

Meeting on Chinese MT

Massachusetts Institute of Technology

October 17, 1964

TABLE OF CONTENTS

Presentations and Handouts

Massachusetts Institute of Technology

V. H. Yngve	6
T. R. Hofmann	11
Tables for Manual Conversion between Systems of Mandarin Transliteration (5 pages)	
B. K. T'sou	18
Sample Sentences Generated by Computer (2 pages)	

University of Texas

L. W. Tosh	30
Report Memorandum RM 64-DA 0110, Linguistics Research Center (with assorted output) (11 pages)	

ITEK Corporation

D. Markus	39
Five Steps to Chinese Mechanical Translation (with Glossary and Symbology) (4 pages)	

IBM Corporation

D. Lieberman	60
(first part)	
F. Wong	64
S. S. Soo	73
D. Lieberman	80
(second part)	

TABLE OF CONTENTS (continued)

Bunker-Ramo Corporation

P. Garvin	88
J. Mersel	97
F. Peng	100
A Syntactic Recognition Routine of Standard Chinese (10 pages)	

Ohio State University

B. Wang	105
Project on Linguistic Analysis (3 pages)	
L. Meyers	123
Chinese Grammars and the Computer at the Ohio State University (4 pages)	
A. Yue Hashimoto	124
Revised Rules, July 1964 (32 pages)	
A Condensed Account of Syntactic Analysis of Mandarin (24 pages)	

Yale University

S. Lamb	130
Handout (5 pages)	

University of California, Berkeley

D. Johnson	148
C-Y. Dougherty	154
Chinese Syntactic Rules for Machine Translation, with S. E. Martin (61 pages)	
The Lexeme de as a Syntactic Marker (15 pages)	

National Science Foundation

R. See	164
--------	-----

The Chinese MT Meeting convened at the Massachusetts Institute of Technology. Cambridge, Massachusetts, on Saturday, October 17, 1964. at 9:20 A.M., Chairman Dick See presiding.

Present:

Berkeley: Doug Johnson, Ching-Yi Dougherty.

Ohio State: Bill Wang, Leroy Meyers,
Ann Yue Hashimoto, Itiroo Sakai (visitor).

Texas: Dr. DeCamp, Wayne Tosh.

Yale: Sydney Lamb, Samuel E. Martin.

ITEK: Dick Marcus, J. Wong, Theresa Lee.

Bunker-Ramo: Jules Mersel, Paul Garvin,
Fred Peng.

IBM: Dave Lieberman, Fred Wong, S.S.Soo.

MIT: Vic Yngve, Ben T'sou, Ron Hofmann,
Frank Liu, Elizabeth Landers (arrangements).

VIC YNGVE: I would really like to welcome you people to MIT. It is a pleasure to have you here. I think we are going to have a very interesting conference, a very profitable one for all. It just occurred to me as I was driving in this morning feeling kind of sleepy that many of you come from further west and maybe we should have started at eleven or twelve and ran to eight or nine this evening. The next time we will do that. I would like to apologize to the people from the West Coast for the early hour.

We are going to have a stenotype recording of the remarks and for this purpose it would be good if you don't all speak at once. We will have these dittoed and sent out to everybody. We are also going to try to photograph the board and interleaf the material.

Some of you may be bringing prepared talks, so there will be no sense to take down stenotypically, so if you will give some indication when you start whether this has already been committed to paper or not. If not, we will take it.

As you know already we will have lunch brought in and this evening we will break up probably about six and at seven those of you who are staying on, I hope it is most of you, are invited to come with us to the Union Oyster House for dinner and if you have wives or husbands in town they are very welcome to come along. If you have any questions about arrangements or anything like that ask as the day goes on. The Union Oyster House is in Boston. You can go by cab. There are a few people who have cars who will be going so there will be room for passengers.

Dick See will be the Chairman today and it is his meeting.

CHAIRMAN SEE: First, I am happy to be the Chairman but it is really both our meeting and it is actually a meeting of the entire group that was assembled at Indiana.

I think we will combine a happy meeting of informality and strict scheduling in order to make sure that every group is able to contribute and at the end of the day nobody has important things left unsaid, and yet during the discussion have a free exchange of ideas.

I think to start off it would be best for our stenotypist if each gave his name and affiliation in a loud clear tone, so that the stenotypist and everybody else will know who you are.

(The conference attendees identified themselves.)

CHAIRMAN SEE: I am going to bring Vic in on all of this because the procedure we adopted was to assume that each group would have up to an hour and fifty minutes or so would be reserved for the straight presentation, I think it best if it is uninterrupted because otherwise we would not get through the day. That will leave ten minutes for questions directed to the speaker and not for discussion. At the end of the day we will have a general discussion period. If any group doesn't

have enough material please don't feel obliged to use up the time. We can convert it into mixed discussion at the end.

I think I ought to say a few words about the history of MT because if I don't say a few words about it in the transcript you might not know and so for the benefit of anybody who reads this document I had better read a few words into the record.

We all know in the history we have a little difficulty in coordinated activity because of the diversity of approaches, the diversity of vocations, and because of the present state of knowledge of linguistics when we started to study the problems of translation. There were many different points of view and there tended to be some isolation among the groups. I think a lot of this is history rather than the present. One of the purposes of being here today is to insure that this is history and in the future will communicate if possible.

I think at this meeting we owe a great deal of thanks to Vic because he has made it possible to organize this meeting without the writing of a single letter, without the transmission of any funds or any other difficulties at all. On this informal basis I think we can have

effective communication. I was asked about press releases. No one to my knowledge had informed the press and I think we had better leave it this way. This will insure that we will communicate among ourselves as best we can. There is the possibility that a volunteer, Ohio State has a possible volunteer, will edit the transcript or produce some sort of article on the transcript. The principal purpose was to have a transcript available for each of us.

I won't say any more because we are already a little late. The schedule we have worked out is tentative. If anybody has any objection please let me know. MIT will lead off, followed by Texas, followed by ITEK, followed by IBM, and this will be either before or after lunch. It is hard for me to say. It depends on how hungry we get how soon. After lunch, or after IBM at any case Bunker-Ramo will follow, followed by Ohio State, followed by Berkeley.

Does the Yale contingent wish to make a separate presentation?

SYDNEY LAMB: We are combining with Berkeley,

CHAIRMAN SEE: I thought so.

Without further ado then let me first say, does

anyone have any reason why this schedule, this sequence, is unsatisfactory? There is no particular methodology here and this is what we choose. Is this satisfactory to everyone? Well, then let's have MIT lead off.

VIC YNGVE: Well, many of you are familiar with the work that we have done. I would like just very briefly to survey the general work at MIT and, the general way of thinking that has grown up at MIT and then let two of the other people in the group have their say. We have a fluctuating number of people in the group. We probably have twenty now. Many of these are part time. Many are students, those who are interested in Chinese, Ben T'sou whom you will hear later and Ron Hofmann whom you will see later, and in addition the two members of the Committee sitting in the rear, Frank Liu and Elizabeth Landers.

However, we have only started very, very recently in Chinese. I would say it has only been in the last eight or nine months that we have done anything at all in Chinese. When Dave Lieberman was with us we also had an interest in Chinese and you will hear from him later.

We very early came to a realization that mechanical translation would not be possible, probably

not possible unless we found out a lot more about language, about meaning, about translation, about communication process, and we decided that the appropriate function of the university group would be to try to engage in basic research, to try to build a foundation on which other people could build a system. That isn't to say that we aren't interested in building a system, we may do that, but we feel it is premature, at least for us, to be working on a system with an end in view of actually using it in the next few years. So our approach, although keeping such a system in mind as an alternate goal, we have felt less involved in trying to get something working that other groups have.

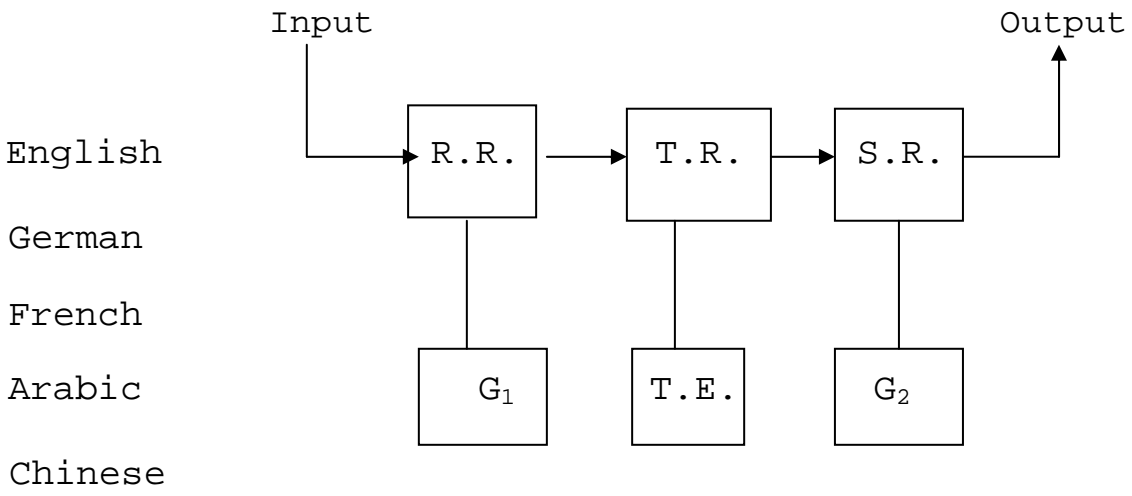
We started with work in syntax a long time ago because we realised that a dictionary would certainly not be adequate for translation and we spent a good deal of time trying to obtain syntactic description, grammatical description of the language of interest.

We spend a good deal of time also on the problem of what is a syntactic description, what we want to do, how can one obtain such a thing. A number of years ago we felt, or I did, that a translating routine, a syntactic translating routine, could be built let's say

in six blocks like this. (At blackboard)* Where you would have an input line and several processing routines and outputs. These would be programs and we would have stored information down here which would be the grammatical information, linguistics, and so on. So we would have a grammar for language 1 over here, grammar for language 2 over here. This would be a recognition routine. This would be a synthesis routine. In here we would have to have some routine that would transfer the syntactic structure you find in the input language in general what you find in the output language. So we put in here a transfer routine. That has to operate by means of some stored knowledge in the computer which would be here. This would be the, let's say, table of equivalencies between the two languages. We have pursued this. We have written grammars which can be plugged in here and here. We have written synthetic routines and recognition routines. In fact, we have a routine now that will take the grammar in a standard format and in a compiler will compile these together and these together so two separate programs, one which would recognize and one which would generate. This is one of our tools.

Now we have produced grammars not complete by

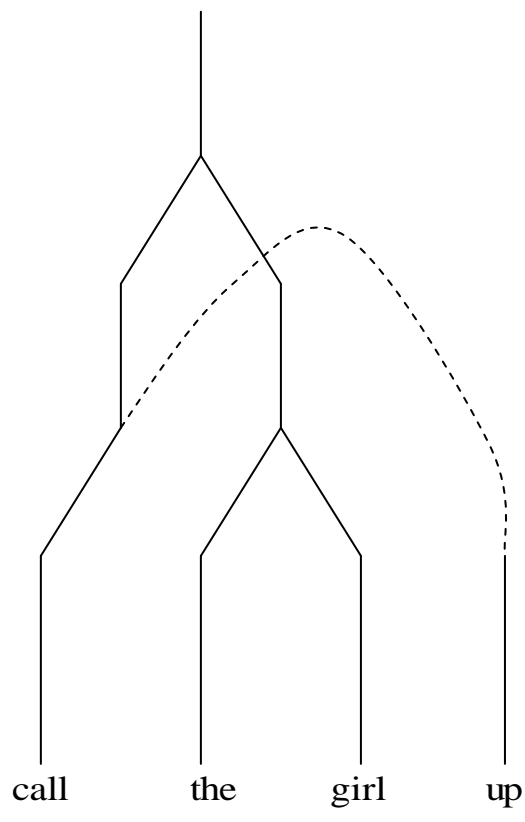
* See p., 8A



any means but we have grammars available in English, German, French, Arabic, and a start in Chinese. These grammars are by and large written in a modified phrase structure format where you are allowed rewrites, expansions, more rewrites, more expansions, discontinuous expansions. So we have in essence three kinds of rules, rewrite rules, which rewrite single symbols and have to do with categories and subcategories, constructions which are not necessarily limited to binary constructions although we have, as a matter of fact, very rarely resorted to higher orders. We do not think there is anything magical about 2. It is that we have found it rather convenient to use binary 1. The discontinuances* or constructions jump over just one node at this next level. This of course may be expanded later into something like that but it jumps over one node at this level. The sentences are produced in a left to right basis in a way that is somewhat familiar to you.

Now we have recognized for a long time that such a scheme could give a translating routine which would be not adequate but possibly could give interesting results. Now where does it lack? It lacks in the area of semantics because it is completely syntactic. Now we have toyed

* See p. 9A



with the idea of putting different semantic classifications into the grammar and toyed with various other ideas. We have been very interested in the various work on semantics done elsewhere. We do not have a proposed answer to the problems in semantics but have been working very hard, have been for years, on the basic problems in semantics. We have people like Eleanor Charney, Jared Darlington, Carol Bosche, Jack Dolan, who are essentially working on the foundation of a semantic theory, certain phases of this. We do not at this time have any way of integrating this work into a picture like this. This is something we don't know how to do yet. The work that we are doing in semantics are several isolated things and we don't feel yet that we have an overall picture.

So instead of worrying about how you would actually do something that you don't know to do, in other words we have resisted building a complete translation system with a big dictionary and a big grammar and then try to run it, we have resisted doing this because it involves trying to make decisions as to what to do when you really don't know what to do, how to resolve translating problems when you cannot resolve on a syntactic basis. We are quite happy if we do build such trial

translating routines. We are quite happy if they do the sort of things they are supposed to do, the syntactic sort of things, but we are not unhappy about the syntactic part. We would rather spend our time working on what we feel is the fundamental approach to the problem of semantics.

I think this gives you a sort of general flavor of the work we are doing. Now our work on Chinese, as I mentioned before, is just beginning and I will cut this short and let, first, Ron Hofmann say a little bit of something he has been doing in the last couple of weeks, and then ask Ben T'sou to tell you about it.

RON HOFMANN: As Vic said we have just started on Chinese and partly as an aid to this conference I made up this transliteration table. It is admittedly incomplete and there is about a page and a half of footnotes that didn't get done in time. However, I think it is a fairly good attempt in capsuling. As it is incomplete I would like any suggestions you may give me during the course of the conference.

As the title said, it is designed for manual conversion between systems of Mandarin transliteration. There are two extremities. One is to do an algorithm where a machine is best, where you can take relative time

to do many steps. One can be quite exact in an algorithm. Mr. Lee at Ohio State I believe did exactly this. However, for the use of the human I feel if the algorithm is more than four or five steps long it becomes cumbersome to do and lo and behold it is hard to translate the characters in the transliteration from one system to another.

There is another extreme, that of taking every possibility in one transcription and giving the equivalent in another transcription. This I feel is not useful for the human being sitting at the conference, say, and somebody writes a word or a sentence in one Romanization and this human being has to run through these six or eight tables looking up the Romanization trying to find a certain spelling and seeing what it is in his own system.

Thus was the motivation for these tables partly as the communication device.

Essentially why I think this system is superior for the use in communicating one person to another is that it is a mixture of the two extremes. One is two short tables and, two, there is a very short algorithm. The algorithm is easily memorable.

Take the Romanization in one system. Take the first string of consonants, call that the initials.

Everything else is the finish. Then one takes up the initial and looks it up in the table and translates to his own system. One takes the final and goes to the final table and translates to his own system.

BILL WANG: That should be ONO.

RON HOFMANN: Right, it should be SHONG. I think my example in the paper is wrong too.

There are various others that didn't get into the text. First, if one has the initial J, for instance, in the Yale system there is a J in the palatal system. One has to look to see if there are two J's in the Yale column. If it is followed by a vowel it is indicated in the first column. Then to translate in the Wade system and Russian. In the retroflex and dental sibilant I have given the finals when there is no vowel final. That is, take the first example there, 4-1, and this turns out to be SHI.

CHAIRMAN SEE: That would be CHI in Russian and SHI in the system.

RON HOFMANN: In Russian?

CHAIRMAN SEE: Isn't the 4 1 supposed to be -- what is that? Okay, I didn't see it.

RON HOFMANN: Excuse me. It is 4, SHI. In

Yale it comes out JR.

The glides are truly part of the initial table. The glides are at the bottom of the initial table. They are put there to merely indicate if you find one of these as your initials you made a mistake. You go to the final table and look there. The final table is organized in generally the same manner but by and large there are two columns for each transliteration system. The first column is the normal spelling and the second is the spelling if it has no initial and is different from the final if it has an initial.

Of course, the national Romanization has four tones of spelling and thus we have four columns and occasionally the final without an initial is spelled differently in which case it is right underneath what it ought to be.

This is not at present in the realm of being synthetically accurate. It was pointed out to me yesterday that one could not synthesize Wade-Giles from this perfectly. This I did not feel for the purposes that I mentioned is important criticism and it is a device mainly for communication so that if this fellow over here uses only Wade-Giles and if somebody writes something

in the Communist Chinese system he can write something closely accurate to Wade-Giles and be able to interpret it which was the only purpose it was meant for.

Any questions?

CHAIRMAN SEE: There are a few discrepancies we can go into later. For example, there is no such thing as IEOU. We can go into this later.

RON HOFMANN: I was getting this out of the character index, I may be wrong.

BILL WANG: Similarly for Item 3 you have under the Mainland column WAI should belong to line 5.

RON HOFMANN: You are right. That is a correction .

SAMUEL E. MARTIN: In line 7 under Romanization the fourth tone should be IAW and not IAU.

RON HOFMANN: Any other typographical errors or otherwise?

BILL WANG: How about characters? (Laughter)

CHAIRMAN SEE: Perhaps it would be better to submit our comments.

BILL WANG: I do have a question. When you do have this algorithm that converts every system so we know it works he did not extend his study into converting

among the four tones of the National Romanization. I was wondering if somebody had extended the work in this direction. That is, we are able to take Yale, Mainland, Wade-Giles, and convert them with the first tone of the National Romanization but not the other three tones.

SAMUEL E. MARTIN: We gave an algorithm in ours.

BILL WANG: Within the fixed limit of the system itself. Do you use the first tone or the bridge?

SAMUEL E. MARTIN: Yes.

SYDNEY LAMB: Weren't there other rules for getting to the other tones too?

SAMUEL E. MARTIN: Yes.

BILL WANG: They changed into the first tone and to another system. Then the question is, is there a way of converting non-first tones? I wondered if somebody had done some work because it would obviously save time. You go direct.

SYDNEY LAMB: I think the system would be simpler if you go to something uniform.

RON HOFMANN: Conceptionally quicker. It is an engineering compromise.

CHAIRMAN SEE: It depends on the nature of the

rules you get when you try to do it.

BILL WANG: I think this question ought to be answered because it would make quite a significant difference in processing input. But it hasn't been explored in connection with your work?

RON HOFMANN: Apparently not.

SYDNEY LAMB: Could I ask a question?

CHAIRMAN SEE: We had originally thought of keeping the questions to the end but --

SYDNEY LAMB: I was wondering about the relevance of the machine translation and maybe it is on the Romanization or alphabetical form for input.

CHAIRMAN SEE: There are several ways it could be relevant. In the past Georgetown's people used two elements, one a specific element for the characters such as the telegraphic code which is what we recommended, and second a pronunciation guide which in effect serves two purposes. It does provide the phonic if you don't know the number and then for scanning you can read the Romanization. Second, it does provide more information because, as we all know, there are characters that have more than one reading so if you want complete information you have to supplement.

SYDNEY LAMB: So, in other words, it seems to me what you want for output purposes is a means of converting from telegraph code to some Romanization, rather than decide on what Romanization to use and use that.

CHAIRMAN SEE: Programs like this exist.

SYDNEY LAMB: I am not clear what purpose a computer program would be put to that it has the ability to compute from one Romanization to another.

CHAIRMAN SEE: Some people use the Gwo ^{國語} Yen Romanization. It would be nice to have a computer program to put down the tone. It is easier to read than a system that uses letters and numbers, rather than compare some using letters and numbers alternating.

RON HOFMANN: I think we should go on to Ben T'sou.

BEN T'SOU: I have already given a report of the work done at MIT at a recent meeting of the Association. What I am about to say here is in addition to what has been said earlier. It is sentences with adjective modifying construction. The handout contains some examples of this. At present I am working on the inclusion of numeral measure words as the classifier. So far about twenty of the common measurable words have been studied

and I hope to include into a grammar in the near future. So far the research I am doing is centered on the basic components of the basics of the grammar concerned. However, we are looking to a more wider horizon and we hope to have development of several types here.

As Dr. Yngve mentioned we have this running program here. What we have is a Chinese grammar here and the overall framework of the system exists. Our problem now is for the translation here and also for comparable English grammar here and also Chinese grammar 1.

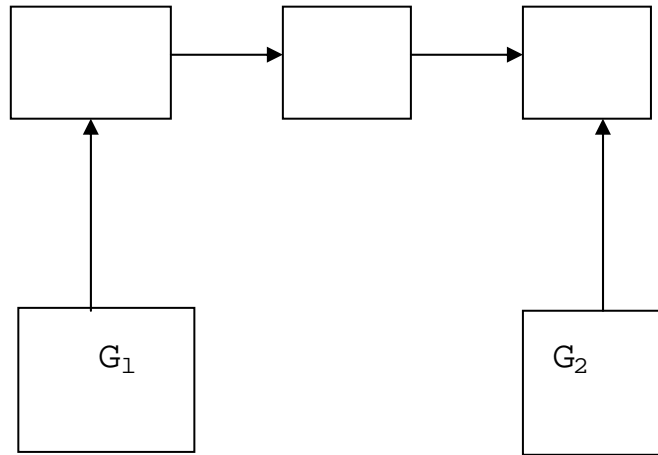
To explain* this a little further what I am saying is from Chinese grammar 1 we can input into this system and we have the running program that can construct recognition routine that recognizes sentences generatable by the grammar.

The second part of the routine we are thinking of would be contrasting and comparing Chinese and English syntheses and utilises the recognition into equivalent English sentences by the third part.

As Dr. Yngve mentioned earlier we have had other experiences in Arabic and English and for this we will probably parallel the effort.

The final part and most difficult part to perfect

* See p. 19A



a system. What I have described is a very restricted type of system and as it is it is not very difficult to construct. However, the ability to translate is confined to the kind of sentences generated by the binary of the grammar. Further work would have to be done to refine this basic grammar to improve the capability of the system. We hope in time to come to successful improvement. I am being very optimistic here as you probably realise.

We have not seriously considered input and output system because we feel these are separate problems requiring other sources.

Are there any problems?

CHAIRMAN SEE: I have a question. You are doing something?

BEN T'SOU: This was about a year ago.

CHAIRMAN SEE: How do you write your grammar?

BEN T'SOU: The grammar is left to right.

CHAIRMAN SEE: How do you write the Chinese?

BEN T'SOU: We do not have the facilities to put in characters. We are looking to ITEK.

CHAIRMAN SEE: The Foundation recommended in I think it was the eleventh or twelfth annual report that whatever else you use if you attach the four digit number

from the telegraphic code to it – it is fairly compact – it is an unambiguous representation of the character involved and this is a great deal more than you get with the Romanization.

BEN T'SOU: This we intend to do. We were thinking, this later publication we have a Chinese typewriter here at MIT.

CHAIRMAN SEE: Publication is another matter than the input that can be changed with other people.

BEN T'SOU: That we intend to do.

J. WONG: In this does the Chinese come first or the English come first? Do you have the example in Chinese first?

BEN T'SOU: The translation is by human.

J. WONG: The example in Chinese is actually the translation or from English?

BEN T'SOU: The Chinese ones are the ones being generated. By the way, the first TA should be T.

FRED PENG: Would you kindly tell us what kind of sentences they generate from for these examples? The first sentence sounds very peculiar. I would like to know.

BEN T'SOU: I think you are questioning the semantic content of the sentence.

VIC YNGVE: Let me say what we are trying to do. If we find some rules of grammar we think are correct and write them in the form we had on the board we then can write a program that will produce sentences where we choose rules at random where we have a choice. This is the result. The Chinese is the result of doing that in the preliminary Chinese grammar. The purpose of doing this is to look at the output, see if we accept it as being the type of thing which we expect our grammar is describing. If we have made a mistake in the grammar we go back and change the grammar. Now since the only constraint that we have introduced on the strength of output Chinese is syntactic or grammatical you will find that the sentences that come out are nonsense, to have no semantic constraint. Does that answer your question?

FRED PENG: Yes.

DeCAMP: I would like to ask a related question. I would like to ask first of all are the fifteen sentences that are here a genuine random sampling?

BEN T'SOU: Yes.

DeCAMP: It seemed to me in the fifteen there was a disproportionate type of sentence in here which are way out of line with what we would expect to find in an

actual text. I wondered if it were random within a certain syntactic type.

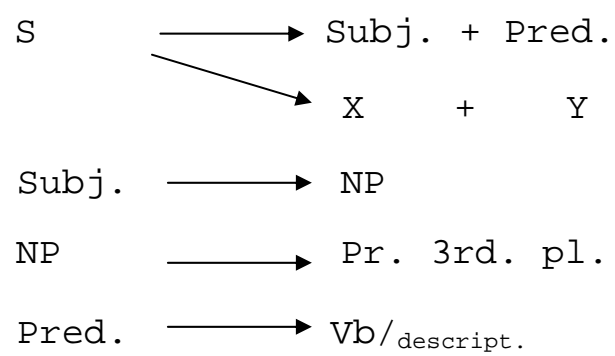
BEN T'SOU: The randomization of Chinese, and as you look through the list, most of the sentences include a modified construction. What I have done is generate the sentences and obviously I can't bring in a code of syntactics so I have selected at random.

VIC YNGVE: These are selected from a random sampling which were generated from a grammar which is a very restricted grammar and is centered about his particular type of grammatical interest.

DICK MARCUS: Would it be possible to take one of these sentences and explain the rules by which this sentence was generated?

BEN T'SOU: Suppose we take a simple one. The simplest one I think is the second last one, fourteen. I am simplifying as I go. Subject* plus predicate and there are various choices. In this case the subject is divided into various types of subjects. I don't go into that. In this case the noun phrase becomes a pronoun, third person plural. This is carried through the generation of the sentence. The predicate is further divided into various categories. It goes into a descriptive verb.

* See p. 23A



Some people call it an adjective. Does that satisfy you with this sentence?

BILL WANG: Did you use the third type of rule?

BEN T'SOU: Not in this.

PAUL GARVIN: Do you call a frame and then get the fillers to fit into the frame or how do you materialize the first arrow subjection predicate?

BEN T'SOU: Well, I suppose that is universal in all languages so we assume that. In this particular language, Chinese. I was thinking of studying certain types of sentences. I just sort of further categorize them.

PAUL GARVIN: When you get the subject you then call a subject routine?

BEN T'SOU: How the program works is quite simple. As it expanded into X plus Y2, then there is a choice. It starts, initiates the program we have. It takes a random, one of these, and then goes on the string. It then comes to a subject and has various choices and takes one of them.

PAUL GARVIN: So each time you have a table of choices and have to randomize and pick one. I was just wondering with the thing that General Precision has.

They did something like this and they call it patterns and then it is patterns similar to what you have and then for each slot in the pattern they have a random selection of permitted fillers. Then they work to keep reducing the permitted number of things and expanding the number of possible slots. That is to say, get to finer and finer subclassifications. So I am delighted to see that there is concurrence.

RON HOFMANN: In connection with the randomization when there is a choice of rules to be applied these rules are chosen with equal probability rather than exactly what you normally expect in the language itself.

BEN T'SOU: You have two choices here for the first one. If you run this program a hundred times fifty per cent should be one and the other fifty per cent the other.

VIC YNGVE: It is probably true that the statistics from this is probably not a fundamental property of language, except as an experimental device. If you want to study a particular type you can eliminate the other easily and reduce the choice. But we are not interested in discovering these frequencies to study them,

DeCAMP: Wouldn't it be true that that random

selection of rules would be proportionate to a random selection of sentences out of a text? If you wanted to generate out of sentences that would be roughly proportionate to the text.

CHAIRMAN SEE: The only thing I know of is Bill Harwood's work in Tasmania where he took children's speech and for every rule that weighted the things in the current of the corpus. Of course, this is on the assumption that the rules themselves reflect something real about the language. There is hardly any point in weighting them unless you know these are the rules. This is the heart of the problem.

VIC YNGVE: It was Bill that said the frequency of aspirin doesn't reflect anything on the incidence of headaches.

DAVE LIEBERMAN: I think you couldn't stop at the statistics of rules but once you have an approximation you have the frequencies of rules. There would be no end to it,

VIC YNGVE: We have speculated on what would happen if we made a word count on the output of a generated program. I have done this. The trouble with it was that our grammar is so tiny that we used for that that we

don't really have a good distribution of words over the various syntactic types and so on but it did come out to some extent similar to the distribution and it is my impression if we had a fairly complete grammar and a selective with any type of probability weight at all, equal probability or any other type you suggest, that it would come out with the distribution. In other words, I think that distribution is sort of inherent in the syntax language.

CHAIRMAN SEE: It is inherent in a lot of things. There is one important point. In the research project you would prefer to generate things that are of interest for further study and most sentences are very uninteresting. The Georgetown's sample of a thousand sentences included quite a few of the type, "The temperature was 180 degrees," and there were quite a number of sentences that were the same sentence with a few changes made. After you have analyzed these there is hardly any point for coming out with these. So really weighted for research purposes is better than of the national language weight at this time.

PAUL GARVIN: I think this raises another interesting point, the choice of sentences that remain

random. What I have in mind, if you have a crude program and the rest is left to random selection. Now as you improve your rules it means in fact that you get to be more and more specific and the area of random choice gets to be more and more reduced. So I think this is an interesting question. Theoretically it would be interesting to know at what point you have to stop at rule making and leave to random. Ultimately you might have a generating system which is linked to a perception device where the perception device governs further selection and you get to a point then where you can tease the behavior assertion and have responses to stimuli.

CHAIRMAN SEE: Well, I guess as Chairman, unless we have further questions, we will leave the discussion, interesting as it is.

BILL WANG: I have a specific question towards the grammar that underlies this. There is a reason for the discontinuous type of rule to take care of the character in the deep structure but you allow only one constituent to come through. I remember in previous meetings you said this turned out to be inadequate. Do you still hope?

VIC YNGVE: Yes.

BILL WANG: You haven't found this for Chinese?

BEN T'SOU: I have considered it.

VIC YNGVE: I might comment. You have to conceive of that in the context of left to right generation where you always go down the left branch first and complete the left branch down to lexical items for words and work back up. You have to conceive of it in that framework. So it is a very high node.

BILL WANG: I think you would be forced to assign too many sentences if you make that restriction.

VIC YNGVE: I am open to changing that if I find the language is different. I started in the beginning thinking that it goes to the end. Maybe it goes over one, two, three, four and you have to have a subject and indicate how many. You might have several kinds. I think we would be quite willing to do that if we find in fact we need it but for English, and Arnold tells me for Arabic also, one seems to be enough. We are very surprised to find this.

BILL WANG: Dick and I were talking of a particular sentence in Chicago, "Have you ever been in Chicago?", where the "ever" is definitely a partner. "Have you never been to Chicago?" Intruding between these two parts of a discontinuous constituent you have to have "You have

never not". That is a little off the topic. I just wondered if you had modified that position.

SYDNEY LAMB: An even simpler example, "He has called her up."

BILL WANG: "Called up" can be regarded as one constituent.

SYDNEY LAMB: But it is only after the call and not called up.

VIC YNGVE: You mean the "ed" thing.

SYDNEY LAMB: It doesn't have to go after a whole constituent. It only goes after part.

VIC YNGVE: First of all we go only to words.

SYDNEY LAMB: Then you have complexity.

VIC YNGVE: Secondly, if we did go to morphemes I am not at all sure it would work.

CHAIRMAN SEE: Any more questions? If not, let's proceed with the Texas group. I should point out, as you all gathered, there is a change in the expected participation. Wayne is here and Madie Gray is not here. So the list is partly revised.

WAYNE TOSH: I apologise for not having enough of these handouts to go around but they are not terribly important. They are just a summary of some of the

statistical data we have. The thing of primary importance to you is the amount of Chinese data we have.

Now on the first page you will notice a statement, "There are 3600 dictionary entries and 3400 syntactic entries." This requires a little bit of clarification. The grammars we are writing are of the context free type. So I am referring to it.

Now the Linguistic Research Center has a staff of approximately thirty people. This too fluctuates largely due to the fact that a good part of our linguistic staff is made up of graduate students and as they take their degrees and move on the staff changes. Approximately a third of the staff is linguistic and the rest split between theoretical mathematical and staff.

You will find details of the organization and the approach that we are using, the theoretical foundation, spelled out in a document bearing this number, LRC63-SR1, and the title is "Symposium On The Status Of Research". I won't spend any time on these details. This is a symposium we presented for the National Science Foundation. You will find the outlines, the formal outline, of the theoretical basis we are working on given there and an outline of the program system and an outline of the linguistics approach we are using.

To review it very briefly the approach we are using is a stratification type of description. Our efforts have been confined to morpho-syntactic description in English, Roman, Chinese and Hebrew. We have just the very beginning of a description in the last two languages. I think you will find these listed on the statistical summary that I passed out.

This morpho-syntactic description is intended to take care of the overt characteristics of the language in question such as word order but to exclude semantic equation features. This will be relegated to a later descriptive effort.

Our terminal you will find is second order description. The second order description will take care of transitional and semantic classification problems. The grammars are intended to be bidirectional so again if you recall the outline that Professor Yngve presented on the translation system ours is logically quite similar. The procedural details are different in some respects but the translation process is broken down into three phases of operation, recognition, transfer and synthesis.

Perhaps the thing that is important to emphasize is that the grammar that serves to recognize the

input language can equally well be used to synthesize the output language assuming that one is coming from another language into this particular one. So if we take the English grammar, for instance, it is designed to work equally well as a recognition and synthesis grammar,

There is likewise an intermediate table to state the equivalent between a rule or set of rules and grammar.

A negative point in the first order description we do not have a facility for handling discontinuance. This will be a function of the second order of transformation description.

That in a very brief nutshell is the type of system that we have right now and is similar in general capacity to the system Professor Yngve has outlined. We are not doing any random generation. At the present time the state of the system is still essentially the same as reported at the symposium. We have the programs developed to the point we can do automatic recognition and each of these five grammars but are not yet prepared to do any syntheses in them. This series of programs in recognition is expected to be completed sometime early next year, 1965.

Now as to more specific problems in Chinese the

description that has been prepared in Chinese is largely suitable for recognition purposes and not for synthesis in that the kind of restrictions that have been built into the grammar here are sufficiently well defined to work for recognition but sufficiently ambiguous to prevent the grammar from being used as a generative grammar. So one of the problems for us now is to add in the necessary restrictions and to this end we are undertaking a study of Chinese syntax and morphology. Some of you have undoubtedly heard statements from time to time that Chinese has no morphology but it is simple to ask a native if he can randomize expressions and get a negative answer so there must be more to it.

The steps we are taking to come up with the kind of description we need in this area have started with a study of a concordance of the Chinese texts that we are presently using which is, incidentally, referred in the telegraphic code and to even code expressions taken from the texts as the terminal to some rule, to some class, and to encode as specific properties of this class as if they were parts of the class name themselves, subclassifications, those expressions which the informant considers permissible in concatenation. Right now they are

restricted to the very primitive level of contiguous expressions.

What we will do when this data is completed for the concordance is to sort these rules again looking at this as if it were a unity class name and bring together all expressions having the same set of properties. This is not a very profound thing as far as linguistic research is concerned. It is just a good classical study in distributional analysis. We want an expression of this sort to put in the generative restrictions we think are necessary for translating into Chinese. So far none of the grammars we have looked at have a sufficiently well defined description, set of classifications. More often the classifications seem to reflect a description based on, let's say, English transitional analogues. Of course, this presents difficulties as you well know.

I thought I would leave the summary of our work at that and leave more time for questions on specific details. I have brought with me some displays of the Chinese grammar and concordance. Unfortunately there are only four or five copies here so it won't be very convenient for everybody but we may want to take a look at these later if not now. So I will turn the floor over to

Dr. DeCamp who has a few words about the program at Texas.

DeCAMP: All I have to say is that the program at Texas is divided into two disparate divisions. There is the program in the college on linguistics and then the linguistics center separate. I am in the linguistics program and until recently the people over in the research center have been carrying the ball almost completely on the Chinese, certainly as far as any research they have beginning with the fine work done here. Until this year the Chinese language in the University has been an incentive shot in an exotic language for the graduate students. Trying to get it out of the exotic into a continuing carefully planned language program is what we have started this year with Chinese being offered on the third year level and beginning to find students both on the undergraduate and graduate level. As such our relationship between the research center and ourselves is one which the teaching work is done by ourselves. The specific MT application is entirely by the research center. General research is done at both places.

There has been considerable planning during the past few months as to what areas of Chinese syntax are going to be touched on, what can be profitably dealt with,

and things of this sort. Wayne can show you the display material he has here. Other studies of this sort are planned and certainly will be under way by us and graduate students in the program as well as those over in the research center.

CHAIRMAN SEE: Are there any questions directed towards the Texas people?

SYDNEY LAMB: Did you say there is some work on Chinese structure going on in the college?

DeCAMP: As of now, yes. There are two graduate students interested now. One of them is on campus at the Austin campus and is on comparative structure. The other one, a student who is at present still in Taiwan, is working on complement structures.

CHAIRMAN SEE: Any other questions?

SAMUEL B. MARTIN: I would like to ask a question. You are dealing with single characters and their neighbors, is this what it amounts to?

WAYNE TOSH: In the simple case but a string of characters may be included as well.

SAMUEL E. MARTIN: Then when you have this long list are these only things that occur in your text?

WAYNE TOSH: No, they are limited to things that

occur in the text, in the traditional grammar such as classifiers.

SAMUEL E. MARTIN: So you have done a certain amount of grammatical analysis before you consider this?

WAYNE TOSH: Insofar as we have limited the set of properties to this list but we are not starting with the assumption that we will have adjectival properties because at this point we don't know what an adjective is.

SAMUEL E. MARTIN: I was wondering.

WAYNE TOSH: It doesn't include the language.

CHAIRMAN SEE: Any further questions?

If not, I would suggest the following procedure.

I had planned a break at this point anyway. The group has copies of their output, I believe, for display purposes. There is a large table over there and a large table over here. I suggest we take a break until eleven o'clock which will allow time for those who are interested to group around the two tables and scrutinize copies of the output and ask questions about it.

(Recess from 10:45 A.M. to 11:25 A.M.)

CHAIRMAN SEE: I think we have had a good half hour break so we can resume with the next two presentations, one from ITEK and followed by IBM. I think we can

expect that both presentations will fall short of an hour so if we can make the questions direct ones at the end we can aim for a one o'clock luncheon and other questions can wait until after lunch. So we will now hear from the ITEK group.

DICK MARCUS: I shall try to keep my remarks fairly brief so that there will be time for questions and discussion. I will give a general picture of the work that is being done at ITEK and Dr. Wong and Theresa Lee can answer specific questions about the Chinese analysis itself.

As many of you know, ITEK's interest in mechanical translation got a strong boost about two years ago when Dr. King who was at IBM came to ITEK and brought to ITEK plans for further development for the so-called photo-store computer, an application including mechanical translation and I might add automatic stenotype transcription. We are not quite ready to give it to you today but the next meeting maybe.

ITEK has a program that began just a few months ago with the Air Force at Rome which is basically for linguistic analysis of Chinese leading to application of translation from Chinese to English. In addition ITEK

itself is supporting mechanical translation itself and hopes to come up with a workable system in the near future. So that we are indeed quite oriented towards some of the practical problems and to the whole spectrum of the problem and perhaps not quite as much at this time oriented towards the theoretical or formal aspects of the problem.

I might mention briefly some of the hardware work that is going on at ITEK because it is pertinent to the kind of translation system that we are devising. First I can mention our Chinese input encoder which again is a development of previous work done at IBM for the Air Force. As you may know, this is a typewriter-like device that with basically three strokes on the flexowriter type keyboard one can encode any Chinese character and the Chi-coder, as we call it, can be operated by non-Chinese speaking persons with relatively short training periods. I had hoped to bring our model down here for you to see today but we had transportation problems so the best I can do for those who are very interested we can arrange to see it at ITEK at some time.

The Chi-coder comes out, as I mentioned, basically with this three bit, three digit code, which is what we are using internally in the computer programs. We have

prepared conversion tables from the telegraphic code to Chi-coder code so that if necessary we can make use of material that is in the telegraphic code or we can for output purposes if someone wants to put the output in telegraphic code.

CHAIRMAN SEE: Excuse me, how many characters do you have for the typewriter?

THERESA LEE: Ten thousand. The thing is that we also have simplified characters. We take the regular character code.

CHAIRMAN SEE: How about variance? 0016, some people write that one way and some another. You can write either flat across the top or slanting.

THERESA LEE: The thing is if the Chi-coder character is in a different code but we recall the telegraphic code as the same.

CHAIRMAN SEE: Can you go both ways or only one way, telegraphic code to your typewriter?

THERESA LEE: We have the conversion table from the telegraphic code to the Chi-coder so if you are inputting telegraphic code the output comes out.

CHAIRMAN SEE: The variance mostly go the other way. You have telegraphic code to a typewriter version

but you may not go to all possible variance.

THERESA LEE: That is right.

DICK MARCUS: The computer development at ITEK again is based on previous work both at Telemeter and IBM. The additional development has been and is going on at ITEK which has to do with expanding the memory capacity of the computer and increasing the flexibility of the logic so that on one disc now we are storing approximately 200 million bits of information and with one content addressed lookup which takes on the average of 15 milliseconds you could search through the whole memory to find that particular strain you are looking for if it is there.

In addition to the basic lookup feature there has been additional logic added to the control so as to essentially allow general purpose digital logic capability in addition to the lookup. So for those bookkeeping type operations which previously would require one or more lookups we can do now in a very short period of time we have two fast memories of thin film and a core memory so we can do these bookkeeping type operations or any other arithmetic operation at a much faster speed from the thin film and core.

Now as far as the translation work that we have

done so far it is still in the early stages. As I mentioned our Air Force contract is only about three months old. We have started by making a tentative word class category, the main elements of which are given in the glossary in the handout, and under each of these main word classes there are many subclasses. For the translation procedure itself we divided into five stages and we have given samples of relatively simple examples of what is meant under these five stages.

The word segmentation by longest match indicates that we have stored on the disc or contemplate storing on the disc Chinese words that these groups of characters that are to be treated as a syntactic element. Then in the input sentence would be read into the computer and starting on the left-hand side you would look up in the dictionary the longest string with the longest word that you can find and that provides the word segmentation as you do this successively. The information that you would get from this word segmentation, that is the information that is stored on the disc, is of two types.

(At blackboard) First there are the word class codes, grammatical codes, for the Chinese words. Second, there would be the translation in several forms perhaps,

noun, adjective or verb, if these forms cannot be derived from one form by simple rule. These English words would have morphological tags which tell the way they are inflected. The information in number two is saved for the final stage of synthesis and what we operate on in the next stages, two through four, would be the word class codes themselves.

Now we have distinguished three types of passes. In general we think of these as occurring in sequence but we realize there will be actually a bit of interchange so that we have the parsing rules, the ambiguity resolving rules and the translation rules. The translation rules are in three types, those that would specify number and tense and so forth, those in which English words are inserted into the output string, and those rules in which a word reordering is specified.

I might mention something of the form in which we indicated some of these examples. The fact that they are written with a double arrow, one outside the other, is not necessarily indicative of any particular kind of transformational analysis for the following. It is just a convenient way of explaining the rule.

There is a suggestion though about implementation.

The implication is that these strings on the left-hand side of the rules will be looked up whole by our longest matched table look up method so that we not only use this table of longest match for initial word segmentation but these syntagmas, if you like, or phrases are looked up whole in the later processing. We feel this is a significant saving of simplicity and speed over trying to analyze strings of codes purely by algorithmic approach.

What we have done so far is build up a fairly extensive, although still tentative, set of parsing rules, rules of recognition if you like, and have made some preliminary stabs on ambiguity resolving rules and translation rules. We hope to soon gather many more ambiguity resolving rules and translation rules with this preliminary set to actually try out translation and on the basis of the kind of results we get decide whether to emphasize what changes to make in the word class characteristics and what particular aspect of linguistic analysis to emphasize.

That very briefly is the scope and direction of the work at ITEK and we would be glad to try to answer any specific questions.

RON HOFMANN: There is one question I have about your symbols. It is not clear whether the CMM is to be

的

interpreted as conumeral and 4104 is indicated as conumeral

的

or conumeral plus 4104?

DICK MARCUS: That CMM is the symbol for that particular conumeral.

RON HOFMANN: I see, and not for the other.

CHAIRMAN SEE: Which you don't have listed here.

THERESA LEE: We just have listed for the material in this.

DICK MARCUS: What is listed in the symbol page is just for the succeeding pages and is not complete.

ITIROO SAKAI: Longest match, do you have any means for getting rid of the linguistics especially in analysis?

DICK MARCUS: Yes, if we find in this syntactic analysis that we are having difficulty we can go back and try to re-segment. We hope, we haven't done too much experimentation yet, but we hope this kind of difficulty will not occur very frequently,

CHAIRMAN SEE: But it does. (Laughter)

SYDNEY LAMB: In Chinese it comes too often.

CHAIRMAN SEE: There are two kinds of situations. There are standard cases with very common connectives which fall repeatedly into the same thing which I suppose you

中

could make up a special rule for. For example, 0022 can easily occur at the end of a phrase and if some nice lit-

國

tle word like 0948 is the next one then you have China coming out where it wasn't intended, I notice in the paper I read every day from New York in the beginning it always

日

winds up the date with 2480 and then proceeds immediately

(東)?

without a break with 2639 which gives you Japan. It is

(2609 本)

quite common so that if you are committed to that, the sentence if you go back and re-examine it I wonder if you have a program in mind that summarizes this.


DICK MARCUS: I don't have the specifics for that case.

J. WONG: We might use the initials in the box to avoid that. Initial the beginning, then the terminal. We use the marks to indicate this is the initial and this is the terminal.

CHAIRMAN SEE; You mean you pre-edit the material. I suppose you get a sentence which has these words in it. It could be you are referring to China or it could be it is segmented between the pieces,

J. WONG: In other words, China would be one word.

CHAIRMAN SEE: Suppose it isn't?

J. WONG: No, on the disc we encode it and it would be one word, Japan. ^(?) TUNG is used otherwise. 

CHAIRMAN SEE: I suppose it is followed by 0948.

THERESA LEE: We would hope there would be something in between.

CHAIRMAN SEE: There often isn't.

DICK MARCUS: The general solution where it indeed is as you say would be that in doing the parsing we would hope again that the wrong interpretation is ungrammatical. It would make the whole sentence ungrammatical.

PAUL GARVIN: This requires a criteria for automatically determining whether the output of your lookup is grammatical or not. This is not easy.

DICK MARCUS: That is what is done basically in the parsing. If the parsing does not succeed to the extent the parsing is right you have made a mistake perhaps in segmentation.

PAUL GARVIN: It is hard to find where the mistake is. We have a problem in a somewhat different area which is the matching of English sentences in a comparison program. Jules Mersel developed a program of sentence extension where you cannot match one Russian sentence with one English sentence you try matching two Russian

sentences with two English sentences and then try to match one Russian sentence with one English sentence which is one way of trying to find if your match is correct. This is where if the shortest way is correct you are okay. The criterion is in the sentence if there is a word match. Could you conceive of a simple lookup where you try going either left to right or right to left and get the longest match and if you don't get a suitable longest match on the next few characters you go back and revise the few characters and then go on. I think this might be one simpler way rather than wait until the whole thing is parsed.

DICK MARCUS: Then you would have to have the rules for segmentation later on. You would never come up with an impossible segmentation because any individual character could be a word.

CHAIRMAN SEE: Would you care to comment on this problem?

SYDNEY LAMB: Yes, we have given consideration over a number of years. When we started out we were working on Russian, We use the principle of longest match which works pretty well for Russian actually but does get you into some trouble in some cases, so we gave it up. In Chinese I think you get into trouble more often than

you do in Russian. The only way to do it is to get all possible segmentations.

DICK MARCUS: I would be interested if anybody had any actual statistics on it. I would like to make one comment on statistics in general though. I personally have a slightly different outlook than maybe other people do. It seems to me when you try and divide syntax up into things that are grammatical and things that are not grammatical it is not just a binary choice, some things are and some things aren't. It seems to me that this is often a gradation of common things that are obviously considered correct in grading off to other strange constructions that may occur very indirectly and yet they do occur, so since we are trying to devise a working system and do as well as we can on our first try we have to consider statistics very much. We have to try and do the most frequent things first and perhaps leave some of the exceptions to later on.

S.S.SOO: When did ITEK start to work in Chinese?

DICK MARCUS: Well, as soon as Dr. King came is when we first considered it. This Air Force contract, as I said, is only three months old, ITEK's work has been

continuing since Dr. King arrived.

PAUL GARVIN: That is two years now. Time flies.

DICK MARCUS: Not quite two years.

CHAIRMAN SEE: I have some information concerning this business of the occurrence of this phenomena. Early in the game when there were only a couple of groups looking at Chinese I circulated several of them on this lookup of the Yale text which is called 青 天 雷, 7230 1131 7219, and this is 雷 7225, I am not sure. Anyway, that is a Yale book which has been run against the McGraw-Hill dictionaries on the principle of the longest match with some after thoughts. So the next to the longest match was also found. There is an asterisk in the margin. These things are marked where there are alternative segmentations as far as the dictionary would reveal them. So one could go through and look at a fairly large example of these. These were not all bona fide examples because the dictionary includes things that some people would not consider words but phrases.

SYDNEY LAMB: Do you gather from looking at this how often you get a wrong segmentation?

CHAIRMAN SEE: I don't think it is very infrequent. As a general impression I wouldn't be surprised

if every page or so you have a serious one.

BILL WANG: To support it take the example of
 中 國 人 多
 0022 0948 0086 1122. Suppose you want the longest match.
 The longest match definitely ought to contain that which
 means there aren't many Chinese. The segmentation you
 would miss in a case like that is where it is in place
 of the verb.

CHAIRMAN SEE: There are two segmentations, one
 with the first three characters and followed by the fourth
 one as separate, and the second analysis with two charac-
 ters and two characters. Both are possible.

PAUL GARVIN: Wouldn't it be possible to enter
 this in your longest match dictionary as an ambiguity?
 You have two alternatives.

CHAIRMAN SEE: Enter the whole phrase?

PAUL GARVIN: In these cases.

SYDNEY LAMB: In the Russian when we decided to
 use the principle of longest match, in places where you
 could get ambiguities where one might have been correct
 you can put the necessary information into the dictionary
 into the longest match. However, that didn't work in some
 cases. In many cases it is a little more complicated.

J. WONG: Excuse me, in the example by Mr. Wang

I think as far as semantic differences rather than segmentation because as far as segmentation I think the same way you would segmentate the phrase. Assume the American Chinese or you say the other meaning, "there are so many people." Then the same way you segmentate the phrase, the same way.

THERESA LEE: We would segment them after the first two and then --

CHAIRMAN SEE: You split them up in each case?

THERESA LEE: Yes.

FRED WONG: We have the experience that there are many, many cases of this kind of problem. The first time I tried to solve some of the problems and see what can we do based on the material you gave me. I think there are about twelve per cent of the problems in the whole text and I can solve about five per cent of those problems, five per cent of the whole text. So I think it will be a very big problem.

There are two types of difficulties. I don't believe we can segment them as much as possible. One told me, "Why don't you go by a string of morphites?" I found it more difficult. We can see for instance 0031
 要 是 *
 6008 2508. This one you can do this way or this way or

* See p. 53A, 53B, 53C

I. a. 0031 6008 2508
 b. 主 要 是
 c. zhu3 ya04 shi4
 d. zhu3 ya04 shi4

a. mainly
 b. main thing is
 c. Lord if
 d. Lord want be

a. 6638 2876 0155 0031 6008 2508
 這 樣 作 主 要 是
 zhe4 yang4 zuo4 zhu3 ya04 shi4
 this way do mainly

3634 0055 4675 0830
 爲 了 簡 單
 wei4 le jian3 dan1

for the purpose of to be simple.

b. 0948 1367 0031 6008 2508 0086 3046
 國 家 主 要 是 人 民
 kuo3 jia1 zhu3 ya04 shi4 ren2 min2
 Country main thing is people

c. 0031 6008 2508 7100 4395 2053 0226

主 要 是 降 福, 我 們

zhu3 yao4 shi4 jiang4 fu2 wo3 men

Lord if

bless

we

1432 1779 2405

就 得 救

jiu4 de2 jiu4

then get save.

d. 0031 6008 2508 1015 4206 1778 4104

主 要 是 基 督 徒 的

zhu3 yao4 shi4 ji1 du1 tu2 de

Lord want the ones who are christians

0006 1131 1016 6008 0008 2508 1015
 上 天 堂, 要 不 是 基
 shang4 tian1 tang2 ya04 bu4 shi4 ji1
 go up to Heaven, want the ones who are

4206 1778 4104 0007 0966 3739
 非 徒 的 下 地 獄
 du1 tu2 de xia4 di4 yu4
 not christians go down to hell.

this way, all possible segmentations. So what do you do? Then you have to go by more what we know now. Then these rules won't help us.

DAVE LIEBERMAN: I think we should distinguish between what is a linguistic problem, that is what is the structural problem, and the question of how we get these structural determinations by machine. I think the first is logical prior and is far from solved. This problem is not unique. Everyone has been plagued with it in Russian and Arabic and anything they have been in. The only thing possibly unique, it extends down to what you call the lexical level. Otherwise the question of how to segment by machine and several possibilities is whole question. As everyone knows, it would be very nice if we could try them all out but again it becomes a question of principle than practice whether you can do it or not.

SYDNEY LAMB: I don't think it is that complicated. The proper solution is the most elegant phrase and the way you do it is to have a segmentation that finds all possible segmentations and it solves the problem. It gets you out of all of these problems.

PAUL GARVIN: Out of all the segmentations how do you select the one?

SYDNEY LAMB: This is the same principle as used in general. At the end of any stage, as a general statement, in any stage you get all possible solutions for that stage and send them to the next stage, then send them on to the next stage, parsing or whatever you call it, and at this stage those that are not correct are thrown out.

PAUL GARVIN: How?

SYDNEY LAMB: Because some of them won't fit. If more than one fits you have two possibilities for the next stage.

BILL WANG: Except sometimes the dependencies are across the stages. In one case it is a logical break and on the other a syntactic break.

SYDNEY LAMB: As long as you get every possibility at every stage there is no possibility of missing anything.

DAVE LIEBERMAN: But why should someone do that? Why should someone write a computer program that would do that?

SYDNEY LAMB: I don't see how you can ask such a question, the answer is so obvious.

DAVE LIEBERMAN: No, it is not obvious.

SYDNEY LAMB: Then why should one not do that?

DAVE LIEBERMAN: Some people do it because they are interested in trying to develop approximate analysis procedure to try to get translation. These people can't do the complete job. As for doing the complete job --

SYDNEY LAMB: I don't mean a complete job in the same sense you refer to a complete job. All I am saying, one must have a strategy that allows one to have all the probability. That doesn't mean that you have to do a complete job in analysis.

DAVE LIEBERMAN: Have you considered the strategy is something that can be motivated by the linguistic theory?

SYDNEY LAMB: Yes, sure.

S. S. SCO: Strategy for getting all possible combinations.

SYDNEY LAMB: No, it is not the strategy for doing that. I am saying to adopt a strategy that enables one to do that. It is a separate question on how to do it.

S. S. SCO: It is extremely easy but to distinguish which is correct is a big problem.

SYDNEY LAMB: No, that is the next stage, the syntax.

PAUL GARVIN: Then you have the difficulty in the next stage.

SYDNEY LAMB: That is where it belongs.

PAUL GARVIN: I agree that it belongs somewhere.

SYDNEY LAMB: You don't approach it by saying we have two things and let's see which fits. You don't have to do that. The point is you accept everything from the last stage and send it through. Those that don't fit will fall out.

DICK MARCUS: Why bother to consider all these cases?

PAUL GARVIN: Because you don't know which are correct.

DICK MARCUS: If ninety per cent of the time you are going to come out with the right answer why not do it that way?

PAUL GARVIN: The ten per cent is important.

DICK MARCUS: The ten per cent of the time you are wrong you should come out with the answer that the sentence is not grammatical and then you go back.

PAUL GARVIN: Well, talk about it from the point of view of efficiency. If you have to go back, once out of every couple of sentences.

DICK MARCUS: It is a question of how often you have to go back. I have a question about this twelve per cent figure. Did you mean twelve per cent of the number of words?

FRED WONG: Well, the number of the syllables in the corpus that Mr. See made.

PAUL GARVIN: Is it twelve per cent of running text or twelve per cent of the dictionary?

CHAIRMAN SEE: Running text.

J. WONG: This explanation you gave, Fred, I understand the first two. The third one you divide into three. What does that mean?

FRED WONG: The Lord.

J. WONG: Oh, yes.

FRED WONG: 主耶穌
To ya shur.

FRED PENG: Then the segmentation should be ya shu.

FRED WONG: No, it has to be a complete sentence to understand.

VIC YNGVE: We have the alternative of putting it on the board.

FRED WONG: I don't think it is very difficult to see. You can have this here and then you can have

another thing in here which can very easily make a sentence.

J. WONG: Of course it changes the grammar.

CHAIRMAN SEE: For the record, you make a grammatical construction involving the third character as connected by itself to the remaining part of the sentence,

(未?)
the 2608.

(是 2508)

BILL WANG: What about the ^是shu?

FRED WONG: You probably would have an argument there. I think some of it the shu is one or two.

BILL WANG: I can see the three interpretations in the following way. In the case of 111 you have subject, auxiliary verb and copulatory. In the case of 21 you have perhaps a phrase with the head deleted. You still have a copulative phrase which is amorphous in this case with an adverb. In the third case if you have a break you have a subject and an adverb but I don't see the third one.

FRED WONG: The third thing is an adverb.

CHING-YI DOUGHERTY: The sentence, this one can be solved by the rest of the sentence by the context.

FRED WONG: This can be solved by the context also.

CHING-YI DOUGHERTY: This is already a full

sentence.

FRED WONG: Not necessarily. You can enlarge it.

SAMUEL E. MARTIN: It might be a full sentence in a text and then it would be ambiguous.

CHING-YI DOUGHERTY: I think it can be solved by taking the context.

FRED WONG: You always try to solve the ambiguity by the context.

CHAIRMAN SEE: Any more questions on the topic? If not, I think we had better continue with the IBM group if they will take over.

DAVE LIEBERMAN: I think I would like to start by thanking Vic or Dick for inviting us and also the organizers because it comes at an opportune time for IBM where the Chinese program is being reconsidered and plans completed, so it is very nice to hear what is going on to help in these plans.

Now for perspective, or at least the kind of perspective I use when I hear a plan described, I want to hang it in a framework that I have for classifying. It is the one I use and I hope you won't object too strongly.

(At blackboard) First of all is level of

analysis, word for word substitution. Then local context. This is physical context, not syntactic environment, but something like two words to the right or left, that kind of thing. Above that can be sentence-wide context but still no explicit analysis. Finally we get to the whole class of the Yngve type system, the type that fits that framework. I have left a space.

Given the grammar we will get all types of possible readings from the input of that grammar. Now we can't talk of synthesizing output until we get to the point of the analysis of the input. Anything below specific analysis I think is still a word-for-word translation, with modifications. It is possible to have something in here which I don't know as anyone has done yet and that is an analysis. That is an explicit structural analysis but not attempt at synthesis.

Now of course within this general kind of framework you can have everything ranging from transformational grammar to Vic's own type of grammar or what have you. I put this here to describe what has been done at IBM and what will be done.

At the time the Chinese project was started at IBM which was about three years ago there was a vigorously

going Russian program. That program was just about leaving that stage and going to this stage, some place around here, word-for-word substitution.

Now it was natural that the Chinese work should be envisioned to begin with as possibly an imitation of the Russian work but it became immediately evident that there was no sense of even thinking of working at this level. This may not be a great shock to many of you but it is interesting that even in an atmosphere where a word-for-word translation of Russian was considered of some potential use it was immediately obvious this could not be done with Chinese. So from the very beginning the Chinese work was aimed at this level, this level being sentence-wide analysis, analysis using sentence-wide analysis but no essential constructual description and not too much attention to recording of ambiguities. It is usually accompanied by some implicit or explicit ambiguity of the rules. I think that this characterizing the Ramo-Woolridge and the Georgetown work, I am not saying you can hang things exactly but I wanted to put this in its place. So this was what was done.

So I would put the Chinese work at IBM in the class of the Paul Garvin work at Ramo-Woolridge in the

level of analysis.

Now I should mention that we did not copy Paul Garvin's method. Now I have to say this because when we were reviewing what we were going to say he said, "Everybody will think you copied from Paul Garvin," and I said, "No danger," but I thought I had better say this.

PAUL GARVIN: Imitation is the sincerest form of flattery.

DAVE LIEBERMAN: Now the presentation will be as follows. Fred Wong will talk about some considerations that were relevant during the early days when the problem was being formulated, when the decisions were made as to what exactly to do, and then S. S. Soo will describe in a cursory way but in as much detail as you may want. Afterwards I will come back on future plans for what we are considering doing in Chinese although no decisions have been made.

There is one other point I think should be mentioned to avoid confusion. The linguistics group at IBM is separate from the machine translator group. We have the same ultimate boss but not the same immediate boss. The linguistics group is not concerned with production. Soo has been oriented towards the production system. Fred has been

a consultant on the project and Soo the ramrod.

Now we will start with Fred and then Soo and then back to me.

FRED WONG: I think everybody knows about the problems from the general linguistics point of view and in Chinese I would start from the problems of the word Chinese or Chinese language. I think before we go into any operational research we have to decide what kind of Chinese we are doing. There are so many kinds that we have to limit ourselves to make a comprehensive Chinese and try to work out some kind of grammar. That one that was on the board, three characters, I can derive two more if I include the ancient Chinese. So we have to decide to limit ourselves to the contemporary writings but the contemporary writings are in a very complicated form. Dr. Lu Chih-wei, who was the president of the ancient university, has been working on the Chinese language since the conquerors took over in the mainland and he described the modern contemporary writings as not Chinese and it is not ancient, not modern, and is not classical and is not vernacular style. So it is a kind of mixture we have there. We cannot eliminate any one of those possibilities.

If we look at it from the complicated point of

view we couldn't do anything about it. So we have to try to make it work somehow but I understood that there is nothing more right than Dr. Lamb's stratification of grammar or that kind of analysis. But we don't know that much. We know far less from that. Even if we can, not everybody is talking about syntactical analysis, even if we can exhaust the syntactical rules we still have a lot of problems left at hand but I hope those problems will be solved sometime according to the type of phonetical structure.

However, whenever I have to make rules for the machine I have decided this much. I know that any one of the rules developed I can break myself. I can make many, many examples to break it but why do it, because we don't have enough in here but the rules that are given I may break. If I have more rules in here that rule will still stand. It is not a wrong rule but just one of those rules. Whenever the higher analysis is done this rule will fit in there. Those are the kind of directions that I have been trying to do. I don't know whether it is possible or not. But I do find problems in the kind of syntactical structures or, I don't know how to put it but I have enough examples that if you don't examine every possibility of

the forms you cannot decide or make a decision in which you will not make a mistake. It is very, very easy to make mistakes if you don't examine the complete sentence, every one of them.

*

The example which may be interesting is something like this, for instance. (At blackboard) If you have a string of characters, 3769 or 6638 0005 2116 0270 6344 0055, a very simple one probably. So we probably would like to find out what the structure is. Then I would go to the sentence through the grammatical markers. This is difficult to describe but every one of those has carried some kind of grammatical information but there are elements which may carry more grammatical information. For instance, if we find out here, 2116 in the third box, we want to know what kind of BA it is. Let's limit ourselves. If you go too far you can't describe it. We have only two. One is the first and one is the measure or classifier. We know enough grammar how to handle this. We have to decide which one it is, this or this. We have to make a decision on it. If it is this we know there must be a verb somewhere along here. We will try to ask, can I find a verb? Yes, in the fifth box. We must find whether this will agree with this.

* See p. 66A, 66B

II. 2116 2116 2116
 把 把 把
 ba3 [measure] bai3 (take) ba3/bai3

		2116 把 ba3/bai3			
		2116 把 ba3/bai3		6344 賣	
		2116 把 ba3/bai3		6344 賣	0055 了
		2116 把 ba3/bai3	0270 傘	6344 賣	0055 了
	0005 三	2116 把 ba3/bai3	0270 傘	6344 賣	0055 了

(three)

[measure for objects with handles]

(take)

(umbrella)

(sell)

[particle]

a.

這	三	把	傘	賣	了
zhe4 (this) (these)		ba3			

These three umbrellas are sold.

b.

王	三	把	傘	賣	了
wang2 (surname)		bai3			

The third Wang sold the umbrella.

We have to examine that. After we examine that it is agreeable so we may think this is a possible kind of BA. Then we go further in examination. 6344 in the fifth box and 0055 in the sixth box. These all agree. It seems that we are all right with the BA and we cannot rest too hard on it. So I try this one. This one I will find a noun, I hope. I find a noun which is 0270 with the umbrella, fourth box. If I know enough grammar I will see whether this noun will be suitable for the object of this verb. I have to find that out. Yes, I found that information. Yes, you can. All of that is right.

When all of a sudden I find out a number in front of BA, which is "Three umbrellas was sold" or something like that. There is a possibility that this BA is a measure and then I have to check whether this will agree with this or not, whether the 2116 will agree with the 0270. This and this agree and then what do we have here? Suppose I find a character, 6638. This one goes along with this very fine but all of a sudden I find a 3769 here. If I have the 0005 it turns out to be something else. So you cannot leave any loopholes. You have to examine every one of them and then you make your decisions so that you may or may not make mistakes.

This is not new. You will have the same problem in other languages. This is very, very complicated, as in Dr. Charles' paper which everybody read and now Mr. Samuel Martin. The first line is only to show the complicated grammar. It is what we talk about very often like other examples by William Wang and Mr. Peng I think talked about some of the ambiguous structures like something like this kind of thing.

(At blackboard) (N₁ Adj. 的 * N₂) If you have this structure, as Dr. Charles illustrated, there are two kinds of structures, I think everybody knows the examples in Arabic that Dr. Charles showed all the time, but there are more than that. This is not the end of it. There is a possibility that you can do this, this being the segment of the first one. As pointed out both of them are in the same IC analysis.

There is another IC analysis that would produce the other kind but this is where I would like to show you the way I am thinking, whether I am right or not. Suppose some of our good interests can produce some kind of semantic structures. What I mean by semantic structures, it has to be formalized. If we may have some kind of structure, this is a part of this, or this is part of this, the

* See p. 68A

III. N_1 Adj. ⁴¹⁰⁴ 的 N_2
de

1. N_1
 N_2

N_1 Adj. ⁴¹⁰⁴ 的 N_2
de

0361 0954 1170 4170 4104 2885

公 園 好 看 的 樹
gong1 yuan2 hao3 kan4 de shu4

park(s) good looking (de) tree(s)

(the good looking tree(s) in the park(s))

2. N_2
 N_1

N_1 Adj. ⁴¹⁰⁴ 的 N_2
de

2885 1170 4170 4104 0361 0954

樹 好 看 的 公 園
shu4 hao3 kan4 de gong1 yuan2

tree(s) good looking (de) park(s)

(the park(s) with good looking tree(s))

first element part of the last element, part of the last element part of the first. Then we can have some kind of light and at least eliminate some kind of ambiguities.

So the example I may show you is something like this.

公 園 好 看 的 樹 裡
0361 0954 1170 4170 4104 2885. In reading the LI is quite

possible to be eliminated. So "The good looking trees in the park" could be interpreted.

But if you find it in the other then you will have, just replacing the first element with the last, I don't think you can as far as these two kinds of ambiguities are concerned it can be seen how little we can resolve this kind of ambiguity.

As far as the other one that is a structure that looks something like this. There are two translations. One is "The child who loves Mrs. Chung" and the other one is "To love the child of Mrs. Chung." So you have two segmentations or you may have this or this. Am I right?

With a verb that is all right because you only have two ambiguities. There are more than two though. With a verb "to like" there are still more than two.

*

But another one which is "to fry chicken in oil or fat" you will find that in this kind of cooking

* See p. 69A, 69B

IV.

V	N ₁	的 de	N ₂
3498	7179	4104	3111
炸	雞	的	油
zha2	ji1	de	you2
fry	chicken	(de)	oil / fat

1.

V	N ₁	的 de	N ₂
to fry	the fat	of	the chicken

2.

V	N ₁	的 de	N ₂
---	----------------	---------	----------------

2.1.

3498	(6638)	(7156)	7179	(3938)	4104	3111
炸	(這)	(隻)	雞	(用)	的	油
zha2	(zhe4)	(zhi1)	ji1	(yong4)	de	you3
(zhe4)	(di1)					

the oil which is used for frying the chicken

2.2.

(6638)	(7156)	3498	7179	4104	3111
(這)	(隻)	炸	雞	的	油
(zhe4)	(zhi1)	zha2	ji1	de	you2

the oil of the fried chicken

IV.

V	N ₁	4104 的 de	N ₂
3498 炸 zha2 fry	7179 雞 ji1 chicken	4104 的 de (de)	3111 油 you2 oil / fat

1.

V	N ₁	4104 的 de	N ₂
to fry	the	fat	of the chicken

2.

V	N ₁	4104 的 de	N ₂
---	----------------	-----------------	----------------

2.1. (6638)(3336) 炸 (這) (隻) 雞 (用) 的 油
 (這) (這) zha2 (zhe4) (zhi1) ji1 (yong4) de you3
 (zhe4) (di1)

the oil which is used for frying the chicken

2.2. (6638)(7156) 炸 雞 的 油
 (這) (隻) zha2 ji1 de you2
 (zhe4) (zhi1)

the oil of the fried chicken

2.3. (6638) (0222) 3498 7179 4104 3111
 (這) (個) 炸 雞 的 油
 (zhe4) (ge) zha2 ji1 de you2
 the oil of one(s) that fry(s) the chicken

2.4. (6638) (0222) 3498 7179 4104 3111
 (這) (個) 炸 雞 的^(Pause) 油
 (zhe4) (ge) zha2 ji1 de you2
 the one(s) that fry(s) the chicken ... oil

2.5. (6638) (0143) 3498 7179 4104 3111
 (這) (位) 炸 雞 的 油
 (zhe4) (wei4) zha2 ji1 de you2
 the oil of the person(s) who fry(s) the
 chicken

2.6. (6638) (0143) 3498 7179 4104 3111
 (這) (位) 炸 雞 的^(Pause) 油
 (zhe4) (wei4) zha2 ji1 de you2
 the person(s) who fry(s) the chicken ...
 oil

verb it could be used as a modifier of this and then you will get instead of "to fry the chicken" it becomes "fried chicken". So it will work out some more problems in the kind of different classes of the verbs.

So here what my problem is I think I believe that the finer and finer grammar will help us to see the light. How can we solve them? The classes of the lexemes, for instance. Now I think that work is already developing into classes which up until now is quite sophisticated but I think this will be more than that to solve both of the problems. The more classes that develop, subclasses, the more you develop and you will have the problem of one belongs to one, two, three or more classes. I think this will become very, very complicated problems that will be solved but the little I can see we have to work on this in technical structures as much as possible and then we will need semantic structures to help us but the line between that I cannot see very clearly. Some of the problems I don't know where to put it. I think that is all.

J. WONG: Fred, may I interrupt? "The oil with which you fry chicken" and in the other case it is "oil that has been used for frying." Is that the difference?

FRED WONG: That is a very, very complicated

thing. I will talk with you later.

J. WONG: I just want to understand on the surface.

FRED WONG: On the surface I think this one you may or may not accept, "to fry the fat of the chicken." You may not accept that.

J. WONG: No.

FRED WONG: Okay. "To fry the oil of the chicken" or "the oil of the fried chicken" or "the oil is used for frying the chicken." Another possibility is "the one who is frying the chicken in his oil." This could be a person who was doing the frying of the chicken.

CHAIRMAN SEE: It is not "the man who stole the chicken" but "the man who fried the chicken in his oil."

FRED WONG: It could be interpreted as "The oil belongs to the one who is frying the chicken." Or you may produce some other kind of structure, "The one who fries the chicken his oil has run out."

J. WONG: Thank you very much. I understand now.

FRED PENG: Related to that "Three umbrellas are sold." There is something about the 2116. ^把

FRED WONG: I started with two possibilities.

FRED PENG: Ambiguities can be solved if we set up more than two criteria. We know that by eliminating one possibility and following by 0008. The second criterion is if BA is processed by any numeral beyond 10 then the second ambiguity is partially solved because you can say 1-JO or 1-BA as a short form. Up to 9 you can have ambiguity but beyond 10 it is partially solved. If you go to 1120 or something like that then ambiguity doesn't exist. So I think this is a commentary on this.

J. WONG: Zampa zampa.

FRED PENG: This is a double.

CHAIRMAN SEE: Did I understand that a number over 10-- let me give you an example. "He bought a \$13 umbrella." He took it and did something with it. "He took a \$13 dollar umbrella and gave it away."

FRED PENG: If BA is preceded by a numeral beyond 10 then the ambiguity doesn't exist.

FRED WONG: Could be.

FRED PENG: A person may be 1-JO.

FRED WONG: There are many, many markers which can help us to solve the ambiguity.

BILL WANG: May I submit there is a problem more to birth control than linguistics. (Laughter)

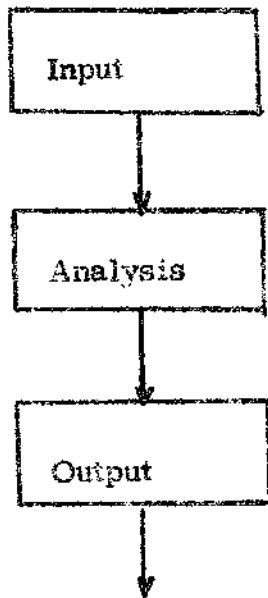
S. S. SOO: Mr. Wong has described the basic general thinking behind our process. I am here to give you a general description. Our work can actually be divided into two phases basically. Phase 1, is what we have done up to June last year under our RDC sponsorship. Phase 2, is what we have done since then utilizing what we have learned from our first stage of endeavor. This will be discussed very briefly.

The work can be divided again into two different parts, software and hardware. I will cover hardware in a very quick manner. In the computer which is basically a photocomputer your entire program is on the photodisc. For Phase 2 we are using the research language processor which is a much more sophisticated machine which includes control by said photodisc and this allows us to perform our analysis much faster and also much more sophisticated programs.

Software I had certain presentations in mind but somehow or other I had to change it at the last minute. I will go ahead with it anyway.

(At blackboard)* We can divide our work into three stages mainly. From this we proceed to analysis, input to analysis, and then we proceed to output. Now

* See p. 73A



(1) Resolution of POS. Ambiguities

(2) Function Word Analysis 把 了 2116, 0055

的
(3) 4104

during input Chinese text which has been reduced to machine-able form in this Sino-writer the three digit code -- for that matter any 1 to 1 would do -- is introduced into the machine and by means of longest match -- there would be dispute as to relativity of longest match but I would leave that aside for a moment -- we get an intermediate machine word and we have essential grammatical information such as parts of speech. We would also include the reference such as the English equivalent and inflection and the input is reflected automatically.

We then proceed to analysis. This conception can be divided into three parts. One is resolution of possible parts of speech ambiguities. Much work has been done on ambiguities. Other work has been performed on ambiguities such as noun-adjective, adjective-verb ambiguities. That is one part of analysis.

The second pass is function word analysis. During this part words such as BA, ^{把 (3)}LU, we look at them because they are syntactic markers of sorts, ^把2116 and other related words are examined. We then proceed to a sentence in an attempt to isolate the related phrases that belong to these words and thus attempt to work out a structural sentence.

的 The third part we place special emphasis on 4104. As Mrs. Dougherty can point out, it is one of the most frequently used words in the Chinese language and it does help tremendously at least as a starting point for analysis.

At the end of this we hope to have the machine words with information, the word order in a sentence, the position in a sentence related to others, and also necessary English inflection if any. Such as for a time we decided if a verb is followed by ³0055 this verb would then be in the present perfect and this would be tagged accordingly.

Then we go to the output stage. During this two things are done. First, the word order is rearranged such as the Chinese word order. The sentence is not translated in Chinese word order but rather in the equivalent English word order, let's put it that way. During the second stages we automatically inflect the conative form as dictated by the syntax, and from this the entire sentence output. This operation applies more to the stage one work rather than the more recent development.

In the more recent development instead of having fixed passes we go through the sentence and wherever

necessary call in subroutines which may or may not take in additional parsing. The amount of time you spend on a sentence is not dictated by a preprogram but is rather dictated by the complexity of the sentence you encounter. Thus we can add to our analysis or routines as many subroutines as we think we can use and develop the scheme further and further.

We have also attempted to define certain phrase structures for a basis for further working wherein we hope the analysis step is followed by a synthesis step. In other words, we hope to have an integrated grammar and then the other grammar. Stage one has been running on a selected process. Stage two is on paper. There has not been any sort of accuracy or elegance of translation.

That in brief is the program at IBM.

BILL WANG: The rule that interprets the noun after the verb is not sufficient, is it?

S. S. SOO: I agree it is not sufficient.

BILL WANG: For instance, $\overline{0005}$ ^(?)_(0055?) it could be the imperfect. It could be "I have gone" or "I am about to go now."

S. S. SOO: Our problem, as Mr. Wong has pointed out, it is indeed rather complex to attempt to resolve our

problem. We have decided that it is more possible to use it as past perfect and therefore we have done it. We realize there will be certain errors introduced.

CHAIRMAN SEE: Are there any more questions before the aromas attract us over that way?

You say it is running. You mean on the disc? What equipment are you running in phrase work?

S. S. SOO: Phase 1 can run on the computer.

CHAIRMAN SEE: I didn't say can but what do you run on it?

S. S. SOO: That machine at present is at Wright Patterson Air Force Base. We have not attempted to move the disc over to the machine because at this point having Phase 2 which I think is a more sophisticated program it would obsolete Phase 1.

DAVE LIEBERMAN: There is a simulator that can be used. There is an actual program. You put it in the program and take it out on the 44 simulator. It is a real simulation of the mark.

CHAIRMAN SEE: I understand. So you can run it on the 7044?

DAVE LIEBERMAN: In a demonstration, for example, it could be put on.

PAUL GARVIN: Have you retained the mark terminology? Is it going to be called Mark III or --

S. S. SCO: I cannot answer that but I personally refer to it as a research language. What the higher-ups decide I am in no position to know.

DAVE LIEBERMAN: I will tell you again. This is just a different name for the machine that follows the mark. Some people call it the Mark III because it is easier.

PAUL GARVIN: That is not the same as the RLP?

DAVE LIEBERMAN: You will find a lot of names and they more or less refer to the same thing.

PAUL GARVIN: In other words, IBM Research Language Processers are more or less the same thing, okay. I am asking the question to know what it is when IBM people are talking about these different pieces of equipment.

DAVE LIEBERMAN: Mark III is a slang term.

PAUL GARVIN: It is not in your lexicon but it is in those of others. It is a dialex, am I right? If there are diophones why not have diollex?

JULES MERSEL: What is the improvement? XW-2 is at the end.

S. S. SCO: Very briefly it allows us direct

addressing in the first place. It also has a more sophisticated search. Well, I could give you an outline but I don't think it has a place in this meeting. It is more of a pure program technique.

DAVE LIEBERMAN: I will give you an analogy. Take the main deficiencies in the Mark II and list in order and go down in order and you will have the Mark III. Or speak to people in ITEK and I will bet you would be pretty close to the Mark. This is a natural development.

SYDNEY LAMB: Has IBM licensed ITEK to use the patents?

DAVE LIEBERMAN: I think the patents are owned by everybody.

JULES MERSEL: I think Ampex.

SYDNEY LAMB: Ampex? Didn't Telemeter own it? Then everybody has got it.

S. S. SOO: The Air Force has the rights so anybody who does work for the Air Force.

CHAIRMAN SEE: Vic has some announcements to make concerning the proceedings.

VIC YNOVE: We will move over the chairs that are in that cart in the back of the room and put them around the table and I think probably about two-thirds of

us can sit around the table and the rest will have to use these chairs at this table and some people can go back in the corner. After lunch we will reconvene at two o'clock.

(The meeting recessed for lunch at 1:00 P. M. and reconvened at 2:15 P. M.)

CHAIRMAN SEE: We will resume with the IBM presentation for which we will allow ten minutes and allow five minutes for questions.

DAVE LIEBERMAN: First, a slight correction. There is a disc. The Chinese program as described was put on a program and run and the reason Soo said it can run on the ABCDEF0 as in Dayton but it doesn't matter.

I thought I would finish by describing what our future plans were and what we might expect might come out of IBM. Firstly there is a possibility of going into operational machine translation. This I think will depend on whether there is a customer for it. If somebody says they want to expand the lexicon to do the usual stuff and is willing to pay for it I am sure that IBM wouldn't say no. Whether they would push as hard as they did Russian I don't know.

The other is whether to continue in the experimental way. I speak in operations only. Even if you know

beautiful rules that will work that might cut down the processing time you can't use them. However, with the same operational in mind it is whether they can be done. In other words, whether we go as fast tomorrow or not there is a possibility this might be done but again I don't know. I don't have much to say about it.

Now the third thing would be what we do in the linguistic group itself. There we are in a way the same as similar groups as Texas and MIT in that we have a hopper so to speak that has been developed for processing English and possibly Russian and we hope to throw Chinese in the same hopper. The environment for Chinese, we are working on a comprehensive dictionary following the MIT theory very closely. Ed Climer of UMAT is our consultant. If you know what the current MIT theory is you know what we are doing but the aim is to make the grammar comprehensive, to include a large lexicon, something of the order of 15,000 or 20,000 entries, to eventually get to the point where we can measure coverage. In fact, we are beginning to measure coverage right now. By measuring coverage I mean to take random text and construct correctional analysis for each sentence.

Now along with this we are developing a

preprocessing program language. Again there are a lot of considerations that go into it as goes into soft hardware. We considered using Comet but felt it was not worth-while putting into the kind of things built into elementary operation. The language is specified now. One programmer is working on it now and he estimates nine months. That means a year and a half but we may get more support.

One other comment on this language. I wanted to see a language that was designed specifically for linguists and in fact I wanted to go even further and say it was only for our use at IBM. It might turn out to be a great general purpose language but I wanted that to be for the future. I fought with the guy who designed it and I almost won. Then IBM gave \$1,000 to about ten people who were taking part in developing FORTRAC and after that they couldn't control it. So it is not as close to a special linguistic program language as I would like but, on the other hand, it is a more general presupposing language. It was meant specifically to be for linguists' use but it has developed into a preprocessing.

Now in addition to the work on the transformational grammar from English we are also working on an approximate grammar from English. The idea behind it is

that it is not an unusual thing for the theory not to be usable in its pure form for any kind of experimental application, not to mention practical application. In fact, it is unusual if it is the other way around. Even in hard theories like physics approximations are necessary. So I think the notion of having some approximate grammar is a reasonable one.

By approximate one I don't mean an approximation to English but I mean as a system that is approximate to the theoretical form. Now I don't mean as good as the 1965 meeting at what I was talking to some of you on at the Denver meeting. It allows categories, restrictions, that is about it. That in a way it can express some of the deep structures, that is it can express some of the deep structure relations and some of the surface structure relations and in an ad hoc way the meaning between them. But the main purpose of this thing is to allow us to have a sentence recognizing routine.

We are also developing the analysis by syntheses method as talked about but that is a very long term job. We needed something with a much shorter payoff. So we have the formalism for writing grammar and a sentence parsing procedure which operates on the formula and not a

particular language.

Any work we will do in Chinese will be in terms of this approximate formalism. It will be motivated as far as possible by other work that is done, particularly transformational work that comes out of Bill's area but the formalism will be the approximate formalism.

Now before I asked Syd why we should analyze questions completely and he said it was a ridiculous question. He said why should we get all analysis.

SYDNEY LAMB: I was talking about something entirely different. I was talking about getting all segmentation.

DAVE LIEBERMAN: Leading to analysis.

SYDNEY LAMB: That is a different thing.

DAVE LIEBERMAN: Okay, either way. The parsing program I spoke about will get all readings for the sentence and will characterize the sentence as grammatical or ungrammatical. That is relative to a given grammar it will give readings or won't and if it can't we will say it is an ungrammatical sentence.

The reason I made the comment, even though there was a misunderstanding I think it is necessary to give all readings for the sentence but that is not enough.

That is perhaps the minimum unit of need. But there is no sense for getting 500 readings for a sentence because you have a grammar that is to read and there is no sense in getting an algorithm that will take twenty years plus the next generation of machines coming through. I think the association with recognition procedures is as important as the procedures themselves and we are paying a lot of attention to this. Again I am not trying to sell you on it but letting you know what our thinking is on where we are putting our efforts. I think that about concludes it.

SYDNEY LAMB: Since you bring up that point let me ask, do you consider the question of getting all possible syntactic analysis? It happens to be the case there is an algorithm for doing this. There is no practical question of taking too long for the machine to get all possible analysis.

DAVE LIEBERMAN: If I saw it running I would use it. I am much more interested by what we mean by structural. I am interested by what we mean by it and how we can use it. I am not that fascinated by getting it by machine. Yes, I would be happy to use it. However, it might turn out that it would work for one kind of

formula and get you all possible analysis from this point of view.

JULES MERSEL: What is your general modified translation written in?

DAVE LIEBERMAN: FLPO.

JULES MERSEL: What machines will that run on now?

DAVE LIEBERMAN: It will run on the 94 at first, FLPO. Why are we doing it that way?

SYDNEY LAMB: My point there was that this practical matter of the possibility that the machine might take too long is no consideration because it doesn't turn out to be the case.

BILL WANG: Is there anybody working on the experiment?

DAVE LIEBERMAN: Yes, right now Fred Wong is the only one working on Chinese.

BILL WANG: What is the magnitude of effort thereabouts?

DAVE LIEBERMAN: Well, he is working as a consultant. Approximately it will work out to about one day every two weeks for the rest of this year. It isn't ethical I suppose to proselytize now or I would tell you

what effort he would support.

PAUL GARVIN: You can always mention it afterwards in socializing. We have no counter-offer, unless ITEK has a counter-offer. (Laughter)

DAVE LIEBERMAN: I said it is being thought over. The people who did the work on the Chinese work are still there. So at the moment is not working on Chinese but it could be easily resurrected.

CHAIRMAN SEE: Any further questions? I have just one question. You mentioned the use of Matthews' analysis by syntheses. You mean through the full transformational route?

DAVE LIEBERMAN: Yes.

BILL WANG: But the impression usually one gets is that it is very complicated.

CHAIRMAN SEE: I should have said part one by the old pre-Indiana or the post-Indiana model.

DAVE LIEBERMAN: It is not that the structure makes it simpler than the deep structure. But still I don't agree that it is simple and I don't expect that we will see any agreement floating around in the near future.

CHAIRMAN SEE: Bill, I don't think, is planning to analyze that way, not that I heard, but he thinks it is

hard. Dave thinks it is very difficult but he is going ahead to do it for early solution.

SYDNEY LAMB: One point, it is extremely difficult and therefore wrong.

BILL WANG: I don't think it is easy.

CHAIRMAN SEE: But you think it is easier than he does.

DAVE LIEBERMAN: Analysis by syntheses is simply a cover term. It means any kind of recognition is an analysis by syntheses even if the syntheses was as simple as a table lookup.

SYDNEY LAMB: Why call it syntheses? It is to make people think what you are doing or that you won't have to admit you have changed your opinion.

DAVE LIEBERMAN: The same reason we talk of predictor analysis. That is the name that Hugh Matthews has given to what he is doing. Until other people begin to write and the terminology gets bad enough to remove it I think we will have to leave it.

CHAIRMAN SEE: Well, the time is now ripe for the Bunker-Ramo group. Paul, are you going to lead off?

PAUL GARVIN: I want to engage in the sincerest form of flattery and do what other people have done,

namely thank Dick, Vic and everybody else for the gracious invitation and now I have the opportunity to thank you for the marvelous lunch and the organization in doing it.

Speaking of organization I thought I would give you an idea of how we are organized or rather disorganized because unfortunately we are losing Jules Mersel who used to be our Department Manager and I think at midnight tonight he becomes our consultant.

JULES MERSEL: No, midnight last night.

PAUL GARVIN: I have already lost you. Some of you may know from corresponding with you that our department has the strange and wonderful name of Synthetic Intelligence Department. Previously language and language analysis was going on but there wasn't a corporate designation or box for it. Now there is a box which is called Language Analysis and Translation on a translation chart. I thought this would be a welcome change from these other boxes that are up here. It is designated as something within the Synthetic Intelligence Department. There are lots of other branches and then there is ERC which is the designation for Bunker-Ramo Corporation. Down here is my name now which goes to show you that I am to all intents and purposes manager of a box which is, however, not filled

with personnel. This is a functional area. It is not a section and I have been so informed. This means in fact that I don't really have to make management decisions about, say, washrooms or office locations and it is all to the good.

Now to turn from organization and disorganization to intellectual confusion let me just say that all of you know that the activity, for the sake of simplicity I will refer to ourselves as Canoga Park because we have gone through a number of names. Some of you will remember there was at one time Ramo-Woolridge. Then there was a division of Thompson. Then Ramo-Woolridge Division. Then it was known as Woolridge Computer Division. Now there is a Bunker-Ramo organization and unless I am mistaken we are a definite division unless they changed yesterday. The thing that has remained constant is 91304 which is the Zip Code for Canoga Park. This is a town, a community in the City of Los Angeles. As you know, or if you don't know you might want to find out, there are cities and communities in the greater Los Angeles area. It is part of the City of Los Angeles, whatever that means. However, Canoga Park has remained constant where the name of the company has not. Consequently we like to say that out in Canoga Park this is

what we do.

Out in Canoga Park there has been machine translation going on before they moved to Canoga Park. Before that it was in El Segundo, a city. We have been demoted from a city to a community, as you can see. I have been associated with the activity since 1959 and now I suppose one could say that the activity is associated with me which is a change in my status although not in that of the activity. This has been concerned with machine translation and other things and in general we like to keep the two fairly separate in terms of contractual and other obligations. That is, we have some non-machine translation activities that are under reasonably separate headings and then machine translation headings where the aim is in fact to produce translations. This has been the aim ever since there has been an interest in this in Canoga Park and I was fortunate enough to find a positive response to my own convictions in the matter which are that in order to do machine translation one would have to employ certain variance on certain empirical linguistic methods under the name of fulcrum, whatever that means. In fact, the purpose is to develop an algorithm that will process one sentence at a time by looking for hotbeds of

information and branching out from there.

In order to reach this hotbed certain preliminary steps have to be taken in the algorithm and this has led to a pass method where we have a number of preliminary passes to establish the major searches. The major searches are those that everybody is interested in what we like to call the clause members for components, such as the predicate, subject. I am now talking of the Russian-English activity. We have a dictionary with a grammar code and we look at the grammar code and go through various passes and hopefully establish a reasonably tolerable parsing.

We have used a term that was invented by a computer man from Detroit, Charles Briggs. He has used the term "sentence image." Nobody else but I use it and I use it infrequently so I thought I would bring it in as a new term.

We use a collection of codes indicating what have been completed successfully and what have been unsuccessful. As a matter of fact, we have some indications of failure in our output which I understand is the so-called fail safe device.

Now we have decided to go into Chinese and he made this decision two years ago when we were fortunate

enough to be under an NSF contract where we could explore our approach to other language pairs and Chinese suggested itself as an interesting possibility. At that time I was able to spend time with a Chinese speaker and use what I consider reasonably good field methods and work out a very preliminary survey of the problem areas.

Since then we have been fortunate enough in finding support from RADC, the Royal Air Force Development Center of the Air Force, the same as is sponsoring the Government portion of the ITEK effort. We are now proceeding in this area of developing an approach to English-Chinese translation.

Now we are operating on a small scale. The main reason for this is I believe that any activity should start small and then grow if necessary. I think that in the past mistakes have been made in this direction which should be judiciously avoided. In order to proceed on a small scale we have decided to map out a particular area of interest and on this syntactic interest and our approach to this. That is to say the fulcrum approach combined with a pass method. Our philosophy of research is one for giving in to and learning by doing. So we have begun by drawing flow charts. I think that other people might have preferred to

go through some basic linguistic analysis and develop flow charts on that basis and then revise the flow charts. We thought we would derive our analysis into the flow charts as we have done in the case of Russian and we feel that we are probably not going to be any further wrong than other people who do differently and at least we will be on the safe ground of doing it our own way.

This means we visualize a succession of passes beginning with a dictionary lookup of some sort and ending up with some sensible way of outputting it to printing equipment such as may be available in the future. Then we want to pick out in this whole large series of processes that which we will immediately be concerned with and that, as I said, is the syntactic portion.

So we pretend that we have a dictionary lookup. We don't bother writing one either on paper or in a program and we are justified in doing it since there are groups working who do in fact have dictionaries and do in fact have fairly good dictionaries. To our knowledge there are at least three Chinese-English dictionaries in existence and since we have begun working on it we are flexible enough to adapt to other people's grammar code if necessary. What we are interested in at this point is not to

do anything. We do want to give our own approach a fair trial but we do feel it is flexible enough to adapt to components available. Therefore we assume there is a dictionary lookup.

We further then have to face the problem of the different portions of the syntax that we sort of visualize. We visualize that at the end of it there will be some kind of what we call major syntax where we dig out the main sentence components which I mentioned before, the subject, object, predicate and so forth. Preceding this there will be prior passes that will package up the portions of the sentence before the main portion of the clause members. Prior to this most likely there will have to be some resolution of word class ambiguities.

There are two ways then for going about this thing. One is to start from the beginning and then work your way down to the end, and the other is to jump into the middle. We decided to jump into the middle because we feel that making some assumptions for what has been accomplished previously and leaving the rest until later we can then work out a nice middle portion in such detail so we can do both the beginning and end more efficiently afterward. In fact, we are making the totally unwarranted

assumption that all the words we have processed are from unambiguous grammar codes. Somewhere along the line there will have to be amorphous illusion thus making up a necessary grammatical package.

Presently we have the good excuse of only being at it a month and a half. I was gone for the summer and we have only had Fred for about six weeks. Consequently we do have a short span that we have been working at it.

CHAIRMAN SEE: There was the previous brief go around and it also was not directly connected with this.

PAUL JARVIN: Well, I think it is only connected to the extent that it formed a basis for our proposal and laid out, I thought, some general principles. So far we haven't seen any reason to reject it but it was too broad to be compatible to the present. Now we have flow charts and arrows and so forth which we did not have in the first survey. We merely said when you look at the Chinese it seems as though you could have a pass method. When we actually do it we don't do all the passes. We make some unwarranted assumptions such as the homographic illusions completed what would be the next thing to do and that kind of thing. We feel these are reasonable working assumptions and then we are interested in ascertaining the boundaries

of syntactics on the phrase level, that is what are word phrases and recognition. I will not go into the details. I think that Fred will be the one to most logically talk about that.

I will only say that our basic principle at present is the following. We are working on two passes in fact. One is designed to ascertain the components of phrases that we call expressions. These are word groups or groups of elements, if you don't want to use the words, which are more than one element which are not a whole phrase. Then the second pass is concerned with phrases.

The second thing, we found it worth-while to direct our search from right to left in the sentence. This is based on the observation that most commonly the head of a construction is the right-most member. So to start with the right to left seemed to be more efficient than left to right. I will leave that to Fred. As a matter of fact, I should leave it to Jules to go on next.

I think we should answer questions as a group rather than individually. Thank you very much.

JULES MERSEL: As Paul indicated I am now consultant with the group. We haven't changed our name as recently as Bunker-Ramo. We have had the name for about

two years. It is a group of essentially software people with a high percentage of Bunker-Woolridge people at Bunker-Ramo. I will be available for consultation as long as he feels he needs me which I expect won't be very long.

The area I am supposed to talk about today is program but, however, when we sketched out the Chinese-English some years ago we learned we had learned something from the Russian that there was no point of sorting at the beginning on the input problem and then going on with the dictionary and lexicography problem. Others who were assigned to this and the activity is in full swing. Syd Lamb and Ching-Yi Dougherty were building a fine lexicon at Berkeley and IBM was also creating equipment for dictionary lookup. We were going to leapfrog to the syntactic analysis and completely ignore computer problems.

Now as you listened to Paul I think that you found that something had been learned in the whole period we had been dealing with Russian. I think that after six weeks any of the group in Russian would speak with the attitude they had solved old problems and were ready with definitive solutions. Nobody has been speaking that way here. One thing I think was the inadequacy of building our own computer routines. Vic was speaking of UCLA.

I got up and said, "Wait until he gets it running and see." It may be a year and a half. I don't know how long it took to get the thing running. The point is that COMIT has been running for some time now and you have it running on at least five different computers. I would like to urge for the sake of avoiding some of the mistakes we made in the Russian-English translation that now there is too heavy an investment in computer programs, that we adopt some common program language. It is going to be a lot more important in the Chinese-English than the Russian to English. The Russian at that time had 709 around and we were faced with 709 and 704. I think you are going to see most of the computers you have been programming for disappear. IBM is bringing out the 706 which will have competition from ITEK and a number of other companies which is going to make it really difficult for us to communicate with each other and use each others' routines. It will be a lot easier to adopt some language now so we can run other routines on whatever computer happens to exist at the research center we are at. I think I would really like to urge at this point because it is designed language. I think this was the base language to the CTV that this would have a possibility for running on other people's

computers.

Now this has not been a description of what we have been doing but rather an exhortation that, as a matter of fact, we haven't been doing a program on the Chinese itself but Fred has been building the flow charts that will eventually be programmed. Now if there is no correction from the Chairman I would like to throw this exhortation open for debate or acceptance or rejection.

FRED PENG: Before I begin the detail discussion of this portion of the presentation I have a couple of corrections to make in the handout and please turn to page 8. On the last diamond on the second column, the middle column, the last diamond please correct A into F.

The second correction is on the right column you have two boxes there and on the bottom one it says "assing". It is a typographical error. It should be "assign".

Now I am going to mention a few things in this portion of the presentation. First, in addition to what Fred Wong said about the complexity and difficulty of the Chinese language I will have to add one more thing I strongly feel is very essential. The difficulty I have experienced is this, that if you take two characters

together it seems to me very difficult to identify whether you are going to take this as a unit or take this as two units.

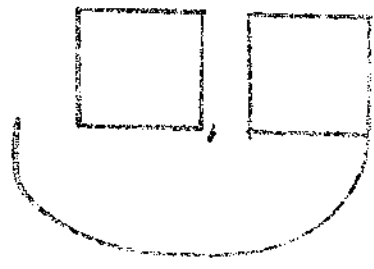
Take an example, ^{以前} 0110 0167. You can have a list of substitutions over here and the question arises then whether this is a unit or separate unit and make notes out of this. This is extremely difficult in Chinese and unclear to most of us.

You can go like this. Is this a unit or two units? This has been my concern. This goes into other parts too.

For example, ^{*} the first one is ^{(22) 以} 2395 0110. The second one is ^{總統} 4920 4827. ^(20% 所)

So we have sort of arbitrarily chosen this as one unit and we may be wrong and will change later if we are wrong and work sort of a flow chart of system for our flow chart. The first time I did it, as Paul suggests, do it sentence by sentence, I started from the left-hand side of the sentence and going from left to right. As soon as I started I was confronted with difficulties so I brought this up to Paul and asked to discuss the problem with him. Then we decided to do it backward. If you input a sentence and start from the right and going to the

* See p. 101A



以前
新以
然純

left, so our flow chart we have in the handout is designed from the direction of right to left.

Now the reason why we do this is twofold. First, as Paul suggests, we are following the fulcrum technique. I intuitively feel that if you take this as two units you don't know which is the head but if you take this as a head you know this is the head intuitively and this is the attributive. So we do this, 1129^{*} is the attributive.

So in order to overcome this difficulty we have designed a sort of system going from right to left and I would like to have you take a look at your handout and recall this now as an element instead of a word because we don't want to confuse some experts by telling them this is a word or this is a phrase or what have you. This is an element and this would be two elements.

Another thing we strongly feel about is the very fine taxonomic classifications for these units and besides Canoga Park I think that Berkeley is the only group that has a very fine and very good taxonomic classification of this type of unit. I will leave this to the group and I have studied their presentation carefully and find it interesting. We are prepared to adopt their system if necessary and we are ready to accept any

criticism and any contributions that could make our program run more easily.

Now on page 4 you have a sort of grammar code and this is based on my taxionic classifications. I could go finer than this but this is the basic code. If you want the telegraphic code it is on page 10.

The grammar code for the first pass, as Paul mentioned, we have at the present moment a two pass system. The first pass identifies the elements which is hand simulated. We are doing the manual work at the present moment. After we identify the units on the first pass we come back doing the same sentence on the second pass. The second pass we are identifying two elements to become another kind of a unit which can be called a noun phrase or a verb phrase. At the present moment we are doing this.

Now if you want to follow this flow chart you may want to borrow this IFEK handout and take an example out of this and see what will come out. For example, in IFEK's handout you have just taken the number 2.

Now in the parsing rule we have a phrase like
 農村 的 社
 6593 2625 4104. If we go from right to left 4357 will be
 合 作
 an element to us. 0678 and 0155 together will be another
 element and we hand simulate them and put them down as

two elements and then go on to 4104^的 which is another element. Then we go to another one, the next two elements, which is two characters which are 6593 2625^{農 村} together will be another element. After we go through this we come back again and do the second pass which combines the first three characters which is 0678 0155 3457^{合 作 (組)?} together as another unit. So it goes. This is the way we do.
(社 土 4317)

At this point I have only accomplished a little bit. This work is subject to any criticism and we will be very delighted to hear your reactions.

CHAIRMAN SEE: I would suggest we first have specific questions on Bunker-Ramo and then later on we could respond, if you want, to Jules' proposition that we use an appropriate language, or perhaps it would be better to wait until the very end to get into that. Now are there any questions directed to the presentation per se?

What is the sense of the group? Vic, do you think we ought to hold that question on common language to the end?

VIC YNGVE: It is not specific on Chinese per se.

JULES MERSEL: I intended to.

CHAIRMAN SEE: I think it might be better to get through all the groups since this isn't a particular group

but all together.

Having said that, who is next? I believe it is Ohio State would be next, Bill.

VIC YNGVE: While this is being handed out, when we handle the transcript we would like to interleaf the handouts in the proper place. Please send us enough copies, namely about thirty, of the handouts so we can interleaf them with the transcript in the proper place. Send a few extras if you have them, please.

CHAIRMAN SEE: Thirty-five, say.

VIC YNGVE: It will make it much easier to read the transcript if the handouts are right there.

BILL WANG: I won't take up much time. I will just briefly go over this first part on the project on linguistic analysis. Most of the time will be taken up by the person who did most of the work which will be Ann.

Page 2 you have a breakdown of the people who participated in the project. Of the people listed here most people are not budgetary working on the project but are intimately related to its activity.

Of the faculty members Fillmore, Langendoen and myself are in the linguistics division. Langendoen is a recent Ph.D. from MIT and Meyers is in the Department

of Mathematics.

Of the research associates Hashimoto is a very accomplished Sinologist. Professor Lu came from Taiwan this year and Sakai came from Japan.

Of the students listed some are in mathematics but most of them are in linguistics. On the next two pages you will see a selected list of the things we have written. Many of these are in the form of work papers and not publicly available as yet. There are two types of items that are available on the list of items on the next two pages. One is the articles in the FOIA report which probably most of you received, and then there are the things that are available in the Oakland journals. For instance, Fillmore's Indirect Object Constructions will be coming out as a monograph published by Mouton and Company which is the chief publisher of linguistic things.

Perhaps it is kind of difficult to reconcile the titles for the work we have on this bibliography with the notion of Chinese machine translation. I think the fact that machine translation, as pointed out before, is quite misleading. It at once says too much and too little. It says too much in the sense that it is the feeling of giving something imminent or even possible and I think these

are controversial points. On the other hand, it says too little because the type of research that is stimulated in this general area is almost conterminous with linguistic research in general. I think certainly the type of information for any kind of mechanized translation possible requires more than all the information that the linguist can present in form.

Of course, the type of information that the linguist is culling from language in the form of linguistic analysis is useful. It is the only field that provides scientific fields for linguistic behavior.

I say these words to try to relate the type of work that we have been doing. The type of work that we have been doing since the latter part of 1961 is specifically devoted to the discovery of regularity in Chinese and in English and to formalize these regularities in terms of rules.

I believe that this type of information is a necessary though insufficient part of any kind of an MT routine.

Of course, when you are working with particular languages you approach the languages with a prior set frame of mind, that is you approach it with a theory.

As you interact the data with the theory both are bound to change. The theory will become more systematic as the result of your organizing the result of the theory and the theory will change because as you try to account for more and more data you find that the theory has become inadequate.

In Ann's presentation she will give you the sort of precise and specific discussion for applying a theory to a language, Mandarin.

I would like to take two or three minutes to say something about the theoretical framework. It has been mentioned a few minutes ago that there is such a thing as a post-Indiana model. I think this characterization is inaccurate. What we now take to be our model says the structure of a language is being divisible into two levels, a surface structure level and a deep structure level. This particular distinction certainly does not date as recently back as this past summer. In fact, it was specifically mentioned in Harker's textbook in 1958, although he didn't pursue it much and only did a few paragraphs on it. He said the deep structure actually portrays the regularity and components of a language whereas the surface structure we encounter either on

paper or in the form of acoustical ways. This essentially is the most crucial distinction that underlines our model.

Actually it goes back much further than 1958. It goes back farther than Harker's book. It has been, for instance, especially pronounced in the work of the late French linguist Lucien Tesniere whose published book "Elements In Syntaxe Structurale" is probably the most detailed type of analysis of the translation from this point of view. The terminology is different. Tesniere talks about logrations dual and logrations initial which has a very good correspondence with deep structure and surface structure respectively.

So in using this particular model to analyze these languages I think this model has been arrived at partly through the effort of our group working on the theory of grammar, let's say, in its improved form over the earlier model for transmission of grammar. Much of the reason for coming to this new conception of the theory of grammar can be found in Fillmore's article "The Position of Embedding Transformations in a Grammar" in which important observations were made on the grammatical rules. For instance, it was found there is no order in relation

on generalized transformations. This was never specifically stated before but because there is no order in relation among generalized translations it is possible to remove them.

Also it was observed there whenever there is a situation of embedding one sentence into the other, embedding the constituent sentence into the major sentence, if something happens it always happens to the constituent sentence so that the rules actually operate inside out from the constituent sentence across the embedded sentence and if this sentence becomes embedded it is an embedded sentence of a larger sentence and so on. This was made specific and imparts a new understanding of the language which now is much more in conformance than the language of Lucien Tesnière.

I guess that those are the words I wanted to provide as a framework for our presentation and perhaps there is a small part of our research that is involved, actual mechanization of what we found, that is actually try to program some of these rules. This has been done under Professor Meyers and after his presentation perhaps Ann can read this condensed account of Mandarin Syntax which is some sort of a summary of the bulk of the work

we have been doing in the last couple of years.

SYDNEY LAMB: When people refer to the post-Indiana model, of course that isn't the correct term. It is a few months older. What they are referring to is the new version of transformational theory because outside of transformation it has been known for years.

BILL WANG: Syd, you and I have differed on this for a good time. It is certainly not a new idea.

SYDNEY LAMB: When people talk of the newness what they are talking about is the newness of this idea into transformational theory.

BILL WANG: On the other hand, the doing of this is very important in transformational theory.

SYDNEY LAMB: Not only transformational but any theory.

BILL WANG: For instance, if you have a pair of sentences "The man saw the book" and "The man played tennis" and you embedded one into the other, depending on what you take to be the constituent "The man who played tennis saw the book" or "The man who saw the book played tennis." In order to discover the relation you need a marker that preserves the history of the rules implied in the final sentence. This is mechanized because

of the large research that you have to do repeatedly. This new notion between the deep structure and surface structure difference you actually don't need the history of the rules.

SYDNEY LAMB: This has been known for sometime and now what people are remarking on is that it has been a discovery.

BILL WANG: I don't agree.

SYDNEY LAMB: In other fields of linguistics, paths of linguistics, it has been known as long as there has been linguistics.

VIC YNGVE: Specifically there is no order relation in generalized transformation. It is this very fact that led to the left to right structure. But this is the very thing that led to the left to right phrase structure that we have been using which eliminates any ordering between the things we have in embedded transformation. This again is a very old thing in linguistics. I think we should welcome the fact that this is now being recognized by transformation. It is a sign of advancement.

BILL WANG: The study of language of course has been an old thing. In fact, it is older than American linguistics or any tradition but if we do not exempt,

let's say, a particular theory including the difference between deep structure and surface structure I think there will be several types of syntactic structure that cannot be made.

SYDNEY LAMB: Bill, you don't have to argue for this because I don't think there is anyone who characterizes.

BILL WANG: I think in your general presentation of the framework you have met to a very large part and we feel very close, we have close feelings about this, but there is a very crucial difference and that is in the format of the rules that we want to use. You have, I think, somehow arbitrarily and unreasonably imposed excessively severe restrictions on the format of your rules. I think with that kind of restriction you are not able to exhibit the underlying circumstances. On the other hand, if you allow your rules to go around more than one I think other things become more regular. So in answer to your objection I was trying to say that to a large part we feel with MIT that linguistic analysis is very important but we don't agree on exactly what would be the nature of the rules for a proper linguistic description.

VIC YNGVE: I think we come closer together if we think about no single theory of syntax or grammar, at least none proposed so far, will exhibit all the regularities that exist in language and the reason is quite simple. The regularities that exist in language are of a much different sort. Certainly semantic regularities are a different sort of thing from grammar regularities and phonetic regularities are of a different sort. To insist that all of these regularities be exhibited explicitly in a particular theory is, I think, asking too much. I don't know of any current theory that can exhibit all the regularities that different theories differ in which regularities they exhibit and people can certainly differ in which regularities they think are important.

BILL WANG: I am just saying I can exhibit more than you, that is all.

SYDNEY LAMB: A different kind. The last thing I said a minute ago I have to take back because as soon as I said it I knew I was wrong. I said I didn't think there was any school that didn't recognize the difference between deep surface and subsurface. There are schools that don't make that distinction and Vic's maybe is one.

BILL WANG: Except for that wrap around rule.

SYDNEY LAMB: But that is in the beginning.

You can't make a clear separation into two levels such as surface structure and deep structure.

VIC YNGVE: We recognize.

SYDNEY LAMB: But you don't have really two entirely different structures as Chomsky does and I think that is a mistake. Your last remark is relevant to this point. You say there is no one theory that can exhibit both the regular. This is the whole point for having different levels. There is one that is the semantic and others regular and the other one phonetical. Not to say there is one theory that does exhibit three separate strata. It is all one theory. That is one of the reasons we have to recognize it.

VIC YNGVE: When I say one theory I would say one in particular. I think probably there is a question as to what is a theory. You know you can write the encyclopedia and say this is a theory or have something very compact which deals with some very small facet of language of a particular type and say this is a theory. I was taking the more narrow view of your theory.

BILL WANG: Both of them said it is hard to get

a theory that covers everything. I don't think that is saying very much. What you have to do is look at concrete cases and a thing that you intuitively know is correct, and taking the sentence we had this morning I don't think your theory is correct, Vic.

VIC YNGVE: Well, it gives to me the intuitively correct analysis and what I haven't shown is how I would handle this methodology.

BILL WANG: Do you know how to do that?

VIC YNGVE: Yes.

BILL WANG: Do you know how to break off?

VIC YNGVE: It requires in the narrow sense a different theory and in the broad sense a strata.

DAVE LIEBERMAN: I would like to make a comment. On the history of the motivation of the new form of theory in addition to what Bill mentioned the generalised translation is in order, a big part was played where a semantic component was built in the theory. You would just have to describe an infinite number of strings that they could work on. So it became clear that a part could not be left in the translation part. The semantic component would have to come before the translation part.

BILL WANG: The definitive thing I think you are

using the wrong argument. The bringing about a simplicity with this new model is attributable that the semantic rules need to operate on the line of demarcation and not the deep structure. There is a very different argument.

DAVE LIEBERMAN: The other point that Syd raised. This is often said, well the notion of deep structure and surface structure you can dismiss with a wave of your hand.

SYDNEY LAMB: No, not dismiss.

DAVE LIEBERMAN: No, but you are dismissing what people now say is a new theory.

SYDNEY LAMB: No, maybe I don't make myself clear. The point that people are making is, people are remarking on, that Chomsky discovered this --

DAVE LIEBERMAN: It isn't a point of Chomsky discovering but he has given a concrete statement on the relation between deep and surface. This has not been done before. But there is something very new there. He has given one version of a way to look at deep surface which is a different matter. Tesnière has a different method.

CHAIRMAN SEE: I think unless there are specific questions directed toward Bill or unless Bill wants to say something to follow up the argument we could do that, I think we are bringing out the different points of view.

BILL WANG: I would like to say a few words about semantics now you have brought up the general argument within the general theory of grammar. I think we are working in a direction that is promising in semantics. In the translation of grammar the first article on semantics that is reasonably precise is the Katz and Fodor article that everybody knows about. We haven't done very much in this direction but it is very suggestive. I got this idea when I was listening to Weinrath talk at the Linguistic Institute on semantics. I think somebody said later after the talk, "He is pretty articulate for a linguist." Now nowhere is it stated in that sentence, or probably if you found this in a text, that linguists are usually inarticulate but obviously that is what is implied. If you add a little word and say, "He is pretty articulate even as a linguist," you have completely reversed it. This means usually, "Linguistics are articulate," and "Even among linguists he is still to be considered articulate."

I think Fillmore, while working at Columbus during the summer, was working on the same problem and gave it a name. I was calling it just semantic inferences and so on. He worked out an exploratory set of rules in unpackaging sentences in a way that brings in sentences

not originally in the text but the meaning was implied by the sentence in the text. Apparently a lot of instant sentences -- he called these entailment rules -- require a relation of positive-affirmative versus negative-affirmative and so on. I say this just to show that within our framework of semantic research it doesn't end with the Katz and Fodor article. We are going in a different direction. That direction is to be found in this set of articles that were presented which is on the first page of the bibliography called Entailment Rules in a Semantic Theory which was presented just a few months ago before the International System of Languages in East Germany. I think very parallel entailment exists in Chinese and probably in many languages. This is an area for very exciting research.

VIC YNGVE: Eleanor Charney has been working in this area, as you know, and she has done a considerable amount of work.

BILL WANG: We have studied the writings.

CHAIRMAN SEE: As a matter of fact, I had to smile because she got through explaining this very article to me yesterday.

VIC YNGVE: This is somewhat related to our

previous topic it seems to me and I would think to Eleanor that the phenomena which you might call entailment sentences is not a syntactic phenomena. I think that if there is such a thing put into a syntactic theory one has to be careful.

BILL WANG: No, we wouldn't do that. Let me contrast. "He is pretty articulate for a linguist." There is a good relation between the negative and the other. We would want to build this type of relation but the entailment from the semantic relation falls outside.

SYDNEY LAMB: These things I don't know the distribution.

VIC YNGVE: Well, see me privately about this.

SYDNEY LAMB: Couldn't you put on your regular distribution?

VIC YNGVE: Eleanor is preparing some things for distribution. There have been some preprints but it is complex. I don't have it in my mind. If you are interested in this now you are at Yale come and visit us.

DeCAMP: This idea of entailment sentences, although I haven't heard it called this, is the serious one. One example where you run into this great difficulty of getting across to Chinese the distinction in a sentence,

"He left the party before eating some ice cream," or "He left the party before eating any ice cream." Now in the latter one there obviously is an implied negative. Where does it come in the words "any"? On the other hand, on the structural there is an argument for the presence of the negative. Where does the negative exist, not in the deep structure, shallow or anything else. The negative doesn't exist in the structure at all but in the entailed sentence.

VIC YNGVE: That is another of Eleanor's structures. You can say "They left before they had eaten" or "after" where "before" and "after" is interchangeable. If you said "They had left before they ate any ice cream" is fine but if you say "They left after they had" there is something wrong.

DeCAMP: But both are possible and both identical sentences.

BILL WANG: I believe this is important for translation because this is part of the task.

DeCAMP: In Chinese there is no way for expressing this in Chinese without adding an additional clause in which the negative or positive is actually expressed. If you are translating it into Chinese you must use a

negative to get the idea.

BILL WANG: I think you are right in that. Actually the negative is very strange in Chinese. In some cases there is no negative but implied negative. In some cases where having a negative or no negative gives the same meaning. "Before he went to high school he played the trumpet." "Before going to high school he played the trumpet." These have to be built into a theory of translation. It is unfortunate that heretofore students were interested in the philosophy of linguistics rather than this. I think now with greater insight we are ready to attack the problem.

SYDNEY LAMB: Can you really say "They left before eating the ice cream"?

DeCAMP: Yes, because it gives a very positive statement that they ate it after.

SYDNEY LAMB: This kind of thinking I would reject.

VIC YNGVE: My sentence was different. "They left before they had eaten any ice cream." In this case they had left before.

DeCAMP: I deliberately tried to make it ambiguous by using the "before leaving".

CHAIRMAN SEE: This is extremely interesting

but I think for the sake of continuity we had better continue with the Ohio presentation.

LEROY MEYERS: I will first describe some of the handout. This is the little one labeled "Chinese grammars and the computer at Ohio State University." It has the one page on the three different programs. Let me mention just at random what was generated and chosen, the sentences for each of the outputs. The grammar with that review is just merely the same as, well let's say, the revised rules which were made up to exactly the total by Ann Hashimoto earlier this year. What happens is that what we may have had, just the revision in the beginning, the program is slightly easier, the rules have been somewhat modified since then anyway.

Now I will give you one minute to read that one page outline of what we have done.

(Short recess at this point.)

SYDNEY LAMB: Is there anything about SNOBOL that makes this more convenient? I could ask the same about COMMIT the other way around.

LEROY MEYERS: I think it is the other way around. One of the troubles I found with COMMIT there is only one work space. Essentially we can have as many

work spaces as we need and as many shelves as we need.

Let me mention that the program isn't quite running yet. We hope only one more trial will make it run.

CHAIRMAN SEE: Any more questions on that? Vic, do you have any comment about this work space?

VIC YNGVE: Well, this again gets into a discussion of the similarities and differences between two programs. I would be glad to do it if anybody is interested. Perhaps now is not the time.

CHAIRMAN SEE: Any further questions? Okay, it is up to Ann now.

(Ann Yue Hashimoto delivered portions from her prepared paper.)

CHAIRMAN SEE: Any questions about the paper? I have a question. It didn't strike me that "The music being high" is ungrammatical. I am merely throwing that out. Can't you use this in the sense of sharp? He is singing high or low? Couldn't you conceivably say that "The music is high?" Couldn't you say that the music is being played at a pitch which is higher than the pitch he is singing and continue this down? Anyway, it seems to be the semantic problem. Anyway it doesn't seem

ungrammatical but purely semantic to say whether the music is high.

BILL WANG: If it is to be accepted it isn't as equally acceptable. In one case you are deleting all sorts of things if you wanted to say if you wanted a full grammatical statement. In the other case that is the source.

CHAIRMAN SEE: I can think of other interpretations. Any further questions?

FRED PENG: On page 10 there are no telegraphic codes. You have resultative verbs and I take it they are meant to be resultative verbs. Which are the resultative verbs? You have ^{吃飽} CHI BAO.

ANN YUE HASHIMOTO: ^{吃飽} CHI BAO.

FRED PENG: The whole thing is resultative. I ^{吃飽} see. The CHI BAO is a resultative verb and the whole thing is resultative.

ANN YUE HASHIMOTO: A kind of.

CHAIRMAN SEE: Any further questions?

J. WONG: I wonder if in your comparative structure where you include the ^有 2539 would that be also considered as a comparison?

ANN YUE HASHIMOTO: Yes. Actually this kind

of positive degree of comparison.

BILL WANG: This is on page 12.

DeCAMP: In the first place on page 5 1.3 I am pleased there is work done since the earlier set I saw in straightening out the problem. But saying that the NP after YOU^有 must be indefinite still doesn't satisfy me and I wonder if anything has been done on the indefinite. This question of money, "He is wealthy and has a great deal of money", would you call this indefinite? There are many things that are definite that are indefinite grammatically and could not occur in that sequence.

ANN YUE HASHIMOTO: I think that is a different YOU^有. It is not the same YOU^有 I am talking about. Here it is something like you have YOU^有 IN CHU FAN, that kind. That is a different YOU^有.

DeCAMP: Is that still different from the YOU^有 you are talking?

ANN YUE HASHIMOTO: Yes.

DeCAMP: Oh, I thought this was the differentiation. In that case I think the problem still stands in the constituent constructive rules only with the stated verb or VP subject N but not the VP subject in SP where you still have the problems providing the NP that follows.

ANN YUE HASHIMOTO: I have considered not as presented but as essential.

DeCAMP: There are so many of these and they are so difficult to describe. You can't say TAI UN YA.

BILL WANG: If you have an abstract noun it is preceded by a possessive noun, but if you have a concrete noun then you cannot do this.

DeCAMP: What is concrete and what is abstract? Concrete in a special sense that would apply only to this one contrast because this is not the usual semantic definition.

BILL WANG: Take the distinction which is very abstract which is table. You can say, "He has a table and he has not."

DeCAMP: On the other hand, I have had several informants that you cannot say, for instance, brotherly love.

J. WONG: That would be adjectively.

DeCAMP: There are many things very abstract and otherwise synthetically abstract you can't do that.

CHAIRMAN SEE: Bill, if you would summarize that this ^有2589 YOU has many variations in the discussion.

BILL WANG: Well, there are at least four

different YOUS in Chinese, not all of which are realized in the surface structure. In the first part of the paper there is. In Mandarin this is no longer utilized. It gets transposed and becomes ^有YEOU, and in the negative since that aspect stays ^有YEOU. This is one ^有YEOU.

Then there is a possessive verb YEOU which we just discussed and this is one of several that seems to combine with several abstracts and seems to act like an adjective.

The ^有YOU that Ann was discussing was another. There are two. There are some people and ^{(有人)?}URN. There are people where ^{(有的人)?}URDURIN is more or less definite but ^{(有人)?}URN is definite. In the case of ^{(有人)?}URN you have in mind sometimes who you refer to but in the case of ^{(有人)?}URN there is no implication of this sort at all. So there are at least these four different ^有YOUS.

CHAIRMAN SEE: Some can take the variation and some can't.

BILL WANG: For instance, the possessive YOU followed by a concrete noun cannot be preceded by an emphatic and cannot enter into a comparative construction, whereas followed by an abstract noun --

J. WONG: I have five here.

FRED PENG: There is another one.

J. WONG: Used as an adverbial phrase.

BILL WANG: That is the possessive. 也有 YA YEOU

where you have a complex verb.

J. WONG: You must proceed systematically.

CHAIRMAN SEE: You must proceed with plans.

BILL WANG: That comes from two sentences where the first one gets embedded and the second is a modification. There is one more 有 YEOU and that is the 有 YEOU that is comparative which means at least ask. So that is the 有 fifth YEOU.

CHAIRMAN SEE: Unless there is some specific question I think we ought to get on with the Berkeley presentation.

DeCAMP: With a qualification we can call these abstract and concrete but the borderline is by no means too easily defined. It does not correlate. You almost have to have a separate list.

BILL WANG: There is a correlation between these and others grammatically. Most concrete nouns can take the generalized but abstract nouns do not as a rule. You cannot say a table with 個 0222 and a table with 張 1728. This is a rather uniform thing.

CHAIRMAN SEE: What can't you say?

BILL WANG: You can't say with the same construction.

CHAIRMAN SEE: We will take a short recess.

(Recess from 4:20 P. M. to 4:50 P. M.)

SYDNEY LAMB: Now, as you know, I was recently at Berkeley and now am at Yale. Since my separation has been so recent the Berkeley-Yale presentation will be one presentation. It will really be the Berkeley presentation.

First, I would like to repeat what many others have done and express my thanks to Vic Yngve and Dick See for the arrangements and lunch which was most appropriate.

We have taken the position in this field that one can't do machine language translation unless someone has some understanding of what the language is about. We have been arguing this ever since we have been in the field. First there were a lot of people who disagreed with this but I think we now have plenty of empirical evidence as well as theoretical evidence that that will not work so I don't think I have to apologize any more over linguistic theory.

I have some handouts to pass out but we are not

quite ready yet. I hope we have enough.

Before that, as an introduction, one way to go about designing a linguistic machine system that is very helpful is to isolate recurrent partial similarities with a resultant simplification. There are various ways this can be done. This is a very general proposition. It involves separating things from one another and simplifying. I am going to give two or three examples.

To give one example that is becoming better and better known, it is very economical to separate the program from the linguistic information. Vic made that point this morning and Gene Pendergraft at Texas has, the simplification you can achieve by separating the program from the linguistic information.

I just realized that I have four types of simplification written now. First it allows the linguist to write his rules as rules, that is in some convenient linguistics rule writing rather than programming language. Second, when he wants to revise some of the rules he can very easily without the need for reprogramming. Third, the various basic operations that must be carried out by the machine to do decoding, or whatever part of the process is involved, have to be written only once in a program if

it is separated from the linguistic information, whereas in an integrated information where it is written in with the program these operations have to be repeated over and over again with the linguistics information which are subjected to the same basic operation. Fourth, and probably most important, the program since it is written to operate with rules of a specified form rather than with specific linguistic information, can operate on such rules not only for one language but any language, so new forms don't have to be rewritten.

So you have tremendous simplification once you have this separation. You have one program for doing syntactic decoding which will work for any language with the rules in the format. I won't say anything more about that.

Another type that I do want to concentrate on is the general type that is concerned with what I call stratification in language. Now we are ready for the handout.

On the front page of the handout there is an abstract diagram that is intended to show the type of simplification that can be achieved. I will give one simple example first. This is another we have been talking about for a long time. This involves a bi-lingual

dictionary as opposed to having two separate dictionaries. If we are talking about translating from Chinese to English in the integrated system you are talking of a Chinese to English dictionary. In the other separate approach you have a Chinese-English dictionary on the one hand and English-Chinese on the other. This is talked about by us and Texas and MIT but it hasn't been accepted by all in the field.

The situation in number one on the left in the first page that is the abstract that you have in the un-separated dictionary as opposed to the one where you have a separated dictionary.

Let's let the capital letter N generally let's think of them as lexical items. If given a lexical item there will be more than one target equivalent. So let the little letters a and b represent the target. Let's take a hypothetical language and be concrete for a moment. Suppose capital A represents a word to be translated in English as search or look for. Let a and b be search and look for. Capital B is a verb meaning to look for or examine. C is something that means something to examine or investigate. You see, if you have a Chinese-English dictionary un-separated you have to repeat the letters

C, D and B and so on.

Suppose that the average lexical item of the source language has three targets in the unseparated dictionary. You are going to have on the average each lexical item of the language repeated each time. It is not the case that it has every target likewise as three times the equivalent. This is easy to demonstrate. Since the source and target language have roughly the same language if you have three targets predicted it has three times or else you have a lot of inefficiency. I think you have to accept that as your proper exchange.

In the separate dictionary you achieve what is in Diagram 2 where the b, c and d only have to be listed once.

The way this is achieved, well, you have a scheme like let us say something like Vic diagramed on the board this morning. At some point in the decoding process you have these lexical items but they are not associated with the target equivalent. They just come in their proper place in the decoding process and the target equivalents are not dealt with until you come to the syntheses part. Let's say each part of the lexicon only has to be included once.

The type of illustration there also applies in single language in the case of the bilingual dictionary but that same principle applies when we deal with different strata in language. Since this stratification is not generally understood that is what I want to talk about in the rest of the time allotted.

This same diagram will apply when we are dealing with two neighboring strata within a given language. I will give one example where they are dealing with English in the lexemic and linguistic stratum.

(At blackboard) Go, go crazy and undergo. Three are enough. I suppose there would be others which have go as one of the components. In a stratification analysis you would consider each to be a single lexeme even though two from the morphemic point of view are complex. Technically these combinations are noted for the components. So in the lexical part of the system you will want to treat these as three separate units. In that part of the system it doesn't make any sense to consider this as making any sense for the part of the component and as a single. But in the amorphic you are going to get dis-economy of the type illustrated on the left of the first page because what we observe moves logically if we form

the past tense of these three the same morpheme applies. The past tense of go crazy is went crazy and the past tense of undergo is underwent. We have to supply the rule for the proper past tense form in each of the three. If we did it will treat as to go and it applies only once and applies automatically to each of the lexemes.

I don't know if these diagrams help or hinder but it is supposed to be another illustration. The same is true between any two strata.

Now the strata, the one that one must recognize in a linguistic structure are the sememic, lexemic, morphemic, phonemic. I am speaking of written languages since spoken languages are derived from the written. One cannot understand to look at the structure of the written language without knowing the structure of the spoken language the basic structure simply wouldn't be revealed.

I want to look at the structure of the spoken language first. On the second page of the handout I give the representation of a particular clause, "The farmer killed the duckling", on each of the four strata. On top is the sememic network. One calls these networks. It is not a complete sememic representation there. The capital letter notation is supposed to suggest that further analysis

of these components is possible but has been omitted here. The DECL is declarative. AGT is agent. GL is goal.

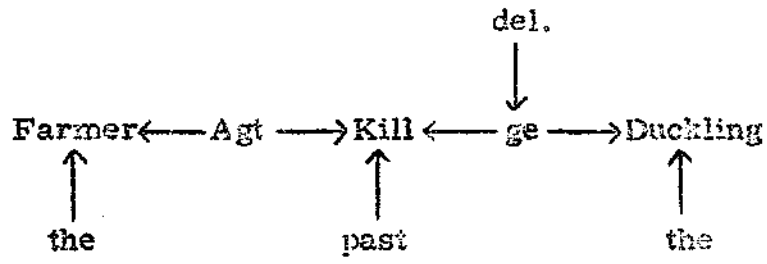
I can take you through the way this became generated. One starts generating a sememic network by empirical, declarative. If one chooses declarative one can choose another agent, goal or attribute. By the way, the arrow means presupposition. So agent presupposes two things, a thing and a deed, something the thing does. In this case the farmer is the agent and kill the thing that the agent did. The goal is duckling. So "The farmer killed the duckling."

Let me just pause briefly to indicate how certain other types of things are closely related to this.

(At blackboard)* Now I didn't mention that the way I have written this thing, the order, the position these are relative to each other is completely nondistinctive. The only things that count are the arrows and the things they are connected with. So if I had written duckling down here that would be no different. I could write it anywhere as long as duckling is connected with an arrow that goes from goal to duckling. It is just a matter of esthetics how one spreads these things out.

This is the basic structure of the clause and

* See p. 137A



this is the basic structure that applies to the one I have given here, "The farmer killed the duckling", the passive or interrogative or the interrogative passive, "Was the duckling killed by the farmer", or the relative class, "The farmer who killed the duckling", and other things such as "Having killed the duckling the farmer", so on and so on. For all of these things one has the same sememic structure. The declarative is over here.

In the case of this network someone is declaring that someone got something done to it, namely the duckling got killed. Whereas here someone did something, the farmer killed the duckling.

Now the next one down is the lexemic. What is shown there is a dependency tree of the type that Tesnière talks about in that book that Bill Wang has already referred to. Now the dependency tree is a simplified notation for what I believe really exists on the lexemic stratum. I believe something more slightly complicated than the dependency tree. The dependency tree is close enough.

Notice there is one feature that is not usually on a dependency tree. It is the little arrow on top of the ED. That little arrow is a special introduction to the morphology.

Notice also in farmer and in duckling there is a little space in the center which is a lexeme, two components. Duckling is also two components, duck and ling. The lexemic stratum you have the larger unit lexemes and as the components from them lexons. It works the same way on each of the other strata. On morphemic you have morphemes which are composed of morphons and the phonemics are composed of phonemes. So the farmer is a lexeme of two. Duckling is composed of two. Kill is composed of just one lexon, kill. Of course, the notation for kill uses four letters. That is just notational. Structurally this is just a single lexeme. Whereas in duckling there are eight letters and there are two elements, two lexons.

Then on the next one we have the morphemic and here I have put somewhat larger spaces to separate the morphemes from one another. The first morpheme is The. This is a morpheme composed of two morphons. The next is farmer and so on.

On the bottom we have phonemics. Each column is one phoneme and each composed of two or more phonons. Those two phonons constitute the phoneme. The next phoneme is composed of a single. When it is not accompanied by anything else is the neutral vowel a. I won't go into

detail.

Notice that each strata each has its own characteristics. The sememics occur in lexons, the lexons in trees. On morphemes the phonemes are in a string.

There is one revision since last year I could call your attention to. That is I have the lexemes occurring in trees where we used to believe they occurred in strings. This made the system a little more complicated because it was necessary to get from phonemes to strings in one fell swoop. Those operations are relatively easy.

Now what I have shown on page 2 and have just been talking about is the structure of the text where we have on the phonemic strata something close to the surface and each of the others is underlying the surface structure relative to it. I am recognizing here not just one surface structure underlined by a surface of strata. The actual surface of spoken language is the phonetic at the real surface and can be directly observed. The phonemic is the underlying for phonemes, and morphemes and lexons under morphemes and sememics or lexemes. The sememics correspond pretty much to the lexemic and the lexemic to the morphemic.

On the next page 3 it shows the structure for a

language as opposed to a structure for a text. For a text we can talk about its surface structure and underlying but we have to consider the text in the language. The text gets produced by the underlying structure which is the language.

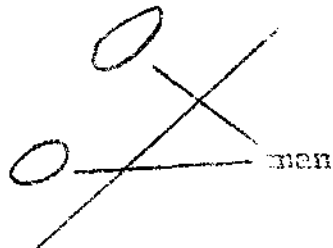
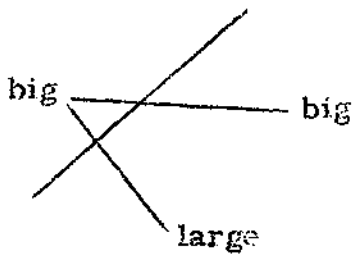
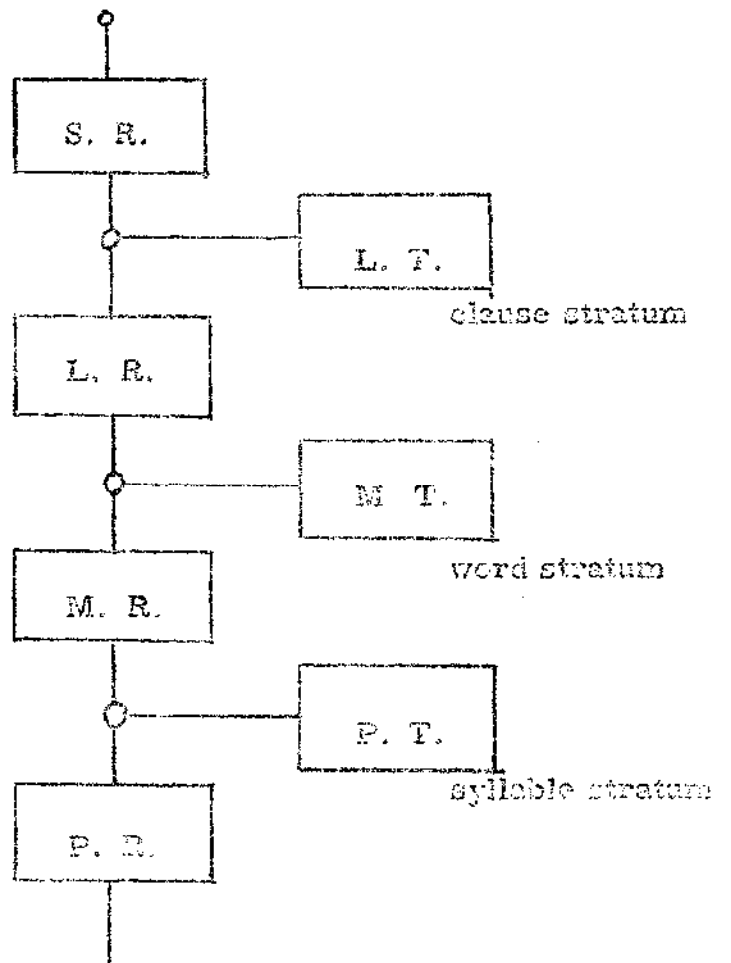
On page 3 in my haste I neglected to put dotted lines connecting those labels on those things. There is supposed to be a dotted line from sememics over to that little circle there.

(At blackboard) Here is a box and the little circle represents the sememic network. So the dotted lines got omitted in my haste. That goes over to a little circle.

Now as shown here a structural language as a whole consists of five parts, semeiology, lexology, morphology and phonetics. The relation between the structure of the language and text is somewhat related. For example, one such network would be for "The farmer killed the duck." The lexeme would be the one I showed on the preceding page and so on.

Now the next point I think would be, as I haven't said anything else I say what is inside those boxes, and that is next. They each have a structure which

- Sememic
- Lexemic - Lexeme - Lexon
- Morphemic - Morpheme - Morphology
- Phonemic
- Phonetic



is somewhat similar to the structure of each of the others. That is each can be divided into two main sets of rules. On the one hand a set of tactic rules which specify whether the arrangements are well formed at that level and, secondly, a set of realization rules which give you the realizations of the surface structure elements which are realizations of the deep structure elements that they realize. I will give you a few examples.

(At blackboard)* Let me draw here another diagram which makes it a little more overt that each of those consist of two parts. Let me tell you what the labels are and in the case of semeiology we can talk of the semeiologic tactics and realization. Eventually people may think of simpler terms. Let me start with the morphology there. This would be the morphological tactics and the morphological realization. We could show a sort of output from that going here and an output from this. Then we would have the phonological tactics and the phonological realization, and so forth. I don't think I will take the time to draw the whole thing. You can see how it goes. It is all parallel.

Now the thing that is interesting to observe, if you look at it in this way, which is vital and also a

* See p. 142A

little perplexing until you figure out what to do about it, is that you have two things that lead to each of these little circles and one should ask what is the meaning of this? That is, we have got realization rules. I think it won't be clear to some of you until I give you an example of realization rules.

The example of morphological realization rules is good in English. This is a basic or realization morpheme or lexon morpheme which can be realized or actualized or represented in either of three ways. Either good or in the comparative and then it is realized as best, because you don't say gooder, you must say better. If it is the superlative you get better. What is that? It gives you the good, better, best, taking the morphological realization rule. For example, sane and sanity. The same as you see in vain or vanity, nation or national, things like that. These are the rules that are often called morphic-morphilic rules.

Similarly up here you have the lexology rules that get you from the basic lexemes. For example, the things that get you from what I call the big sub 1 to big or large as representing this basic underlying form. So the rule that provides that is up here. The rule that

deals with things like sane or sanity are in here and the rule that deals with the alternate realization of good is in the morphilic rules.

Now over here we have the tactic rules, the functions for any kind of tactic rules is by what combination these may form. To make it a little clearer the formal tactic rules could be called syllable structure rules. The morphologicals are word structure words which tell you what structure of basic morphemic rules.

Up here semelological rules. It is in these rules we take the semantic rule possibilities.

Now I have to get to this basic point which is how can you have two outputs leading to the same thing here. To indicate the nature of the problem I show you arrows, the direction things are going. This is in the production process. Of course, you also have the decoding process in the opposite direction. What happens here? What is going on? Well, this diagram is a little bit simplified in one respect but it is like this. Recall what we were talking about after Vic Yngve's presentation this morning. The question came up about weighting these rules where you have a choice or not weighting, assigning various statistics.

Now keeping that in the back of your mind consider what we have here. Let's take a morphological tactic rule. Well, if we generate combinations of tactical elements, say for example verb versus suffix. One chooses the particular verb to go in. If one is doing it the way Vic and others have done you have a random number which selects the particular verb but, of course, when people speak they don't make their choice by random number. Instead the selection is made by the next upper strata. So this is what this thing is coming down here. This provides which of the alternatives to select whenever the tactic is faced with a choice. Whenever there is an alternative the choice is determined by one up here. So it does take both to determine a single element. So a lexon, one of the components of a lexeme will be a specification from which member of a morphological classification we will choose.

I have a further breakdown of the structure on page 4 which I think I had better skip and not go into any detail although it has the necessary features. I will just make a claim and substantiate later for those interested. It is this. Once you organize the structure of a language in this way that the grammar is so organized can

be used without change for both production and decoding. You don't have to organize it in two different ways, one for recognition and one for production, and the diagram on page 4 shows how this can be done for decoding. Let me explain rather than go into details in vague terms, which I can do with the aid of page 5.

Basically when one is faced with alternates, as one generally is in the decoding process at any of these stages, what one does is not try to resolve ambiguities but let all the possibilities go up to the next stage. If a particular element has possible realization, for instance, if we are faced with "man" and we don't know whether it is man the noun or man the verb, you just go ahead and suppose it might be either one and carry both possibilities up to the next stage. But at that stage all of the possibilities that come with it have to go through the tactic rules and the rules will automatically eliminate all those not well formed. So if you get something like man the rules will rule out that this could be the plural of the word "man".

I am afraid that is not particularly clear without the example. So if you think I am talking of something magic you will just have to think that until

some future date. What one does to decode, you take all the possibilities allowed by the rules but on each stage the tactic rules are gone through for all of the possibilities and they will have the function in the decoding process of eliminating all those possibilities that don't work out.

To get to Chinese it happens that Chinese is a rather simpler situation than other languages in one respect, that we have less structure from the English to deal with since the telegraphics are basically morphemes so we don't have to deal with any morphology. And this is, I think, the way we have set it up actually is like this. We take the characteristic of the morphemes and go immediately to the lexeme rules. This specifies what the lexemes are. This is the stage one is segmenting the string of characters into lexemes. As I said this morning, we get all possible segmentations and then the basic forms of the lexemes are to be considered in the class structure. It will automatically eliminate all the segmentations that didn't happen to pan out; that is, all the segmentations that give you syntactically ill-formed clauses.

That is all that I need to say or all that I am allowed to say, so I can turn it over to Doug who will tell

you a little more about how the program works that has been written so far for the first part of Chinese decoding.

DOUG JOHNSON: Maybe I had better first characterize the grammar in a slightly different way. It does fit into the scheme but for generality and simplicity I think I had better say that the grammar that is now being written and will be operated on by the programs is, well, it is almost true that it is a constituent structure grammar whose terminology symbols are Chinese characters. We will have two parts, of course. There will be a dictionary which consists of rules whose right-hand sides are strings of Chinese characters. These strings are what we call lexemes. They are the unit of entry in the dictionary. In the other set, of course, are the rules whose right-hand sides are strings of non-terminal symbols for tactic codes. The program we are writing together is to assign the structural description to strings of Chinese characters that are implied by the grammar. It assigns all the structural descriptions that are implied by the grammar.

Now this system consists of a series of, well, three programs, the first of which is not very interesting. It is simply a preliminary pass which converts a string of Chinese telegraphic codes into addresses used

in the next stage. The next two stages can be characterized perhaps first of all generally by what information is used. In the second stage the dictionary is in code storage and the third stage the binary rules.

Now what is done in the second stage is this. The program goes through position by position and each position it will list all of the lexemes which match a segment of the sentence that ends at that position.

(At blackboard)* If it is at position 4, say, and if b, c, d and e are lexemes in the grammar then in the section of the list pertaining to position 4 it will list b, c, d and e or, as a matter of fact, what actually it will list is the location and address referring to the location where these lexemes were found in cold storage. In addition it will list for each of these a code which refers to a list of all the non-terminal symbols that generate this list. Of course, what this output then implies is what can be deduced from this if you want a list of all the ways of segmenting a sentence by a series of lexemes.

The third pass then takes this kind of list as input and then the binary non-terminal rules are brought into storage and the sentence is then parsed. The parsing

*

See p. 149A

a b c d e

4

b c d

d

program is again left to right, position by position, listing all of the trees which can be formed which ended that position. As you can see, the two kinds of output from the second stage and the output from the third stage are similar in many ways. The structure from the list is very simple and this makes the program fairly easy. Since all the structures are found over any one stretch of the sentence all the structures for the whole sentence are found. I think that in general describes the programs.

S. S. SOO: What language do you use?

DOUG JOHNSON: This is programmed for the 794.

DAVE LIEBERMAN: Is it running?

DOUG JOHNSON: We use it for structural grammar but we have the program. What we have now are the adapters which turn the grammar into machines that will form.

S. S. SOO: When you say components what do you mean?

DOUG JOHNSON: This takes the grammar written by linguists and turns it into something for machine programs.

SYDNEY LAMB: For example, the computer will work with the Chinese dictionary on the one hand but the linguist works better with the Chinese. It is really a

Chinese dictionary with English equivalents for the aid of the linguist. One of the things the machine has to do is convert this into the two dictionaries. This is not too easy. I mean it is not too easy to write the program.

VIC YNGVE: We talk about a separation between program and grammatical information, linguistic information. We talk about separation in these various strata. There is a further separation which I think is extremely important which is recognized some places and not recognized in others. There is the separation between a research program, that is a research computer program, and a final operating version which would be efficient, fast, cheap, and so on. Now when one talks about program language I personally feel very strongly if you are doing a research job you should use a computer language, one which makes the program as simple as possible, as quick as possible, so you can get it over with and do your research. Then if it turns out that it doesn't work you can redo it without the expenditure of too much effort. If it does work you can re-program it in an efficient way and tackle that as a separate engineering problem. For this reason I have this objection to program machine language when one is dealing with a research or developmental

program.

SYDNEY LAMB: The presentation isn't done yet so maybe you had better wait. I think this is a good point.

DAVE LIEBERMAN: The reason before I said that I thought that parsing programs were not a trivial matter is the examples we have had in the past, Harvard, Jane Robinson's programs were pretty slow at first. This I assume is much faster than that.

DOUG JOHNSON: I don't know as it is faster than the Jane Robinson program.

SYDNEY LAMB: It is faster than the Jane Robinson program of a year ago. There may be a faster one now but it is fantastically fast.

DAVE LIEBERMAN: Could you characterize where it is fast? Is it the remote or use of bits?

DOUG JOHNSON: I think it is clever use of bits in the machine.

SYDNEY LAMB: Wait a minute. I think the answer is between the two. It is the strategy of the procedure. It is also very clever in the way the bits are used in the machine. It happens that the basic strategy is the main factor.

DAVE LIEBERMAN: You haven't considered the use of characteristics left to right or scanning.

DOUG JOHNSON: We haven't attempted to include any characteristics. We just attempt to find all the structures just as the Jane Robinson program but we use a different strategy. The result is the same.

SYDNEY LAMB: John Cox makes multiple passes through the string. This one makes a single pass. At each position in the sentence it finds all possible trees at that position. It is a single left to right pass.

DOUG JOHNSON: I am not sure this would characterize its being more efficient.

SYDNEY LAMB: This is part of it.

S. S. SOO: How big a list because the speed is in a way related to the amount of size of the list.

DOUG JOHNSON: Sorry, I don't know.

S. S. SOO: Initially you said your sentences were reduced to a, b, c, d and then you go in an attempt to discover the various structures, in your initial pass I mean.

DOUG JOHNSON: I think what we are talking about now is not either of the first two passes but the third pass that assigns the trees.

CHAIRMAN SEE: I think if there are any further points you can take them up with Doug on a one to one basis because we want to finish by six.

CHING-YI DOUGHERTY: I have brought copies of the papers we put out. I hope everyone got copies of each.

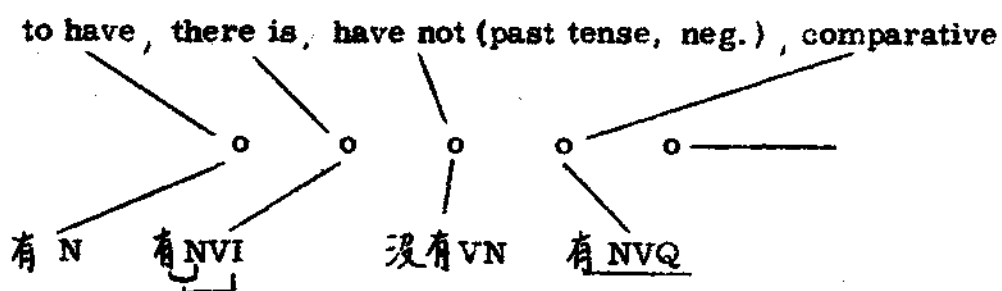
I just want to give you a complete example of the method we are using here so that in a way it will make it clearer. For instance, we were talking about the character YEOU so sometimes I use the same example and the character YEOU is one more type. We just mentioned before there were four or five meanings but in our case there are four or five meanings in this network. How do we know that this character has four or five meanings? Usually the different meaning is realized by the different syntactic construction.

(At blackboard)* This would be on the morphemic level and this on the syntactic rule and the syntactic rule would apply here. On page 34 the verb YEOU is discussed and under that I have several rules. One is "to have" and when YEOU is followed by a noun in this case it usually has this semantic "to have". YEOU is 2589.

Also YEOU in early discussion Bill says there

* See p. 154A

有 2589



	VY	have
-	NVI	there is
-	NBY	VQ

are a group of nouns following ^有 YEOU or if ^有 YEOU is followed by a group of a certain number of nouns the construction would become an adjective. I have that in that Rule 87. This group of nouns can be listed. There are not too many, not too few either. Like the lexeme YEOU "to have money" that is a concrete noun. It is not an abstract noun as I listed in here. In this case I would list that YEOU change as a lexical item.

In here this meaning could be "there is". Usually in this meaning it was followed by this and followed by that. That is a transitive verb plus object. Like the Rule 86, ^{有人抽煙} YEOU REN CHOU-IAN, the subject follows the predicate.

So that here this meaning, also ^有 YEOU has the meaning of past tense and usually occurs with the word ~~沒~~ MEI so therefore I have a Rule 83 that ~~沒~~ MEI becomes a combination and this combination if it is in front of, if it is ^沒 MEI, you have the ^有 YEOU occur with ^沒 MEI and you have a verb that usually in the sense means "have not", past tense, negative.

What other four meanings do you have? You say there are five meanings. There is the comparative ^有 YEOU. The comparative ^有 YEOU usually you have the ^有 YEOU, a noun, as an adjective. In this way this could become comparative.

And the other meaning, the fifth one, I combined one of them, ^{有人} YEOU REN, because ^{有人} YEOU REN is different.

BILL WANG: Only in one of the senses.

SAMUEL E. MARTIN: The other sense would be taken care of elsewhere.

BILL WANG: One is negative and the other is not.

CHING-YI DOUGHERTY: As you can see, the differences in meaning are realized by the difference in construction. This is quite regular except that you can say that this is like this and could be the same. Actually this usually has another subject. In this case the YEOU is always preceded by a space with the initial.

As you can see, this is the way we are doing that. Therefore, in the dictionary the entry 2589 ^有 YEOU we have different meanings. "To have" is the verb ^有 YEOU. Also this is followed by a space noun VI. For instance, space ^有 NB_y, this will be group, 87, ^{有 意思} YEOU YIH-SY. That will be in this group. In this case the combination from this becomes VQ. Our dictionary will be something like this. You can see what follows this word and precedes the word and we can make the right choice.

The coding could become quite complicated.

*
 For instance, we divided the transitive verbs into many classes. VT can only take one object and VO take one object. VS takes this kind of object. VL can take this kind of object. VP will take this kind of object. The difference between is the object of this. VTO SHOOO VSTVVT VLNVVT VFNVTN.

We divide the classes as we can but there are some properties we do not want to include them in the classes. We add them to properties. Such like as VT with this can be duplicated. Then we have VTR, you see it can be duplicated. Such as ^吃0676.

Whether the object of this verb can be inverted by the word ^把2116, some can use ^把2116 and some cannot. If I don't have any sign it means it can be inverted. If this verb cannot be inverted by ^把2116 then I put it in there.

VT. I make the distinction whether this verb, usually a verb has to be measured by a human being but some circumstances can be by an inanimate object. Suppose the verb "to eat" has to take an agent that can eat, either human or animate object, the reason I do that is that when I analyze the biochemistry and come to a lot of sentence structures like this I have like cells, wash,

* See p. 157A

VTO

VOOO

VS(TVVTN)

VLNNVTN

VPNVTN

Cells washed 3 times

three times. Cells washed three times. You know very well this cannot be the subject and Chinese grammar cannot tell you whether this has no way to tell. You have to devise a limit, this as a subject. By doing so I have to say the wash has to take a human agent since the cell is not human this cannot be the subject. This has to be the object.

BILL WANG: How about rain?

DeCAMP: How about machine?

CHAIRMAN SEE: This takes care of the old --

CHING-YI DOUGHERTY: Rain cannot wash, can it?

DeCAMP: A Maytag can wash the clothes.

SAMUEL E. MARTIN: Only when personified.

CHING-YI DOUGHERTY: An animate noun like machine, like the meeting, like the conference, has the ability of the human being.

SAMUEL E. MARTIN: Tide washes clothes cleaner than any human soap.

VIC YNOVE: What kind of cells can that cover, just by logical cell, if you have a popularization for clean.

CHING-YI DOUGHERTY: I limit my grammar to this generalistic.

CHAIRMAN SEE: In slang people say, "His cells jumped. Every cell in his body was jumping up and down."

SAMUEL E. MARTIN: They don't say it in bio-chemistry text.

J. WONG: You make that passive voice to solve your problem.

CHING-YI DOUGHERTY: How do you know this is the passive? You have to put the passive in here.

J. WONG: Because this noun cannot assert any action.

CHAIRMAN SEE: That is what she is saying.

J. WONG: But make it passive voice.

CHING-YI DOUGHERTY: Either that way or making this the object. I used the 2116 in the same way.

DeCAMP: Excuse me, when you still have ambiguities in Chinese also. You say "The clothes are washed clean" and if you take the sentence as the right term for the washing machine. Now could that have both meanings? "This washing machine washes the clothes very clean" or "The washing machine itself has been washed clean."

CHAIRMAN SEE: I think the point is there are no ambiguities left in this particular context. The cells are not going to do the wash. Maybe the machine will.

DAVE LIEBERMAN: I think if we don't admit the ambiguities we will keep going in circles. I am not saying we can't get everything clear but in some things there is a deviate.

VIC YNGVE: Is there a higher stratum you can go to?

CHAIRMAN SEE: Is this really semantics?

VIC YNGVE: She has an ambiguity there.

BILL WANG: I think in English there would be no distinction.

DAVE LIEBERMAN: It doesn't hold here because of personification.

BILL WANG: You substitute 被 BEY in some cases.

CHAIRMAN SEE: This problem comes up in Russian when you worry about inanimate and animate and so that this question has to resolve. When you make any elaborate fantasy around a word then I suppose it does begin to form a new class.

BILL WANG: To show it is a grammatical matter there are two words for corpse in Russian, one is animate and one inanimate. This distinction would be a grammatical one which would not make any semantic sense.

CHAIRMAN SEE: What coding?

BILL WANG: It was in lexology when it was not as high up as now.

CHAIRMAN SEE: This would be roughly what we call syntactic classes?

SYDNEY LAMB: Yes.

PAUL GARVIN: In Russian there is one as an inanimate and one as a corpse. In English there is a distinction between corpse and dead man.

CHAIRMAN SEE: I guess it is a question where the soul is. Do you have more, Ching-Yi? We interrupted you.

CHING-YI DOUGHERTY: I have lots more but no time.

CHAIRMAN SEE: You can wind up in a few minutes.

CHING-YI DOUGHERTY: I think each one of you has a copy. I would appreciate it if you would examine carefully and let me know of the inaccuracies and inconsistencies. I am sure this is not the finish. Some rules are very fine. Some are just skimming the surface. So I would appreciate any comments. All of this has to be tested by machine I am sure. Then we can discover more details.

FRED PENG: Before we dismiss the group I think

other people and they all agreed with me. I don't know how many others agree with me.

BILL WANG: On this discussion of the YEOU I think it is interesting that an ambiguity that is resolvable with these rules when it is coupled with an ambiguity that Fred mentioned this morning, renders this ambiguity irresolvable. Take a pair of sentences such as "He didn't go" and "He has no box." You would know in one case it is an aspect and the other position. You know this by the word class in the following item. In one case it is a noun and the other a verb.

Take Fred's example this morning, "The oil of the fried chicken." Once you have that ambiguity then you can no longer tell which YEOU it was. Then one interpretation, "He has not fried chicken yet." In the other case, "He doesn't have any fried chicken," and the ambiguity of YEOU becomes irresolvable. In a given sentence you can't tell. You can have a different tree form but you wouldn't know which is the correct tree.

SYDNEY LAMB: I wonder what it means to say tree in that context. How can we say one is correct and the other not?

CHAIRMAN SEE: What does it mean? The sentence

under discussion is 他 沒 有 炸 雞
0100 3093 2589 3498 7741.

It is ten after six. If people aren't too tired I would like to spend five or ten minutes on the telegraphic code. I can tell you some of the ways to memorize it. Each person has a different mental set that leads to a different approach just as people study languages by different methods. This is a kind of language.

I thought I could give a brief history. It is something like an archeological garbage heap that has built up. The importance of the telegraphic code is that it is actually used so we have to study it. In its broad outline it is very systematic but there are certain fringe phenomena to show how this built up stratum by stratum.

In the first place, when the telegraph was introduced in China in the past century a rather bright diplomat, who was quite an active person, went over to be an ambassador in Paris. His wife didn't want to go and he brought his number one concubine and all the Europeans thought it was his real wife.

In any case he was a well traveled man and he recognized there was a need for a telegraphic code. He drew up a book which consisted of 100 pages with 100 squares

on each page. So on the first page which might be called page 00 he started out with the number 1 logically enough. Any mathematician would have told him that was a mistake, that you start with zero. So he started 0001. Finally the last number for the 10,000 left off the zeros. It became obvious that it was inconvenient to end the page with a number different from all the rest which was 00 and suddenly have 0100. So at a period of time this was abandoned and the first number was made 0000 on the first page and that is left blank for the reason it was blank in the first place. So the 1 starts out in the second square. So that is the very first point.

I will symbolize this by saying that 0000 became the first entry in the book. The construction of the book was very simple. It ran for 80 pages which could be described from 00 to 79 which accounts for 80,000 numbers and a section of words from the standard dictionary were taken and inserted along this long axis, 8,000 positions long, and the standard dictionary collection was spread across this axis eliminating certain obvious unused words. Some of the bigger dictionaries included 40,000 or 50,000 characters. So this man I told you about made the preliminary selection reducing to about 8,000.

Unfortunately he either eliminated or the dictionary didn't have a very few common words. Some words like steam, which is now ³⁰⁸⁶, was not to be found anywhere in the 8,000 and yet steam is a fairly important concept.

* So for this reason and others it became necessary to put some of these other words overlooked into the remaining 2,000 positions. So these were put in in radical order. I haven't mentioned radical order but, of course, that was the conventional dictionary use, just the same as the index for the Matthews' dictionary.

So that the supplement began to develop and the first supplement which I might call Supplement A is the biggest one in the old style book and it consisted of words like steam which had been omitted and ran for most of the 2,000.

I should mention in the first part of the list exactly one space was left after each radical. So after the radical for wood where there are literally hundreds of spaces there was one space before starting the next radical. Since under some there are only one and under others are hundreds the ones following the highly used radicals were filled up and ideally one would have left a space in proportion to the new radicals. So the reason

* See p. 166A

for the supplement was that he didn't leave a lot of space for some of these things.

So then these spaces were filled up to a certain extent but there was only one to be filled and you had to go to the supplement.

The supplement wasn't planned well either so the metal radical as more metals came that filled up so they began to fill up a second supplement and the thing became rather chaotic because it wasn't well enough planned in advance. But it still remained the overall structure from 0000 to 7900 plus which was fairly simple and straightforward and followed the dictionary order.

Now in addition to inserting characters at the end of the radical sequence where there was a space in a certain number of places the space was used in front of the radical for a character from that sequence. For example, I couldn't find the word "fry" a moment ago for this reason. This was the last character on page 34 way down in the right-hand corner. What they had done, the previous radical 85 is ^フWAT so all these are ^フWAT characters. For some reason they stuck in the character for "fry" just before because they needed it. I guess he wasn't a cooking man so he didn't have to worry about this. The word "fry"

could become important so they inserted this word prior to the radical rather than at the end. This again caused confusion and that is why I couldn't find it a moment ago and had forgotten it was inserted in this abnormal manner. So therefore we have a few of these still descending to this day. Another one is the "mountain" radical. There is a character before that that should be following that.

So we have supplements. We have this shift that I mentioned which is not very important really and then we have this one phase. Then the next phase which is the supplement phase A and B, and this third phase filling in the slots between the radicals.

A fourth phase could be said to be deletions. He wasn't very careful about it and there were some close duplicates or duplicates where the characters were put in twice. One example would be "each" which was originally in under 0028 and 0416 with variations. They deleted the first one. So we have a certain number of deletions that took place.

Now when the mainlanders took over, the Communists took over, they found this to be a rather inconvenient tool and they revised it completely and this turned out to be very useful, this revision. What they did

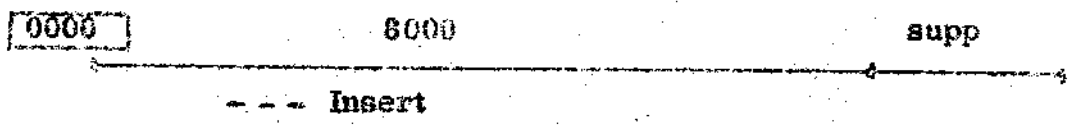
essentially was to eliminate all this confusion about supplements A and B. They completely abolished the arrangement that had grown up over the years and stuck to the basic sequence and straightened out completely the supplements. So we have supplement revisions which took place. That was somewhere in the early fifties.

This was a very constructive step because in the first place these are very rarely used characteristics. So the fact that most of these were rearranged had very little impact for daily use for sending telegrams or whatever use people wanted to use it for. So it resulted in greater efficiency because it is an index to rarely used characters.

They did do something beyond sorting out and making it rational. They removed a large number of rarely used characters from the main body of the book and put them over into the supplement and took a number of more frequently used characters and put into the main body, not always the right place. One example was "steam" removed from 8655 to 3086.

So what do we have then at this point in time? We have the 8,000 sequence plus a single supplement.* We have a few inserts that took place before and still retained.

* See p. 169A



I will simply ignore for the moment the slight revision and being out of stroke order that took place in inserting characters that I just mentioned. But we still have a large number of spaces due to the fact there were two sources. One that they were already there between radicals and then that they were put in. Second, that some of the characters were eliminated because of duplication. So there were spaces in the 8,000.

This is the final chapter of the story that it was around 1958 that they announced in a Peking newspaper that they were going to fill all these in except for 0000 and they published in the newspaper, and this is the important thing, an alphabetical list, which sounds like a nice trick, of what you might call inserts. Now what I mean by alphabetical list it is arranged by the ^{拼音}PINYIN System and many of these modern ones. They simply put them in right on through. So if you realize this is a little easier to understand some of the things that happen.

I will just take a second to show an example to make this more or less abstract, more or less the first page, to show you the way it is now. I think this will help. In my mind I divide each page up into four quadrants and I find it helps to actually remember the position of a

character on a page. After a while if you remember three or four characters on a page a pattern forms in your mind.

Now I will quickly sketch in a few of the radicals. Radical 1 is 0001. Radical 2 is 0019. Now here is where the radical had been eliminated. So putting in alphabetically the first one was put in here at this space. I don't remember what this chemical is for 026 but it is a high numbered chemical. The next one went in here simply because of alphabetical order.

So knowing this I think it makes it easier to understand how the thing has developed. Once you know this radical is here it is not hard to remember that there is likely a space in front of it and this is the logical place to put this. If you remember a few of these things everything ties together.

I will just sketch in a few of these things. This is some of the stuff on the first page. I wanted to get to this portion because this illustrates a few more. This is where Radical 4 occurs at 0434 and here we come to another one which is another way of writing arsenic I think. It is an element anyway. That is the end of Radical 3. A little further along we come to the end of Radical 4 and here comes eucalyptus. So you see it is

going on in a rather simple-minded fashion. I think that is enough to illustrate the point.

One can follow through. You can determine what pronoun suggests certain obscure characters must have had by the place where they put them. The list is a little longer than the holes in the structure and they wound up with about a dozen after 7902 which is the last one in the original batch, 7908 for example.

That is a rough sketch of the way this thing has grown up. We have the original 8,000 more or less untouched and now we have a completely systematic which is easy to find your way around in and we do have this sort of irksome feeling in and there are about thirty or forty or so if you know where to look for them. Fortunately the University of California has come up with a fine set of indexes which permit you to find your way around. I think it makes the telegraphic code a useful tool to deal with Chinese characters.

This is a kind of hurried explanation but I thought I would go through it. I have memorized a few of these numbers just for fun. It isn't too hard. It becomes a game after a while. If you know a couple of hundred you don't need to know any more because you can find

your way around in the list. If you know where a certain radical is the structure of the system tells you where the others will be.

I think that winds up the meeting. I think the last speaker takes the most time. I think we should again thank Victor Yngve and Miss Landers and Ron Hofmann for the fine arrangements and Frank Liu also. I think we owe them at least a round of applause for the fine effort.

(Applause)

I think everything has been stated. The transcript will be distributed. We are now open for adjournment to go to dinner.

(The Chinese MT Meeting adjourned at 6:30 P. M.)