[From: *Proceedings of the International Congress on Technology and Blindness*, vol.I (William Byrne Press, 1963]

AUTOMATIC MACHINE TRANSLATION: POTENTIALITIES FOR BRAILLE ENCODING

VICTOR H. YNGVE
Massachusetts Institute of Technology, Cambridge, Massachusetts

The automatic transcription of contracted braille from uncontracted material shares many of the problems of the translation by machine of such languages as German and Russian. To the extent that the braille spelling rules refer to the conventional spelling of the original the problems are minor. But to the extent that the braille spelling rules refer to pronunciation, grammatical function, or meaning the problems are severe and can be attacked only with very sophisticated methods. In other words, going from a code (ordinary inkprint spelling) into braille is really a problem in translation.

There are available source documents in the form of punched paper tape that are by-products of the printing industry. It has occurred to many of us that these could perhaps be used to produce braille copies automatically. There is also the possibility of providing, from a typewriter keyboard, pulses corresponding to ordinary spelling, and then translating these into the correct contractions in braille. There are various other possible ways of tying the two systems of representation together.

The problem is made difficult by the nature of braille. Braille is in a sense based on spelling, but many of its rules refer not to the spelled form but to the underlying language. In other words, in a very real sense braille is a direct representation of the spoken language rather than a direct representation of the spelled form that we find in books.

In order to make the rules of braille easier many of the rules of ordinary spelling have been adopted, so that many words in braille are spelled exactly the same way as they are in a book. Since the purpose of braille is still to transmit information, however, I think it proper that the rules have been stated in terms of the underlying language. This is true as long as a human being transcribes the braille, because he can understand the language that is being encoded into braille and can very easily use rules couched in terms of the underlying language. However, given the problem of translating into braille from a representation equivalent to inkprint, or from the output of a typewriter keyboard, one faces a different problem. First of all, the spelling system of English is notoriously poor.

I think I can illustrate the sort of problems one faces by stating a couple of rules from the braille standard of some years ago. Rule 34 for Grade 2 braille has to do with contractions; it says, "Contractions forming parts of words should not be used when they are likely to lead to obscurity in recognition or pronunciation, and therefore they should not overlap well-defined syllable divisions." This rule is stated in terms of syllable divisions, something that is not explicitly represented in inkprint. "Word signs should be used sparingly in the middle of words unless they form distinct syllables. . . . Special care should be taken to avoid undue contractions of words of relatively infrequent occurrence." It goes on, ". . . when words occur at the end of a line, they must be at the end of a syllable." Here we have a rule for contracted braille stated not in terms of the inkprint spelling, but in terms of syllables, which are a feature of the underlying language. The question arises: Can one syllabify a word automatically? I think most of you know that this is very difficult. I have to look words up in the dictionary in order to separate them correctly at the end of a line. The difficulty is partly due to the traditional spelling of English, a heritage from past eras, and does not in many cases conform exactly to the pronunciation.

The next example is contained in Rule 23. According to this rule the contractions "to," "into," and "by" are always to be written close up to the word or that word which follows. It goes on, ". . . in such phrases as 'it was referred to yesterday,' and 'he was passed by when others were noticed,' the 'to' and the 'by' should be written in full and not contracted, as they refer to the preceding verb and not to the word that follows them."

In other words, if “to” or “by” are prepositions, as in “to the house,” or “by the table,” then one would contract the “to” or the “by” according to this rule, and write the contracted form without a space immediately preceding the next word. However, if “to” and “by” are adverbs, as in the case “. . . it was referred to yesterday . . .” and “. . . he was passed by when others were noticed,” they are not contracted. There is no clue in the inkprint that these words are in one case prepositions and in the other case adverbs; they are not marked explicitly. One needs to have an understanding of the sentence in order to make that distinction, or else one has to have a method of grammatically parsing the sentence so that he can determine whether these words are prepositions or adverbs. This is a problem that has been faced in the mechanical translation of languages, and I shall say a little bit about it below.

The following sentence is actually syntactically ambiguous: “It was referred to the other day” (or, “It was referred, to the other day”). The first inkprint makes no distinction; one can *say* it either way, however, using a different tone of voice. The “to” in this sentence (as I read the rules of braille) would in one case be contracted and written next to the word; in the other case it would not be contracted. The resolution of such ambiguities is relatively easy for the person who reads the material, if he understands it. The resolution of such ambiguities would be very difficult, however, for a machine. It is also the sort of ambiguity resolution that the people working in mechanical translation of languages have been facing.

I shall give a brief summary of this work. The first hope was that one could put a dictionary into a computer. The computer would simply look up the words one at a time in the dictionary, finding equivalent words in the other language, and print them out. Such a dictionary would be easy to mechanize; the problems were involved primarily in the large size of the dictionary as compared with the relatively small size of memories. The most promising method of implementing such a thing would be to put the dictionary on magnetic tape available to the computer and arrange to look up the words in batches. I say this was a hope; there were many problems with it. First of all, especially in languages such as Russian which was given a lot of attention by people working in mechanical translation, it was realized very quickly that the size of the dictionary could be greatly reduced by not storing a whole word complete with its ending, but storing instead its stem separate from the ending. Then a program in the machine would take each Russian word, examine it letter by letter, split off any inflectional endings there might have been (e.g., case endings, verb endings), and look up the remainder in the dictionary. Then, having found the stem of the word in the dictionary, the machine could go ahead and interpret the remainder of the word as an inflectional ending and give it its appropriate meaning. Programs of this type have been written at a number of universities and at a number of industrial firms that have been working on this type of translation. I can report that the problem is effectively solved.

However, our hopes were really too high. The result of writing out just the words from such a dictionary look-up process was completely inadequate as a translation (and I mean *completely* inadequate). There are two main reasons for this. One was that if one looks up almost any word in a dictionary, one finds that it has several renditions in the other language. The other reason was that even if one could select the correct meaning of each of the input words and string these meanings together the word order would be wrong and in general a grammatically correct sentence is not obtained. In some cases the “translation” is so badly garbled that one cannot make any sense out of it even if the correct word is there. The problem was, what next to do?

The next step was to look at the output and see whether something more could be done. Certain rules were set up, ad hoc rules, which worked perhaps 80 percent of the time. Let me illustrate such a rule. The letter sequence d-e-r in German can be an article in front of a noun; it can be a relative pronoun; if it is an article it can be nominative, genitive, or dative. The translation of this three letter word would depend on its grammatical function. Thus, a very simple rule of thumb is: “If d-e-r follows a noun without a comma translate it ‘of the.’” This rule will give the correct answer about 90 percent of the time; perhaps even 95 percent of the time. It is wrong when “der” is dative (and it could very easily be dative), but it is dative perhaps only 5 percent of the time. Thus the translation is wrong about 5 percent of the time. It sounds impressive, however, to have a rule that works 95 percent of the time. This is the type of rule that I call an “ad hoc rule.” The rule is not really based on the structure of the language. In other words the case is not determined and the part of speech is not determined.

Many of the mechanical translation groups looked for and discovered a large number of such rules of thumb; they were able to make a fairly reasonable improvement in readability. Another example they found was this: “If there are three meanings for a word, and one is very frequent while the other two are not as frequent, then print the frequent meaning and forget about the others." Again the quality is improved because it is very difficult for the reader to be faced with three alternatives; he can read much more easily if he has only single words to consider. Choosing the most frequent meaning is more often right than not. This is also the kind of rule that is not really a "correct rule." On the average, however, it will work.

Mechanical translation people were quite optimistic about this procedure; they thought, ". . . it's just a matter of finding more and more of these rules; fixing up the order; eliminating more and more of the problems." Unfortunately, one can't go all the way with this approach. It becomes much too complicated, rules conflict with rules, and one never really knows what one has when it is done. I want to emphasize, however, that such rules will take care of perhaps 80 percent of the problems involved. This first 80 percent of the problem is easy to solve. It is the remaining 20 percent that is extremely difficult and that cannot be solved by such rules of thumb.

The next approach was to try and do it right: to find out what are the actual parts of speech of each word in the sentence. Let us try to find out whether it *is* a preposition or an adverb. Let us find out what *is* the subject of the sentence; what *is* the verb; what *is* the object; and so on. In other words, do a complete parsing of the sentence. Programs of this sort have been written and they are fairly successful, but a new batch of problems has shown up.

In general it is not possible to parse a sentence without knowing its meaning. This we found through experience. I suppose if we had thought about it we would have known, but we hoped that a simple parsing of the sentence would give us enough improvement in the output of the translating program to be useful. Take the sentence we used above: "it was referred to the other day," or, as it may be read, "it was referred, to the other day." It would appear that parsing would help. However, this sentence is ambiguous; it has two different parsings. In any given text this sentence would be unambiguous because of its context, because the person who reads it would understand what was meant very readily, and it would never enter his mind that the sentence was ambiguous. Unless he can understand the text too, he cannot do this. So the limitations on automatic parsing of sentences is just at that point where we need to understand the meaning of the sentence in order to resolve ambiguities. I can report to you that such ambiguities are a very frequent occurrence. A very large number of sentences are really ambiguous from this grammatical point of view. We are not bothered by such ambiguity when we read because we understand the meaning and it is this understanding of the meaning of the sentence that carries us through the ambiguities.

Our hopes have been dashed again. The essential limit of a program for parsing a sentence is just in this area which I like to call “semantics.” A number of the groups working on mechanical translation are now facing up to the problem of semantics. This problem appears to be orders of magnitude more difficult than the syntactic problem. We have a few hunches, but I don’t think we have the foggiest idea, really, of how to solve this problem. Nevertheless, most of the groups are working at it. They are trying ad hoc rules, and they are trying various other schemes. They have also tried schemes such as the following.

You all know that you can *row* a boat. Now, it turns out that there aren’t very many other things that you *row*, other than boats. The word r-o-w is ambiguous: it could be a row (a brawl) or a row (of objects). In other words the meaning of this word or the solution of this ambiguity can be found partially but not completely. In the general case one must also take care of the meaning of the sentence. One way of doing this is to list in the dictionary that it is *boats* that you row and not other things. Much information of this kind in the dictionary might be quite useful in resolving ambiguities. There are other methods that have been proposed. One is to order the words in the dictionary in much the same way as they are in a thesaurus, by meaning categories with indexes and connections between words, and putting them into fields of knowledge and fields of interest much the same way Roget did in his Thesaurus. There are several other such schemes. In other words, we are taking the first faltering steps into the area of semantics.

Now, as to braille, I think that the complete and correct transcription of contracted braille, according to the currently accepted official rules of standard English braille, is not currently feasible. I want to be very clear about this; it is exactly what I mean. I say, “It’s not feasible,” but on the other hand it is. Attend very carefully to the qualification: it is more than “not feasible” vs. “feasible.”

Automatic transcription is feasible if certain of the rules are compromised. The real question is, what is the degree of compromise that is necessary? I suggest that we work out the best compromises and standardize them into a new type of braille specifically for machine transcription. All the rules should be phrased in terms of the conventional spelling of the original text with no reference to pronunciation, grammatical function, or meaning. This “machine transcription braille” should conform *as closely as possible to the current practice* so that it could be read interchangeably with hand transcribed braille. Now this is precisely what is being done, except that we have not standardized our usage. The braille programs that we have now *do* operate with rules stated in terms of the traditional spelling. In other words, a pronunciation rule would be restated: “you will do such-and-such,” instead of saying, “you do such-and-such except when you would pronounce it some other way” (you do such-and-such and list the exceptions). This is tantamount to restating the rule in terms of the inkprint spelling.

If we devise machine programs that actually are used for transcribing braille, a little thought should be given to stating these rules the way we really want to use them, while realizing that machine programs can be very easily changed to conform with any set of braille rules one might wish to use. I think it would behoove the people who are interested in what the machine produced braille is going to look like to look at the rules *as they are stated now,* and to the problem of restating these rules in some way so that machine programs can be written that will give the kind of braille they want. They must realize that it is impossible to program a machine to transcribe braille according to the rules as they now stand, because the criteria now put down have to do with the pronunciation, with the grammatical structure, or with the meaning of a sentence. These are problems that have not been solved even in the mechanical translation of languages. They are in fact extremely difficult problems.

I have one other comment. It would be a good idea to capitalize upon the rather wide availability of punched tape from the printing industry. I imagine that this has been suggested before. I feel that this material should be placed in a central repository so that people who want to make braille editions would have it available. There are other groups that are also interested in a centralized repository for this material. I would think it would be very wise to contact these groups and work with them. The other groups are primarily concerned with mechanical translation (who would like to have the material for translation) and the groups associated with information retrieval (or the automatic library). I don’t know where the best place would be for such a center; possibly the Library of Congress. Publishers send copies there anyway for copyright purposes. Perhaps they wouldn’t mind sending their punched paper tapes to the Library of Congress. I don’t know; I presume that the Library of Congress is not set up for this kind of thing, that there would have to be something added. Perhaps it is unsatisfactory as a repository for other reasons; but I certainly feel that this should be explored, and it should be explored concurrently with other groups that are also interested, particularly the information retrieval people.

In concluding, I should like to consider a number of specific questions having to do with problems in applying the caveats I have discussed.* Among these I would include those dealing with (1) paper tapes, (2) contractions vs. syllable boundaries, (3) the anticipated or possible contraction in a revised and “computer-oriented” braille, and (4) the argument for complex translation programs versus the generation of a modified braille.

THE PROBLEM OF OBTAINING PUNCHED PAPER TAPES

This is an extremely difficult problem. I personally feel that the best solution, which perhaps is not feasible, would be to update the procedures in the printing industry. After all, the printing industry was mechanized about 50 years ago, when Monotype was really the last word in automation. It uses a player piano-like roll which is not quite as wide as that for the player piano. It is read the same way, with compressed air; it huffs and puffs and chugs along, and is not really in line with modern automation techniques. Monotype has served the printing industry admirably; I could imagine they would be loathe to change unless a very real advancement were achieved. There are people who are thinking in terms of computer programs to help correct errors. In fact there are some at MIT who are doing this sort of thing. If one can get the material that is to be published on tape and into a computer, then it is possible to write a program that will correct this tape to order. This has a very great advantage, namely that one does not have to proofread the material carefully once more. Once it is set, once it has been proofread, once it is correct—it is there, it is done. I feel that there is a great deal of room for advancement in this area.

CONTRACTIONS, SYLLABLE BOUNDARIES, AND THE COMPRESSION OF COMPUTER-ORIENTED BRAILLE

I think that there is no doubt that contractions across a syllable boundary could tend to slow a person up. I think the problem here is to state the rules in such a way that a machine can follow them; in other words, to state mechanical rules that will give braille that is readable. Probably a statistical approach here would do. If there is only one word out of ten pages that is going to slow up a reader, then it is not going to slow him very much overall. If one can state the rules in such a way that they do the right things

* The material in this section was prepared from the question and answer period following Dr. Yngve's paper—Ed.

effectively most of the time, then if a mistake is made once in ten pages, or a contraction is made across a syllable boundary, the risk is worth taking. Let us make a standard to do the contracting so that different people who have different programs can still produce the same braille. I think that once the reader is used to the results they might not slow him up very much.

My guess is that the degree of compression in braille would not be changed appreciably. This is only an impression, for I have not made a study of this. The size of a braille book is not likely to be increased by very much; perhaps by one page out of 100, or something like that.

There are two problems here. One is the physical problem of storing all the words. One would have for example the word “hothouse,” in which the “th” should not be contracted, presumably. There are several ways of approaching this problem in a machine. One is to list all these words. This means merely looking up the word in the dictionary, seeing what list it is in, perhaps storing only one list, the smaller one. If the word is not in that list, then do the job the longer way. There is one problem here in that the list might require a fairly large storage. One way out is to store only the words one expects to run into frequently and not the others. Then the rule would be correctly followed most of the time. Another approach would be to look at the spelling and to make such rules as “ ‘th’ after ‘s’ should be contracted” (assuming that we find that this is generally the case), and for “hothouse,” the “th” after a vowel perhaps would not be separated. In other words, it might be possible to state the rules in terms of spelling and yet have a fairly satisfactory result. Whichever way it is done, there is not too much difference from a machine point of view, except it is out of the question to list all of the words involved in some of these rules.

The other problem is that there are many of these words; with vocabulary one is dealing essentially with an open class. People can invent new words, for example, and when they have invented new words one wants the program to deal with them correctly. It is not feasible, however, to list words that haven’t been invented or used.

LARGE COMPUTERS *VS.* MODIFIED BRAILLE

First of all, I agree 100 percent with the statement made here that the machine should serve man and not vice versa; this is in part my own motivation in working towards mechanical translation. Communication between different linguistic communities now goes entirely through people who are to some extent bilingual. If we had some machine aid in this area we could, I think, do something by machine which is quite a burden to people. I don’t want to be misunderstood on this point: from one point of view, one must change the rules of braille if the job is going to be done by machine. From another point of view, the rules need not be changed. It depends on just what one means by the phrase "changing the rules of braille." If one lists all the words, and indicates how they are to be contracted, this is a rule. This is a different rule from the kind of rule that tells us, "You must not cross syllable boundaries." The result may be precisely the same, which is to the good if it is judged that the braille as currently written is the best.

In other words I am not proposing to alter the braille codes as they are currently written unless there are good reasons to do so from the point of view of the reader. But I am proposing that the rules be restated in a machine-usable form, as they are in fact now being applied by working programs. The other comment I would have is that such rules as the use of "to2 contraction, and "by" contraction (in the case of preposition and not in the case of adverb) is something that is rather difficult to mechanize; it is not out of the question, but it would take a rather sophisticated computer program. We don't know quite how to do this completely adequately. If this rule were restated in some other way that would give the result intended (or very close to it), then I think we should do so, and we should say to ourselves, "This is machine braille that we are using."