reports published by that committee, which during that period comprised 12 publications in all.

A momentous achievement was his organization and management of the Second International Study Conference held at Elsinore in September 1964. The pertinent proceedings, appearing under the title "Classification Research" as a 563-page volume edited by Pauline Atherton, were published in 1965 and are still regarded as classic.

All the publications mentioned bear the unmistakable stamp of RMH, who had a genius for expressing and presenting ideas and experiences in a most intelligible way.

These publications are monuments to his contribution to science. Their contents are still of considerable information value and will continue to serve as source material.

These monuments do not bear RMH's name, however. As the humble servant par excellence of his branch of knowledge, he was far too modest a personality to claim credit for them.

RMH was an august humanist, too.

Recorded on tape should have been the golden moments when he, in so brilliant a fashion, let his audience share in his philosophical views of life.

They made a deep impression on us and left an unforgettable memory of a wise man.

All honor to his name.

Dan Fink