DISCUSSION — Panel III, Saturday morning

OSWALD: May I ask, Mr. Belmore, what the speed of the key punch operation is?

BELMORE: This is done gradually. It is similar to a typewriter. The material to be translated could be punched days before actually putting it on the machine. All the punched material for one Russian book or journal could be put on the machine and then run in a few hours.

OSWALD: Have I called the correct time on how long it's going to take to do the key punch, whoever does it? 60-75 w.p.m.?

BELMORE: It would depend upon the speed of the typist. It is strictly a manual operation.

GARVIN: The key punch will not necessarily remain part of this operation. It could be replaced by one of these reading instruments when perfected.

BELMORE: Exactly. If we use a reading instrument of the type mentioned yesterday, ... when perfected.

MUELLER: I have a question to ask Mr. Lehmann. I would like you to elaborate a little on these categories. There are limiting adjectives and descriptive adjectives based on analyzing their formal behavior. We should list them somewhere else. Under II-B and II-C you have "sein Dunkles" and "das recht Dunkle". You have there the first item, the "sein" and "das" listed as limiting adjectives, but they behave quite differently. Why don't you categorize them differently?

LEHMANN: Well, the term "limitation" is one that would include both the DE and the KE. Let us take that as a cover term.

MUELLER: But then you have to subdivide them. Somehow or other you have to indicate that one is a limiter that conditions a secondary meaning and the other is a limiter that does not.

LEHMANN: This may be a bit too free, but that is what was done up under I-B. Then you go over to *a,* then you go over to *alpha* and *beta.* That same subdivision is presumed. You see, *a* is limiting, then *alpha* and *beta* are the DE and the KE words, and that would

be assumed to carry all the way through the chart. The chart would become rather long, of course.

MUELLER: I see now. It wasn't completely clear to me from your listing here.

BUCKLAND (USAF): About this question of print reading and the idea that we shouldn't worry about it now and that we won't need it until we have a translator. I'm not sure I go along with this, but I think that you linguists sooner or later are going to have to use machines to look through a lot of running text and I don't feel that looking through a million words of running text is out of the question. However, if you exclude the factor of speed, key punching 1,000,000 words of text costs about $200,000. I was wondering if there wouldn't be some possibility of using monotype or some of these foreign typesetting machines to provide this source of data. I take it that monotype is something similar to linotype machines except that they produce this type from punched tape, and then it would be simply a matter of converting this into code, into a form suitable for use with our own machine. Does anyone know anything about this?

THOMAS: I believe McGraw-Hill make paper punch tapes of all the materials they print. This is for their own use. It would be very nice if everybody did this. It would solve some problems.

KIRSCH (Bureau of Standards): I think it is important to emphasize here that in going to the trouble of preparing a lot of data in mechanical form, — in other words, in typing a lot of data, — it costs no more to type it in such a form that it is ultimately usable by a computer— that is, typed and produced on punched paper tape copy at the same time, — and I think that the concern over what form this data should be in in punched paper tape form, needn't worry you for the time being, because when you finally have something ready that you can feed into a machine, this large corpus on punched tape will be usable by most machines, or certainly convertible by mechanical procedures to be used by large machines. Consequently, I think that Mr. Buckland's point that it is worthwhile to prepare this data in such a form that it is feedable into some machine eventually is a particularly important one, — one that will permit considerable saving.

PAPER: I would like to bring up a small terminological point with respect to Mr. Zarechnak's paper. I wonder if a term something like "morphographemic" wouldn't be better than "morphophonemic", since

happily at this stage we are not interested in spoken input, but rather simply printed input. Of course, the morphographemic shape reflects a morphophonemic difference.

ZARECHNAK: As you know, I have used morphographemic in all my papers, but only in one place one has to be acquainted with the morphophonemic processes in Russian to qualify and classify properly morphographemic symbols. And there you have to be sure you are doing it by morphophonemic processes before you classify your morphographemic signals. I agree with your statement.

PAPER: There is one general question I have. I wonder if the people who have been involved in MT research have come up with a specific definition of the term "idiom" for their purposes?

DOSTERT: I think I can formulate one which may be acceptable. An idiom is a combination of separate words, each of which loses its individual signification in the context. For example, in "right away", in which "right" and "away" no longer mean "right" as "right" and "away" as "away", then each of the components has lost its original signification to acquire a new one when combined. Now, the number of items that can be so combined is variable, obviously. When in French we say "tout à l'heure" or "tout de suite", we don't think "tout", "à", "l'", and "heure", we think "tout à 1'heure" as a single sign signifying a single concept. Does Mr. Joos agree with that?

JOOS: Yes. That is a perfectly reasonable way to start the discussion. One wonders just how one can draw the line between the point when we are going to know that we have finished finding all the idioms. I mean to say, how are you going to draw the line between idiom and construction?

GARVIN: I will mount my positivist horse and say that whenever I can translate an item in a context individually, it's not an idiom. If the translation of each individual piece gives us gibberish, it is an idiom. And then you will find out how many idioms you have when you have covered a large enough corpus.

JOOS: You could put it this way then: An idiom would be whatever your machine is not prepared to handle by its primary routines.

DOSTERT: We will have to store idioms as lexical units. In other words, you will have to take the complex which we call an idiom and

store it as a single unit, giving it its 'idiomatic' equivalent in the largest language and handling the whole as a single lexical manipulation rather than an analytical one. I think that has been fairly well agreed.

JOOS: I think it's a perfectly sensible engineering approach. If you want to give time to a linguistic discussion, you could find out an opposite possibility. You could say that it is possible to define an idiom within one language at a time; only then you won't find so very many. For English I have found very few. For example: "Get your own breakfast", which doesn't mean "get your *own* breakfast", but get it without help.

DOSTERT: In a case like that, I think you would have to cue the word "own" when it operates in the manner you have indicated.

JOOS: The way it operates in this particular idiom is in contradiction to the general rules of English grammar. Here is another contradiction: "I can't seem to find it". According to the *rules,* that *ought* to *mean,* "I can't give the appearance of finding it". We have here a syntactic-semantic shift resulting in another thing that I would call an idiom without comparison with another language. However, once you do make that kind of survey and find these idioms in English like the two I have just mentioned, you will find that you have only a tiny fraction of what needs to be taken care of as idiom in a translation program.

YNGVE: I would like to say that the papers by Mr. Zarechnak and Miss Pyne come very close to the approach that we have been using for German. What they are apparently trying to do for Russian is very nearly the same thing that we're trying to do for German. Naturally, I agree wholeheartedly with this.

DOSTERT: In this connection, I would like to address a question to Mr. Belmore. He has put down as Step No. 1 on Page 1 of his handout the linguistic analysis of the source and target languages. Now, that is a very broad statement indeed. I have struggled with the idea that it may not be necessary to go through a complete analysis of the source language as such and of the target language as such, on the theory that there are some parts of the analysis that are not pertinent to the translation process. It seems to me that there is an area of analysis which you might call your "transfer area" between your S and T languages. The question is, before you can focus on the transfer, must you do an exhaustive analysis of S and then an exhaustive

analysis of T, and then look for the transfer pattern? That seems to be the trend with your group. Mr. Yngve, we would welcome comments on this, because it seems to be rather basic to the whole problem of MT.

JOOS: I wouldn't mind commenting on that. I think that if you do it in the MIT fashion, analyzing the two languages in question exhaustively first, you will find that for your MT purposes most of your results are trivial. I think that if you seriously want to do this as an engineering proposition, the empiric approach of Georgetown looks more reasonable to me, and should not be very disappointing to a linguist either, because by this empirical approach he will now get the complexities of each language *in order of their importance,* from at least one viewpoint. It may be that if he does this with a number of languages he will get them in order of their importance generally, which would be a very nice thing for a linguist to know.

YNGVE: I wouldn't want to be pushed all the way to the very wall and say we are going to do an exhaustive—completely exhaustive—grammar of English and German first, but I will say this much, we certainly ought to have a more exhaustive analysis of German and English than we have at present before we go and look at the specific coding problems.

DOSTERT: Isn't it also not only a matter of the extent of the analysis, but also the type of analysis which is specifically suited for the transfer process? It seems to me that the formulation of the results of your analysis should constantly bear in mind that you are faced with the problem of meaning transfer between one set of signs and the other, and you are interested in the internal structure of each set of signs only to the extent that they relate to the job of transfer.

YNGVE: Yes, but even this we don't have yet.

DOSTERT: You mean that we don't have enough now to start looking at the transfer?

YNGVE: No, I'm talking about the structure of English or German or Russian as of now.

DOSTERT: Well, we can do it with segments, and out of the segment analysis and the cumulative results we can arrive at broader formulations.

PYNE: I would like to ask Mr. Joos if in referring to the empirical approach at Georgetown he referred to an approach to the languages in terms of their relationships with each other. That is, English as viewed only in terms of its difference or similarity to Russian?

JOOS: When I said empirical I meant to use the word in the sense that Mr. Dostert introduced it in yesterday. Namely, you work out a routine for translating the first sentence that you randomly encounter from one language into another, then you work out your routine for the second sentence and usually you're lucky enough to be able to use some of what you used in the first, and so on, thus gradually accumulating notations, procedures, etc. That is what I meant by the empiric approach. I think that's what Mr. Dostert meant.

DOSTERT: That is one of the techniques that one of our co-workers has developed, working on French. He took a corpus of chemical text in French and worked on it sentence by sentence. In other words, his idea was eventually to look at the forest by looking at enough trees, one after the other.

JOOS: I think I'd like to add another comment here on the other side. It may well be that the result of this kind of empiric approach, if it ever does result in a successful machine translation, will be a set of routines so complex that they will cry aloud for simplification by symbolic logic methods, and they may have gotten it so complicated that it would take much labor to simplify them.

OSWALD: I just wanted to go on record as agreeing with the statement you made, Mr. Dostert. There are three kinds of analysis involved, really. There is an analysis of English, let's say, as a source language; an analysis of, let's say, German as a target language; and the third body of information concerns entirely the relation of these two languages and constitutes a new realm of discourse, in which the formulations can be quite different from those normally used to analyze the languages as such. This I take it is what we are all agreed upon as meant by the empiric approach. It does constitute a new sort of formulation, a new series of statements, in a quite different realm from the normal type of linguistic analysis.

DOSTERT: I don't think, Mr. Oswald, it means exactly this, and I wouldn't try to put the tag "empirical approach" on the process I tried to describe and with which you agree. That I would call the analysis of the transfer pattern, rather than the empirical. Now let me reassure

Mr. Joos. Our group is working with three different techniques, the empirical, that is the one you described; the other is Mr. Garvin's method where he tries to find through an analytical process the formal cues within a segment and tries to develop a theory from an analysis of a given number of sentences. It is not as empirical as the one of Mr. Brown. It is more a combination of the analytical and empirical, with a focus on formal cues. Then the third is an attempt to formulate a rather broad theory—that is what Mr. Zarechnak, Mr. Pacak and Miss Pyne are doing—and then to actually test it on a text. A fourth group has followed what we have called the internal coding system, or internal coding plus bilingual matching of code to arrive at translation. One word of clarification is needed. As evidenced by the paper given by Mr. Belmore, we are trying as we formulate our linguistic steps, or our steps in terms of linguistic analysis, to bring the discipline of symbolic analysis to their rendition so that we will not run the danger of being confused to a point beyond retrieval. The reformulation would become a very complicated task.

GARVIN: I just wanted to elaborate on this last point. I think that in the procedure of developing rules and getting a large body of rules there will have to be intermittent periods of stock-taking and compression. What I mean is that, for instance, at the present time we find that in our glossary there are recurrent sets of entries which have each the same ten diacritics. In a case of that sort, provided we have a large enough body to do this with a degree of statistical and other validity, you can then replace this set of ten diacritics with a single one thus to reduce the problem of size of special entries. Another thing I visualize as one of the future operations in compressing the corollary codes is that equivalent steps in various rules can be spelled out more simply and by a simple reference, rather than to have to start out each time with the instructions. In other words, Rule 22, Instruction A, B, and C, and then Rule 23, Instruction C, B, and A, and so on down the line. This is another form of compression. The third thing is that we are already beginning to see that by itemizing one rule per situation we have a tremendous amount of redundancy because it turns out, for instance, that out of our 100 rules, a set of rules covering suffix translation, there are so far about a half a dozen of them which have exactly the same steps and exactly the same key entries. Now if this is borne out by a larger body of data, you could then take these six rules and lump them into one, and use the same set of diacritics for six times in these situations instead of six different sets of diacritics, and so on down the line. I think that if this is not

done as you go along, then confusion will result. I agree that some awareness of this is present.

LEHMANN: It seems to me that Mr. Thomas' paper illustrated very well that we will need a better analysis of English such as Mr. Yngve suggested before. If, instead of using the analysis you did, you would begin with a linguistic analysis based on immediate constituents, I think you would get far fewer types of English, and I think you indicated that in your talk as well.

ZARECHNAK: We are approaching here the same problem on a different level. Still we have one common denominator, namely, we do not forget the data which we are facing. We believe that if the same data are approached from different points of view, we would get information which we would mutually exploit later on. One comment more: When we are explaining or taking some attitude, we never can forget transfer from source to target; this transfer is actually information on two codes, and specifically what is common in between. Here is the place where we so appreciate cooperation with programmers.

DOSTERT: Yes, the fact that after all we are all working with the same data means that there will have to be a measure of concordance in the results, even though in the initial phase there seems to be considerable divergence. The fact that the data with which we are working are finite and systematic is another encouraging factor. Another thing I think that should be pointed out is that languages of relatively similar cultures and of the same family present much less difficulty than languages of different families and of widely divergent cultures. Since at the beginning we are focussing our efforts on languages in relatively similar cultures and relatively similar systems, we are trying to tackle the more modest problems before the big ones.

I had asked Mr. Crossland yesterday if he would not be willing to take the floor this morning to explain to us the method and objectives of the Cambridge Language Research Unit. Mr. Oswald took Mrs. Masterman somewhat to task yesterday for her formulation. I understand that a further exchange of views between Mr. Oswald and Mr. Crossland rectified what may have been originally an erroneous impression. I would not want the group to be uninformed about the Cambridge Unit's efforts, and since Mr. Crossland is with us, I would like to ask him if he would come up here and tell us about some of the things they are doing.

CROSSLAND: I don't think I wish to give a long statement and explanation about what we're doing, because I think that was well put forward at the MIT conference in October. I did feel, though, that Mr. Oswald had considerable uneasiness about Mrs. Braithwaite's close reliance on a thesaurus, and for the sake of informing this meeting, I thought I would mention our method in a broader sense. That is, we are prepared to deal with specialized vocabularies, idioglossaries if you wish to call them so, though perhaps we think that is a little too precise a term, to explain that we have to cut up the vocabulary into technical vocabularies, but we are certainly prepared to play with the idea of specialized vocabularies, although we now think the thesaurus method is the most promising for solving this problem of language and exact semantic correspondence between words in different languages. Another point I wanted to make is that our experience has not been that the amount of general vocabulary of ambiguous terms is very wide, from random experience on each side. We are at the moment working with restricted language. We are trying out our technique on a certain amount of botanical literature in Italian. So in a sense we are doing perhaps what you suggested in an early stage, although relying on the thesaurus method, or giving it a try for solving problems of this sort for semantic correspondence.

OSWALD: It may be that Cambridge linguists working with Italian and botany have had a different experience from myself and Lawson working with German and brain surgery, because of the temperamental differences between the Italians and the Germans on the one hand, and also because of the fact that botany, I should think, probably lends itself to a more metaphoric way of expression than brain surgery. As Mr. Dostert has suggested, Mr. Crossland and I have had no difficulty in ironing out our differences. I would like to add that I spoke to Mr. King yesterday and asked him whether in his super-gigantic millenial computer he could dispose of idioglossaries—whether this meant that the whole principle of idioglossaries could not be discarded as obsolete and belonging to the primitive stages of our investigation—and he said "no". He felt quite definitely that no matter how large his storage system would be, the idioglossaries would have to be in there in some form, and suggested only that the question of where they should go—he put it quite bluntly—we should leave that to the hardware people and just produce them and not worry about where they would go in a given system.

DOSTERT: The fact that they should be produced is not controverted at all.

AUSTIN: I'd like to ask Mr. Crossland if the thesaurus method he has in mind is anything like that outlined by Mr. Parker-Rhodes in a recent paper?

CROSSLAND: I would give a qualified yes to that.

AUSTIN: Well, there were two categories in this. As I remember there were around fifteen or twenty that he listed, and one of these categories was "objects", and another was "things to think about". Can you tell me any possible use for those categories?

CROSSLAND: I think it's a good thing to try out, that's all.

AUSTIN: Wouldn't this about cover the vocabulary of any language?

CROSSLAND: That could be taken in a broad sense, I suppose.

KING: As I understood that paper of Parker-Rhodes, this was only a trial idea more or less along the lines of "Twenty Questions", and how you organize a thesaurus. Well, we have Roget's, but that isn't necessarily directed toward mechanical translation.

GARVIN: I'd like to change the subject back to idioglossaries and their use in translation. I think that a compilation of idioglossaries is at least a necessary preparatory step. They will either have to be left each separate or lumped together in one big glossary depending on the particular storage method and what-have-you, that a given scheme contemplates. But I do think that this is merely a preparatory step because you still have the problem of choice in a text, namely, whether to go to a general glossary or to an idioglossary, and the problem of ambiguities within a technical terminology. The point is that even a chemist in a chemical article occasionally departs and uses a word from some other field or uses a chemical term in a non-chemical sense.

OSWALD: I was very careful to say yesterday that the totality of my experience with idioglossaries, which is not very great, nevertheless indicated an equivalence only up to about 80%. I haven't the faintest idea how you solve the other 20% of ambiguities, of unpredictable words, and brain surgeons suddenly using unpredictable adjectives. This clearly is something that has to be gone into and it is precisely in this area of choice, as Mr. Garvin calls it, that that 20% falls.

GARVIN: Well, I think there is another problem here. That is, whether or not the choices were necessary, because one of the things that came out in the very, very small piece that we have done so far, is that apparently you don't have to solve every single problem, in order to get a viable translation. And from the linguistic point of view, this relevance criterion which Joos brought up I think is tremendously significant to the whole problem. In the "transcendentalist" Prague School tradition, it used to be called functional load, and now "functional yield", according to Martinet. And I think that this is one way of getting at it in a purely numerical, pragmatic and precise way. At the same time, of course, we can make the decision on the following basis: if I don't bother to translate this properly, what will happen to the rest of the text? Will it be unintelligible, or will it be intelligible enough and elegant enough even though I have left out this property? For 23½% of the instances you would have to have a very extensive routine to resolve the ambiguity between ether and esther. After that Father Sohon came to me and said, "Well, this is really not very important anyhow, because as a chemist I have the impression that the two terms are roughly equivalent and that it is just a stylistic difference". Now if this is the case, then I will leave my 23½% sit there and put out a translation ETHER/ESTER or ETHER/ESTERS and not worry about it and say to myself that the labor I have saved might better be used for resolving an ambiguity without the resolution of which the machine is going to blow a fuse, and so on down the line. So that you can get down to these engineering criteria if you want, or you can use some other criterion of efficiency, and it's just a continuous process of decision-making in detail. Those little details each require a very specific decision. Now, for instance, you have eight possible translations of the suffix "oi". Now at this point we would theoretically include all eight of them and devise a routine for it, and it's extremely extensive and very complicated. I assume that after we have gone through 65 or 100 or 200 sentences, if we find that out of those eight translations, the one requiring the most complicated routine occurs once in a text of I-don't-know-what magnitude, then we might just drop this altogether, and say when we get to that particular point we will simply take translation A. Or we might say we have already developed a machine that reads fast, and it costs only three bits to leave this in there, so we leave it in. But this sort of continuous bread-and butter decision-making, I think, is what has to be kept in mind in order to solve your problem of choice.