[From: Internat.conf. *Methodology & Techniques of Machine Translation*, British Computer Society, Cranfield, 13-15 February 1984]

Automatic Integrated Dictionary Systems

J. Jelinek — University of Sheffield
J. Hawgood — University of Durham

0. Introduction (The index numbers in this Introduction refer to chapters of this report)

The AidTrans Consortium[1] already has, in the form of a set of books[2], a left-to-right (or top-to-bottom) multiple-path[3] syntactic analyser[4] which renders Japanese sentences in crude English[5]. This system has evolved from a continuous research since the early '60s and has been successfully applied to teaching Japanese-English translation. It has also been emulated for other pairs of languages[6].

In the first phase of our work, now completed, we have formulated a comprehensive Japanese script I/O keyboard applicable to PERQ or similar machines[7]. We have also formulated an interactive program[8] which elicits lexical and grammatical data and organises them into an Automatic Integrated Dictionary[9].

Our present effort is directed towards the creation of transferable personal software facilitating Japanese-to-English translation[10], a Japanese-English teaching machine[11] and a scientific & technical Japanese-English data bank[12].

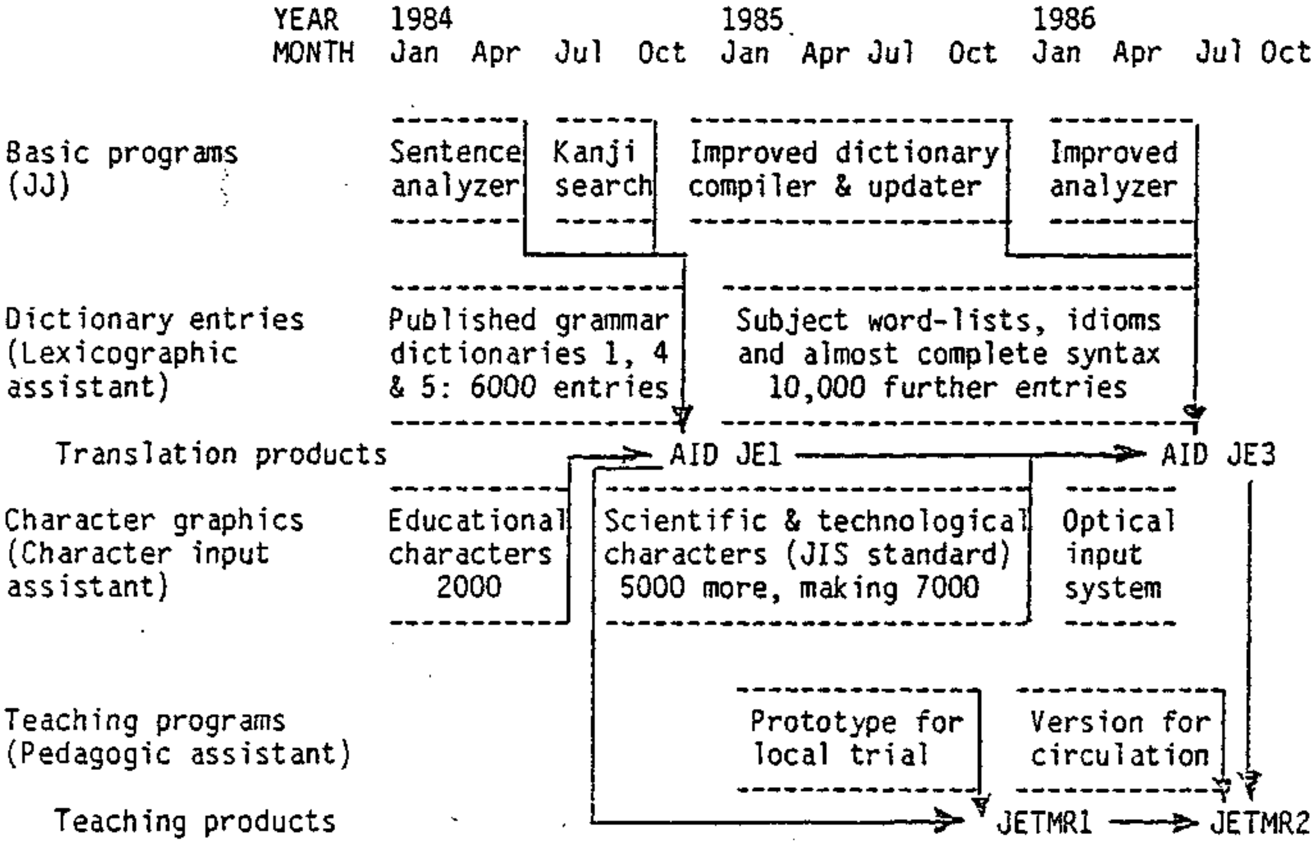
In the third phase, we intend to formalise and incorporate into our system the principles of depredicational analysis[13] to capture the formal links between units larger than sentence[14] and thus allow a coherent translation of texts. At that stage, in an increasingly refined form, we hope to be able to produce a Japanese-to-English translating machine[15]

1. AidTrans Project

Diagram 1 summarizes our development plans both for computer-aided translation and for computer-aided instruction in the second phase of the" project. The first phase, covering the basic dictionary-generation program (see Chapter 8) the AAA-ZZZ code list for Japanese characters (see Chapter 7), is now complete. The dates given for Phase 2 assume two dedicated PERQ systems and three research assistants are available from January 1984.

Diagram 1:

ESTIMATES OF EFFORT AND COMPLETION DATES FOR SUBPROJECTS IN PHASE 2
(Assuming availability of 2 dedicated PERQs and three research assistants) •



YEAR 1984 1985 1986
MONTH Jan Apr Jul Oct Jan Apr Jul Oct Jan Apr Jul Oct
Basic programs (JJ)
Sentence analyzer
Kanji search
Improved dictionary compiler & updater
Improved analyzer
Dictionary entries (Lexicographic assistant)
Published grammar dictionaries 1, 4 & 5: 6000 entries
Subject word-lists, idioms and almost complete syntax 10,000 further entries
Translation products
AID JE1
AID JE3
Character graphics (Character input assistant)
Educational characters 2000
Scientific & technological characters (JIS standard) 5000 more, making 7000
Optical input system
Teaching programs (Pedagogic assistant)
Prototype for local trial
Version for circulation
Teaching products
JETMR1
JETMR2

Our ultimate (and perhaps unattainable) ideal is a fully viable translating machine (see Chapter 15) to convert modern technical Japanese into passable English without human intervention. The practical long-term aim (Phase 3) is a comprehensive computer-stored dictionary of modern Japanese (AID JE4), with programs to aid human translators by providing alternative paraphrases not just for sentences but for paragraphs and longer texts, using Dr Jelinek's concepts of depredication and read-forward to unravel implicit cross-references between sentences. We have no doubt that this aim is attainable, but acknowledge that Phase *3* may take 10 years. An intermediate "large-dictionary" version might appear by 1991.

In Phase 2, corresponding to the development of an "Alvey" Demonstrator Project, we would expect to be able to offer a versatile translation aid(JE3) and its teaching version TMR2. These would contain basic dictionaries, including idioms, and a choice of special subject word-lists. They would propose alternative paraphrases of complete sentences, handling all syntactic problems, but the user would require to look up some characters (mainly nouns) in printed dictionaries.

The very first product (AID JE1) would correspond exactly to the printed grammar dictionaries already published by Dr Jelinek and used by students taking his course in reading technical Japanese (now totalling over a thousand). It should be possible to produce this by late 1984. As a translation aid, it would only be useful to people knowing Japanese at the level of this course. By contrast, the use of AID JE3 would be available for people knowing no Japanese beyond basic characters, but with enough subject knowledge to turn the crude sentence-by-sentence translation into good English. The prototype teaching machine TMR1 should be ready by late 1985, but would only be used in Sheffield under close supervision, to gain experience of points requiring improvement for TMR2.

A subscriber to the Automatic Integrated Dictionary System could build up from the first elementary product AID JE1 to all the others just by receiving files and programs on floppy disks. The comprehensive dictionary JE4 would be issued incrementally as different segments of existing printed and card-index dictionaries were added to the master file over a period of years, and would still need continual updating even when "complete". We would offer financial inducements to subscribers to report new meanings and frequency-counts (made automatically by the system), to help in keeping the dictionary up to date.

Diagram 2:

PRODUCTS	PHASE 2
AID JE1 Late 84	The computerized version of the existing Grammar Dictionaries, requiring basic knowledge of Japanese to use as translation aid.
JETMR1 Late 85	Prototype teaching machine using AID JE1.
AID JE3 Mid-86	Standard translation-aiding machine, usable without knowledge of Japanese beyond some training in finding characters; some use of printed dictionaries required for specialized and proper nouns.
JETMR2 Late 86	Self-contained learning aid, usable without teacher but needing occasional reference to printed dictionaries, like JE3.
	PHASE 3
AID JE3.5 1991?	Containing equivalent of large dictionary (Kenkyusha) so much less recourse to printed sources required. Rudimentary depredication.
AID JE4 1996??	Including fully comprehensive dictionary and full-text analysis.

2. I.D.S. Japanese Reading Course

In the 1960s J. Jelinek (then of Charles University, Prague) and Dr K. Novak (of the same) jointly developed the Automatic Syntactic Analyser of Japanese[1] , which amongst other features contained the first consistent formulation of valence-based word classes and was a strictly contrastive study with English as its yardstick. As its publication for various reasons was confined to the decimal machine code in which it was developed, it remained largely unknown in the circles of Japanology.

This analyser was brought to the U.K. in October,1968 in the hope that the research might continue, but it was soon found that all MT research in the U.K. had by then been stopped, and it became necessary to think of applying this analyser to teaching, if it were not to be entirely abandoned. Although no such application had previously been envisaged, it only took 2 years to convert the analyser to what has since been known as a Grammar Dictionary, and to prove that it allows absolute beginners to translate from Japanese to English in 8 weeks of intensive instruction. Since the first successful course in Summer 1970, many revisions have been effected both in the analyser itself and in the teaching method, resulting in shortening the course to 7 weeks.

This course([2]), now run permanently at the Centre of Japanese Studies of Sheffield University and also annually at Nanzan University of Nagoya, works with four books[3] published in Sheffield and two dictionaries available on the market[4].

The Course has now worked with some degree of efficiency for many years, but we are painfully aware of a great and growing number of possible improvements, which can no longer be made by the hitherto methods, because of the cumbersome bulk of published materials. Each alteration in one of the books results in rendering the rest out-of-date, and only a complete computerisation of the whole course (see Chapter 11) allows further progress.

3. Multiple-path Predictive Analysis

This was the name of the method used by what, until the big crisis of MT in the mid-sixties, rated as the most advanced Machine Translation establishment, namely the M.I.T. (Cambridge, Massachusetts) Oettinger-Kuno group[5].

The main advantage of this method of tackling Automatic Analysis of natural languages was, in our view, its linearity, proceeding (more or less) consistently from one end of the sentence and having all grammar formulated and stored in terms of "predictions" of possible continuations from any given point to the full stop. This feature was seen by us, staunch neo-structuralists, as distinct advantage, since it could reflect the sentence perspective(6) neglected by other schools of grammar, and by the same token, it was regarded as a damming disadvantage by the then ascending linguistic school of Chomsky-type Transformational Generative

Grammar [7], who believe in the so-called ‘deep structure’ and insist on a global hierarchic approach to sentence analysis.

While we have fully shared the strictest requirements upon the task of grammar with Chomsky[8], we see no advantage in fulfilling these requirements by a hierarchic device, and believe that a linear progression carries a deeper significance in natural languages than merely having to be that way in order to accomplish the "surface" structure. We believe that the overall order in which information is presented in a text, reflected also in the order in which constituents form a sentence, is, in itself an essential part of the information contained in the text. In other words, no matter what logicians and semanticists may think, "a = b" is not at all the same thing in a natural language as "b = a". It is in fact closer to the reality of natural languages to regard linear progression as something relevant to the same extent as it is, e.g., in PASCAL, to get the sides right in "a:=b".

We have therefore seized upon the multiple-path predictive analysis and purified it by overcoming the few global-type procedures still left in its original version, thus arriving at our I.D.S. (= Integrated Dictionary Systems) method([9]).

The main inherent technical problem, much discussed in the M.I.T. publications of the early sixties, was the tendency of this method to generate uncontrollable branching-off, which at a certain length of sentence results in the need for introducing sequentiality, and beyond a certain length resulted in utter unfeasibility.

For one thing, some of today's micros are more able to handle this problem than the huge M.I.T. computers were in the early 60s. Moreover, PROLOG and similar high-level languages allow a much easier job for the programmer in simultaneous branching. The scale of the problem has been considerably reduced by twenty years of hardware and software development, while the complications of natural languages have remained the same.

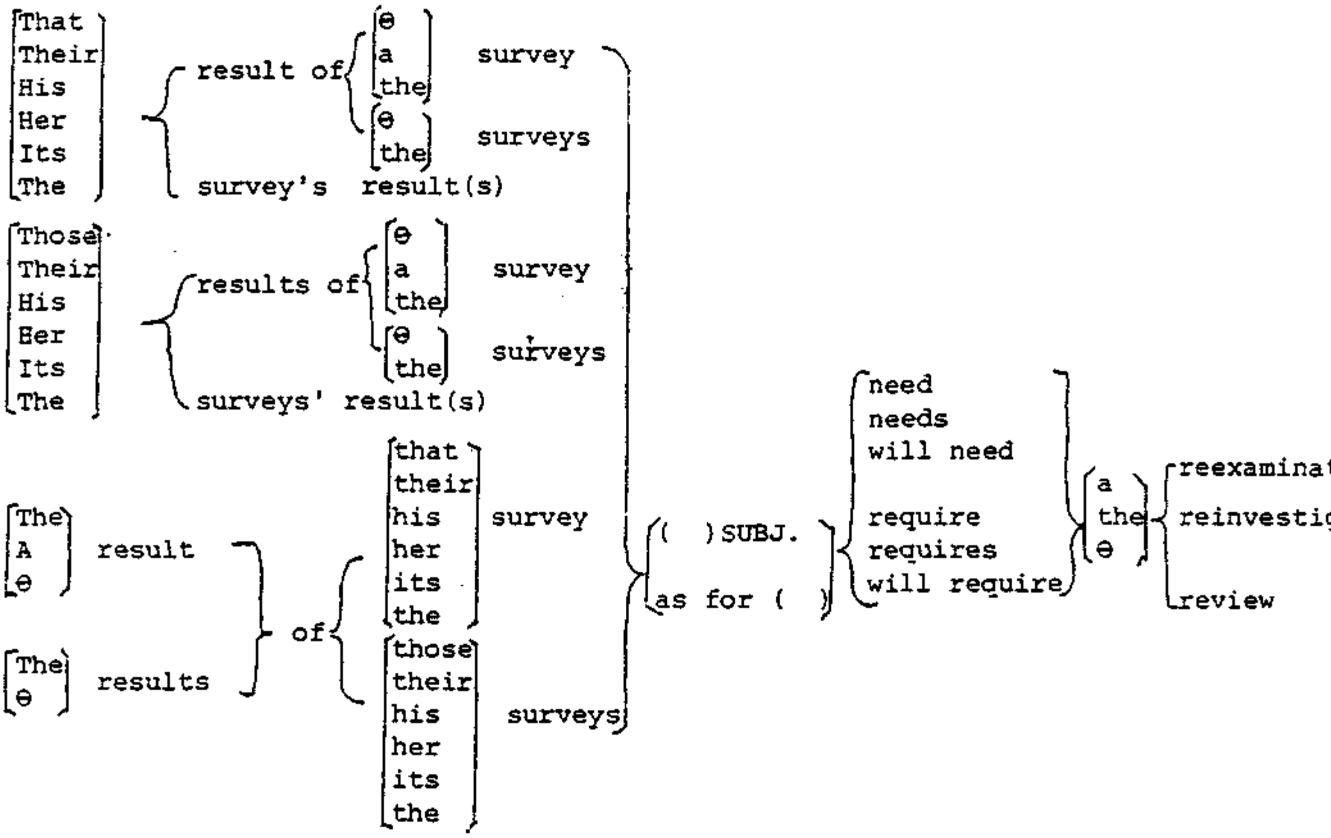
Even so, no-one knows quite how long a sentence of a natural language is allowed to be, and the problem remains in principle. While on the one hand we now have at our disposal various programming devices, such *a* "relaxation", it is hoped in the near future that a machine designed to combine a "declarative" (PROLOG-like) principle with a "functional" (LISP-like) principle may emerge([10]), solving our problem altogether.

4. Sentence-for-sentence Analyser

As stated before, the sentence-for-sentence analyser (Jap. to Engl.) was in our possession in 1968. During the last 15 years in Sheffield, it has not been possible to continue work on the analyser directly but, as a result of applying it to teaching and of continued linguistic research, numerous improvements have been achieved.

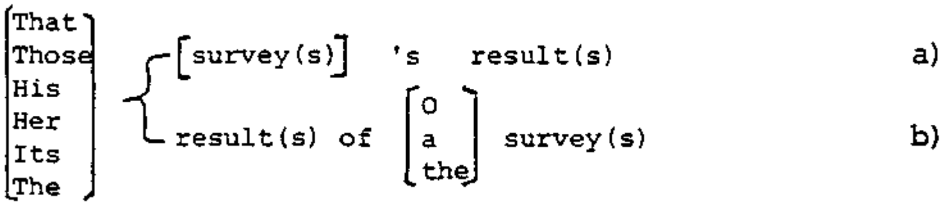
The analyser in its present form accepts an unpreedited Japanese sentence and returns a crude "English" output, as a rule in the form of a number of alternatives, e.g. as follows:

その調査の結果は再検討を要する。



If all the possible neat English alternatives were to be listed from the beginning to full, stop, we would obtain 4752 alternative English sentences, rendering this device too cumbersome for words, in its present form, as a fully independent translating machine.

However, when user-machine interaction is allowed at each stage of the decision-making, affecting the range of choices carries, a very much manageable output can be achieved. Thus, after



the user is given the chance to select the a) or the b) line, to decide between "survey/surveys" and "result/results" and also to select one out of "that-those/his/her/its/the". If the user exercises his choices to the full at this stage, which he may be able to do in the light of the prior context of this sentence, he will reduce the overall number of alternatives from 4752 to 36. Deciding subsequently whether "require" or "need" is better will halve this number to 18, and choosing between "DO(es)" .and "will DO" down to 9. Picking one each out of

$$\begin{bmatrix}\theta\\ \text{a}\\ \text{the}\end{bmatrix} \quad \begin{bmatrix}\text{reexamination}\\ \text{reinvestigation}\\ \text{review}\end{bmatrix}$$

will result in just one English translation for the Japanese sentence.

It is obvious that as a man-aided translating machine our analyser is now fairly adequate, although there is great scope for reducing the need for this interaction. One obvious (and large-scale) reduction will be achieved by bringing to bear the frequency count accumulated after a period of extensive interactive use. At each of the decision-making junctures, users' preferences will be registered in terms of frequency, and the most frequently preferred versions will be put out first, with the option available at the terminal of suppressing lower-order frequency choices. Only when the resulting output seems wrong or incomprehensible, the user may activate the suppressed choices to obtain an improvement.

This subterfuge, however, will only represent a substantial improvement in the burden imposed upon the interactive user, and will not amount to more than a marginal step ahead towards a fully independent translating machine. Where the choice represents the user's taste with both the available alternatives being equally correct (such as "need/require"), our translating machine may be allowed to reflect the prevailing taste of our users. But choices such as

$$\begin{bmatrix}\text{That}\\ \text{His}\\ \text{Her}\\ \text{Their}\\ \text{Its}\\ \text{The}\end{bmatrix} \quad \begin{bmatrix}\theta\\ \text{a}\\ \text{the}\end{bmatrix}$$

have little to gain from knowing their respective frequencies, as each of the options is bound to rank high compared with other words of the language. An automation of such choices awaits the incorporation of the theory of depredication.

5. English Language Output

As must be clear from the preceding Chapter, the hallmark of imperfection is too much choice. As we are not able at this stage to produce a single perfect or even acceptable English language version for the Japanese input, we have the problem of deciding how to represent the open choices for the user to avail himself of.

While, on the one hand, it is our consistent policy not to simplify matters and not to obscure any potentially eligible alternatives, we do not wish to drown the user in an unmanageable multitude of choices.

Having taught over a thousand people with the device in the book form, we are aware of the fact that some people can efficiently handle multiple choices (and those would resent having the range of available choices reduced). Others prefer a less defined smaller range of choices for inspiration (and those would be inconvenienced by a large number of defined choices, as well as resenting any reduction in the range of real choices). Others still (hard working, if less flexible) can only be presented with a very small number of choices, each of them well defined.

We can proceed in any of the three modes, and it would seem that *a* good machine should ideally allow the user a choice of "modes", and knowing of the existence of a large number of students whose inclination changes between Monday morning and Friday afternoon, there should also be a flexible and organic way of switching over mid-stream from one mode to another.

In this respect, some of the problems fully involved in the Teaching Machine overlap with the "user-friendliness" of the Man-aided Translating Machine, and there is advantage in pursuing both in parallel.

6. Other Language Pairs

The I.D.S. method has now been adopted for the creation of similar analysers, mainly for teaching purposes, between other pairs of languages: Russian-to-English(11), German-to-English (12) ,productive English-to-German for business correspondence(13), Latin-to-Dutch(14), Korean-to-English and Czech-to-English. Other applications are being considered.

It is important to realise that an I.D.S. analyser is not simply reversible, and a Japanese-to-English device cannot easily be turned into an English-to-Japanese one without a lot of additional research and rearrangement. (This is, of course, not to deny the usefulness of a shared dictionary data-bank.) Also, a "receptive" device such as our Japanese-to-English analyser is meant for speakers of English who do not know Japanese and would not be very useful to a Japanese native who wishes to produce English.

Moreover, one cannot necessarily make e.g. a Japanese-to-French analyser by just translating the English of the

Japanese-to-English device into French. Such a conversion, as has been found, requires much larger modification than might be expected.

All these points really inevitably follow from the fact that the I.D.S. method is strictly contrastive.

7. Japanese Script Input/Output

The original analyser was developed representing Japanese script by integers, thus avoiding the painful question, to which there was no better answer in the 60s.

Since then, especially in recent years, a variety of Japanese script keyboards have become available on the Japanese market, but for our purposes none of these are the ideal answer. All of them are either finite or contain few "free" slots for user-defined characters, whereas we require a very large number of characters (ca 7000) to start with, and carefully distributed reserved slots for later needs.

There is also a very important consideration to do with the fact that, for our purposes, the user cannot be assumed to have any knowledge of Japanese characters beyond a very elementary grasp of a "looking-up" procedure. The selection procedure for the purposes of input must therefore allow the user to identify the precise shape which occurs in his text on a keyboard arranged not on any phonetic principles - he does not know which sounds correspond to the characters - but on the steps of the looking-up procedure which he is familiar with.

Such a keyboard in any hardware form would be too unwieldy. On-the-screen selection is therefore called for, and a speedy well-defined dot-matrix retrieval system is the answer.

PERQ combines these features very well, indeed and we have formulated an AAA-ZZZ file of 32x32 dot matrices adequate for our purpose.

Quite apart from the application in our system, this I/O file can be regarded as a new type of Japanese script typewriter.

8. Automatic Integrated Dictionary Compiler

This is an interactive software eliciting new entries from the user and merging them into the existing body of the Integrated Dictionary.

This compiler, now available as the first product of our consortium, performs the following tasks:
a) decide whether a proposed new entry already exists or not,
b) if the entry does not yet exist, create it in its appropriate place, and do d)
c) if the entry already exists, compare each section to identify if the proposed entry contains any new information. If so, incorporate in its appropriate place.
d) from each alternative way of writing the entry word in Japanese script, seed a new entry in its appropriate place, if this does not yet exist.

e) monitor maximum length of entry, maximum length of each entry word, maximum number of alternative entry words, maximum number of alternative translations, maximum length of translation, number of entries in the dictionary.

Entries are packed into the dictionary text-file in the most economical way possible, with special symbols used for separating and quickly identifying each section.

This compiler exists in four different versions, of which only the first is a full-scale interactive program capable of initiating and compiling an Automatic Integrated Dictionary. The second version is intended for adding entries after all the grammatically unique entries are already formulated, and requires less decision-making about points of grammar by the user. The third version assumes that all the grammar-sensitive words of Japanese are already in and asks the user to make a choice out of only three word-classes. The fourth version, very simple and intended for the use of outside subscribers, accepts only new nouns, which of course represent 90% of the so-called scientific and technical terminology.

9. Automatic Integrated Dictionary

We use this term to describe a file of at least six thousand entries, each of which a) is accessible by inputting the appropriate sequence of Japanese graphemes either through their AAA_ZZZ representation or through on-the-screen selection of their dot-matrix images

b) contains at least one "meaning" defined by the substitutability of the entry word by a (possibly empty) set of alternative sequences of Japanese graphemes

c) contains in each "meaning" at least one "submeaning", defined by having one "entry code", which is one of four hundred numbers, each defining the grammatical acceptability of this submeaning at that particular point in the sentence

d) contains in each "submeaning" at least one "translot", which is defined as all the translations of this entry word in this submeaning which share a "continuation code", i.e. which generate the same predictions of how the sentence may continue from that point onwards.

e) contains in each "translot" at least one "translation", which may be either a word of English, a group of words of English, an instruction to reshuffle or modify what has already been obtained or what will follow, or it may be empty

f) contains for each entry, each of its meanings [1..16] , each of their submeanings [1..20] , each of the translots [1..20] and for each translation ([1...30] in each translot) a set of five "comments", which have to do with semantic, lexical and grammatical eligibility. One of the five "comments" is reserved for depredicational analysis (see Chapter 13)

g) also for each of the above, a frequency count is being stored.

10. Man-aided Translating Machine

The first version of this machine (JE3), which we could have in 1986, is a combination of the Automatic Integrated Dictionary with the on-the-screen selection Japanese script keyboard and a versatile HELP system.

When the still cumbersome inputting of the Japanese sentence reaches a point of some predictable continuation, the machine will offer such a continuation on the screen for the user to accept or reject, thus immensely speeding up the input process. By the same token, when an unexpected symbol is input half-way through a predictable stretch, a warning of possible error will be given.

In further stages, resulting from predictable developments in technology, this machine will be able to profit from direct optical input as soon as this becomes available, and the most cumbersome human operation will then be eliminated.

The HELP system will be available in several modes, partly selected by the user himself and partly resulting from the user's interactive behaviour. The most flexible area of this system will be in guiding the user to polishing the crude English output, by offering examples of similar type decisions. Guidance in back-tracking will be triggered off when the user loses himself half-way through a sentence, and the sentence may be presented to him in various stages of reduction, i.e. the "bare sentence" only, or that with only its immediate expansions, etc. When a "translation" of a sentence is achieved but the user pleads incomprehensibility, the machine will back-track over the textually most sensitive decisions, such as choices of anaphorics and ways of putting sections together.

Each time a translation has been completely approved of by the user, the machine will raise by one the frequency counts of all the nodes activated in that translation, as well as up-grade the steps in the HELP system which have proven useful.

This machine will improve by the incremental growth of the vocabulary stored in its AID as well as by a steady improvement in the order in which alternatives are presented, resulting from the current frequency count, and from the steady improvement in the HELP system.

The next generation, however, cannot be achieved without a new quality added to the analytical power of the machine, allowing it to draw links beyond the borders of sentence.

11. Japanese-English Teaching Machine

This machine (JETMR 1,2) consists of a programmed course built around the Automatic Integrated Dictionary and aims at eventually emancipating the user from this dictionary. The set of rules which governs the I.D.S. search (see App.1) is imparted to the student through a sequence of graduated exercises, until he behaves to all intents and purposes like a translator, referring to dictionaries only for unusual lexical items.

For the last 13 years, we have been running a seven week Japanese-to-English translating course, both in Britain and in Japan, with a considerable measure of success. Being an application of our syntactic analyser to teaching, this course has undergone numerous revisions, incorporating every improvement achieved in the analyser itself as well as all the experience we have gained on the pedagogic front.

The course now comprises four bulky publications and five further volumes, although compiled and printed, can no longer be handled by our departmental "Publishing house". Any further revisions or additions have now become technically unfeasible in the hitherto fashion.

For the above considerations alone, it has become necessary to computerise this course. When the complete analyser becomes a computer software, updating will no-longer be an insurmountable problem of time and expense, and size will be measured by very much more acceptable criteria.

In particular, the computer-based course will cut down on the student's time and labour by eliminating the traditional dictionary search. An efficient HELP system will facilitate the student's compliance with the rules of search and help him in making decisions.

The eight specialised panels (chemistry, shipbuilding, electronics, geography, economics, food industry, metallurgy and linguistics) which we already have and another dozen of panels partially completed could not so far be properly incorporated in the course because of the shortness of its duration and the cumbersome bulk of printed hand-out. In a computerised version of this course, an increasingly tailor-made personalised set of exercises can be offered to each student, producing higher motivation and better results.

The development of a teaching machine is one of the most promising and worth-while immediate developments and inspires much enthusiasm amongst our collaborators.

12. Japanese-English Scientific & Technical Data Bank

The lexical limitations of the first Automatic Integrated Dictionary are manifest by the fact that it contains only some six thousand entries, of which over a half are grammatical entries rather than genuine lexical items.

While the dictionary remains limited in its coverage of vocabulary, particularly that of scientific and technical nature, the user has to look up a certain percentage of lexical items, mainly straight-forward nouns, in traditional dictionaries. Although the analyser does provide guidance as to where best to find the missing item and how to place its translation in the resulting "English" sentence, the need to refer to outside data slows down the process and can cause errors.

We are planning for an incremental growth of the AID, concentrating at first on all inflected items of Japanese and trying to incorporate all items of idiomatic nature, but gradually covering more and more purely lexical items. All the known productive principles of word formation and derivation have now been incorporated, thus significantly enlarging the lexical powers of the tool. This works fairly well on Japanese, which tends to abide by well-defined principles of forming scientific and technical terminology, and the "English" rendering of such terminology tends to be comprehensible, if not directly usable. English is often haphazard and eclectic in scientific terminology, relying heavily on greco-latin word stock.

The ultimate solution can thus only be achieved through a massive analysis of and excerption from Japanese texts and their verified translations, as well as the incorporation of all the existing specialised word lists.

We are lucky in this respect, since probably the world's largest Japanese-English scientific and technical word list (= the Gerr File) is in our custody, contributing ca. 500 000 terms.

However, it is to be expected that by indefinitely expanding the number of lexical entries available to the analyser, we are bound to slow down the process of analysis and increase the multiplicity of homonymy which needs to be resolved.

The answer to this problem must be provided at a rate matching the increment. Partly, a speedy hardware will help. A frequency count on each item handled during the analysis and a presentation of results in the order of their hitherto frequency is another partial solution. Also, a detailed set of "conditions for acceptance", which are carried by each entry and each English translation, will help to weed out some unwanted alternatives.

13. Depredicational Analysis

The principal methodological feature of our model of grammar is that it is based on the notion of valence. This will be in our case "syntactic valence" and compared with those (Benveniste, Wenck, Rickmeyer)[15] who had used this term, a considerably broader and farther-reaching notion.

In our system, every syntagma (see below) has a valence which is identifiable as one of the finite set of those possible in the given language.

A syntagma is a coherent construction (or: formation) on any level of the syntax of the language. It can be anything from the lowest meaning-endowed unit (i.e. morpheme) up to the whole sentence, so long as it has autonomy of explication (: in Chomsky terms, so long as it can be generated as a whole from a non-terminal symbol).

The syntactic valence of a syntagma is defined by:

1.the precise set of all the positions which this syntagma may occupy in a correctly formed sentence (and one should really say all the correctly formed higher-level units) of the language vis-a-vis all the other syntagmata,

2.for each (type) of these positions, the precise set of all the functions which this syntagma can fulfill (= which explications it generates).

This is why the field research method which yields the necessary descriptions for IDS purposes is called the "Distributional and Functional Analysis".

Valence is thus the syntagma's complete set of possibilities within the grammar, or the syntagma's "syntactic prophile". Apart from that, of course, each concrete syntagma has its unique meaning.

Syntagmata divide into classes according to their valence. A special set of such classes are those of all the terminal (i.e. units of input, i.e. those not divisible further in terms of the given grammar) symbols in syntax. These are called the word-classes. We have found empirically in the case of Japanese that the word-classes cover the whole range of the existing types of valence. In other words, all syntagmata larger than words/terminal symbols can be assigned a valence of a single word/terminal symbol.

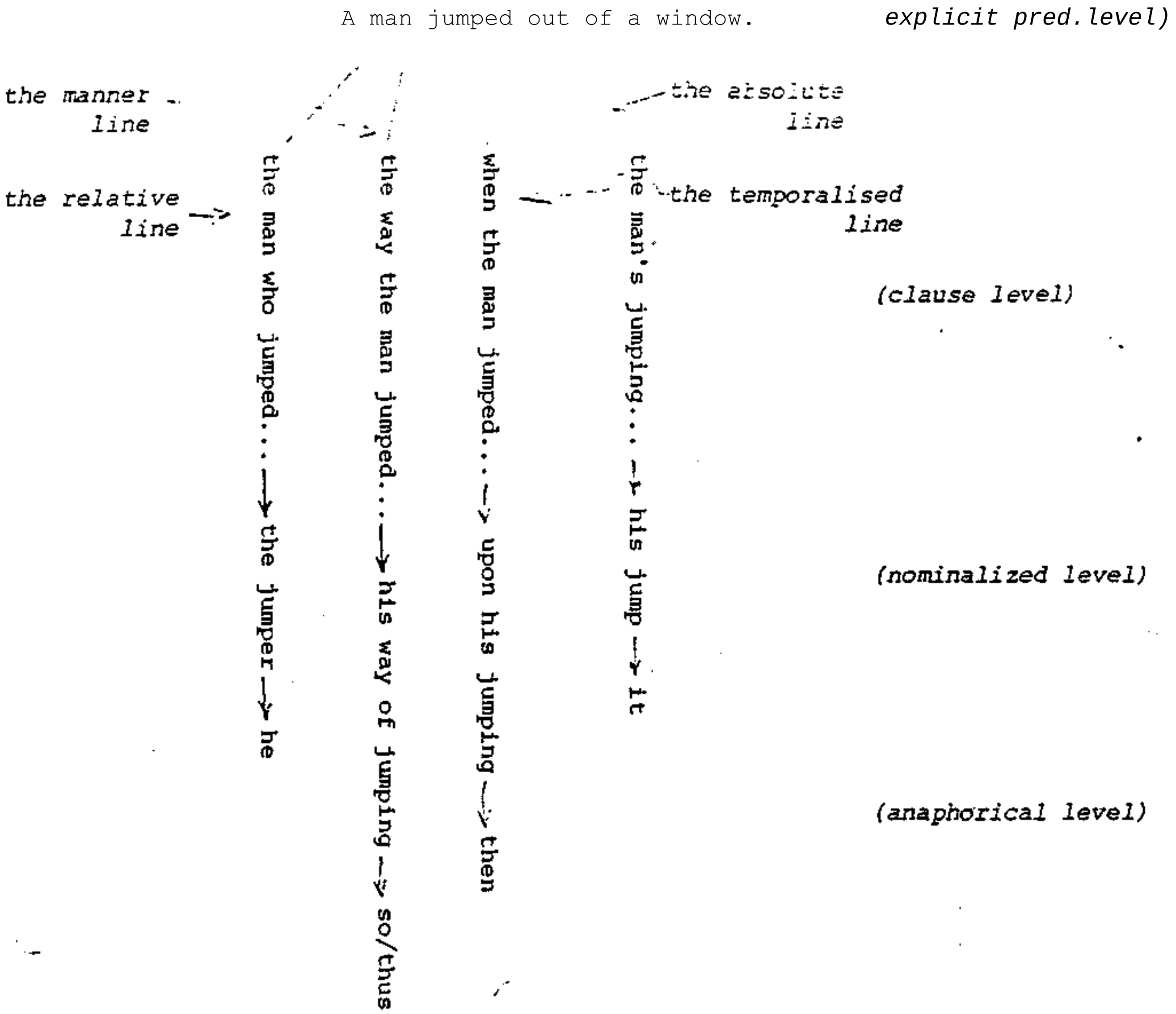
There is a purely grammatical way, no different from sentence grammar in its ability to succumb to formal rules, of handling units larger than sentence: in fact the larger the better. This is by formalising the notion of depredication(16) and thus allowing valence to apply to this area.

A. Let us first suppose that every thought which is ever conceived in the form of language is first conceived explicitly, before it can be "packed" into implicitness. (Note that this assumption does not necessarily entail any insistence that each thought must first be stated or formulated explicitly.)

B. Further suppose that once a thought is conceived, whether explicitly formulated there and then or only tacitly passed over, it has an irresistible tendency to become "packed" so as not to clutter up the store.

To use linguistic terms, a newly conceived thought, when expressed, takes the form of an explicit predication, i.e. associates all the required members of sentence with an actualized finite verb. In such a predication the largest range of grammatical distinctions is available, e.g. tense, person, voice, aspect. As this becomes older, the predication - while still retaining its original reference - gradually becomes deprived of certain internal grammatical distinctions, i.e. becomes depredicated.

The valence of explicit predication is that of the full sentence, and as it becomes depredicated it is reduced to the valence of one of the terminal symbols (i.e. a word-class).


A man jumped out of a window.
explicit pred.level)
the manner line
the absolute line
the relative line
the temporalised line
(clause level)
(nominalized level)
(anaphorical level)
the man who jumped... → the jumper → he
the way the man jumped... → his way of jumping → so/thus
when the man jumped... → upon his jumping → then
the man's jumping... → his jump → it

And there are dozens of other lines of depredication.

In all natural languages, depredication takes the form of grammatical regularity and predictability to the extent that it can be formalised almost as tightly as sentence-bound syntax. If this were done, the computer would "know" for every 'it', 'they', 'his' etc, what this refer to in a previous text. Then we should be paraphrasing connected texts rather than isolated sentences.

The slight snag implied by the word "almost" lies in the possibility of smuggling in synonyms, which is more fashionable in some languages (e.g. English) than others (notably Japanese or, even more so, Chinese). It will take very much more work before depredication can be formally captured including the possibility of "A man jumped out of a third-floor window."

↓

"Such (recklessness) is to be deplored."

14. Text-wide Integration

By extending the multiple-path predictive analysis beyond the borders of sentence, that is to say by carrying predictions for all possible linkages from source to every "it", "his", "the" etc., a text-wide integration can be achieved.

Such an integration would result in a drastic reduction of the choice of anaphorics even before the semantic rephrasement snag is overcome, and would also offer a glimmer of real hope for automatic text comprehension, scanning etc.

Even before we can establish exactly how far-reaching such tactics promise to be, we have of course a very real practical problem of processing capacity. Even on PERQ, the extent up to which previous text could be brought into account could not go as far as a full page, and possibly much less than that.

Initially, using the reserved slot in the "comment" register of each section of each entry (see Chapter 9), we hope (in the 3rd phase of our project) to introduce text-wide integration over the length of a paragraph, or a defined maximum number of lines, whichever is smaller. From that initial stage, serving as a prototype, we hope to discover a possibility of a "sliding range" of text integration, which seems to us nearer the way depredication operates in real life.

15. Japanese-to-English Translating Machine

A machine accepting standard Japanese texts and producing their finished and acceptable English equivalents without human intervention would merit the above name. Whether this can be achieved in every sense of such a trade description is still only a matter of faith.

It has however proven useful to have precisely such a machine as our research target for the last twenty years, and useful "by-products" have resulted from aiming at the perhaps impossible.

What is beyond challenge is the fact that we are now able to make a good man-aided machine, and we know of a number of ways in which this machine can gradually be made less dependent on human interaction.

At the stage when text-wide integration begins making inroads into the human monopoly of choosing anaphorics, the AID will be one of the largest existing dictionaries from Japanese to English and will continue approximating to comprehensiveness. When the actual shape of achievable output at that stage is known it is more than likely that, as always happened before, the next step or two will become obvious.

It is intuitively clear, although the formulated theory does not go that far, that depredication analysis is the answer to text-wide integration, and that rephrasement too could be fully incorporated into this analysis if a semantic descriptions calculus could be developed which is fully subordinated to depredication analysis.

The existing approaches to semantic analysis of language fail to be of use because they are based on deductive principles, divorced from text analysis. What we need for our purposes is a semantic calculus based on the possibility of textual substitution. Such a semantic calculus would of course result in full incorporation of semantic descriptions into grammar, complementing our notion of valence, which is itself based on substitutability.

An inductive process, slow as it may be,would be a safe basis for a gradual construction of a fully integrated text-wide analyser.

It should be borne in mind, that once a Japanese-to-English Translating Machine has been achieved, it would for the first time be reversible, allowing at very little extra effort the formulation of an English-to-Japanese device, and also allowing automatic synopsis, automatic text comprehension and other dreams of the future.

It seems to us that this future is not very far, so long as there is to be any future at all.

Bibliography:

1. J Jelinek: A Syntactic Analyser of Japanese, Joho shori gakkai shiryo 67-1, Tokyo 1967

   J. Jelinek: Sentence-for-sentence Analyser of Japanese, Prague Bulletin of Mathematical Linguistics 8, Prague 1968

2. Final Report to H.M. on the Development of Japanese Reading Course, Centre of Japanese Studies, University of Sheffield 1973

3. J.Jelinek: Japanese-English Grammar Dictionary (001), Sheffield 1974

   J.Jelinek and P.A.Heron: Reading Japanese (002), Sheffield 1975

   J.Jelinek: Integrated Japanese-English Grammar Dictionary (004), Sheffield 1976

   J.Jelinek: Reader in Scientific & Technical Japanese (O11), Sheffield 1978

4. Kenkyusha's New Japanese-English Dictionary, ed. by K.Masuda, Tokyo 1974

   A.N.Nelson: The Modern Reader's Japanese-English Character Dictionary, 2nd ed., Tokyo 1974

5. M.I.T. Machine Translation Project Reports (8 vol. between 1959 and 1964), dir. Oettinger, Cambridge, Mass. (In U.K. available only at National Physical Laboratories in Teddington)

6. See V.Mathesius, K otázce tzv. aktuálního členění větného, Čeština a obecný jazykozpyt, Praha 1946. An analysis of English, German and Czech, based on the idea of theme-rheme segmentation, is given in a series of papers by J.Firbas, e.g. Thoughts on the Communicative function of the Verb in English, German and Czech, Brno Studies in English I, 1959; On the Communicative Value of the Modern English Finite Verb, Brno Studies in English III, 1961

7. N.Chomsky in Syntactic Structures (1955) and more specifically e.g. in On the notion "Rule of Grammar", in Structure of Language and its Mathematical Aspects, Proceedings of Symposium in Applied Mathematics, AMS 1961

8. N.Chomsky: Aspects of Structural Syntax, Mouton, Hague 1968

9. J.Jelinek: Distributional and Functional Analysis, in Sheffield Studies in Japanese I., Sheffield 1979

10. Declarative Systems Architecture, ed. R.Kowalski & J.Darlington, in Vol.2 Section 1 of Intelligent Knowledge Based Systems, SERC-Dol 1983

11. P.A.Heron: Integrated Russian-English Dictionary,
Aston University Press, Birmingham 1974

12. P.A.Heron: Reading German, enquiries to Dpt. of
Modern Languages, Aston University in Birmingham

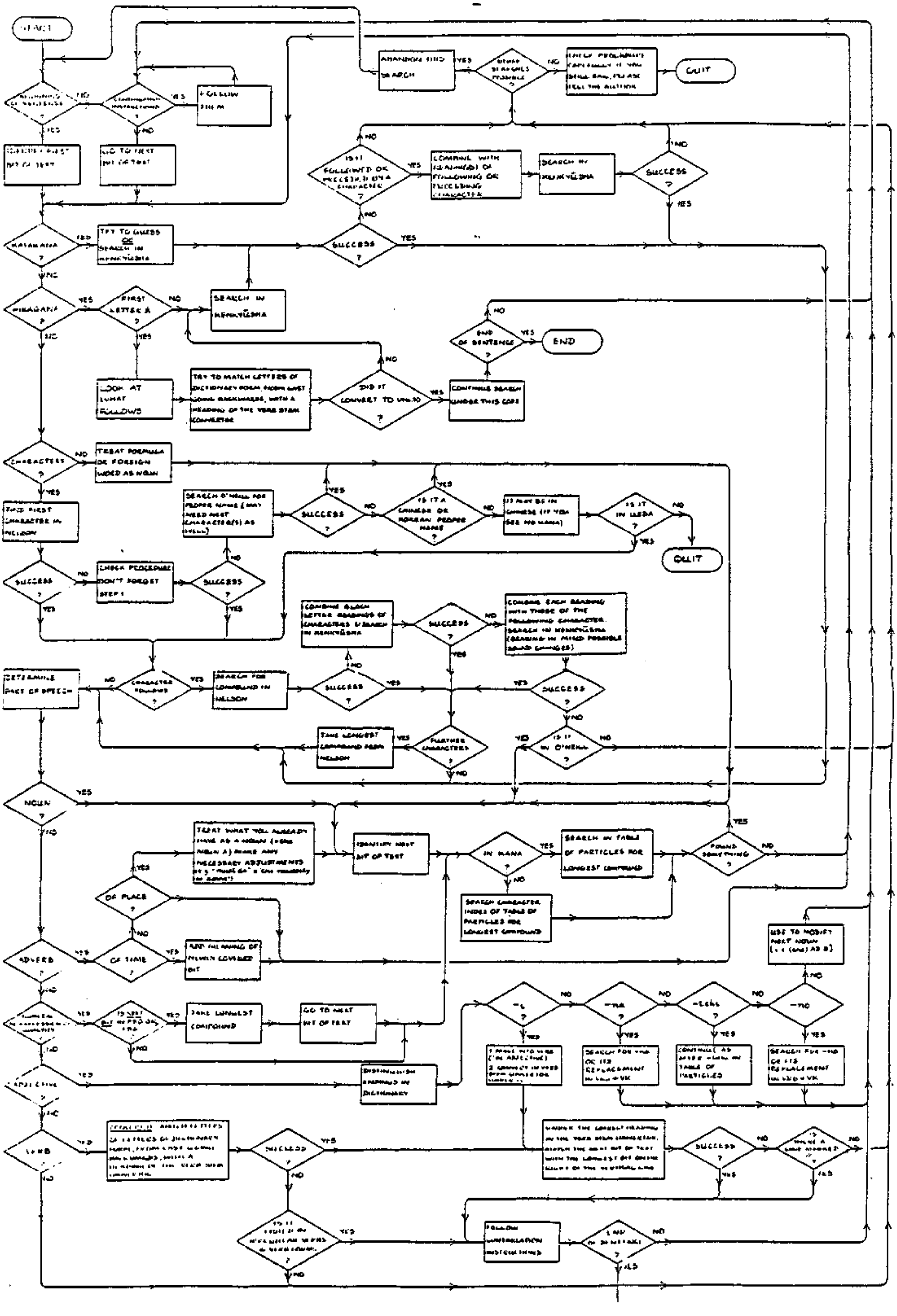
13. unpublished, refer to Sheffield Polytechnic Dpt. of
Modern Languages

14. refer to Drs Groen at Dpt. of Latin, Katolieke Hogeschool
Eindhoven, or Drs Meijers at the Taalcentrum,
Katolieke Hogeschool Tilburg

15. E. Benveniste: La phrase relative, problème de syntaxe
générale, Bulletin de la Société de linguistique
de Paris 53 No. 1
Gunther Wenck: Systematische Syntax des Japanischen
(I, II, III), Wiesbaden 1974
J.Rickmeyer: Kleines japanisches Valenzlexikon,
Hamburg 1977

16. J.Jelinek: A Linguistic Aspect of Transformation
Rules, AUC-Slavica Pragensia VII, 1965
J.Jelinek: Construct Classes, Prague Studies in
Mathematical Linguistics 2, 1966

THE INTEGRATED DICTIONARY METHOD FOR JAPANESE→ENGLISH TRANSLATION


START
BEGINNING OF SENTENCE ?
CONTINUATION INSTRUCTIONS ?
FOLLOW THEM
IDENTIFY FIRST BIT OF TEXT
GO TO NEXT BIT OF TEXT
ABANDON THIS SEARCH
OTHER SEARCHES POSSIBLE ?
CHECK PROCEDURE CAREFULLY. IF YOU STILL FAIL, PLEASE TELL THE AUTHOR
QUIT
IS IT FOLLOWED OR PRECEDED BY A CHARACTER ?
COMBINE WITH READING(S) OF FOLLOWING OR PRECEDING CHARACTER
SEARCH IN KENKYŪSHA
SUCCESS ?
KATAKANA ?
TRY TO GUESS OR SEARCH IN KENKYŪSHA
HIRAGANA ?
FIRST LETTER A ?
SEARCH IN KENKYŪSHA
LOOK AT WHAT FOLLOWS
TRY TO MATCH LETTERS OF DICTIONARY FORM, FROM LAST GOING BACKWARDS, WITH A READING OF THE VERB STEM CONVERTOR
DID IT CONVERT TO VERB ?
CONTINUE SEARCH UNDER THIS ONE
END OF SENTENCE ?
END
CHARACTERS ?
TREAT FORMULA OR FOREIGN WORD AS NOUN
FIND FIRST CHARACTER IN NELSON
CHECK PROCEDURE DON'T FORGET STEP 1
SEARCH O'NEILL FOR PROPER NAME (MAY NEED NEXT CHARACTER(S) AS WELL)
IS IT A CHINESE OR KOREAN PROPER NAME ?
IT MAY BE IN CHINESE (IF YES SEE NO NAMA)
IS IT IN LLEDA ?
QUIT
COMBINE BLOCK LETTER READINGS OF CHARACTERS & SEARCH IN KENKYŪSHA
COMBINE EACH READING WITH THOSE OF THE FOLLOWING CHARACTER. SEARCH IN KENKYŪSHA (BEARING IN MIND POSSIBLE SOUND CHANGES)
DETERMINE PART OF SPEECH
CHARACTER FOLLOWS ?
SEARCH FOR COMPOUND IN NELSON
FURTHER CHARACTERS ?
TAKE LONGEST COMPOUND FROM NELSON
IS IT IN O'NEILL ?
NOUN ?
TREAT WHAT YOU ALREADY HAVE AS A NOUN (=THE NOUN A) MAKE ANY NECESSARY ADJUSTMENTS
IDENTIFY NEXT BIT OF TEXT
IN KANA ?
SEARCH IN TABLE OF PARTICLES FOR LONGEST COMPOUND
FOUND SOMETHING ?
SEARCH CHARACTER INDEX OF TABLE OF PARTICLES FOR LONGEST COMPOUND
OF PLACE ?
ADVERB ?
OF TIME ?
ADD MEANING OF WHEN COVERED BIT
USE TO MODIFY NEXT NOUN (i.e. (one) AS B)
NUMERAL OR EXPRESSION OF QUANTITY ?
TAKE LONGEST COMPOUND
GO TO NEXT BIT OF TEXT
ADJECTIVE ?
DISTINGUISH ENDINGS IN DICTIONARY
-i ?
-na ?
-teki ?
-no
SEARCH FOR THE OR ITS REPLACEMENT + VK
CONTINUE AS AFTER IN TABLE OF PARTICLES
SEARCH FOR -NO OR ITS REPLACEMENT + VK
VERB ?
MARK THE CORRECT READING IN THE VERB STEM CONVERTOR. MATCH THE NEXT BIT OF TEXT WITH THE LONGEST BIT ON THE RIGHT OF THE VERTICAL LINE
IS THERE A LINE MARKED ?
FOLLOW CONTINUATION INSTRUCTIONS
END OF SENTENCE ?
YES
NO