[From: *Meta* 15 (2), June 1970]

pidgin translation

During the brief history of Computerized Translation (CT) so far, its practitioners have had to ask their sponsors to accept something less than «good translation» in the traditional sense. The sponsors have generally remained dissatisfied, although there are CT systems which do function and their admittedly imperfect output does find a few users [1].

Traditionally, a good translation — we need hardly say it in these pages — should read as though it were written by a native speaker of the target language. Furthermore the translation ought to have the same level and excellence of style as the original. Some of the early researchers in CT, realizing the limitations both of forseeable computer techniques and of current linguistics, as well as the inherent limitations of computing machines themselves, did indeed question whether the traditional ideal of translation was appropriate to CT. We shall review some of their efforts below. Nevertheless, virtually all CT research since 1960 has been conducted with a view toward translating from one natural language into a fair approximation of another such language. This is true, for instance, of the project at the Université de Montréal, where the aim is to translate government publications from English into « good » French, using a minimum of post-editing (Dugas, 1969). In the jargon of CT, « post-editing » means putting a skilled writer of the target language to revise rough translations produced by computer.

---

1. The systems referred to are those of : i) The U.S. Air Force Foreign Technology Division, at Wright-Patterson Air Force Base, Dayton, Ohio, ii) The Atomic Energy Commission, at Oak Ridge National Laboratory, Oak Ridge, Tennessee, iii) Centre européen du traitement de l'information scientifique, at Ispra, Italy.
   The most widely used term for translation done with a computer is «machine translation» (often abbreviated MT). However, MT has become associated in many people's minds with attempts to produce a particular type of translation from which we dissociate ourselves
   in later paragraphs. For this reason we introduce « computerized translation », or CT, as a broader term to cover *any* mode of translation that can be implemented on a computer. To readers interested in the « brief history » of CT, we recommend Pendergraft, 1967, even though it makes scant mention of research done outside the U.S.

Although one may sometimes lose sight of the fact, even translations done by humans do not always reach the high standard that professional translators set themselves. In wide areas throughout the world, speakers having something less than mastery of a useful second language nevertheless manage to make themselves understood in it. Whole communities of these imperfect polyglots have developed mongrel tongues called generically « pidgins » (Hall, 1954[2]). Not only is the grammar of a pidgin English simplified from that of standard English, but a linguist can always detect considerable influence of the speaker's mother tongue on the pidgin. Thus the Chinese who asks for *two piecee apple* (2 apples) or the French speaker who says *me, I am not in agreement* (personally, I don't agree) is recognizably carrying over features of his native language into his « English ».

However, we are not concerned with spoken pidgin in this article beyond showing why the term pidgin can be loosely applied to translations already produced by CT, and to artificial language intended for CT production in the future. Masterman observed pithily that « machine translation always is a pidgin » (Masterman, 1967). She meant that the output from CT systems shows all too obvious vestiges of the language being translated. But this kind of pidgin is involuntary and unprincipled. It is therefore essentially different from what we have in view.

Great problems still face anyone who aims to make CT rival traditional translation; to some of these problems no early solution can be expected (Bar-Hillel, 1964; Hofmann, 1968a). Consequently it is still worth discussing and investigating simpler kinds of translation. Taking a look first at traditional translation, we consider that it attempts a twofold task :

a) to convey the cognitive meaning of the original;

b) to re-express this meaning in a version of the target language that is stylistically acceptable, without losing any of the emotional and evaluative overtones imparted by nuances in the original.

CT has typically been sponsored for the sake of obtaining speedy translations of scientific and technical documents. For this purpose, task a) is far more important than b). We assert therefore that the prime requirement for CT products that their *content* should be understood by the readers for whom they are intended; one's feeling about the style of *expression* is, to say the least, a secondary consideration.

In the past, CT researchers have not expected their readers to undergo any special training at reading the translations in spite of some unaccustomed features in them. In some of the early systems, for instance, problems of ambiguity were partially overcome by the expedient of printing out small sets of alternative translations and leaving readers to choose among them. This did not prove popular, presumably because « a reader is less confused by a text containing occasional vague equivalents than by one containing all the possible equivalents of

2. The distinction between pidgins and Creoles does not concern us here.

every word » (IBM, 1959, Vol. 6, p. 3). Such success as CT has had, has been due to the nature of the texts selected for translation: the choice has ensured that the readership be limited to people with a good background knowledge of the subject treated. It has been shown that such readers are capable of extracting some worthwhile information even from very defective CT translations [3]. Even so, why do readers accept inferior translations which they can only understand in part ? The alternatives open to them are:

1) To press for human translations. These ought to be better, but they take considerably longer to produce. Speed is the very *raison d'être* of CT.

2) To learn the language of the original. This, however, would practically oblige the researcher to learn as many languages as are used to publish original articles in his field — an awesome task even for linguists. Moreover some languages important for a researcher, for example those with different writing systems from his own, present exceptional difficulties to the learner in a hurry.

These objections lead us to pose a question that is all-important if one is to appreciate the usefulness of translation into pidgin-like languages. Is it better to delay for months, perhaps years, in order to learn a foreign language; or to spend a few hours, even days, getting used to the peculiarities of pidgin English or pidgin French? Many people in pidgin-speaking areas of the world have already found that, for practical purposes, pidgin will do.

On the other hand, not all CT researchers expect their translations to be handed « raw » to readers. The U S. Air Force system uses extensive post-editing, while Booth describes her research at the University of Saskatchewan as « machine aided translation with a post-editor » (Booth, 1967b). The drawbacks are obviously that skilled human intervention is still required for the post-editing, and that some of the time saved by CT is lost again. Rather than spend large sums on CT research, would it not be better to spend it on training more translators, who could make a translation from the source just as quickly as they could revise a poor CT one? The answer to this criticism lies in the present shortage of translators coupled with the ever growing volume of translation to be done. There are grave shortcomings even with English and French in Canada, where there is a reserve of bilinguals who can be used as translators. Consider then the problems of translating Chinese scientific reports. Yet Chinese currently presents one of the most urgent translation problems in the U.S. (See, 1967). When bilingual people are scarce it should be particularly valuable to have a CT system providing adequate raw translations for polishing by monolingual post-editors, who ought to be more readily obtainable.

---

3. Carroll, 1966; Orr, 1967. The latest evaluation to come to hand is the following claim for the system now in operation at the Euratom centre at Ispra in Italy : « *A Ispra, les traductions sont livrées aux clients sans révision. Bien que la qualité des traductions ne puisse être consideree comme parfaite, elle semble néanmoins adéquate aux exigences de l'information, puisque les clients n'ont pas fait usage de la faculté qui leur est offerte d'une traduction « humaine » des textes, dans le cas où ils ne seraient pas satisfaits de la traduction automatique.* » (Perschke, 1968).

However, it is always difficult for a post-editor to understand a rough translation which does not convey all the content of the original. Interpretive help may be provided by background knowledge of the subject matter or by sheer professional experience as a translator, but there is the constant danger of misinterpreting. For this reason, we consider that a satisfactory rough translation intended for post-editing must conserve all of the semantic and grammatical information contained in the original. We pursue this ideal below in the proposals for « specific pidgin » (see also Hofmann, 1968b).

As for the current output from CT systems, Masterman's complaint that its « characteristics *per se* are never investigated » is as apt today as when she first made it ten years ago (Masterman, 1967[4]). True, there have been investigations in which « consumer opinion » was tested; but the test material was a product with the arbitrary merits and defects that resulted from an existing system. So far as we know, the problem has never been looked at the other way round, which is to test readers with deliberately designed « simplified » translations before designing the system that would produce them.

Sponsors of future CT projects would be well advised to avoid the « hit or miss » approach typical of most CT projects to date by attending first to the goals: what is minimally acceptable, and how much tolerance readers show for various ways in which CT output may deviate from a man-made translation par excellence. With these goals duly specified, CT could become an engineering problem to which the criterion of the cost-quality relationship could be effectively applied. We shall claim below that human readers can tolerate considerable lack of quality provided certain conditions are fulfilled[5].

II

Mark Twain said in a speech to the Vienna Press Club :

> I am indeed the truest friend of the German Language — not not only now, but from long since — yes, before 20 years already... I would only some changes effect. I would only the language method — the luxurious, elaborate construction compress, the eternal parenthesis suppress, do away with, annihilate; the introduction of more than thirteen subjects in one sentence forbid; the verb so far to the front pull that one it without a telescope discover can. With one word, my gentlemen, I would your beloved language simplify so that, my gentlemen, when you her for prayer need, One her yonder-up understands.
>
> ...I might gladly the separable verb also a little bit reform. I might none let do what Schiller did; he has the whole history of the 30 Years' War between the two members of a separable verb in-pushed. That has even

4. Most of Masterman's article was actually written in 1960 : see her footnote, p. 197.

5. Numerous writers have criticized man's natural tendency to try and make computers simulate man. One could cite Arthur C. Clarke, Wayne Danielson, and the Sedelows. Meanwhile the few remaining sponsors of CT still hope that the computer will produce translations in the same sort of language as we are used to. « To require computer-generated language to conform in detail to all, or most, of a human's linguistic conventions is analogous to requiring early printers to make their output resemble manuscript writing. Initially, just such « magic realism » may be demanded, to demonstrate the machine's true virtuosity and intelligence as well as to fully achieve communication. Later we may learn to accept as satisfactory very different output language from the computer. » (Sedelow, 1967, p. 210).

> Germany itself aroused, and one has Schiller the permission refused the History of 100 Years' War to compose — God be it thanked! After all these reforms established be will, will the German language the noblest and the prettiest in the world be.
>
> (quoted in BB&N, 1966)

This passage was meant to satirize the extremes of dissimilarity between German and English syntax, yet it is quite easy to understand; and it illustrates how quickly a reader can become accustomed to a very different word order. In the first section of this article we said that people can and often do make use of « compromise languages », i.e. pidgins, when the only goal is the exchange of practical information. And reviewing the difficulties encountered in trying to translate by computer, we suggested that the machine might convert a text written in one language into a pidgin-like language similar enough to another language for the pidgin to be read with little training by speakers of that other language. Furthermore such a pidgin might be suitable for rewriting into a stylistically acceptable form of the target language by a post-editor who did not know the source language. In this section, we want to propose a definition of an ideal pidgin which we shall call « specific pidgin », and from it we may draw certain conclusions about the feasibility of pidgin translation.

We can speak of the process and the product of pidgin translation apart from the computer implementation of it. Let us then delay any further mention of the computer until after we have taken a closer look at the product and its virtues, and have related these to the work of our precursors.

Specific pidgin translation is characterized by three constraints:

1) Very nearly all of the meaningful elements in the text after translation are composed of target language words.

2) A single target language word (or phrase) is used in all contexts to translate each source language word. For a source word which has several distinguishable meanings (polysemes), one tries to find a target language word which comes close to covering all of them; as a last resort one might create a new word out of morphemes of the target language and assign just these meanings to it (as suggested in Masterman, 1967).

3) No information present in the syntax or the lexical contrasts of the source language is lost in translation.

Natural pidgin languages exhibit characteristics which, though they are much less rigid, correspond to our proposed constraints; that is why we call translation of this nature « pidgin translation [6] ». We further qualify our proposed variety as « specific » to distinguish it from the unspecific or universal pidgin translation proposed by Richens and Booth (Richens, 1955) and later « sophisticated » by Masterman (Masterman, 1967). Their versions were « unspecific » in

6. Note especially that the syntactic structures and lexical contrasts in the pidgin will be those of the *source* language. So Caribbean *créole* usually sounds degenerate to French speakers from other areas because they cannot help recognizing most of the words in it as French but they are ignorant of all that is imparted by its Afro-Caribbean substratum. Likewise any of our pidgins may seem degenerate, but only until the reader comes to acquire its source structure by using it.

the sense that these earlier workers translated from a number of source languages into a similar number of English-like pidgins and then compressed all the pidgins into a single language. They thereby destroyed a considerable amount of information carried by lexical contrasts in the source languages. As an example of lexical contrast in English, we may cite the choice of *ask* instead of *demand* or *request;* we would insist that these three be translated by three distinct words in a specific pidgin. In addition to losing lexical information, our precursors also destroyed most of the grammatical information: the grammars of different languages are sufficiently different that if several grammars are conflated the result is no grammar at all, just strings of unrelated words. Even in the case systems, much of the distinctive information was destroyed: for instance, in making pidgin English from German and Russian, which both have a dative case, both datives were represented in the pidgin by *d* in spite of the fact that they are used quite differently in their respective languages. The remaining grammatical information was further reduced by discarding important grammatical suffixes altogether if they were not easily disambiguated [7].

Richens and Booth employed the « most frequently used equivalent » to translate lexical items. We would criticize this on two counts: first, the translation so chosen for one source word (its « translate ») could be also chosen for another source word, thus losing information about the lexical choices made by the author; and secondly, it often happened that a word's translate did not have all or even most of that word's meanings and could thereby prove quite inappropriate in some contexts. To avoid this deterioration, we impose on our type of pidgin an additional but less rigid requirement that:

4) A translate need not be a frequent word in the target language, but its range of meanings must match as closely as possible the range of meanings of the source word it stands for.

This constraint is nicely exemplified by the French word *langue.* It has two principal meanings, one translatable into English as *language,* the other as *tongue.* Even though the word *tongue* has an archaic flavour when used to mean « language », the more important consideration for our purpose is whether the resultant translations can be understood. Thus, using *language* as the translate allows correctly *The study of languages is interesting* but also the horribly incorrect *He stuck his language out of his mouth.* Using *tongue* as the translate for *langue* admits unusual translations, but they are always understandable: *The study of tongues is interesting. He stuck his tongue out of his mouth.*

Richens and Booth used multiple translates to stand for a single source language word which had widely divergent meanings, i.e. they presented alternative translates to the reader, who had to chose among them. While this seems a reasonable thing to do and probably does aid the novice reader, it is ruled out by our conception of pidgin translation. Moreover it has been found detrimental to understanding, as we mentioned earlier; and in any case, after reading a

7. Richens and Booth reduced all such information to *v* (vacuous) or *z* (unspecific), while Masterman went further and deleted these marks as useless and confusing.

few paragraphs a reader is no longer a novice. Masterman realized how clumsy multiple translates were and avoided them by several techniques[8].

Yet in spite of the objections made here, the translations by Richens and Booth were surprisingly readable, and Masterman's were even more so. With the constraints we propose, we do not expect readability to be so good for a person who is scanning a specific pidgin for the first time; but for reasons we shall expand on later, his performance should go on improving indefinitely with practice. Briefly the earlier pidgins lacked that fidelity to a natural source language which in our case would allow the reader to decode more with experience. Consequently he was unlikely to get much more information out of them than at his first encounter.

We give great weight to the view that natural languages are subtly and intricately structured in such a way as to afford the human intellect both easy communication and easy learning. But given a biologically inherited language capability, what is « easy » for that capability need not be what seems simple or regular to our rational thought processes. This orientation leads us to formulate aims and techniques that differ significantly from those of the earlier pidgin proponents. We feel that a specific pidgin translation must be constrained to match natural language structure closely until we receive experimental evidence that proposed deviations do not hinder communication or acquisition. Thus we resist the notion that simplifying the surface structure of a communication is necessarily an advantage; and we prefer to reproduce all the intricacies — even the « redundancies » — of the language of origin, although it may seem to English readers obviously expedient to clear away certain marks such as gender or case endings.

III

Like the earlier pidgin attempts, specific pidgin (henceforth just « Pidgin », with a capital P, for short) grew from the idea of constructing artificial languages which it would be realistic to consider translating into by computer, yet would be easy for English speakers to learn. Bearing in mind this idea of Pidgins as new but easily learned languages, we shall now explore some of their possible applications.

Pidgin translation might be employed in the following situations: when a researcher wants to read an article relating to the matter under investigation; when an expert wants to keep abreast of all the foreign literature in his field; and in the case of an agency interested in producing rapid translations at minimal cost.

Such agencies often call for preliminary rough translations which are then polished by the most experienced translators. Pidgin CT might profitably be used to produce the rough drafts. In such a system, we would expect an interesting development: the translator, having been given both the Pidgin for polishing and

8. She proposed the manufacture of « pidgin variables », words constructed out of English word-material to incorporate the semantic and syntactic properties of source words — usually grammatical function words — that were impossible to match otherwise. She also employed micro-glossaries (special dictionaries for technical jargons) and phrase-for-phrase translations. We reserve judgement on the usefulness of micro-glossaries, but commend her invention of pidgin variables. We also believe that translating phrase for phrase is sometimes justified.

the original text to refer back to, would find himself looking less and less to the original text until he completely ignored it and worked entirely from the Pidgin. The reason for this prediction is that Pidgin text must by definition preserve all the information contained in the original. The translator would be bound to learn more and more of the correspondences between Pidgin words and the vocabulary of the source language, a process likely to be subconscious in the main. If he were pressed for time — and what translator is not? — he would certainly begin to exploit this knowledge and not look so frequently at the original text. In the French to English example given earlier, when he came across the Pidgin word *tongue,* he would have a strong expectation that if he looked in the original text he would find *langue.* This informal learning would continue until the translator did not look at the source text at all. But if an experienced translator can eventually dispense with source texts, is it essential or efficient to insist that newcomers should be able to read text in the source language in the first place? Indeed if, as we shall argue, it is easier to learn Pidgin than an original language, then it ought to be cheaper and faster to train a translator to translate from Pidgin English into English (or from Pidgin French into French) than to do real translation. And it being cheaper and faster to train people to translate from Pidgin, one would expect there to be an easier supply of them [9].

At the introduction of any new technology, people can fear the worst (« being made redundant by the computer ») or they can profit from the innovation. There is, of course, no immediate danger of redundancy; Pidgin translation would be most profitably applied to translating languages for which there is the most serious dearth of translators. And even taking the long view, there is little cause for worry, because in the forseeable future Pidgin is unlikely to be of value for translation involving literary merit: this will continue to require a person with excellent control of the target language and an acute sensitivity to the nuances in language of origin. It is a serious question whether a mechanical device can ever attain this; and even if it were possible, nobody yet has the slightest idea of how to build such a machine. However, we see Pidgin CT as offering to remove much of the drudgery of hack translation — the translation of material that is of indifferent stylistic quality, but which is necessary for everyday use. If this proved to be so, Pidgin translation would free the translator to apply himself to more rewarding tasks.

We pointed out in Section I that any serious attempt to replace or aid human translators must begin with a realistic assessment of what aid translators need and of the needs of their readers; and that this essential first step has been overlooked by mechanical translation projects. As a contribution to this study, the concept of Pidgin offers a well-defined and reproducible standard level of translation — considering it for the moment to be translation, which it is after a fashion. It can be deviated from in measurable ways so as to determine the

9. One might ask: if it is so easy to learn to read Pidgin translations, why have them post-edited at all? Indeed we do believe that for specialist readers post-editing would be a waste of resources. Nonetheless, we accept that many documents, especially those for wide and general publication, have to be translated into normal language.

effects of the deviation, and thereby allow measurement of the cost-effectiveness relationships for the various mechanizable processes of translation.

In the first situation mentioned earlier, suppose a scientist has learned that a particular article in, say, a Chinese journal bears on the investigation he is undertaking. If he has recourse to the traditional translator he has little chance of satisfaction, for even with « common » languages like Russian or German he can expect a three month wait if not more (See, 1967), by which time he may either have duplicated successful research or wasted great effort in making mistakes that the report was intended to warn against. He is even more handicapped if an article is in a language for which there are few translators available, because not only does the delay increase but the quality usually decreases — to the point where the translation may be rendered worthless.

Our specialist has an alternative which is far better if he has the time, namely to learn the foreign language himself. Universities have long recognized this desideratum and required a reading knowledge of several languages, but if the language is unfamiliar, then the practising scientist will have to delay his researches by the length of tune it takes to learn to at least read it. This may be as short as a few days for mathematics in a language similar to one he already knows, since mathematics do not involve so much textual reading and have a large international notation and vocabulary; or it may be as long as several years for a language as different as Chinese is from English. Most scientists, for obvious reasons, do not take this alternative unless there is a great deal of other interesting literature in the same language. Moreover, there is a strong tendency today to write in one's national language, and this threatens to increase scientists' language learning loads. For instance, there are linguists of repute writing today in Italian, Danish, Dutch, Russian, Polish, Bulgarian, Chinese, Swedish, Spanish, Japanese, Hebrew, etc., all in addition to the traditional French, English and German. Learning half the languages on this list would be an admirable accomplishment: in practice, scientists need an alternative both to learning numerous languages and to reliance on human translators.

A more recent possibility, that of obtaining a mechanically translated (MT) version of the article, has yet to prove satisfactory, notwithstanding preliminary successes in the U.S. and Europe. Even with existing computers, MT is faster than manual translation and competitive in cost, but its products have been of value only for grasping the « gist » of an article [10]. It is true that with experience, a reader probably becomes familiar with the idiosyncracies of the translating machine and so his reading becomes easier and more accurate. He is inevitably

10. This assertion about the cheapness of MT is disputable. The strongest evidence for the opposition was gathered in the « ALPAC Report » (ALPAC, 1966, Appendix 9, « Cost estimates of various types of translation », p. 54ff). But besides opinions which others have expressed counter to the Report, we doubt its present worth because computer hard-ware developments are all the time tending to make MT cheaper irrespective of its programs and linguistics. For example, one of the major expenses until now has been preparing the texts for input into the machine: usually it had to be punched on cards. Already optical scanners are coming onto the market which enable the computer to read directly from a typed page. The latest computers calculate about four times faster than the best five years ago.

limited, however, by the lack of whatever information was lost in the translation process. This is not to deny that MT is being improved by research. But the improvement is slow and requires very sophisticated efforts from linguists, from computer scientists, and eventually from researchers into artificial intelligence and information retrieval.

Pidgin offers an alternative which is more accessible than either of the foregoing. It would cost less than MT because it would use less computer time. This is because the algorithms needed are comparatively simple: it only needs a dictionary look-up and perhaps a rough parsing to reduce the homographs, whereas effective MT must incorporate more complex transformational grammar for deeper semantic analysis of the source language, and also a synthesizer of the target language. No less important is the economy that Pidgin offers in the human resources needed to write the grammars and program the systems. The programmers for their part can already provide all that is needed for Pidgin CT. Pidgin removes the major obstacle to effective CT, namely that if a computer is to simulate conventional human translation, the machine has to *understand* the original text.

In comparing Pidgin output with that of MT, we concede that it is less perspicuous to the beginner. We would therefore propose that he be given a few hints, as we have done for the Arabic example to follow. But after he has been helped to read and understand the first paragraphs, he ought to be able to go on by himself and get the gist of the article with a little effort. Thus, subject to some extra effort at the very outset, Pidgin translation should be at least as readable as MT output.

To learn a Pidgin is clearly simpler and faster than learning the original language, if only because the task of learning the thousands of content words in a foreign vocabulary is mostly removed by drawing a Pidgin vocabulary from the reader's own language. So he already knows the meanings of most of the Pidgin words — not their exact semantic ranges but approximations. This follows from the constraints which we have stipulated. This much and no more already allows him to read roughly. What he must then learn by way of vocabulary is how people who speak the other language put these already familiar lexical signs to different use; it is a process similar to learning how British and Americans use English words differently, and it requires almost no formal training. As he finds words in contrast or in positions of synonymy in the Pidgin, he hones his perception of their meanings, just as he does while reading his native language (indeed nobody knows all the meaning of most words even in his own language). Finally, after reading a number of articles he can expect to read the Pidgin with an ease approaching a native's competence in reading the source language. To sum up with some insistence: benefit would continually accrue to the Pidgin user from the fact that a Pidgin is a language in its own right with a borrowed but natural syntax and lexical structure to retain a maximum of information.

In the second situation which we envisioned, that of an expert desiring to keep abreast of the literature in his field, Pidgin translation offers more advantages than in the case of an *ad hoc* researcher wanting to read an occasional article;

for once the Pidgin has been assimilated by reading several articles written in it, the user could read other Pidgin articles with little more effort than is needed to read articles in his own language.

IV

We have spoken of « translating into a Pidgin » and « learning a Pidgin ». By way of illustration, we shall first imagine a Pidgin English which is derived from Chinese. Perhaps we ought to remind readers that with few exceptions a « word » is represented in Chinese writing by a single character instead of by a combination of letters from a small alphabet as in English. This means that in Chinese there are approximately as many characters (so-called ideograms) as there are words, which is to say that there are many thousands of them. To proceed, we number all the characters in the Chinese dictionary, never assigning the same number to more than one character. For a subset of the characters, the most frequently used ones, this has actually been done in the code used for sending telegrams in Chinese. Now consider a Chinese text in which every character has been replaced by its number. Anybody can « read » the coded text, since the symbols of the telegraphic code are the familiar international (« arabic ») numerals. So it would be easier for an English speaker to learn to recognize Chinese telegraphic code than to learn the Chinese characters, because not only are the symbols already familiar to him, but so too is the way of concatenating them into numbers of any length. However, it still takes a Chinese telegraphist to *understand* a text in this code, because only a telegraphist knows what Chinese character each number stands for, and only a Chinese speaker knows the meaning of the characters. For our English speaker the task of learning the *meaning* of each number would still be formidable, especially as numbers are harder than words for most people to remember. With numbers replacing Chinese characters, the burden of learning would be lightened but not enough. So suppose that instead of assigning a unique *number* to each Chinese character, we represent it by a unique *word* drawn from the vocabulary of English. Furthermore let the word be so chosen that it has roughly the same meaning in English as the character does in Chinese. Thereby the English reader will be relieved of much of the learning necessary to understand Chinese words transcribed in this code. The result is Chinese Pidgin consistent with our definition of a Pidgin. Notice that a text in this Pidgin has not been « translated » in the sense that the meaning of its sentences has been re-expressed in English. All that has been done is to *encode* the Chinese text with special characters, and it could be done by any Chinese telegraphist merely by changing his usual code book for our Pidgin dictionary. So « mechanical » is this encoding process that it could be done by existing computer programs.

The grammar of the Chinese encoded in the Pidgin just described would still be Chinese. If it were some other language than Chinese that had been coded in the same way, then the resulting Pidgin would have the grammar of that other source language. There are therefore potentially many varieties of Pidgin English, and each one will be different both from English and from other Pidgins[11].

11. Conversely, if one Pidgin were really Italian encoded with English words, and another were Spanish encoded in like manner, the Pidgins from these two Romance languages

Because Chinese Pidgin is really only Chinese with English-looking code words substituted for the Chinese characters, it is correct to speak of learning it as a language, of translating into it, and of translating from it into English. And because we have chosen our code symbols from the English lexicon, a person who can read English needs the minimum of learning effort to read this Pidgin; which is tantamount to saying that he needs the minimum of effort in order to read Chinese, notwithstanding the difficulties he will still have with grammar and usage. There is perhaps another advantage in Pidgin, one which would be less immediate but which is fundamental to good translation. We are of the opinion that good translation (or good understanding of a text in another language) cannot be bad without fluency in *speaking* the language of origin. Because human linguistic ability is based on speech, even when someone is only reading in a language he necessarily pronounces it — perhaps silently, perhaps only in his imagination (Shillan, 1968). A person learning only to read a second language is thus forced to pronounce it in some fashion, and of course he pronounces it according to a language he already knows or an imaginative modification thereof. After some reading, he is likely to establish pronunciation habits internally which are difficult to correct and may later hinder him from speaking the language well. On the other hand, if someone first learnt to read another language via its Pidgin he could go on to learn the original language for fluency in speech without this hindrance, since the Pidgin words that he read would be words from his own language which he might pronounce according to his native phonology without this eventually « interfering » with his pronunciation of the source language [12].

---

would be noticeably similar in their grammars and perhaps to some extent in their word choices. Therefore a tyro might well mistake one for the other. Striking similarities would arise between Pidgins from every group of closely related languages (e.g. Dutch, German and English) because such languages tend to differ most in pronunciation and their pronunciation differences would be totally hidden by encoding them in Pidgin English. Indeed « Chinese » is often said to be really a whole family of related languages, the differences between them being bridged over by the ideograms. Even within English most dialect differences are obliterated in the written language by its none too phonetic orthography.

In the case of Chinese characters, their interlingual usefulness extends beyond China. Due to China's long preeminence in the Far East, her writing system was adopted by the Japanese, Korean and Vietnamese cultures. Both Japan and Korea devised phonetic writing systems, but these were not used much except to supplement the characters for case endings, or for words which were not in Chinese. For a thousand years or so, material published in any of these places could be read in any other of them in spite of the fact that Japanese, for instance, is widely dissimilar from Chinese. At present this koinized writing system is on the decline, Vietnam having almost completely abandoned Chinese characters, North Korea trying to, and Japan having gone a long way in that direction. But even today an educated Japanese can take a Chinese book and get the essential meaning out of it just as a Chinese can with Japanese book. Though chauvinism is ousting this natural pasigraphy, it is still remarkable how the use of Chinese characters has facilitated written intercommunication between speakers of at least four widely different languages without their needing to speak Chinese. A member of this « writing community » composes a text in his own language but employs Chinese characters

(Kanji) to write it down; another member in some other country and speaking a different language reads the text in what for him is effectively a Pidgin, pronouncing it in his own tongue but inevitably following the word order and registering the lexical choices of the writer's language. This historical note is our final example in support of our contention that Pidgin translation is not only theoretically conceivable, but that it has already been practised on a large scale for centuries.

12. The pedagogical use of Pidgin translation is another avenue worth exploring. It is well known that learning a complex behavior pattern, such as a language, is facilitated by splitting the desired behaviour into simpler components and learning them in separate

The following example is of an English Pidgin from Arabic. We have chosen Arabic because it is quite a difficult language for English or French speakers to learn. It will be made abundantly clear that Pidginized Arabic is much easier for an English speaker to acquire than normal Arabic.

The Pidgin is interlined between the Arabic so as to bring out the word-for-word correspondence between the two. We have copied the printed Arabic text meticulously in matters of punctuation and vowelling; only the long vowels are indicated in most Arabic texts. Our aim is to give readers an idea of how much they would have to learn if they wanted to read the original, or at least as accurate an idea as is possible from a romanized text. In reality romanization is hardly ever used in the Arabic-speaking world, and consequently learning the Arabic alphabet remains a considerable initial obstacle to foreigners [13].

Content words of native origin in Arabic are derived from a 3-consonant (occasionally 4-consonant) radical. The simplest form is traditionally the third person masculine singular of the verb in the aorist tense. This is the form which is printed as rubric in Arabic dictionaries. The persistence of the radical consonants through derived forms and meanings is well illustrated by the following example from Yushmanov, 1961:

> The root QTL which per se cannot be pronounced will denote everything concerning « killing » and appears in the words *qatl* « murder », *qatil* « killer », *qatil* « killed one », *qital* « battle », etc. [14].

Of course there is no hope of making Pidgin represent this morphological structure of Arabic words. We have tried, however, to reproduce the lexical connections that the consonantal triplets symbolize. One example should suffice. In the text the word *nhãr* occurs, and the corresponding word in an English translation would be *day.* But *nhãr* is not the common Arabic word for « day » : that would be *ywm.* So we are obliged to look for another translate lest we lose the lexical contrast between *nhãr* and *ywm.* We take into account the derivation of *nhãr* from the radical NHR, which expresses the idea of « flowing like water ». Hence we have proceeded to create a new word in Pidgin, *dayflow,* which conserves both the meaning « day » and the connection with « flowing » without asking too much of the reader.

Many Arabic affixes and enclitics would be separate words in English, e.g. *'l'wl* « the first ». These agglutinates are hyphenated in the Pidgin. Furthermore we have reduced to subscripts some grammatical affixes, e.g. case endings, which

---

stages. The usefulness of Pidgin would lie in teaching the syntax and the lexical structure separately from the phonology and forms of the words. The Arabic example that follows shows some of the insights that Pidgin can provide into the two former aspects. It also demonstrates, incidentally, the importance of pronunciation in learning : it is all the more difficult to remember the correspondences between the Pidgin and the Arabic words because written Arabic is shorn of the vowels which an English speaker needs in order to give it some sort of pronunciation. One remedy is to say a shwa after every unvowelled consonant.

13. We have used the standard transliteration, except that we have distinguished by parentheses the special *t* grapheme used for the feminine noun suffix, e.g. *zlm(t).* Even native speakers have trouble with written Arabic because of the paucity of vowels (Cowan, 1968).

14. French readers interested in the structure of Arabic should refer to Blachère, 1952, especially p. 13 ff.

we preserve on the principle that no grammatical information should be abandoned, but which might hinder the eye of the unaccustomed reader.

The passage chosen is a Biblical one (Bible, 1912) so familiar that we have not bothered to give an English version. In comparing this Pidgin rendition with English translations, one should bear in mind that we are faithful to the *Arabic*, which in turn was translated from a Greek version and not from English. The reader's previous acquaintance with this passage should nevertheless enable him to understand most, if not all, of the Pidgin text[18]. In any case, we have placed after the text a few « hints » of the kind we think the student of a Pidgin ought to be given. Once again we would like to emphasize that Pidgin is not misbegotten attempt to translate into English but is intended to be read as a language in its own right[16].

Arabic and Pidgin text

altkwyn
the-do-be-ness

'l'ṣḥãhḥ 'l'wl

the-make-correct-ness the-first

1 fy lbd'i khlq 'lllãh 'l'smwãt w'l'rḍ.

1 in the-begin-ness create the-God the-heaven-s and-the-earth

ind she she

2 wkãnt 'l'rḍ khrb(t)an wkhãly(t)an w'lã wjh

2 and-be the-earth debris and-vacant and-on face

she she she-acc she-acc

'lghmr ẓlm(t)un wrwh 'lllãh yrffu

the-submerge-ness dark-ness and-inspire-ness the-God he-be-twinkling

she-nom

'alã wjh 'lmyah. 3 wqãl 'lllãh lykn nwrun

on face the-waters. 3 and-said the-God lo!- let-be light-ness

he nom

fkãn nwrun. 4 wr'ã 'lllãh 'l'nwr 'nh

then-be light-ness . 4 and-see the-God the-light-ness verily-he

nom nom

ḥsnun. wfṣl 'lllãh byn 'lnwr w'lẓlm(t).

fine . and-divide the-God between the-light-ness and-the-dark-ness

nom she

15.It is not our purpose here to provide a fair test of one's ability to read the Pidgin and understand it. That would require a far longer text. We are only attempting to illustrate what a Pidgin is like, and to demonstrate how it may be at least much easier to learn than its source.

16.One person on whom we tried out the following example objected at first that this Pidgin did not make any sense, that it was « translated from some primitive language ». When he was informed that the source language was Arabic he then found our text to be quite readable because, so he said, it was like Hebrew, a language he knew and presumably respected.

5 wd'ã 'lllãh 'lnwr nhãran w'lẓlm(t)

5 and-call the-God the-light-ness dayflow and-the-dark-ness
acc she
d'ãhã lylan . wkãn msã'un wkan ṣbãḥun ywman wãhdan .

call-her night . and-be evening and-be morning day one
acc nom nom acc acc
6 wqãl 'lllãh lykn jldun fy wṣt 'lmyãh .

6 and-say the-God lo!- let-be firm-ness in middle-ness the-waters .
he nom
wlykn fãṣlan byn myãh wmyãh . 7

and-lo!- let-be divid-er between waters and-waters . 7
he acc ind ind
f'ml 'lllãh 'ljalad wfṣl byn 'lmyãh 'lty

then-produce the God the-firm-ness and-divide between the-waters which
she
tḥt 'ljalad walmyãh 'lty fwq 'ljld . wkãn

under the-firm-ness and-the-waters which above the-firm-ness . and-be
she
kdhlk . 8 wd'ã 'lllãh 'ljalad smã'an . wkãn
like-that . 8 and-call the-God the-firmness skyhigh-ness . and-be
msã'un wkãn ṣbãḥun ywman thãnyan .

evening and-be morning day 2nd
nom nom acc acc

Hints to the reader of Arabic-based Pidgin

1) The usual phrase order in Arabic is *verb-subject-object.*

2) The copula is optional in Arabic. Thus *verily-he fine* means « verily, he was (or is) fine », and so on.

3) The « base form » of the Arabic verb is often called the *aorist,* because it is like the Greek aorist in that it does not connote any specific « time of action » a priori (cf. the English infinitive), though it is often used in context for past or terminated actions. We have translate it by the « base form » of the English verb, i.e. the infinitive without *to.*

4) *The-do-be-ness* means « the creation ». The auxiliary *do* is used to represent the emphatic form of the Arabic verb (traditionally called the « second form »), which is often, but not always, used as a causative. The causative of « being », the « bringing into being », is of course « creation ».

5) The case endings are here represented by the subscripts *nom* for « nominative », *acc* for « accusative » and *ind* for « indirect ». The last is a case that combines the functions performed in some other languages by genitive and dative. Most of the case endings are dropped both in spoken and written Arabic.

6) The genitive relationship is usually indicated by the sequence *indefinite-noun* + *definite-noun.* Thus *face the-waters* in Pidgin means in English « the face *of* the waters ».

7) Arabic adjectives follow the noun modified. Thus *the-make-correctness the-first* translates into « the first chapter[17] ».

8) The only native Arabic punctuation is the period. On the other hand, *w* (and) is used very freely and performs the function of English commas, semi-colons, etc. The Arabic alphabet has no capitals.

V

We do not claim that Pidgin as we have defined it is the best form of pidgin translation possible. We have already remarked that technically speaking it is not « translation » at all: it is merely encoding or — to coin a word — « transcoding ». We drew up its specifications to show the potential of a type of *computerized* translation and to point the way for CT experimentation which might prove more fruitful than what has been tried hitherto. In practice it might become expedient to deviate for our preconceptions considerably in whatever way aids the reader.

The claims we make for Pidgin, subject obviously to far more in the way of experimental confirmation, are that: *a*) It is feasible to transcode into it by computer without loss of meaning; *b)* Consequently Pidgin transcoding is an adequate substitute for « translation » in many circumstances, and one which can be turned out very much more rapidly than conventional manual translation; *c*) Pidgin transcoding is *prima facie* cheaper than other kinds of CT; *d)* Pidgin uses the grammar and vocabulary of natural languages, hence it possesses the advantage for human beings that it can be *learned* like natural languages; *e*) It is very much easier and quicker for adult foreigners to learn than the source languages.

The crucial component of a Pidgin transcoding is the Pidgin dictionary, which must be governed by the constraints laid down in Section II. The authors' experience with Arabic and Chinese, besides less « exotic » languages, leads them to believe that a full Pidgin dictionary can be compiled [18].

TH. R. HOFMANN and BRIAN HARRIS

17. By rendering *'l'shãh* as « the-make-correct-ness », we have drawn a strained connection between *'shãh &* the root SHH « soundness, health ». It is true that *'shãh* is formed on the pattern of the verbal noun of the causative (4th) form of the verb, yet it is a special Christian-Arab term used only to head chapters of the Bible (cf. English *capitulum).* Are Arabic readers aware of the connection with the radical in the case of words that contain its consonants but whose meaning is far removed from it? This is a matter for psycholinguists which has yet to be researched. Indeed the connection may be quite fortuitous in the case of loanwords from other languages. The Arabic tradition of listing a form under an apparent radical may sometimes be only a lexicographic convenience (Wehr, 1966, p. x and xii ff). Much the same question can be asked about some English words that we connected by morphology: whether, for instance, English speakers are aware of the link between *grave* « serious » and *gravity* « terrestrial gravitation ».

18. At the time of concluding this article (February, 1970) work on transcoding into three Pidgins is being done at the Université de Montreal, viz. French Pidgins from English and from German, and English Pidgin from Arabic. Some computer output has been produced. This research is being generously supported by the National Research Council of Canada as a marginal activity of the Projet de traduction automatique. Some work is also being done at the Université d'Ottwa on English Pidgin from Chinese.

We would like to take this opportunity to thank Professor A. Querido for his encouragement; and the editorial board of META for their criticisms, many of which have been heeded in our final draft.

REFERENCES

ALPAC (1966): *Language and Machines: Computers in Translation and Linguistics,* Washington (D.C.), National Academy of Sciences/National Research Council, Publication 1416.

BAR-HILLEL, Y. (1964): « Some Linguistic Obstacles to Machine Translation », in *Language and Information,* 1964 (q.v.), p. 75-86.

BB&N (1966): *Report p66-MM-20,* Cambridge (Mass.), Bolt, Beranek & Newman Inc.

BIBLE (1912): *Bible (Ref.), 2nd ft., 281,* translated from the Greek, Beirut, American Press.

BLACHÈRE, R. & M. GAUDEFROY-DEMOMBYNES (1952): *Grammaire de l'arabe classique,* 3e éd., Paris, Maisonneuve & Larose.

BOOTH, A.D., ed. (1967a): *Machine Translation,* Amsterdam, North-Holland Publishing Co.

BOOTH, K.H.V. (1967b): « Machine-Aided Translation with a Post-Editor », in Booth, 1967a (q.v.), p. 51-76.

BORKO, H., ed. (1967) : *Automated Language Processing,* New York, John Wiley.

CAB: *Current Affairs Bulletin,* Sydney, NSW, Tutorial Classes Dept., Univ. of Sydney.

CARROLL, J.B. (1966): « An Experiment in Evaluating the Quality of Translation », in *MT* (q.v.), Vol. 9, Nos. 3-4, p. 55-66.

COWAN, W. (1968): « Notes toward a definition of Modern Standard Arabic », in *Language Learning* (q.v.), Vol. 18, Nos. 1 & 2, p. 45-60.

DUGAS, A., M. GOPNK, B. HARRIS & J.P. PAILLET (1969): *le Projet de traduction automatique a l'Université de Montréal,* Preprint No. 55, International Conference on Computational Linguistics, Stockholm, KVAL, 1969.

Endeavour (édition française), London, Imperial Chemical Industries.

HALL, J.R. Jr. (1954): « Pidgin English », in CAB (q.v.), Vol. 14, No. 12, p. 179-192.

HOFMANN, T.R. (1968a): « Problems and Possibilities in Computational Linguistics », in Rondeau, 1968 (q.v.).

HOFMANN, T.R. (1968b): « Specific Pidgin Translation », in *Rapport trimestriel* (q.v.), No. 10, p. 62-74.

IBM (1959): *Final Report on Computer Set AN/GSQ-16 (XW-1),* Yorktown Heights (N.Y.), IBM.

Language and Information (1964): Reading (Mass.), Addison-Wesley, and Jerusalem, Jerusalem Academic Press.

Language Learning, Ann Arbor (Mich.), University of Michigan, 1951-

LOCKE, W.N. & A.D. BOOTH, eds. (1955): *Machine Translation of Languages,* Cambridge (Mass.), M.I.T. Press, and New York, John Wiley.

MASTERMAN, M. (1967): « Mechanical Pidgin Translation », in Booth, 1967a (q.v.), p. 195-227.

Meta, Journal des traducteurs/Translators' Journal, Montréal, Les Presses de 1'Université de Montréal, 1956-

MT: *Mechanical Translation and Computational Linguistics — An International Journal,* Chicago, University of Chicago Press, for the Association for Machine Translation and Computational Linguistics, 1954-

ORR, D.B. & V.H. SMALL (1967): « Comprehensibility of Machine-Aided Translations of Russian Scientific Documents », in MT (q.v.), Vol. 10, Nos. 1-2, p. 1-10.

PENDERGRAFT, E.D. (1967): « Translating languages », in Borko, 1967 (q.v.).

PERSCHKE, S. (1968): « Traduction à la machine — la seconde phase du développement », in *Endeavour* (q.v.), No. 101, p. 97-101.

Rapport trimestriel: Recherche sur la traduction automatique, rapport trimestriel (devenu semestriel)/Mechanical Translation Project Quarterly (later Semi-Annual) Report, Montréal, Université de Montréal et Conseil national de recherches du Canada, 1966-

RICHENS, R.H. & A.D. BOOTH (1955): « Some Methods of Mechanized Translation », in Locke, 1955 (q.v.).

RONDEAU, G., ed. (1968): *Linguistique et mathématique,* Montréal, Les Presses de 1'Université de Montréal.

SEDELOW, S.Y. & W.A. Jr. (1967): « Stylistic Analysis », in Borko, 1967 (q.v.), p. 181-213.

SEE, R. (1967): « Machine-Aided Translation and Information Retrieval ». To be published in *Proceedings of the 2nd Conference on Electronic Information and Handling, Testing and Evaluation,* Pittsburgh (Pa.), April 12-14. Pre-publication copy issued by OSIS, May 8.

SHILLAN, D. (1968): « Phrasing and Meaning », in *Meta* (q.v.), Vol. 13, No 2, p. 47-51.

WEHR, H. (1966): *A Dictionary of Modern Written Arabic,* edited by J. Milton Cowan, 2nd printing, Ithaca (N.Y.), Cornell University Press.

YUSHMANOV, N.V. (1961): *The Structure of the Arabic Language,* translated from the Russian by Moshe Perlmann, Washington (D.C.), Center for Applied Linguistics.