Japanese-Sinhalese** machine translation system Jaw/Sinhalese

Samantha Thelijjagoda*, Yoshimasa Imai and Takashi Ikeda

Department of Information Science, Graduate School of Engineering, Gifu University, Gifu, Japan.

Revised: 05 October 2006 ; Accepted: 16 March 2007

Abstract: This paper describes a machine translation system, Jaw/Sinhalese, that translates Japanese into Sinhalese (Sinhala). This is the first Japanese-Sinhalese machine translation system. Both Japanese and Sinhalese are agglutinative languages. The Japanese language contains bunsetsu (Japanese basic linguistic units), which consist of a content word (a lexical root) with one or more function words. Sinhalese also has bunsetsu-like units, but the grammatical structure is not necessarily fully revealed. The paper proposes a method of analysis for the bunsetsu-structure of Sinhalese which shares features with Japanese, and translation solutions for Japanese function words after predicates and nouns. Function words after predicates express tense, modality, conjunctions, etc., while function words after nouns express the case, topic, thematic roles, etc. This paper clarifies the bunsetsu-structure of Sinhalese. It also distinguishes case marker (a leading group of function words after nouns) correspondences using three types of pattern-based translation rules and solves the multi-layered problem of the translation of function words after a predicate by means of translation rules in a table format. Translations were implemented on the pilot machine translation system, Jaw/Sinhalese. As an experiment, 200 sample sentences were evaluated. The results (72% rate of success) indicated that this approach is within an acceptable accuracy range.

Key words: Bunsetsu, case marker, expression elements, function word after predicate, Japanese-Sinhalese machine translations, Sinhalese bunsetsu, Sinhalese grammatical structure, transfer rules

INTRODUCTION

Machine translation (MT) developments in Japan over the past decade have increased extensively. Most projects consider translations between Japanese and English. Other Asian languages such as Chinese and Korean have also been considered, but languages like Sinhalese (Sinhala) have received less attention. Jaw/Sinhalese has been developed as a pilot machine translation system to translate Japanese into Sinhalese. This is the first attempt at a MT system for this language pair.

Japanese and Sinhalese are typologically classified as Subject Object Verb (SOV) languages and both are agglutinative languages. (In agglutinative languages some function words agglutinate to a content word. Function words after the noun, work as case markers, topic markers, etc, and function words after the verb, express tense, modality, etc.) Japanese has bunsetsu (Japanese basic linguistic units) that consists of a content word with one or more function words. The function words of bunsetsu may be divided into two types. The first type includes the function words that come after a predicate, which express grammatical categories like tense-aspect, modality, conjunctions, etc. The second type includes the function words that come after nouns, which express grammatical categories such as case, topic, focus, etc. Sinhalese also has bunsetsu-like units, and for the case of comparison, they will be referred to as Sinhalese bunsetsu.

In this paper, a method is proposed to analyze the Sinhalese bunsetsu structure for machine translations. More than 20 kinds of verb inflection words (base parts of the verb bunsetsu) are defined and a set of function words that follows them is reorganized. Likewise, for the case inflection word of a noun, a new organization of function words is proposed. The case inflection is formed by a combination of a noun stem and a function word for case marking. 27 function words are isolated for case marking.

Since Sinhalese and Japanese are very similar in this respect, in many cases there are one-to-one correspondences between Japanese bunsetsu and Sinhalese bunsetsu. But at the same time, these correspondences are...
The translation of Japanese case markers (a leading group of function words after a noun) into Sinhalese is ambiguous; that is, one Japanese case marker often corresponds to several Sinhalese case markers. For example, the Japanese case marker “を” “wo”, which is usually used to indicate the object case, corresponds to 16 case markers in Sinhalese {O (to), SDO (kata), SAP (valata), I (va), SDO (kva), SDO (nwa), SDO (en), SDO (kin), SDO (valin), SAP (ak), SAP (ek), O (o), SAP (ehi), Z (ka), SDO (wala), Ø}. These are distinguished by using base-type rules of the system, which are dependent on the fact that the case frame pattern of a verb basically determines case markers.

One Japanese function word after a predicate often affects more than one expression element in a Sinhalese sentence. For example, the Japanese function word “[食べ]た” “[tabe]ta”, which expresses “hope/eagerness” on the part of the subject, is translated into “O(ta)”, “32s? (aky-z5>25te)”, “kannay” and “3E)oQ)SG3<3(awashyayai)” after a subject, after an object, after an inflected verb, and after a verb) in a Sinhalese sentence. This problem is called a multi-layered problem of translation and is solved by describing the translation rules in a table format.

The first stage of the Jaw/Sinhalese system processes the translations of propositional content. The clarification of case marker correspondence is handled in this stage. The second stage covers the translation of tense and modality, which are expressed by function words after the predicate in Japanese. The multi-layered problem of the translation of Japanese function words after a predicate is resolved at this stage. To estimate the accuracy of these rules, 200 Japanese sentences were evaluated. The results (72% rate of success) show this approach is effective and acceptable within machine translations.
The expression tree and generation of Sinhalese:
The system is implemented in C++ and the transfer rule is, in fact, a program stored as a dynamic link library (DLL). The execution of the transfer rules in the transfer tree (TT) produces a network for Sinhalese expressions (expression tree: ET) (Figure 4.). A linearizing function is defined for each object as a class method of C++. The execution of the linearizing function on the ET orders the members of the ET to make a Sinhalese output sentence.
In the case of translating Japanese function words after predicates, the information of the function word is directly provided to the ET from the IT, and this activates the transfer rules for function word translations while executing the linearizing function.

Every object of the ET has its own generation procedure. The linearizing function of each object linearizes the components, which are represented as members of the C++ objects such as subject, object 1, object 2, adverb and verb. Besides them, noun modifiers such as time, numerical, time-era, time-begin, time-end, and the expression components for function words after predicates such as After-Subj, After-Obj, Before-Vb, and After-Vb perform a major role in the linearizing function. The order of the linearization is programmed according to Sinhalese word order. The arrangement of all components for the generation of Sinhalese is shown in Figure 5.

"Jaw/Sinhalese is a pattern-based machine translation system. It performs the translations of propositional contents and Japanese tense-aspect and modality (Japanese function words that come after a predicate). The propositional contents are translated through the IT, TT, and ET. The ET holds the propositional contents in Sinhalese. The translations of Japanese tense-aspect and modality were introduced to the ET through an appropriate method, which will be described later.

Three types of rules are employed for Japanese expression patterns and transfer, a base-type rule (b-rule) and two addition-type rules (a-ruleFW, a-ruleCW). A base-type rule is a case-frame-like rule and deals with the translation of basic propositional contents. Addition types deal with adverbial expressions and conjunctional expressions, which are optionally added to the base-type expressions. Some examples of these three types are shown below (Table 1).

Table 1 also shows the corresponding translation rules for Sinhalese. The classes, member classes, member names, values, and case markers are used to build a C++ program for the translation rule. The C++ program is automatically constructed with this. At present, a database of 370,000 Japanese expression patterns is used. The "Jaw"-editor has been developed to write the patterns and transfer rules.

Sinhalese and Japanese grammatical structures

The outlines of Sinhalese and Japanese. Sinhalese is a member of the Indo-Aryan family of languages and its script bears a close structural resemblance to Thai and Malayalam scripts. The Sinhalese writing system is a syllabary system (a set of written symbols that represent syllables, which make up words) derived from ancient North and South Indian scripts. The traditional literary Sinhalese alphabet consists of 58 symbols. Of these, only 42 symbols (12 vowels and 30 consonants) are necessary to represent the writing system.

Japanese belongs to the Altaic language family and employs a combination of three different types of writing systems hiragana, katakana, and kanji. Hiragana is a system of Japanese words for which kanji cannot be easily provided. Katakana is used similarly for transcribing foreign loan words (other than Chinese) and some onomatopoeic words.

In the agglutinative language structures, Japanese and Sinhalese words are formed by joining meaningful units. The words should be separated by spaces in

---

Table 1: Japanese expression pattern and translation rules for Sinhalese

| Japanese Pattern | Rule Type | Class       | Member Name          | Member Class | Value (Sinhalese) | Case Marker |
|------------------|-----------|-------------|----------------------|--------------|------------------|-------------|
| 1 N1 ga          | b-rule    | CProposition| m_subject            | CProposition | 1                | Ø           |
| 2 N2 to          | b-rule    | CProposition| m_object             | CProposition | 2                | Ø           |
| 3 (N1 associate with N2) takau | a-ruleFW | CpConnection| m_pSubordinate       | CProposition | 1                | Ø           |
| 1 N1 wa          | b-rule    | CProposition| m_subject            | CProposition | 1                | Ø           |
| 2 N2             |           |             | m_object             | CProposition | 2                | Ø           |
| 3 da             |           |             | m_centerW            | CProposition |                  | Ø           |
| 1 V1 temuru to   | a-ruleCW  | CAdjective  | m_adjective          | CAdjective   | 1                | Ø           |
| 2 V2 (Vb2 when Vb1) |     |             | m_centerW            | CAdjective   |                  | Ø           |

---
Sinhalese. If the space is not placed correctly, the sentence becomes meaningless or assumes a different meaning. But the Japanese writing system does not separate each word. The Japanese sentence, as described earlier, can be partitioned into several meaningful linguistic units called “bunsetsu”.

A bunsetsu is a chunk of words consisting of a content word [noun, verb, etc.] accompanied by some function words [particle, auxiliary, etc.]. A function word is based on a prime root that has a lexical meaning. This root word, which is a lexical item defined in the dictionary, includes a verb, a noun, an adjective, and an adverb. A function word consists of post-positional particles and auxiliary verbs. The concept of bunsetsu is common for Sinhalese, too.

In literary Sinhalese, the subject agrees with the verb in number and person, whereas in spoken Sinhalese, there is no agreement between the subject and the verb. Japanese uses neither number nor person agreement. Japanese is a topic-comment prominent language with a basic word order of SOV. Beyond being verb-final, Japanese word order places a modifier (such as adjectives and clausal modifiers) before the noun being modified. The notation of the topic plays a vital part in organizing information to form an utterance or a sentence. The topic is marked by the topic-marking particles “wa” and “mo”. The Japanese language has particles or post-positions that express not only a grammatical relationship, but also interpersonal feelings. Non-specification of the topic, the subject, the object, and the particles is common. Unlike Sinhalese, Japanese has a rich system of respectful and humble forms, as well as a variety of polite expressions.

The main parts of speech of Sinhalese are the noun, the verb, and the particle. Adjectives and adverbs are not considered parts of speech in Sinhalese, as they are derived from nouns and verbs. But in Japanese, all these components are regarded as parts of speech.

The prefix “no” is added to a verb base to yield the corresponding negative meaning in Sinhalese. In the case of compound verbs, “no” is added to the last element, which is a simple verb. In Japanese, adding the suffix “-nai” to the stem of a verb creates its negative form. The negative forms of Japanese verbs become adjectives, which are used to express a condition. Irregular Japanese verbs require different padding vowels.

The Sinhalese bunsetsu structure: The Sinhalese bunsetsu structure depends on a basic Sinhalese grammatical unit. More research is needed to analyze its internal structure. In this paper, the linguistic structure of Sinhalese bunsetsu was observed and a sound description is proposed.

In Sinhalese, a bunsetsu consists of one content word and some function words. Content words are nouns, verbs, adverbs, etc. Function words are added to the content word in noun bunsetsu and verb bunsetsu. Other bunsetsu consist of a content word only. The structure of the noun and verb bunsetsu will be discussed here.

The Sinhalese noun makes a distinction between definite/indefinite, singular/plural, and animate/inanimate. An animate noun has six cases (direct, accusative, dative, genitive, instrumental, and vocative), and an inanimate noun has four cases (direct, dative, genitive, and instrumental). These case forms are distinct for singular-definite nouns, singular-indefinite nouns, and plural nouns. Sinhalese function words such as කිරීම (samaga), සුදුම්පන් (pasuwa), අතරම (unath), දෙයිම (pawa), දෙයිම (herd), දෙයිම (sandaha), and කිරීම (nisa) are agglutinated after these case forms.

A noun inflects its ending part according to case. Instead of the idea of case inflections in nouns, a set of function words for case marking is proposed. The proposal contains a method of constructing a case inflection word by linking a noun stem and the function word for case marking. For this purpose, a linking system has been devised. The linking system handles the changes of the ending letter of the noun and the beginning letter of the function word for case marking. Other function words precede this case inflection (Figure 6).

![Figure 6: Sinhalese noun bunsetsu with function word for case inflection](image)

Figure 6: Sinhalese noun bunsetsu with function word for case inflection

The Sinhalese bunsetsu structure depends on a basic Sinhalese grammatical unit. More research is needed to analyze its internal structure. In this paper, the linguistic structure of Sinhalese bunsetsu was observed and a sound description is proposed.

In Sinhalese, a bunsetsu consists of one content word and some function words. Content words are nouns, verbs, adverbs, etc. Function words are added to the content word in noun bunsetsu and verb bunsetsu. Other bunsetsu consist of a content word only. The structure of the noun and verb bunsetsu will be discussed here.

The Sinhalese noun makes a distinction between definite/indefinite, singular/plural, and animate/inanimate. An animate noun has six cases (direct, accusative, dative, genitive, instrumental, and vocative), and an inanimate noun has four cases (direct, dative, genitive, and instrumental). These case forms are distinct for singular-definite nouns, singular-indefinite nouns, and plural nouns. Sinhalese function words such as කිරීම (samaga), සුදුම්පන් (pasuwa), අතරම (unath), දෙයිම (pawa), දෙයිම (herd), දෙයිම (sandaha), and කිරීම (nisa) are agglutinated after these case forms.

Figure 7: Linking way of the noun “pasela”(school) and function word for case marking “en”

![Figure 7: Linking way of the noun “pasela”(school) and function word for case marking “en”](image)

A noun inflects its ending part according to case. Instead of the idea of case inflections in nouns, a set of function words for case marking is proposed. The proposal contains a method of constructing a case inflection word by linking a noun stem and the function word for case marking. For this purpose, a linking system has been devised. The linking system handles the changes of the ending letter of the noun and the beginning letter of the function word for case marking. Other function words precede this case inflection (Figure 6).

![Figure 8: The form of Sinhalese verb bunsetsu](image)

Figure 8: The form of Sinhalese verb bunsetsu
Table 2: Sinhalese diacritic symbols for vowels

| Notation | Vowel (short) | Diacritic | Notation | Vowel (long) | Diacritic |
|----------|---------------|-----------|----------|--------------|-----------|
| ad₀      | a (♀)         | ∅         | ad₁      | a: (♀)       | ∅         |
| ad₁      | e (♀)         | t         | ad₂      | i: (♀)       | t         |
| ad₂      | i (♀)         | n         | ad₃      | u: (♀)       | n         |
| ad₃      | u (♀)         | u         | ad₄      | u: (♀)       | u         |
| ad₄      | e (♀)         | @         | ad₅      | e: (♀)       | @         |
| ad₅      | o (♀)         | ®         | ad₆      | o: (♀)       | ®         |
| ad₆      | a: (♀)        | a:        | ad₇      | a: (♀)       | a:        |

Table 3: The linking methods of a noun stem and function word for case marking

| Linking of noun stem NZ and function word for case marking | adding | dropping |
|-----------------------------------------------------------|-------|----------|
| (Z is the last letter of noun)                            |       |          |
| In roman                                                  |       |          |
| In Sinhalese                                              |       |          |
| 1  a:ta  ☞                                           |     ad₈ | a:       |
| 2  ekwa  ☞                                           |     ad₈ | e        |
| 3  o    ☞                                           |     ad₆ | o        |
| 4  akin  ☞                                           |     ad₅ | e        |
| 5  ekuge ☞                                           |     ad₆ | e        |
| 6  aka  ☞                                           |     ad₆ | e        |
| 7  en    ☞                                           |     ad₆ | e        |

Twenty seven linking methods are introduced and some of them are listed (Table 3). To explain the linking method simply, the function words for case marking are denoted in Roman letters instead of Sinhalese script. These linking methods are implemented with the help of the linking rules that are employed for Sinhalese consonant-vowel combinations.

Sinhalese characters are generally composed of consonants and vowels. When a Sinhalese consonant is joined to a vowel, a diacritic (a small mark that added to a letter to alter its pronunciation) is added to the consonants instead of the vowel. This is known as the linking rule of a consonant and a vowel. These vowel diacritics will be called “Sinhalese additions” hereafter. They are placed to the top, bottom, left, and right of the consonants. Sinhalese has 12 vowels and each of them has a diacritic or diacritics (Table 2). In total, 16 linking rules were developed for constant-vowel combinations (each vowel u (♀), u: (♀) and e: (♀) consists of more than one symbol, depending on the consonant, and only the vowel “a (♀)” has no diacritic (♀)).

Figure 7 shows the example of a linking method according to the 7th method “(NZ = pasela) + en”, as illustrated (Table 2). When the function word for case marking is ☞ (en), the final letter of the noun ☞ (la), changes to ☞ (le), and the ☞ (e) of ☞ (en) is dropped. The last letter ☞ (n) of ☞ (en) is joined to the changed noun stem ☞ (pasele) and the inflected noun ☞ ☞ ☞ (paselen) emerges.

Sixteen (16) function words for case marking were defined for animate nouns and eleven (11) were defined for inanimate nouns. These are shown in Tables 4 and 5 respectively with the sample nouns “friend” and “school”. In this method, the noun stem is always in the singular-definite-direct case. This noun form will be called “the noun” hereafter. In this way, all inflected forms can be formed for each noun from 27 function words and 16 linking rules.

In Sinhalese, these combinations are formed individually and necessarily listed in a dictionary (1). But with this method, the number of combinations is reduced considerably (2).
Table 4: Linking ways of a noun and function words for case marking on animate noun

| Case forms | Singular | Plural |
|------------|----------|--------|
|            | Definite | Indefinite |        |
| direct     | ᄅ겠습니다 + ᄀ | yahaluwa + ek | yahaluwa + o |
|            | yahaluwa + ᄀ | yahaluwaek | yahaluwo |
| accusative | ᄅ겠습니다 + ᄀ | yahaluwa + ᄀ | yahaluwa + ᄀ |
|            | yahaluwa + awa | yahaluwa + akwa | yahaluwa + awha |
|            | yahaluwa + ᄀ | yahaluwaek | yahaluwaek |
| dative     | ᄅ겠습니다 + ᄀ | yahaluwa + ᄀ | yahaluwa + ᄀ |
|            | yahaluwa + awha | yahaluwa + ekata | yahaluwa + anu |
|            | yahaluwa + ᄀ | yahaluwaek | yahaluwaek |
| genitive   | ᄅ겠습니다 + ᄀ | yahaluwa + ᄀ | yahaluwa + ᄀ |
|            | yahaluwa + age | yahaluwa + ekage | yahaluwa + ange |
|            | yahaluwa + ᄀ | yahaluwaek | yahaluwaek |
| instrumental | ᄅ겠습니다 + ᄀ | yahaluwa + ᄀ | yahaluwa + ᄀ |
|            | yahaluwa + ahen | yahaluwa + ekagen | yahaluwa + angen |
|            | yahaluwa + ᄀ | yahaluwaek | yahaluwaek |
| vocative   | ᄅ겠습니다 + ᄀ | yahaluwa + ᄀ | yahaluwa + ᄀ |
|            | yahaluwa + e | yahaluwa | yahaluwa |

Table 5: Linking way of a noun and function words for case marking on inanimate noun

| Case forms | Singular | Plural |
|------------|----------|--------|
|            | Definite | Indefinite |        |
| direct     | ᄅ겠습니다 + ᄀ | pasela + ᄀ | pasela + (-q) |
|            | pasela + ᄀ | pasela + ak | pasela + (a) |
|            | pasela + ᄀ | paselaak | pasel |
| dative     | ᄅ겠습니다 + ᄀ | pasela + ᄀ | pasela + ᄀ |
|            | pasela + ata | pasela + akata | pasela + watala |
|            | pasela + ᄀ | paselakata | paselwatala |
| genitive   | ᄅ겠습니다 + ᄀ | pasela + ᄀ | pasela + ᄀ |
|            | pasela + eta | pasela + eta | pasel + wala |
|            | pasela + ᄀ | paselata | paselwatala |
| instrumental | ᄅ겠습니다 + ᄀ | pasela + ᄀ | pasela + ᄀ |
|            | pasela + en | pasela + akn | pasel + welin |
|            | pasela + ᄀ | paselan | paselwalin |

a. The irregular plural nouns were not be inserted
where \( n \) = number of nouns in dictionary

In the case of a Sinhalese verb, they are generally conjugated according to person, number, and tense. Sinhalese has conjugated single verbs (Table 6) and conjugated inflected verbs. The inflected verbs are preceded by some function words. If there is a function word after the inflected verb, the function word at the end is conjugated for inflected verb. The most general way to form an inflected verb form is by the addition of suffixes to the verb-base. The function words are optional after them. The verb-base is derived by dropping “nawa” from the general verb form (Table 7).

In the Sinhalese verb bunsetsu, the inflected verb and function words have no clear definition as of yet. The inflected verb and function words are represented as a single verb block. But this has some subdivisions, which must be separated. Here, a solid separation is defined between the inflected verb and the function words after the verb as Sinhalese verb bunsetsu (Figure 8). More than 20 kinds of inflected verb forms emerge initially. The Sinhalese function words can be easily placed after the inflected verb with the help of these separations in machine translations.

All inflected verb forms are listed in a dictionary because of their irregular formations. Table 8 shows only a part of the dictionary. All inflected verbs are derived from either the present form or past form of Sinhalese verbs. Adnominal present, Adnominal past, Formal-request, etc. are the abbreviations for the inflected verb forms adnominal present, adnominal past, and formal request, etc. respectively.

The verb bases of Sinhalese causative verbs are different from general verb bases. It is impossible for the machine to distinguish between causative verbs and the general verb. Another parallel dictionary was developed and used for Sinhalese causative verbs.

Translation methods for Japanese function words

One characteristic of the Japanese language that is shared with Sinhalese is that the meaning of a sentence cannot be understood entirely until one reaches the very end. In an extreme case, it is possible to alter the meaning of a sentence by changing the last word. This is a characteristic of an agglutinative language, which does not possess the prefixes and suffixes of an inflectional language or the strict word order of an isolated language. In Japanese, grammatical functions are performed by function words.

A function word is an inflecting agent that agglutinates to indeclinable words (nouns) and to declinable words (verbs). Here, a solid trial of translations has been made for Japanese function words after the noun and Japanese function words after the predicate while minimizing the ambiguity problems and multi-layered translation problems.

The translation of Japanese function words after a noun: Japanese function words for case marking are considered the leading group of Japanese function words after the noun. This section discusses the translation methods of Japanese function words for case marking into Sinhalese. Japanese function words for case marking such as \( \text{E}(\text{wo}) \), \( \text{E}(\text{wa}) \), \( \text{E}(\text{ga}) \), \( \text{E}(\text{ni}) \), and \( \text{E}(\text{no}) \) often yield several function words for case marking in Sinhalese. For example, the Japanese function word for case marking “\( \text{E}(\text{wo}) \)”, which usually indicates an object, corresponds to no fewer than 16 function words for case marking in Sinhalese: \{ \( \text{E}(\text{u}) \), \( \text{E}(\text{kata}) \), \( \text{E}(\text{vala}) \), \( \text{E}(\text{va}) \), \( \text{E}(\text{ku}) \), \( \text{E}(\text{wa}) \), \( \text{E}(\text{en}) \), \( \text{E}(\text{kin}) \), \( \text{E}(\text{valin}) \), \( \text{E}(\text{ak}) \), \( \text{E}(\text{ek}) \), \( \text{E}(\text{hi}) \), \( \text{E}(\text{ka}) \), \( \text{E}(\text{vala}) \), \( \text{E}(\text{wa}) \), \( \text{E}(\text{no}) \) \}. Some examples of the multi-correspondence of function words for case marking \( \text{E}(\text{wo}) \) and \( \text{E}(\text{no}) \) are shown in Tables 9 and 10 respectively.

Basically, the function words for case marking are determined by verbs. Therefore, they can be distinguished by using the base-type rule (the b-rule).

Examples:

1. \( \text{E}(\text{wo}) \) \( \text{E}(\text{wa}) \) \( \text{E}(\text{ga}) \) \( \text{E}(\text{ni}) \) \( \text{E}(\text{no}) \).

I send a letter to him by post.

2. \( \text{E}(\text{wo}) \) \( \text{E}(\text{wa}) \) \( \text{E}(\text{ga}) \) \( \text{E}(\text{ni}) \) \( \text{E}(\text{no}) \).

He mixes corn flour using water.

3. \( \text{E}(\text{wo}) \) \( \text{E}(\text{wa}) \) \( \text{E}(\text{ga}) \) \( \text{E}(\text{ni}) \) \( \text{E}(\text{no}) \).

I know Mr. Sakai at the Tokyo office well.

* In inflectional languages inflectional morphemes are added to a word, which may indicate grammatical information (case, number, tense, person, etc.) as in the case of English.

** In isolated languages like Chinese, the word expresses just substantial meaning and there is no inflection of word forms. The grammatical relations are mainly expressed by the word order.
In these Japanese sentences, the function words for case marking *wa*, *wo*, *ni*, and *de* are used with different verbs. Corresponding Sinhalese function words for case marking are different in each sentence and can be successfully translated using the above-mentioned base-type rule through the benefits of pattern-based translations.

**Table 6: Conjuation of the Sinhalese verb**

| singular | Non-past | Past | Non-past | Future | Past |
|----------|----------|------|----------|--------|------|
|          | present  |      | present  |        |      |
| First    |          |      |          |        |      |
| balami   |          |      | balamu   |        |      |
| balannemi|          |      | balannenu|        |      |
| beleenu   |          |      |          |        |      |
| Second   |          |      |          |        |      |
| balahi   |          |      | balahu   |        |      |
| balannehi|          |      | balannahu|        |      |
| baluahu   |          |      |          |        |      |
| Third    |          |      |          |        |      |
| balai    |          |      | balathi  |        |      |
| balanneya|          |      | balannahu|        |      |
| beliuha   |          |      |          |        |      |

**Table 7: Examples for Inflected verbs and optional function words**

| Verb  | General form | Verb base | Inflected verb suffix | Optional function words |
|-------|--------------|-----------|-----------------------|-------------------------|
| Give  | गुज्जमः | de-       | गुज्जमः                | वियाहेकि             |
|       | denawa       |           | sitinawa              | wagei                  |
| Go    | गक्षामि | za-       | गक्षामि                | नेय                 |
|       | yamawa       |           | labanna               | padawan               |
| Eat   | गक्षामि | za-       | गक्षामि                | गि                 |
|       | yamawa       |           | labanna               | padawan               |
| Run   | गुक्षामि | go-       | गुक्षामि               | गि                 |
|       | duwanawa     |           | duwamin               | wagei                 |

**Table 8: Inflected verb forms of Sinhalese verbs**

| verb  | Perfect | Progress | Adno -present | Adno -past | Formal -rqst |
|-------|---------|----------|---------------|------------|--------------|
| go    | गिन   | गिन   | गिन   | गिन   | गिन   |
|       | yam    | yam    | yam    | yam    | yam    |
| take  | तक्षान | तक्षान | तक्षान | तक्षान | तक्षान |
|       | gena   | gena   | gena   | gena   | gena   |
| give  | गिन    | गिन    | गिन    | गिन    | गिन    |
|       | deela  | deela  | deela  | deela  | deela  |
| walk  | गिन    | गिन    | गिन    | गिन    | गिन    |
|       | evidala| evidala| evidala| evidala| evidala|
Table 9: Multi-correspondence of the Japanese function word “wo” with Sinhalese function words

| Japanese sentence | Sinhalese translation | Sinhalese Function words |
|-------------------|-----------------------|--------------------------|
| tsuma wo aisuru   | පළකිය ගෝජට පිවේ | 0                        |
| (I] love [my] wife) | brändadu adarakarana | wa                       |
| kodomo wo tsurete kru (accompany the child) | launarya ekka yanawa | va                       |
| hashi wo wataru (cross the bridge) | palamen godawenawa | en                       |
| gohan wo taberu (eat rice) | bath@ kanawa |                          |
| tnu wo kau (a dog is bred) | ballek hadanawa | ek                       |

Table 10: Multi-correspondence of the Japanese function word “no” with Sinhalese function words

| Japanese sentence | Sinhalese translation | Sinhalese Function words |
|-------------------|-----------------------|--------------------------|
| Tokyo no oji (Uncle in Tokyo) | මොළකු ප්‍රශ්නයක් | wala                     |
| oji no te (Uncle’s home) | manege gedara | ge                       |
| fuku no iro (Color of the clothes) | endunmih paata | ehi                      |
| uchi no niwa (Home garden) | gedara midula |                          |
| gakkō no hon (Book belongs to the school) | pasela satu | \(0\)                       |
| haha no tamena (On behalf of mother) | amma wenuwen | \(0\)                        |

Table 11: Japanese sentences without and with function words

| Sentences without function words | Same sentences with function words |
|----------------------------------|-----------------------------------|
| 太郎は道を歩く (Taro wa michi wo aruku) Taro walks in the road. | 太郎は道を歩いていた (Taro wa michi wo aruite ita) Taro was walking in the road. |
| 彼は来る (kare wa kuru) He will come. | 彼は来るでしょう (kare wa kuru deshō) He will probably come. |
| 愛を開く (mado wo akeru) Window is opened | 愛を開けてください (mado wo akete kudasai) Please open the window. |
| 箱はジャンプする (kuma wa jumpu suru) Bear jumps. | 箱はジャンプするらしい (kuma wa jumpu suru rashii) It seems that a bear jumps. |
| 私はワインを飲む (watashi wa wain wo nomu) I drink wine. | 私はワインを飲みたい (watashi wa wain wo nomiita) I want to drink wine. |
| 本を読み (hon wo yomu) The book is read. | 本を読みやすかったでしょう (hon wo yomi yasukatta deshō) The book was easy to read. Wasn’t it? |

Append special meanings to ordinary verbs. They are restricted to dependent usage and always follow independent words or independent phrases. In other words, Japanese function words are postpositional. At the second stage of Japanese-Sinhalese translations, the translations of Japanese function words after a predicate are carried out covering tense-aspect and modality. The multi-layered problem of the translation of a Japanese function word after a predicate into Sinhalese is solved by translation rules presented here in a table format. As an introduction, Table 11 illustrates some examples of various Japanese function words in various environments.

For ease of translation, Japanese function words after a predicate are divided into four groups by IBUKI system: (1) voice, etc., (2) tense-aspect, etc., (3) judgments, and (4) conjunctions (Table 12).
Table 12: Groups of Japanese function words

| [1] voice, etc | [2] tense-aspect, etc | [3] judgments | [4] conjunctions |
|----------------|-----------------------|---------------|------------------|
| られる (rareru) | た (ta) (past)        | だろう (dorō)  | て (te)          |
| させる (saseru) | ている (teru)         | だろう (dorō)  | が (ga)          |
| たい (tai)     | ていた (taita)        | らしい (shōi)  | と (to)          |
| なる (naru)    | だ (da)               | でしよう (deshō)| ように (youni)   |
| 退くれる (tekureru) | である (daeru)       | なければならない (nakerebanaranai) | である (demo) |
| とても (temoru) | です (desu)          | かもしれない (kamoshirenai) | から (kara) |
| やすい (yasui) | してきた (kaita)     | ね (ne)       | ので (node)     |
| ほしい (tehoshii) | ます (masu)        | べき (beki)   | ば (ba)         |
| なくなる (nakaranu) | なかった (nakatta) | な (na)       | ような (youna)  |
| にくい (nikui) | ...                   | ...           | ...             |

For example, the formation of the above groups of function words can be described as follows.

Sample sentence: 本は読みやすかったでしょう。

hon wa yomi yasukatta ta deshō.

The book was easy to read. Wasn’t it?

According to the above sentence, “hon wa” is the object with a case marker and “yomi” is an inflection of the verb “yomu”. The other parts are function words which are picked up from each column of Table 12 (the relevant parts are underlined). The word “yasukatta (was easy)” is a Japanese expression for the past of “yasui (easy)”. This information is collected from column[1] and “ta” in column[2]. And the ending word “deshō” comes from column[3].

A Japanese function word is translated into several expression elements at different places in a Sinhalese sentence. They cannot be translated through a simple one-to-one correspondence. This is identified as the multi-layered problem in Japanese-Sinhalese translations.

Examples:

1. sakana ga kujira ni taberareru. (The fish is eaten by the whale.)

   咲殺な鯨に食べられる (talmaha visin maluwa wa kanu labanawa.)

2. sakana ga kujira ni taberareteirayōda. (The fish is probably being eaten by the whale.)

   咲殺な鯨に食べられている (talmaha visin maluwa wa kanu labuwayagei.)

The translation rules for these function words are described in a table format (Tables 13-16) with expression elements in its each column (Japanese function words were taken from Japanese newspaper articles that were published in the past ten years and listed in a descending sequence. Only the top rows of each table, with rules, are presented here). The expression elements are Vb-Inflection, CVb-Inflection, After-Verb, After-Sub, After-Obj, and Judge-Verb.

“Vb-Inflection” indicates the inflected form of the verb for Sinhalese. A Sinhalese verb has more than 20 inflected forms (Table 8).

“CVb-Inflection” indicates the inflected form of the causative verb. The causative verb means the causative form of the verb (see the example below).

Example for a causative verb:

kare ni wain wo nomasetakatta yōda.

(1) probably have him drink wine.

The “After-Verb” and “Before-Verb” indicate the elements to be put in just after the verb and before the verb respectively.

The “After-Sub” indicates the elements that always follow the subject, while the “After-Obj” indicates the elements that follow the object in the output sentences. The “Judge-Verb” indicates the elements at the very end of the translated sentences.

A simple example for the translation of function words:

kare wa wain wo nomitakatta yōda.

He probably wanted to drink wine.
The Japanese phrase “-takatta yōda” is divided into tai/ta/yōda (want/ed/probably). The Sinhalese translations are generated from the translation rules for “-tai” (Table 13), “-ta” (Table 14) and “yōda” (Table 15), applying the linearizing function for Sinhalese.

The rules for the fourth group, “conjunctions”, are different from the rules of the other groups. Table 16 shows the rules for conjunctions. There are four types for building Sinhalese sentences with conjunctions.

1. Subordinate Sentence [with inflected verb] + Main Sentence

The inflected verb form of the subordinate sentence is indicated in column “V-change (Present)” for type 1 in Table 16.

Example:
taiyou wa higashi kara dete nishi ni shizumu (The sun rises in the east and sets in the west).

2. Subordinate Sentence [with tail-changed verb] + conjunction + Main Sentence

This type is prepared for cause-effect and temporal conjunctions like node (since), youni (as if), demo (although), nagara (while), etc. The special effect is that the tail of the verb in a subordinate sentence is changed according to the tense. The last letter of the verb is dropped in present tense [DropLL] (example: だける) and the last addition of the verb is dropped in the past tense [DropLA] (example: でゅる).

Example:
ame ga futte iru node kuruma de iku. (As it is raining, I travel by car.)

wessa wahina(wa) nisa moter ratayen yanawa.
The “wa” is the last letter that is dropped from the verb in the present.

3. Subordinate Sentence+conjunction+ Main Sentence

This type is for conjunctions such as ga (though), nara (if), to (and), ka (or), etc. In this case there is no change in the verb of the subordinate sentence and it is simply connected to the conjunction.

Example:

kono hon wa nando mo yonda ga, yoku wakarimasen.

{Though I read this book several times, (I) cannot understand it.}

me pota keepavitak ma kiyawuwath terum ganna apahasui.

4. Subordinate Sentence (with inflected verb) + conjunction + Main Sentence

Both the inflected verb and the conjunction are employed for this type. Toshite (and next), made (till), kara (after), etc. are included in this type.

---

Table 13: Translation rules for the Japanese function word group “Voice, etc.”

| Japanese function words | Vb-inflection | After-Verb | After-Sub | After-Obj |
|-------------------------|---------------|------------|-----------|-----------|
| rareru                  | Formal-comm   |            | でる       | でる       |
| saseru                  | C-Present     | -          | でる       |
| tai                     | Formal-rqst   |            | でる       | でる       |
| naru                    | Formal-rqst   | でる       | でる       |
| tekereru                | Conj          | でる       | でる       |
| temoraru                | Conj          | でる       | でる       |
| nanu                    | Formal-rqst   | でる       | でる       |
| tehoshii                | Conj          | でる       | でる       |
| nikui                   | Formal-rqst   | でる       | でる       |
| sugiru                  | Past          | でる       | でる       |
| tekudasai               | Conj          | でる       | でる       |
| saseraeru               | C-Passive     | -          | でる       |
| temoraitai              | Conj          | でる       | でる       |
| temoraeru               | Conj          | でる       | でる       |
| teyaru                  | Conj          | でる       | でる       |
| takunaru                | Formal-rqst.  | でる       | でる       |
Table 14: Translation rules for the Japanese function word group “Tense and Aspect, etc.”

| Japanese function words | Vb-inflection | CVb-inflection | After-Verb | Before-Verb |
|-------------------------|---------------|----------------|------------|-------------|
| ta                      | Past          | C-Past         |            |             |
| teiru                   | Progress      | C-Progress     |            |             |
| da                      | -             | -              |            |             |
| nai                     | -             | -              |            |             |
| teita                   | Progress      | C-Progress     |            |             |
| data                    | -             | -              |            |             |
| dearu                   | -             | -              |            |             |
| desu                    | -             | -              |            |             |
| tekita                  | Progress      | C-Progress     |            |             |
| masu                    | Present       | C-Present      |            |             |
| nakatta                 | Past          | C-Past         |            |             |
| dewanai                 | -             | -              |            |             |
| deoru                   | Progress      | C-Progress     |            |             |
| tensai                  | Progress      | C-Progress     |            |             |
| dekuru                  | -             | -              |            |             |
| mushita                 | Past          | C-Past         |            |             |
| tekku                   | Progress      | C-Progress     |            |             |
| teemasu                 | Progress      | C-Progress     |            |             |
| tekkuau                 | Conj          | C-Conj         |            |             |
| dekuau                  | -             | -              |            |             |
| nanoda                  | -             | -              |            |             |
| tenzamatta              | Conj          | C-Conj         |            |             |
| masen                   | Present       | C-Present      |            |             |
| deshu                   |             |                |            |             |
| temuru                  | Progress      | C-Progress     |            |             |

Table 15: Translation rules for the Japanese function word group “Judgments”

| Japanese function words | Vb-inflection | Judge-Verb |
|-------------------------|---------------|------------|
| ro                      | Will-Vb       |            |
| daro                    | Present       |            |
| souda                   | Present       |            |
| deshō                   | Present       |            |
| rashii                  | Present       |            |
| yōda                    | Past          |            |
| nakerebanaranai         | Ending-NE     |            |
| kamoshirenai            | Present       |            |
| ne                      | Present       |            |
| bekii                   | Adno-Present  |            |
| bekida                  | Adno-Present  |            |
| na                      | Present       |            |

Example:

gohan wo tabete kara gakko e iku. {After eating, (I) go to school.}

“Kewa” is the inflected verb and “ta pasu” is the conjunction.

RESULTS AND DISCUSSION

The MT system from Japanese to Sinhalese, Jaw/Sinhalese, is proposed, including a clear definition of the Sinhalese bunsetsu structure and solutions for Japanese function words that follow a noun and a predicate.
Table 16: Translation rules for the Japanese function word group “Conjunctions”

| Japanese | Conjunctions | Type | Sinhalese | Conjunctions | V-change (Present) | V-change (Past) |
|----------|--------------|------|-----------|--------------|--------------------|-----------------|
| te       | 1            | *    | Conj      |              |                    |                 |
| ga       | 3            | गा   | Past      |              |                    |                 |
| to       | 4            | टो   | DropLL    |              |                    | DropLA          |
| youni    | 2            | यूनी | DropLL    |              |                    | DropLA          |
| de       | 1            | *    | Conj      |              |                    |                 |
| demo     | 2            | डीमो | DropLL    |              |                    | DropLA          |
| kara     | 4            | का   | past      |              |                    |                 |
| node     | 2            | नोड | DropLL    |              |                    | DropLA          |
| ba       | 1            | बा   | IfCondit  |              |                    |                 |
| youna    | 2            | यूना | DropLL    |              |                    | DropLA          |
| ra       | 4            | रा   | Past      |              |                    |                 |

* the system itself generates these conjunctions

Table 17: Japanese input sentences and translated Sinhalese sentences by Jaw/Sinhalese

| Japanese sentences (input sentences) | Sinhalese sentences (translated sentences) |
|--------------------------------------|--------------------------------------------|
| (I) borrowed an umbrella from a friend.) | අංගමක් නමුත් මිය වූනාදුණු කළඟක් බිද කරන්න. |
| (It seems the game was finished in a draw.) | අශ්ඨමුශ්ම මිය මෙම යොදා ගෙන නෙමෙම නිරැමුණක් කරන්න. |
| (She made new clothes for the daughter.) | ඝෙම් නොදු මොළම කොටස් තුළ ලැබේම්. |
| (It seems that he was awfully surprised when he heard the news from the friend.) | අශ්ඨමුශ්ම මෙම මේම විශේෂී සිරදු අවශ්කය මෙහෙයේ ආදියක් වීමේ අතරින්. |
| (It seemed mat grandfather is working on distributions during the fruit season.) | අශ්ඨමුශ්ම මෙම මේම මෙයේ සිරදු අවශ්කය මෙහෙයේ වේශේ බැඳුමෙක් අතරින්. |
| (This medicine will relieve your pain, won’t it?) | මෙමේම කැන්තා මෙම පායි මෙම කෝට්ඨා සිරදු අවශ්කයේ අතරින්. |
| (His name was famous in (his) hometown.) | මෙමේමේම මේම පායි මෙම කෝට්ඨා සිරදු අවශ්කයේ අතරින්. |
| (It grew up to be a beautiful swan from an ugly duckling.) | අශ්ඨමුශ්ම මෙමේමේමේමේම කැන්තා කෝට්ඨා සිරදු අවශ්කයේ අතරින්. |
| (Will your dream come true someday?) | අශ්ඨමුශ්ම මේම මේම පායි මෙම කෝට්ඨා සිරදු අවශ්කයේ අතරින්. |
| (He bought a new car for his only son.) | අශ්ඨමුශ්ම මේම මේම පායි මෙම කෝට්ඨා සිරදු අවශ්කයේ අතරින්. |
| (The teacher is teaching us geography.) | අශ්ඨමුශ්ම මේම මේම පායි මෙම කෝට්ඨා සිරදු අවශ්කයේ අතරින්. |
| (If I went to the dentist and got that tooth pulled out.) | අශ්ඨමුශ්ම මේම මේම පායි මෙම කෝට්ඨා සිරදු අවශ්කයේ අතරින්. |
| (I) would like you to represent my colleague and make a speech on behalf of all. | අශ්ඨමුශ්ම මේම මේම පායි මෙම කෝට්ඨා සිරදු අවශ්කයේ අතරින්. |

Although the usage of Japanese function words in translation is difficult to master, it is an important part of translations from Japanese into Sinhalese. In many cases, there are one-to-one correspondences between Japanese bunsetsu and Sinhalese bunsetsu. But these correspondences are far from perfect: for example, a single Japanese case marker often corresponds to several Sinhalese case markers. The translation of a Japanese function word after a predicate affects more than one expression element in a Sinhalese sentence. This paper has proposed a solution to these problems, culminating in the Jaw/Sinhalese Machine Translation System.

Two hundred example sentences, which cover most Japanese grammatical characteristics, were chosen manually from 1000 sentences that were taken from a Japanese-English dictionary. The translation rules for function words after a predicate and after a noun were
Table 18: Errors at translations

| Japanese sentence | Sinhalese sentence and error description |
|------------------|----------------------------------------|
| (a.1) watashi wa kare ni takusan okane wo kiteiru (I borrow money from him in large quantities.) | (The translation of “teiru” is in present continuous tense. But the simple present tense is required.) |
| (a.2) watashi wa kare wo watashi no nanijin to kangaete iru (I think that he is my companion.) | (same as error description of (a.1) ) |
| (b.1) niku wa mikkakongo ni kusatte shimatta (Meat were rotten after 3 days) | (Translation of “kusatta” should be used for Sinhalese translations.) |
| (b.2) kawa ni ochitefuku ga nurete shimatta (I fell into the river, and clothes were soaked.) | (Translation of “nurete shimatta” is not appropriate. Translation of “nurete” is needed.) |
| (c) hitibito wa isamashii kare wo aishiteiru (People love brave him) | (“wo” makes different translations) |

Table 18 shows some examples for incorrectly translated sentences. The underlined parts are the function words (single underlines show function words after a predicate and double underlines show function words after a noun. Japanese function words with dot underlines have no translation for Sinhalese and dark underlines show conjunctions). For the growth and improvement of the system, a thorough contrast of this language pair is required. To improve the system and its accuracy, the size of the dictionary and translation rules have to be increased.

CONCLUSION

The Jaw/Sinhalese system is described in this paper, including a definition of the Sinhalese bunsetsu structure and solutions for Japanese function words. This study on Japanese function words in a Japanese-Sinhalese machine translation system is the first attempt at constructing syntactical noun and verb formats for Sinhalese. The formation of nouns and verbs in this language pair is virtually unexplored from the machine translation perspective. This work provides the first test data on the evaluation of a practical pattern-based transfer approach. A test was conducted for 200 sample sentences using the Jaw/Sinhalese system, and it achieved a good translation result. The sample is small as it is the first trial of MT for this language pair. The sample size is to be increased in the next development stage, and then the forthcoming problems have to be solved.

For the growth and improvement of the system, the Jaw/Sinhalese system produces a special object for the subject called “ellipses” (Figure 3 and 4). When there is no subject, the verb is kept in dictionary form according to Sinhalese grammar.
Compared to Sinhalese, the Japanese language has a rich set of function words after the predicate and noun. The translations in this study so far have been limited to a sample, which needs to be expanded with more investigations. Furthermore, the grammatical aspects such as definite/indefinite, singular/plural and animate/inanimate in Japanese nouns have to be solved with a meaning (semantic) analysis to determine the correct formation of Sinhalese verbs and case markers.

References

1 Imai Y., Shieh J., Thelijjagoda S., Eiikewala N., Chau M., Bu Z. & Ikeda T. (2003). Machine translation system for Japanese to Asian languages. Proceedings of the 65th National Convention of Information Processing Society of Japan (IPSJ). Tokyo University of Technology, Tokyo, pp. 363-366.

2 Thelijjagoda S., Imai Y., Eiikewala N. & Ikeda T. (2002). Sinhalese generation rule and machine translation system Jaw/Sinhalese. Proceedings of Forum of Information Technology 2002 Conference. Tokyo, pp. 173-174.

3 Karunatilake W.S. (1990). An Introduction to Spoken Sinhalese. pp.6-16. M.D. Gunasena and Company Ltd. Colombo.

4 Sugawara M.(1989). A Japanese approach to Japanese. pp.537-593. The East Publications, Inc., Minato-ku, Tokyo.

5 Herath A. (1997). A study on natural language processings for Sinhalese. PhD dissertation. Gifu University, Japan.

6 Karunatilake W.S. (2001). Sinhala basha viyaranaya. pp.173-210. M.D. Gunasena and Company Ltd, Colombo.

7 Maynard S.K. (1990). An introduction to Japanese grammar and communication strategies. pp.134-148. The Japan Times Ltd. Tokyo, Japan.

8 Thelijjagoda S., Imai Y., Eiikewala N. & Ikeda T. (2003). Machine translation system Jaw/Sinhalese. Proceedings of the 9th Natural Language Processing Conference. Yokohama, Japan. pp. 477-480.