Session 7: THE DICTIONARY

QUESTIONS AND DISCUSSION

G. BROWN: I suggest, if anyone has a question that he would like to address to some specific member of the group, that he simply so specify; otherwise I will just let the panel discuss. Go ahead, gentlemen.

KING: I do not like these figures of 500, 000 words in a *32,* 000 memory and think this ought to be clarified. I think I know what Dr. Lamb means, but people might get the wrong impression about this.

LAMB: I definitely said the vocabulary coverage is 500,000 words, but of course you do not put 500, 000 words in the dictionary, because what you put in are the lexes.

KING: Yes, but it sounds like a great many words, which it is not-- I mean 500, 000 words are not very many words, at least in Russian where any one stem can have from 10 to 100 different endings put on it.

LAMB: The way I make this estimate is to use the same figure which Dr. Oettinger uses--you take the number of morphemically different stems and multiply by 10. So, then we have 50, 000 stems.

KING: That is more like it. Do you think 50, 000 stems are going to get you anywhere in translation?

LAMB: I think indeed it will.

KING: I disagree with you. If you want to stick to organic chemistry, fine; but if you want to do arbitrary text, I do not think you are right.

LAMB: Callaham's technical and chemical dictionary has only about 30, 000 entries, by a rough count which I made. It is quite adequate. In fact, it contains a lot of words which do not exist in Russian. It is a rather bad dictionary.

KING: I disagree with you, because there are a great many different

fields with a tremendous number of words. Look at Webster; it has 2 or 3 million words in it--which are stems.

LEHMAN: I wonder if I could make a pertinent comment on the basis of our own investigations. One of our group, Stanley Werbow, investigated some text on the basis of de Vries's German scientific dictionary, which is very widely used. He found that 40% of the compounds were not in the dictionary. This seems to be a dead end.

I would like to bring up another problem. I think we are agreed that for certain languages segmentation is important; for example, Dr. Reifler used the word Grundgedanke. There are a number of words like this in German which can be made up at will, and you cannot find these in any dictionary. These correspond to combinations of words. In a sense, English phrases often correspond to German full words. Now, suppose that we want to do segmentation and that we begin segmenting English, which has moderately many suffixes. Suppose we take the suffix "ous". I have jotted down "efficacious, tremendous, mucous, enormous, populous, ingenious, ingenuous", and you can cite any number of these. Are we going to list in our dictionary something like "ingen" from which we might then form "ingenuity", and what are we going to do with "ingenious"? We have "ingeni", but can not form "ingeniety". We have "populous", but we can not form "populiety". I think that there are a number of problems on either side of this segmentation issue.

MERSEL: I think the problem here is not so much whether it is best in general to segment or not; I think that the decisions have been pretty much based upon the economics of the individual machines and the purpose of the investigators. Mrs. Rhodes does not feel that for her purposes it is even worth the trouble of creating much of a dictionary; and for those groups whose memory device does not have a large computing capacity coupled to it, it has been much more beneficial to list all the forms. For those of us who have had a comparatively small memory and a lot of computer, it has been much easier to segment. Dr. Lamb is going to a stem glossary; Dr. King is going to a full-form glossary.

KING: That is not true.

MERSEL: It is not a full form? Is it stems?
KING: We can do it either way.
MERSEL: Which way are you doing it?
KING: Both.

MERSEL: Fine. I like this because that same compromise is the one we have settled on for the configuration of the computer we are using. Ours happens to be a form dictionary, but it is not a full-form dictionary nor do we rely completely upon a stematic method.

LAMB: Let me say that we are also putting in some full forms. Wherever we get the type of difficulty that Professor Lehman was just talking about, we would put in the full form.
KING: What do you do about words like "hotdogs"?
LAMB: I think I would put that in as a full form.

KING: Well, you now are not going to have 20, 000 entries. There are hundreds of thousands of these word pairs.

LAMB: If you wish, I will calculate the number of such entries.

KING: But I am talking about it from the practical standpoint. When you really start looking at text, you see many of these new forms that are not truly idiomatic.
LAMB: Well, "hotdog" is not a new form.
KING: It is not idiomatic either.
LAMB: What does idiom have to do with it?

KING: I am saying that there are a great many word pairs that can really use up your memory space in a big hurry if you say that you have only 20, 000 entries.

LAMB: Right now we have, as far as I know, the largest dictionary in existence in the MT field, and it has only 15, 000 entries in it. If we get those additional 5, 000, it turns out that we do not have

enough, then I will come to you.

KING: Who is running this contest of who has the biggest dictionary?

LAMB: I should add that most of our entries are fragmentary, because--as I pointed out yesterday--if you do not have the information, then you should not make guesses. So, many of our 15, 000 entries actually have blanks in the space where a meaning should be provided.

KING: Yes, but on what authority can you say that you have the biggest dictionary?

LAMB: I have a coverage right now of an estimated 300, 000 different graphemic words, and that is a higher figure than I have heard from anybody else.

KING: I think we have a lot more than that.

LAMB: The figure that you gave *in* your paper the other day was about 200,000.

KING: I said 25,000 stems.

LAMB: Stems, yes, but my 20, 000 items are not stems but bases.

KING: When I say stem, I mean the kind of thing you are talking about.

REIFLER: I want to say that the fact that we are using not only full forms but also constituents of full forms is not by any means a compromise. We enter all those full forms which we know. Among these recorded forms there are some which are still productive and which may be constituents of extemporized components. Thus we have procedures of dissecting compounds that have no memory equivalent. Naturally, we have to dissect these and then identify their constituents and translate constituents-wise. The wonderful part about this whole thing is that, as far as extemporized compounds are concerned, the constituents are in most cases not saddled with multiple meanings. If they are saddled with multiple meanings, their number is limited or they can be dealt with in such a way that the machine can determine which meaning is intended when the constituent occurs together with a

certain type of another one.

MERSEL: I think I am going to be accused of changing the subject. We have been discussing different ways of looking up the memory. I contend that to some extent this is a function of which machine you have. I would like to raise the question as to what should be in the dictionary. It is quite obvious that the Russian word or its stem should be in there and that the English should be in there. I would like to hear some of the other members talk about grammar codes and semantic indications. Dr. King, what do you keep in your dictionary besides the Russian and English?

KING: First of all, we would not like to keep English. We might have some dictionary definition and be sophisticated; but since we do not understand how to do syntax, we do not know what to put in our dictionary.

ZIEHE: The RAND dictionary has, besides the Russian and the English, the grammatic descriptions which describe the form morphologically and also has syntactic information. I might add that there is also grammatic information for the English as well as for the Russian.

LEHMAN: We have not compiled any dictionary. I think the dictionary that you have is going to be comparable in many ways to our grammatical rules and glossary, so that some of the grammatical codings which you will have in your dictionary will be purely a set of rules for us.

MERSEL: Our grammatical entries in the dictionary give us the flags for which rules to use. One of the things that bothers me most in our grammar coding is the possibility of misleading ourselves both now and in the future. This became quite obvious to us after going to a bit representation as to whether something falls into a particular class or not. Then it suddenly became obvious that what we really needed was not a binary notation but a ternary notation, because we found that we were using the "1" to indicate "Yes, it is so" and the "0" to indicate "No, it is not so". We had a need for another symbol

that would indicate "I do not know". I think that this feature of "I do not know", and later on in calculation "I do not care", is one of the things the dictionary needs most.

REIFLER: I should like to say that we consider the automatic translation product based on our dictionary not as final translation but as the input for another processing, and possibly afterwards for more processing steps in which the output would be further improved. Some of that work has already been done; our dictionary contains information which would enable a logical device to carry out such processing. For example, in the case of technical terms belonging to one field of science the English alternatives for one particular Russian word have subscript numbers indicative of the field of science in which each particular alternative would be the proper translation. A logical device could, with the help of this information, make the right choice.

JOSSELSON: I think the nature of a bilingual dictionary depends also on the structure of the item that you are dealing with--it depends on the class. I do not see any particular reason for segmenting uninflected items in a language like Russian. It is much more economical to look them up by any other means. On the other hand, when it comes to finding equivalents for a Russian preposition, you have a completely different matter. A bilingual dictionary simply does not state that an item in language X is equivalent to an item in language Y or to so many items in language Y. What you also need is the precise indication of the conditions under which you will use those equivalents. The most important thing is how to organize the dictionary so as to contain as much information as possible, in order to make the matter of semantic or logical operation easier later on. That means that in addition to containing all the information which you need about the form itself, the memory also must have notations of information about what can occur with a form and what cannot occur with a form. You may want to add some other considerations, and 1 do say that before you finally finish your dictionary, you will have to have a section for after-thoughts. It is not a hindsight pool but simply a provision for things that you do not know about. We do not know everything about language yet.

GIULIANO: Many people have advertised their dictionaries as being available. I would like to advertise a very small dictionary that I just compiled here for Russian-English. It has 33 symbols in it, the letters of the alphabet; and I think it will translate any text. Of course, we would need some algorithms to put these letters together. At the other extreme, we could go to the Library of Congress and translate everything and store it. There we would have it, and I think we could adjourn this meeting and forget about syntax; we would handle it by a table. The way in which a compromise is to be reached is relative to our knowledge of linguistics. As we learn more and arrive at better ways, we will have better dictionaries. The natural units with which to start were words. It just seems to work out that way; at least, we are using them. They make good lexemes, but we do not have enough storage. You store pieces of them and then you must provide techniques to put them back together into words.

LAMB: There is something widely misunderstood about what we do. Once we have located these things, we do not put them back together to form words. These are what serve as the bases of translation. It is only after we get to the English that we put the words back together.

LEHMAN: I think it might be pertinent to say that the word is a very deceptive unit. Just what is a word? If you wish, you can define a word on the basis of graphemic units, but then you will find that a word has different definitions on the basis of the dictionary you use.

REIFLER: I should like to say that it is quite possible that in the future a purely structural linguistic analysis and purely linguistic approach may find all the cues necessary to create procedures and logical programs to enable an automatic system to supply perfect translations. If that can be done economically there will be nobody, on whatever side he may now be working, who will be against it. On the other hand, it is quite possible that a purely linguistic approach may not be able to solve a large number of what I usually call multiple nongrammatical meaning problems and that for these particular problems a purely lexicographical solution may be used. Suppose we are going to use in the future purely lexicographical means to solve subgrammatical and nongrammatical meaning problems which logical procedures either cannot do or cannot do as economically. The

question which I myself cannot answer today, but which perhaps some among you might already have some idea about, is this: Will it be necessary, for the solution of other linguistic problems in the remaining text, to reconsider those problems which are being solved lexicographically, which the machine can translate only by lookup, and which therefore no longer present a syntactic or lexical problem? Will it be necessary for the analysis of the remaining parts of the clause or sentence to reconsider these again in the analysis? If one is worried about English word order with Russian as a source language, I think the answer is yes; but if the source language is Chinese, I think one would not have to worry too much. As far as the cost per word of the translation is concerned, I would like Dr. King to tell us what he feels about it.

KING: How good a translation do you want? If you want something readable, the cost is infinite right now in anybody' s scheme.

MARCHAND: The question was raised here by Professor Reifler as to how far we could get with linguistics. He suggested that, since we pretty soon run into our boundaries with structural linguistics, we should work with the lexicon. However, the lexicography is a part of linguistics. What is linguistics is really a problem sometimes. In other words, what is or is not linguistics has not very much to do with it. The answer lies in doing machine translation in whichever way you can do it. If you use a linguistic technique and are not a linguist, or if you use a computer technique and are not a computerman, it makes no difference.

A.F.R.BROWN: A question was asked by a discussant about what should go into the dictionary. One of the glories of the simulated linguistic computer is that you can also put little pieces of program into the dictionary. It is inelegant but lovely, and the program can be anything from 2 words to 500 words. Mr. Ziehe, I feel, is a kindred spirit; he is walking like me, while some people are flying, like Dr. King and Dr. Lamb. How long it takes to read the dictionary once depends on how long the dictionary is: 4, 000 words can be looked up in the time it takes to read the dictionary, and the information can be left in the top half of the core memory rather than being pumped out on a tape to be read back in afterward. Suppose

it takes one minute to read the dictionary. How big a batch can be looked up in one minute of time? Judging by your figures, it spills you over the dictionary reading time. The question is: For 30, 000 words looked up, how many dictionary read-times does the whole process take?

ZIEHE: The figuring we have done has shown that the internal processing that goes on while the dictionary is being read is simply the computing of the random addresses and the maching of dictionary forms with text forms. This we expect to run slightly over tape read-time--probably by a factor of, say, half again as much. So, the buffering effect you get in the IBM 709 permits the two to go on simultaneously.

QUESTION: Do you mean 100, 000 entries in the dictionary, or the entries that take care of 100, 000 forms?

ZIEHE: The number I used was 200,000 dictionary forms, one entry per form.

QUESTION: Mr. Ziehe, what is random about your random occurrence ?

ZIEHE: You can use any formula you like. It is just a means of computing, from the representation of the form, its address within a certain range in the computer memory. This address can be computed any time the form is encountered in text and again when the form is encountered in the dictionary, so that you get to the same location in this region of the memory.

RHODES: We talk very glibly about a billion, but we do not really understand what a billion is. If a child were born at the same time as the founder of Christianity, and if he were given a dollar for every minute that he lived, and if he lived until now, he would still not have a billion dollars. Multiply this billion by 100, 000, and you will get the number of cells within the human brain. These gentlemen are trying to make a dictionary; they are trying to do what the human brain can do, together with our sense, together with all the rest of the experience that we have accumulated. They think they can translate by computing the numbers of stems; but whether it is 50,000 or

500, 000 stems does not matter. What does matter is what are you going to put with that stem? How are you going to put into the dictionary the entire world of knowledge that our God-given brain has accumulated.

If you heard me speak the first time, I said that if I only knew how our brain works, I would not have any trouble with semantics. The trouble is not the coding. The trouble is that I do not know how our wonderful brain works, and the reason I do not know is that we cannot comprehend 10^{14} cells just in our brain, to say nothing about our senses. I feel that in the future we will have to have something very revolutionary, something very magnificent, something very wonderful in the form of our external memory, which will be on the order of 10^{12} bits, not 10^{7} bits.

Dr. King asked, "How would you do it in 50, 000 stems?" We cannot. What we will have to have is different kinds of memories, maybe like a photoscope. One will be labeled "astronomy", another will be labeled "history". Every time we deal with a history book we will hook on the history memory. We cannot possibly have them all at the same time. We could easily make up the right ones, besides a general one. The general dictionary need not be more than 50, 000 stems, I believe. With 50, 000 stems in the general dictionary, plus all these special dictionaries, perhaps in the future--maybe 10 years from now--we will be able to give you something that we will not be ashamed of. Do not let anybody tell you that he can give you anything better today.

REIFLER: Mrs. Rhodes, I want to say that what you have outlined has been why we attempted to collect the general language vocabulary that is current in scientific publications. We did not concentrate on any particular field of science, but thought in terms of many photoscopic discs which would have the technical vocabularies necessary for translation in any specialized field.

LEHMAN: I would like to say that these numbers we have just heard may be right, but I would like to point out that machine translation does not involve the mastery of information which the human brain can master. It is simply a transfer from one code to another, and

consequently we should not need the capacity that is necessary for the human brain to master the sum of human knowledge.

HAYES: I do not want to have the comment of Dr. A. F. R. Brown go without being repeated: namely, that you can store the program to handle a dictionary entry as a part of the contents of the entry, and therefore the handling of some of the deeper linguistic problems can very probably be delayed because of the ability to handle a special situation without the need to encompass it in the general program.

G. BROWN: I can recall sometime around 1945 when one of our colleagues, very well-celebrated, started turning out tables with a high-speed calculating machine. People said, "What on earth do we need tables for?" We are now going to compute things as we need them. It was a very interesting thing. It still is true. There are places for tables, and there are places where you will compute as you go; and that is similar in spirit to this problem, only a little bit simpler actually.

In 1945 Professor von Neumann thought that 4, 000 words of high-speed storage would make him happy forever. He could not get that in 1945. Some of the early machines limped along with 256 words of high-speed storage, and people broke their backs to make things fit. Now, when they have 32, 000 words, they are breaking their backs to make them fit. The next time around we will have a few times 100, 000 words of high-speed storage, and we will still be breaking our backs to make them fit. It seems to me the lesson here is very simple. You have heard about a number of different techniques. They are not at odds with one another at all. The job is hard enough so that before we are through we are going to need all of them and we are going to need all of the progress that is ahead of us in the computing field. I want to close with the thought that the mix with which these things are used will turn out to depend in each case on the nature of the application. Fortunately, we are wealthy. We are beginning to see that we have different resources that have tremendous power relative to where we stood a few years ago. The job is still large, and we are going to need it all.