On the Nonexistence of Articles in Japanese: A Problem in Machine Translation

Robert A. Russell
Few Western grammars of Japanese (descriptive, pedagogical, or otherwise) fail to note the absence of distinct forms in that language corresponding to the articles of English ('the', 'a(n)', and most uses of unstressed 'some'). The Japanese word shinbun, for instance, would be translated variously into English as 'the newspaper(s)', 'a newspaper', 'some newspapers', or simply 'newspaper(s)', depending upon the context (cf., e.g., Alfonso 1974:2).

Fewer still, however, are the grammars of Japanese which provide any meaningful insight at all into the identification and analysis of functionally equivalent (or at least similar) means in Japanese of expressing the types of semantic information normally associated with articles in English. The fact that morphologically distinct articles exist in some languages but not in others has given rise to a number of interesting questions, not only in general linguistic theory, but also in the areas of translation and second language acquisition research. The purpose of this paper is to discuss certain problems related to the situation just described, particularly from the perspective of the machine translation of Japanese into English.

Before we proceed to an examination of specific facts, a few general comments may be in order with regard to machine translation (hereafter MT) and the role of linguistic theory therein. Although, as Bar-Hillel and others argued several years ago, fully automatic high-quality translation is an unrealistic goal (and will likely remain so for the foreseeable future), a number of projects have continued to pursue MT research, motivated in some cases at least by the growing commercial feasibility of machine-assisted translation.

While all approaches to MT are ultimately concerned with source and target languages and with the development of algorithms for transforming structures from the former into the latter, there is, of course, considerable diversity with respect to the specific algorithms so employed. Nevertheless, as outlined by Hutchins (1978), MT design strategies may be broadly characterized as either "direct" (sometimes referred to as "first generation") systems or "indirect" (or "second generation") systems.

As the name implies, "direct" MT systems were designed for particular pairs of languages and generally incorporated only as much analysis of either language as was absolutely necessary to transform structures from the given source language (SL) into those of the given target language (TL). "Direct" approaches have typically been ad hoc, highly empirical, generally atheoretical, and relatively lacking in systematic semantic, or even syntactic, analysis.

More recent "indirect" systems, on the other hand, are (in design at least) characterized by more systematic syntactic and semantic analysis of both SL and TL, by the relative independence of SL analysis from
TL synthesis, and by the incorporation of a more or less abstract, intermediate level of representation—an "interlingua" or "metalanguage" of sorts—to which the SL may theoretically be reduced, and from which the TL may subsequently be synthesized.

Actually, as Hutchins (1978:130) notes, although concepts of universal grammar have a long history, even the most sophisticated system designs have been largely unsuccessful in implementing such ideas. In practice, most "indirect" systems have taken a somewhat more modest, three-stage approach consisting of (1) SL analysis, (2) SL-TL transfer, and (3) TL synthesis. In this approach, for any given language, the SL analysis is substantially the same regardless of the particular TL, and the TL synthesis for each language remains the same regardless of the particular SL, and only the SL-TL transfer component differs significantly according to the particular language pair.

I assume in general that there is a universal component of grammar, essentially semantic in nature, underlying all natural languages, and further assume (with Locke (1975:414-15) and others) that human translation of natural languages is not merely a linear replacement of SL words with TL words, but rather that it is an indirect process involving a common core of meaning at some nonsuperficial level. While an indirect "interlingua"-type MT system might more accurately reflect these general assumptions, I will also assume, for purposes of the following discussion, the rather more pragmatic Analysis-Transfer-Synthesis model of MT referred to briefly above. In any case, the nature of the analysis suggested below is not fundamentally altered on account of the simplification involved in this assumption.

Let us now return to the matter of the absence of articles in Japanese. In developing a Japanese-to-English MT or machine-assisted translation system, algorithms must be designed which insert appropriate articles into the English target text where required, based on information generated during analysis of the Japanese source text. This is a special case of a more general type of problem in MT; namely, that of inserting elements in the TL to which there are few, if any, distinctly identifiable corresponding forms in the SL. Other problems of this same general type in a Japanese-English system would include the insertion of appropriate pronouns in English corresponding to the commonly ellipsized (or "zero-pronominalized") subjects and objects of Japanese, the insertion of information necessary for number inflection in English based on clues in Japanese which are usually fragmentary at best, and so forth.

Given the general absence of unique forms in Japanese corresponding to English articles, a direct, one-to-one mapping approach to the problem is ruled out in principle. Instead, the assumption of a universal semantic base of some kind suggests the following general methodology: (1) on the basis of observation, hypothesize the primitive semantic notions underlying the use of articles (etc.) in various languages (here, we will be concerned only with English), (2) search for equivalent processes in languages which lack articles (i.e., for Japanese in this case, search for linguistic patterns expressing the hypothesized semantic primitives in the surface structure of Japanese), and (3) test the resulting mapping from SL (here, Japanese) surface structure to the universal base
(or, in our simplified model, through the transfer component) to TL (here, English) surface structure. One cannot realistically hope, in a paper of this brevity, to defend or to exemplify the entire process just outlined. Instead, an attempt will be made merely to give some idea of both the computational and the linguistic scope of the problem.

From a superficial, quantitative point of view, depending upon the ratio of actual to potential article occurrence in English target texts, at least two different general approaches may be suggested to the problem at hand. That is, in comparing Japanese source texts with their corresponding English translations, if the number of Japanese forms having English translations with actually occurring articles is a sufficiently high percentage of the total set of those forms whose translations may, in general, take articles depending upon the context, then it might well be argued that the more promising strategy would be to insert articles before every candidate form unless a given candidate could be ruled out on the basis of contextual analysis. This might be referred to as the insertion-default approach; i.e., when in doubt, insert the article.

If, on the other hand, the number of actually occurring articles is a relatively small proportion of the total number of potentially occurring articles, then the opposite approach would seem more promising: that is, insert articles nowhere unless a given candidate form could be shown, on the basis of contextual analysis, to require an article—the so-called noninsertion-default approach, or, when in doubt, leave it out.

Obviously, the matter becomes very much more complicated than this, but in order to get a little better quantitative feeling for which of the two general strategies mentioned above would be preferable, a small preliminary sample was taken, in which several pages of Japanese source text were compared with their corresponding English translations, and the number of articles was compared with the total number of forms that might, in general, have taken articles. Surprisingly, the ratio of actual to all potential occurrences of articles was less than 1:3, suggesting the preferability of the latter, or noninsertion-default approach.

To confirm this, a somewhat more formal sample was taken, in which the following hypotheses were tested:

Null Hypothesis: Proportion of actual to potential articles equal to or greater than 35%;

Alternate Hypothesis: Proportion of actual to potential articles less than 35%.

The sample proportion of actual to potential article occurrences was 596/1828, or approximately .326, which at a .05 significance level leads us to reject the null hypothesis (namely, that the population proportion of candidate forms actually occurring with an article in any given text is equal to or greater than 35%). Stated simply, if the data do indeed lead us to reject the null hypothesis, we are 95% certain that the number of forms requiring article insertion will amount to less than 35% of the total number of article-eligible forms in a given text.
Oversimplifying somewhat, the events of interest, along with their worst-case probabilities, may be represented in matrix form as follows:

<table>
<thead>
<tr>
<th>Article Required</th>
<th>Article Not Inserted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article Inserted</td>
<td>Correct (.35)</td>
</tr>
<tr>
<td>Article Not Inserted</td>
<td>Error (.35)</td>
</tr>
<tr>
<td>Article Not Permitted</td>
<td>Error (.65)</td>
</tr>
<tr>
<td>Article Not Permitted</td>
<td>Correct (.65)</td>
</tr>
</tbody>
</table>

Ignoring for the moment the problem of which article to insert (a'(n)' or 'the'), if we could simply narrow down the number of eligible forms to which the default option must apply, then we might considerably improve the 65% minimum accuracy that would result from no article insertion at all. To a certain degree at least, this may be accomplished in two different ways: (1) by inserting articles where we are reasonably certain they are required, and (2) by blocking article insertion where we are reasonably certain they are not permitted.

By slightly revising the hypotheses given above, we can capture the idea of reducing the number of article-eligible forms with a concomitant reduction in the critical ratio of actual to potential article occurrence overall:

Null Hypothesis': Proportion of actual to potential articles (excluding forms which meet conditions x,y,z) equal to or greater than k%;

Alt. Hypothesis': Proportion of actual to potential articles (excluding forms which meet conditions x,y,z) less than k%;

where the term "conditions x,y,z" refers to those conditions under which it is possible with virtual certainty to instruct the computer correctly to insert an appropriate article where required or correctly to prevent it from inserting an article where not permitted (in effect, removing such forms from the domain of application of more probabilistic rules of article assignment).

As we correctly treat, and thus eliminate article-eligible forms from the proportion of actual to potential article occurrences (by virtue of conditions x,y,z), we realize a corresponding (though smaller) decrease in the percentage k at which it will still be possible to reject the null hypothesis. To take a simple example, if k = 35% and total potential article occurrences are equal to 100, the expected number of actual article occurrences (based on the foregoing) would be 35. If, however, we were able (using conditions x,y,z) to account explicitly for even only 40% of the potential article occurrences, the noninsertion default applied to the remaining 60 potential occurrences (.35 x 60 = 21) would yield an effective reduction in k from 35% to 21% overall (21/100 = .21). That is to say, if no articles are inserted except where explicitly required by conditions x,y,z, the maximum expected number of errors of the type "Article Not Inserted Where Required" might be reduced from 35% to 21%, resulting in an overall accuracy rate of nearly 80%.
The question that remains, of course, is "what are the conditions x,y,z,?" While it is not possible at present to give anything like an exhaustive specification of such conditions, I will attempt in the balance of this paper to describe briefly their nature and to give some indication as to how one might proceed, in general, to treat them within the context of Japanese-English MT.

As suggested at the outset, very little contrastive analysis has been done which sheds any light at all on the relationship of articles in English to any functionally analogous constructions that may exist in Japanese. Such references as have been made to the problem are usually incidental to discussions of the postpositions wa and ga in Japanese, the nature of themes, and so forth. In a general discussion of the nature of thematic noun phrases, for example, Kuno (1973:39 ff.) observes that, in order to qualify as a theme, a noun phrase (hereafter, NP) must have "some specific referent in the universe of discourse," but must also already have been referred to in the present or a previous discourse (i.e., must be in the "temporary registry" of discourse) or must already be part of the "permanent registry" of discourse reserved for nouns of unique reference like 'the sun', 'the moon', 'my wife', etc. The following example illustrates the initial entry of a NP ('a man') into the temporary registry of discourse where it serves as the antecedent of the anaphoric NP 'the man' in the very next sentence:

(1) a. My wife saw a man and a little boy picking flowers out of our garden this morning.

b. The man was wearing a strange hat and an old, worn-out suit.

Since NPs in Japanese must generally be either anaphoric (in the extended sense above), generic, or contrastive in order to be marked with the postposition wa, and since in English the corresponding anaphoric and generic NPs generally require (or, as in the case of generic NPs, at least optionally permit) the definite article in the absence of any other overt determiners, it would be most convenient for MT if we could simply scan for wa-marked NPs and mark them for subsequent insertion of the definite article. (An analogous strategy is, in fact, often suggested in textbooks for native speakers of Japanese attempting to learn the obscurities of article usage in English.) In fact, in the texts sampled, virtually every occurrence of NP-wa involved definiteness (in the general sense).6

There are, however, cases of wa-marked NPs in Japanese corresponding to which the English NPs normally do not take a definite article; as, for example, in the contrastive environment shown in (2) below:

(2) a. pen ya enpitsu ga arimasu ka.
   pen and pencils NOM be/have ?
   'Do you have any pens and pencils?'

b. pen wa arimasu ga, enpitsu wa arimasen yo.
   CNTRST but CNTRST not be/have EMPH
   'I do have (some) pens, but I don't have (any) pencils.'
A further difficulty with the NPs in (2)b is that, while they have been mentioned previously in (2)a, they refer not to a specific object(s) but to a class of objects, and therefore are not anaphoric in the usual sense.

That the NP-wa scanning approach alone would be insufficient is further indicated by the fact that there are, of course, many NPs which are not marked with wa but which nevertheless require definite articles in the corresponding English NPs. Whether associated with wa-marking or not, anaphora is, to be sure, an extremely important factor in determining the requirements for article insertion in Japanese-English MT. Surprisingly, however, only about one third of the definite articles in the sample could be accounted for on the basis of anaphora in the narrow sense (that is, excluding those nouns which were not mentioned previously in the given text, but which are in the permanent registry, or the anaphoricity of which derives from a general understanding of the particular discourse setting). From a total of just over 1800 nouns, just over 400 occurred with a definite article, and only about one third of those were strictly text-anaphoric, which means in turn that fewer than 10% of the total number of article-eligible nouns could be expected to have definite articles assigned to them on the basis of strict text-anaphora alone. There are, of course, other conditions relating to definiteness and specificity, however.

The search for further conditions governing article insertion ultimately requires information from almost every level of analysis—from the lexical, phrase, subordinate and main clause levels, as well as from the discourse level, as required in the treatment of anaphora.

At the lexical level, for example, SL (Japanese) dictionary entries may be flagged to reflect information of the type mentioned above. That is, words in the permanent registry, like hi or taiyoo 'sun', tsuki 'moon', etc., might be flagged as "pre-definite" (or "probably definite," where the term "definite" is to be understood as a feature governing the syntactic requirement for the insertion of a definite article) and unless this status is revised during later stages of analysis, the default procedure during the transfer phase would be to insert the definite article preceding the TL equivalents of words so flagged.

Proper nouns, on the other hand, might be flagged at the lexical level as being "inherently definite" (or, with respect to the mechanical process of article insertion, "unmarked definite") and, therefore, as requiring no definite article unless, of course, subsequent analysis dictates otherwise. The sentences in (3) and (4) below exemplify the two cases just described. The noun tsuki (normally definite as in (3)a) is, in (3)b, reanalyzed at the phrase structure level as an indefinite, quantified expression, and so is transferred to English 'moon(s)' without the definite article. The proper noun Tanaka-san (normally "inherently definite" as in (4)a) is reanalyzed in (4)b as "definite" (i.e., as requiring subsequent article insertion) because of the relative clause modifying it.

(3) a. tsuki ga deta.
      moon NOM appeared
      'The moon has risen.'
b. dosei ni wa tsuki ga kokonotsu ijoo aru.
Saturn DAT THM moon NOM nine over be/have
'Saturn has more than (*the) nine moons.'

(4) a. Tanaka-san wa Shinjuku no chikaku ni sunde-imasu.
Tanaka-Mr. THM GEN near DAT live(s)
'(*The) Mr. Tanaka lives near Shinjuku.'

b. watakushi ga shitte-iru Tanaka-san wa Shinjuku no chikaku ni
I NOM know
sunde-imasu.
'The Mr. Tanaka (that) I know lives near Shinjuku.'

At the phrase and subordinate clause levels, there are several clues that may be appealed to as conditions governing the insertion or noninsertion of articles in the TL. The presence of determiners, quantifiers, certain kinds of adjectives, genitive NPs and relative clauses, together with information from the lexical level, will frequently be sufficient to force a decision as to whether or not an article should be inserted. Consider, for example, the following sentences:

(5) a. watakushi wa sono hon o yonda.
I THM that book ACC read
'I read that book.'

b. Kodomo ga sannin ita. (cf. also (3) b above)
child(ren) NOM three were
'There were three children.'

c. sono sannin no kodomo ni koinu o ageta.
those three GEN child(ren) DAT puppy ACC gave
'I gave those three children a puppy.'

d. yatto saigo no basu ga tsuita.
finally last GEN bus NOM arrived
'Finally the last bus arrived.'

e. Nihon no seifu wa kooshoo o tatta.
Japan GEN government THM negotiations ACC cut off
'The government of Japan (or, Japan's government) cut off negotiations.'

f. musuko ga kooen de atta shoonen wa kyoo mo mata kita.
son NOM park LOC met little boy THM today too again came
'The little boy that my son met at the park came again today.'

Nouns (such as hon 'book' in (5)a above) which are found during phrase structure analysis to be modified by a demonstrative (such as sono 'that')\(^8\) may be flagged at that point as "unmarked definite" (again, only in the sense of not requiring insertion of the definite article in the TL). Nouns (like kodomo 'child(ren)' in (5)b) modified by a quantifier (like sannin 'three') may generally be ruled out as "unmarked definite," unless further modified by a determiner (as in (5)c) or some other defining structure (such as a relative clause). Nouns modified by certain types
of adjectives (like saigo no 'last' in (5)d and saisho no 'first') which imply the existence of a unique referent(s) may safely be flagged as "definite," as may the objects of genitive nouns (e.g., seifu 'government' in (5)e). Alternatively, of course, the objects of genitive NPs may frequently be translated better in the form given in the parenthesized portion of the translation of (5)e (namely, 'Japan's government'), in which case the noun in question could safely be flagged "unmarked definite." Finally, nouns modified by relative clauses (as in (5)f) most frequently (though certainly not always) may be flagged as "definite." It may not, after all, be too much to hope that the probability of correctly flagging NPs with various types of modifiers as "definite," "indefinite," unmarked definite," etc., may be raised to a very high level indeed, by combining lexical, case-marking, phrase structure, discourse, and other information at different levels of analysis.

At the main clause level, in addition to analyzing the relationships between relative clauses and their head nouns, the relationships between verbs and their various arguments, etc., certain types of syntactic patterns tend to signal the introduction of new vs. old information, or vice-versa. In (6)a, for example, the typical existential pattern of (Locative NP - Subject NP - Existential Verb) strongly indicates that the ga-marked subject represents new information and should therefore be flagged as "indefinite" (i.e., as requiring the insertion of an indefinite article). The word teeburu 'table' in both (6)a and (6)b, on the other hand, can safely be considered syntactically "definite," since the principal NPs both in the locative phrases of existential sentences as well as in the antecedent clauses of cleft sentences generally represent old or presupposed information:

(6) a. teeburu no ue ni usagi ga imasu.
     table GEN top LOC rabbit NOM be
     'On top of the table (there) is a "the/*0) rabbit.'

     b. teeburu no ue ni iru no wa usagi desu.
        be PRO THM be
        'What is on top of the table is a rabbit.'

Finally, at the discourse level, it is possible to trace back through the NPs of preceding sentences in search of antecedents for potentially anaphoric NPs, but not without considerable costs in terms of both processing time and storage space. In our sample texts, most of the antecedents of the strictly text-anaphoric NPs occurred on an average of between five and six sentences previously. At an average of between three and four nouns per sentence, one can imagine the potential complications, even if one attempts to trace back only as far as the average distance to an antecedent. Fortunately, many (if, indeed, not most) NPs which might otherwise trigger an anaphoric trace may be eliminated at earlier stages of analysis by some of the kinds of conditions discussed above.

In conclusion, although Japanese lacks morphological articles as such, largely the same types of semantic information (anaphora, new vs. old information, genericity, specificity, etc.) that underlie article usage in English are grammaticalized in Japanese in a variety of (sometimes not so obvious) ways. We have seen that an analysis which hopes to capture that information in the form of conditions on article insertion
within a Japanese-English MT system must be integrated over virtually all levels of linguistic structure. Although there are many relevant questions that could not even be addressed, let alone resolved, in a paper of this kind, some definition has (hopefully, at least) been given to the problem of article insertion in a Japanese-English MT system, indicating in the process, perhaps, possible directions for further, more detailed research.

NOTES

* I would like to thank two of my colleagues in particular, Akira Kurahone and Kerry Wyckoff, for their useful comments and suggestions during the preparation of this paper. Of course, the usual exculpations apply.

1. See especially Bar-Hillel (1960) and (1971).

2. We will not enter into a discussion here concerning the possibility of interaction during the analysis to provide some of the information necessary to insert articles. There may be debatable benefits to such an approach from a commercial point of view, but it is relatively uninteresting from a theoretical point of view, if the theoretical objective is a completely explicit account of the relationship between English articles and forms or constructions bearing the same types of information in Japanese.

3. I am not assuming that all semantic primitives, whatever they may be, will necessarily be grammaticalized in all languages. Specific semantic distinctions made in one language in whatever form may be totally absent in some other language.

4. The sample texts were taken from Natsume Sōseki's Kokoro (1952:7-22) and its English translation by Edwin McClellan (1957:1-15), and from Lafcadio Hearn's Kwaidan (1959:4-33) and its accompanying translation.

5. See especially Kuno (1972) and (1973) for a discussion of the now well-known distinction between "thematic" and "contrastive" uses of wa as opposed to the "neutral description" and "exhaustive listing" interpretations of ga-marked subject noun phrases.

6. Allowance must, of course, be made for the fact that certain nouns (proper nouns, nouns modified by demonstratives, etc.), although definite in some sense, are prevented from taking definite articles in the target text. The occurrence of NP-ga presents a more complex situation. Nearly 50% of the ga-marked subject NPs were definite in the general sense, but most of these were the subjects of subordinate clauses which are normally marked by ga regardless of whether they represent new or old information.

7. There is a rather nontrivial problem which concerns us here in that, as Kuno (1973:235) notes, there are apparently no phonological, morphological, or syntactic distinctions between restrictive and nonrestrictive relative clauses in Japanese. The dilemma is, of course, that (4)b can be translated either as it is in the text (with a restrictive reading) or as follows (in a nonrestrictive reading): 'Mr. Tanaka, whom I know,
lives near Shinjuku.' In the nonrestrictive reading, of course, the
definite article would be inappropriate with 'Mr. Tanaka.' A fairly
sophisticated semantic analysis is required here, one which goes beyond
the sentence level.

8. There are cases, of course, where sono is best translated as 'the', not
'that'. See Lyons (1977, vol. 2:646 ff.) for a discussion of the historical
and functional relationships between demonstratives and definite articles
in English.

REFERENCES

L.L. Center of Applied Linguistics.

Bar-Hillel, Y. 1960. The present status of automatic translation of
languages. Advances in computers.1.91-163.

-----1971. Some reflections on the present outlook for high quality
machine translation. Feasibility study on fully automatic high
Griffiss AFB, New York: Rome Air Development Center.


Hutchins, W. J. 1978. Progress in documentation: machine translation and
machine-aided translation. Journal of documentation.34, no.2.119-
159.

Kuno, S. 1972. Functional sentence perspective: a case study from Japanese
and English. Linguistic Inquiry.3.269-320.


information science, vol. 16, ed. by A. Kent, H. Lancour, and


Chicago: Henry Regnery Co.
