# Proposing a Multi-lingual Translation Scheme Utilizing the Extensible Markup Language XML

# Proposing a Multi-lingual Translation Scheme Utilizing the Extensible Markup Language XML

Bhooshan R. NEUPANE, Shuzo YAJIMA

Abstract

The paper proposes a new idea concerning a Multi-lingual translation scheme utilizing the newly evolving Internet tag language, Extensible Markup Language XML. The data description property of XML can be used to create an effective system to translate documents. First of all the XML tagged document of a source language is prepared manually. These tags are not only for the words but also for the grammatical structure, or for example, the phrase structure grammar, so that the analytical process of the translation can be reduced to a level suitable for many Internet applications. Next, XML document type definitions (DTDs) of grammatical structures of different languages are created. In the translation process the source sentences are broken down into pieces and then categorized to the respective DTDs to which they belong. The sentence structure of each language serves as a main structure tree, the broken elements are mapped with the elements of the DTD accordingly, which consequently maps them with the relative elements of the target language structure tree and the appropriate transformations are made. Among the possible transformation patterns, the most appropriate one is selected as an output.

## 1. Introduction

Developments in machine translation systems can be considered as one of the most important achievements of the modern era. But one hundred percent accurate translation system is still beyond our reach. The Extensible Markup Language XML is appearing as one universal interface with many possibilities for handling data structures. Thus far, the use of XML has been limited to describing data and recording the changes in their structure. The unique property of XML to tag data to their meanings has not been applied in the multi-lingual translation environment. This paper proposes a new machine translation scheme utilizing XML, which is thought to be flexible enough to operate within various linguistic environments, in particular multi-lingual translation within the Internet environment.

The flow chart shown in Fig.1 is the bone-structure of the research planned. There are three phases, namely, "Definition phase", "Analysis phase", and "Proposal phase". As this research area is so vast, the present paper is a proposal for the first step forward. We have tried to use our target languages from three cultural environments, one from the western end; English, one from the eastern end; Japanese and one Nepalese from the center of the Eurasian continent, one of the closest offsprings of Sanskrit, which is one of the oldest languages and commonly taken as the mother of Indo-European languages.

The proposed idea is still a scheme; neither the actual system nor the proposed DTDs

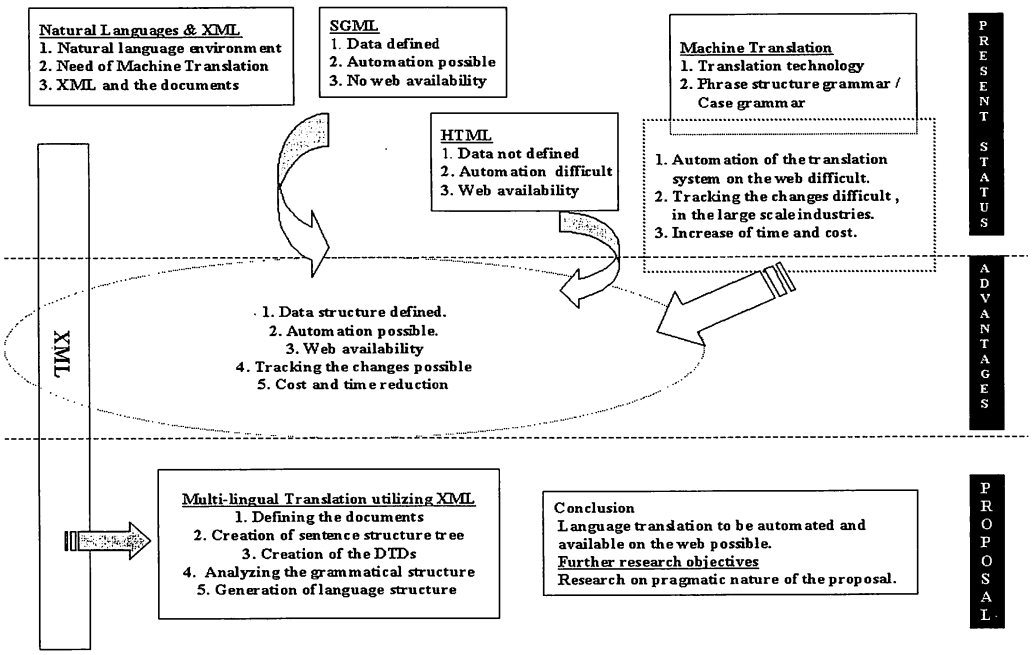


Fig.1. Flow chart of Multi-lingual translation schema; use of XML and the advantages

have yet been completed. Much work will be necessary to develop the system, which actually runs as a universal platform for the multi-lingual translation environments.

## 2. Multi-lingual Translation and XML

The most commonly used translation systems so far are bilingual. The effective Multi-lingual system on the Internet has still not been developed. The machine translation systems developed thus far are not effective on the web. There is a need for manual processing before translating the text extracted from the web. This manual processing is turning out to be one of the biggest handicaps while designing the web efficient multi-lingual systems.

The biggest challenge in the management of today's language translation is to manually keep track of the content what is new, reused, revised, translated, reviewed and post-translated. Industries face the trouble of translating user manuals and catalogues throughout the production time. The translating process must run along with the production process and should be completed together, but changes in products during the production period suddenly complicates the translation job. Machine translation systems generally are not able to produce translated documents in time. Post-editing is usually done manually, which makes the whole system work once again on the same document when some minor changes are to be made. Human translators are always place biased and when the translation is required for some of the documents from different departments and subjects, the specialist of the technical field has to run toward them. It creates panic on the job, and sometimes results in decrease of quality and companies have to enter the market without appropriate translations. The need for fast communication and extremely fast translation processes are still the biggest headache. There are many cases where massive losses were incurred even though the product was excellent but they were unable to provide the correct and fast translation of the manuals and lack of communication.

The arrangement of changes and manual tracking of the content can be recorded with the use of XML. But in so far as the automatic recording of the changes and automated translation system are not reached, the process will keep on going late and the current problems could not be solved.

XML has not been used so far as a system for the language translation of documents. And using XML to the translation process so far is to break up the information into smaller components, so that it is easier to pinpoint the changes, translate only new information and automatically update information reused throughout the document. One of the systems developed for keeping track of the documents is XML's content management system. Without

burdening the author, the system automatically collects information and identifies what is changed. Instead of sending the entire document only modified components need to be extracted and sent for translation. With the use of a content management system, the source and target language of the document are guaranteed to be aligned systematically. Use of the system will help companies by reducing the initial translation costs by reusing common content across documents, improve re-translation by pin-pointing the changes, shorten time to market by overlapping authoring and translation process and minimizing the volume of translation.

Describing the documents, independent of media, gives XML the benefit of using XML documents for print, on the web or any other document medium. This flexibility facilitates information system designers to use XML as they can adopt one set of standards, tools and methods for processing documents, regardless of their various distribution targets. Broadly, it is said XML helps [9] to

a. enable internationalized media-independent electronic publishing
b. allow industries to define platform-independent protocols for the exchange of data
c. deliver information to the agents that allow automatic processing
d. make easy for people to process data using inexpensive software
e. allow people to display information the way they want it
f. provide metadata (data about information) which will help people find information and help information producers and consumers find each other

This paper suggests a method for

g. facilitating a new multi-language environment suited for the Internet

Describing documents can be used to semantically categorize the words and sentence structures, and defining the rules for the sentence structure and analyzing meaning, which in turn could help to create automatic translating machines which could also be used on the web. Unicode, a default coding used in XML is one of the great advantages in creating a system to understand multi languages within the same interface.

## 3. Introduction to the Nepalese language

The Nepalese language is one of the offsprings of the Sanskrit language which is considered as one of the oldest and mother of some Indo-European languages. Sentences in Nepalese languages are divided into four types depending upon their structures, and there are strict rules that are followed when a sentence is created. Analyzing the sentence structure of the Nepalese language and defining them with the computer understandable

rules can lead to the digital analysis of the language which in turn could be a great help for the different set of languages that are Sanskrit derivatives. Moreover the Nepalese language shows certain resemblances to Japanese and English, which also shows the possibility of the Nepalese language emerging as the bridge between the two.

The Nepalese language falls within the Indo-European language group and is a direct descendent of the Sanskrit language that is considered as one of the oldest and still in practice in some parts of the Indian sub-continent. Some of the Indo-European languages are thought to be descendent of Sanskrit and maintain its basic grammatical structure.

The Nepalese language follows Sanskrit's grammatical structure which shows the possibility, that analysis of Nepalese language structure could be the base for the analysis of most of the Sanskrit derivatives. Generally sentences in the Nepalese language can be divided into four forms

a) Simple sentence
b) Compound sentence
c) Complex sentence
d) Mixed sentence

As described by generative grammar, sentences are of two types, basic and derived. Clauses and complex sentences are derived from basic sentences. A sentence having only one finite verb is called a simple sentence; a sentence having only one independent clause and one or more dependent clauses is called a complex sentence; a sentence having only independent clauses is called a compound sentence and a sentence having at least two independent clauses and at least one dependent clause is called a mixed sentence.

A sentence can be defined as follows:

a) Sentence = (subject) noun phrase + verb
b) Noun phrase = modifier + noun
(measurement)
c) Adjective = adjective, possessive / genitive, numeral classifier and adverb
d) Finite verb = modifier + complement + verb
e) Complement = (indirect object) measurement + (direct object) name + complement

From the above five definitions a typical sentence pattern can be defined as

Sentence = {(modifier + subject measurement) | (modifier + indirect object measurement) | (modifier + direct object measurement) | (modifier complement + measurement)} + (modifier + verb)

Using the formulae like above the sentence structure can be defined, and following the grammatical rules (briefly described below), a sentence can be transformed into simple sentences and broken into smaller constituents as shown in Fig.2.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Timro** | **bahile** | **mero** | **bhailai** | **afuno** | **bhai** | **sandhai** | **mancha.** |
| **Timro** | **bhai+ le** | **mero** | **bhai+ lai** | **afuno** | **bhai** | **sandhai** | **mancha.** |
| **Your** | **brother** | **my** | **brother** | **his** | **brother** | **always** | **thinks** |

**Your brother always thinks of my brother as his brother.**

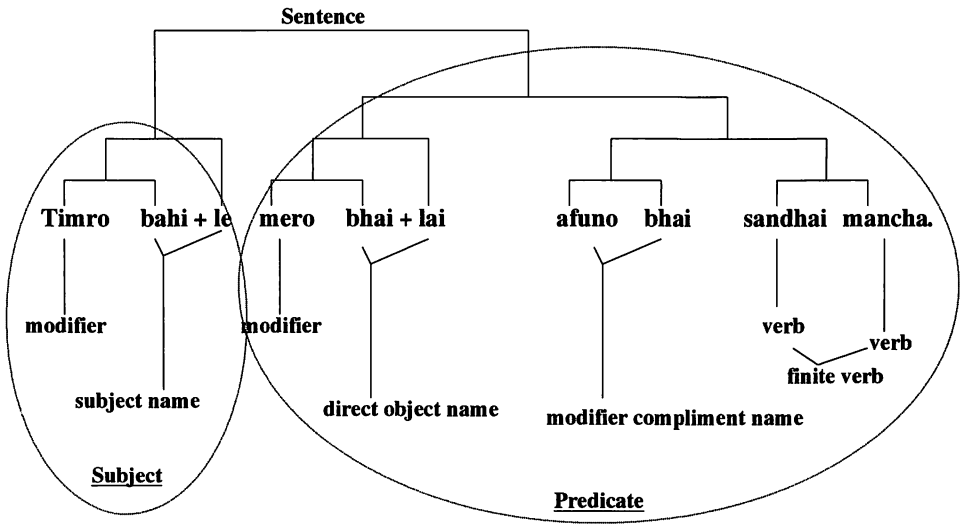


Fig.2. Language structure of a simple Nepalese sentence.

The clauses in a sentence can be defined as Principle clauses and Sub-ordinate clauses. Generally in the Nepalese sentences the Sub-ordinate clauses are on the left side of the sentence, except the proteases, which fall on the right side of the principal clause. The clauses are also defined by the verbs used in the sentence. The clauses having only a participle but not a finite verb can be taken as the non-finite clauses and the clauses having finite verb can be taken as the finite clauses.

One another way to analyze the Nepalese sentence is to determine the relative clause, starting from the pronominal words using "j " (phonetic ja) at the beginning. As a sentence might not carry the interrogative pronominal words even at the beginning of an interrogative pronominal sentence, the Aryan languages, such as Nepalese contains one another clause called co-relative clause, which is used in a sentence to make an interrogative pronominal

sentence. There are also other different types of rules that show the relationship of a clause with different conjunction or verbs. The clauses also depend upon the functions they perform in a sentence. A clause might function as a verb, as a noun, as an adjective or as an adverb.

The analysis procedure requires a complex job of breaking a complex, compound or mixed sentences into a simple sentence. Simplified sentence is then broken into different constituents and the relationship between them is shown to understand the semantic meaning.

The Nepalese language shows some of the resemblances in the grammatical pattern with the Japanese language. The arrangement of the words in a sentence is almost similar, which makes it easier to analyze the word order in the similar way as the Japanese as shown in Fig.3. Whereas, some of the words used might be formed by adding certain components to the actual words. In Japanese fixed articles are generally added to the words so most of the words are independently changed with the verb in different pattern. But in Nepalese or in English the components added to the words like "I" or "you" to make "my" or "your" are not fixed. In other words, the added components do not have independent existence. It shows the need of analysis procedure similar to English to analyze the Nepalese words.

The Nepalese verbs, adjectives and adverbs change when used with masculine and

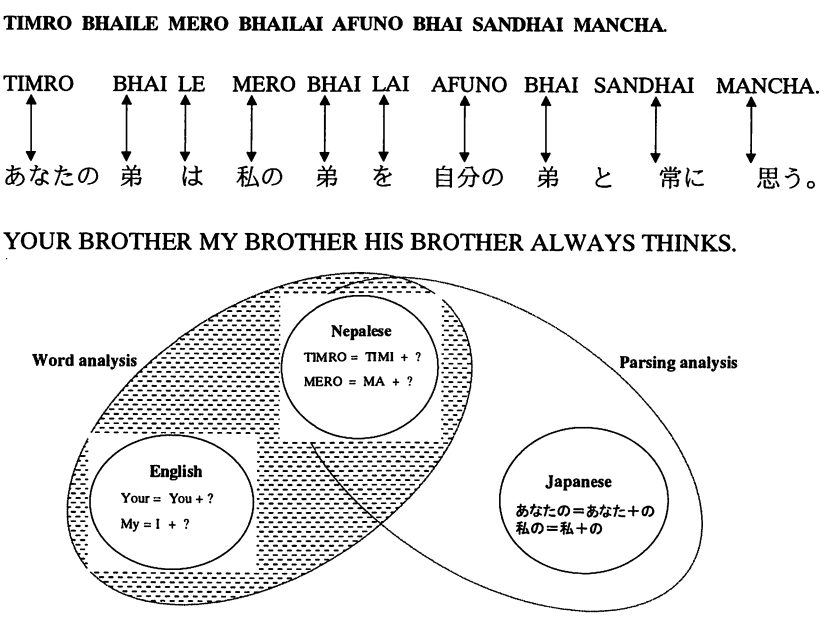


Fig.3. Similarities on the word and sentence analysis of Nepalese, Japanese and English languages.

feminine words. There is a need of a different type of analysis procedure as all the components used within a sentence always refer to masculine or feminine or non-living forms and animals. Also the words are changed when addressing the person above, below or within the same level.

As sentences in English generally do not refer to the levels of the speaker. There might be a loss of information when a Nepalese sentence such as follows is translated in English. For example,

UHAN SANGA RAMRO KITAB CHA.

It is generally translated in English as

HE HAS A GOOD BOOK.

In the Nepalese sentence the word "UHAN" refers to the male person, who is in higher level than the speaker. The translation in English will not show the level of the speaker and his relation to the person talked about.

Words in English rarely change when addressing to various peoples, where as in Japanese the addressing is different. The addressing in the Nepalese language is even more complicated as there are much more words and phrases used and many complex rules to be follows when addressing as person in the different or same level with the speaker. The different levels are shown in Table 1 and Table 2. This type of difference in linguistic environment brought by the cultural difference in various language groups makes the translation complicated when generating a sentence using a machine. The semantic analysis of the speaker's social factor cannot often be translated into another language.

Table 1. Different levels of pronouns used in Nepalese

| | Singular | | | | |
|---|---|---|---|---|---|
| | Non-honorofic | Honorofic | Formally honorofic | Highly honorofic | Royal |
| First Person | ma | | | | hami |
| Second person | ta | timi | tapai | hazur | mousuf |
| Third person | tyo | tini | uha | uha | mousuf |
| | u | uni | yaha | yaha | |
| | yo | yini | | | |
| | ti | | | | |
| | yi | | | | |

Table 2. Different levels of pronouns used in Nepalese, Japanese and English

| Nepalese | Japanese | English |
|---|---|---|
| ma | 私、僕、俺 | I |
| ta, timi, tapai, hazur, mousuf | あなた、君、お前 | you |
| tyo, uha, u, yaha | 彼 | he |
| tini, uni, yini, yi | 彼女 | she |

Note : The Nepalese words in the last line of the table refers to not only the feminine words, they are also used for addressing the masculine gender in some cases.

## 4. Translation Process Utilizing XML

The translation process basically requires the database of the words with their semantic meanings, database of the sentence structure, and the systems (rules) to analyze the sentence (syntactically and semantically), and to the generation of the sentence. This proposal is based on the concept of creating the database and the rules which are understandable to the XML so that it is easier to analyze the source sentence structure, to retrieve meaning and then again to combine the words and create the target language.

Much simply it is the system in which the grammar itself is made understandable using XML so that much of the ambiguity in analyzing the sentences is not faced. The sentence is always understood grammatically and only the semantic meanings need to be analyzed. Different databases are used for explaining grammar. When a source sentence is given to the system, the system starts breaking it into different parts. The suitable translations for the broken parts are then checked with the DTDs and then the translations for the different parts are done and using those words possible sentence structures are generated. The appropriate target sentence is then selected and the output is made. XML uses the Unicode (ISO standard) so that it is easy to handle different languages within one system. It is easy for the interface to understand and process the different languages within the one system, which is one of the important aspects of the multi-lingual translation.

First, the text for the translation is feed into the system. The text might be pre-coded with XML or it might not be. Before analyzing the text, the type of the document is declared. That is, whether the document is of technical matter, a poetry, a fiction or etc is noted. This notation helps the system to look for the suitable dictionaries to find out the appropriate words when the sentence is generated in the target language. However this declaration must be done manually if the document is not pre-coded with XML.

The document (sentence) thus encoded is then broken into smaller components. That is, a sentence is broken into the words, which is done by recognizing the white space between the

words in the languages where there is presence of white space such as English. For the languages where white spaces are not present a morphemic analysis is done. The word structure thus changed is now linked to the DTD where they are checked with the available database, and each word is taken for the translation process. Flow chart of translation process is shown in Fig.4.

Basically the translation process is carried out in two phases; analysis and generation. In both phases knowledge of present translation process is used. But the sentence analysis is carried out using the DTDs so that the process is carried out fast. It will reduce the time taken in analyzing a sentence.

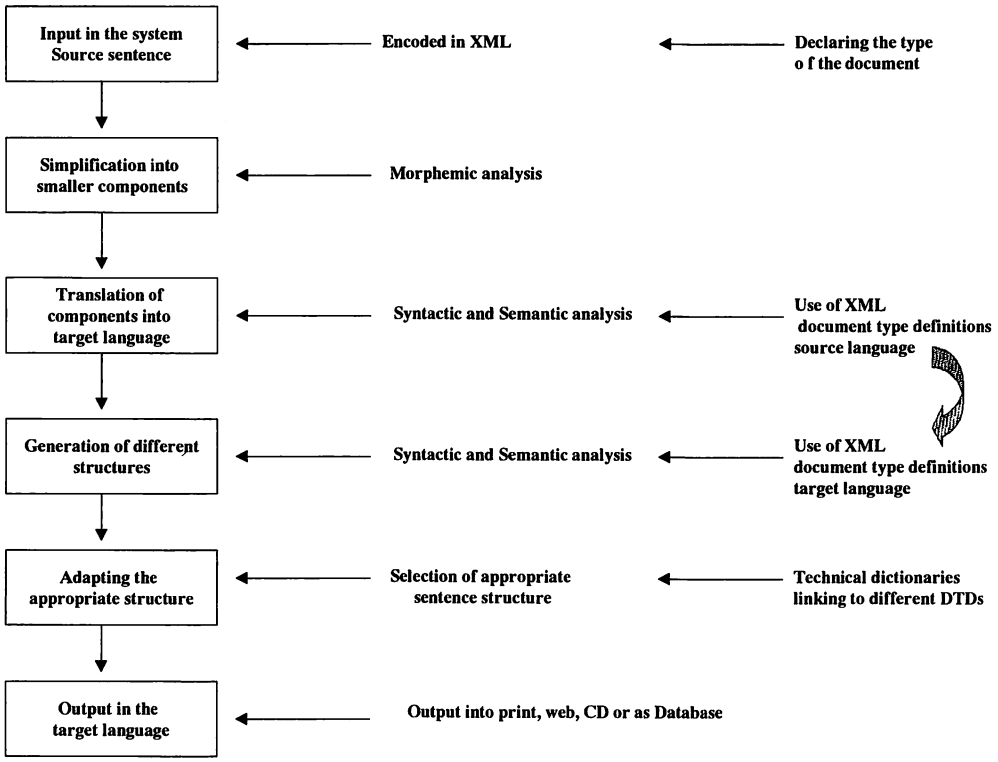


Fig.4. Flow chart of the translation process utilizing XML.

## 4. 1 Translation of the Words

For example, take a sample sentence,

I HAVE A GREAT GOLDEN ELEPHANT

The parsing analysis of the above sentence is done as follows.

The sentence is first broken into words and each word is mapped with the corresponding DTD and the analysis of different levels of the sentence is analyzed.

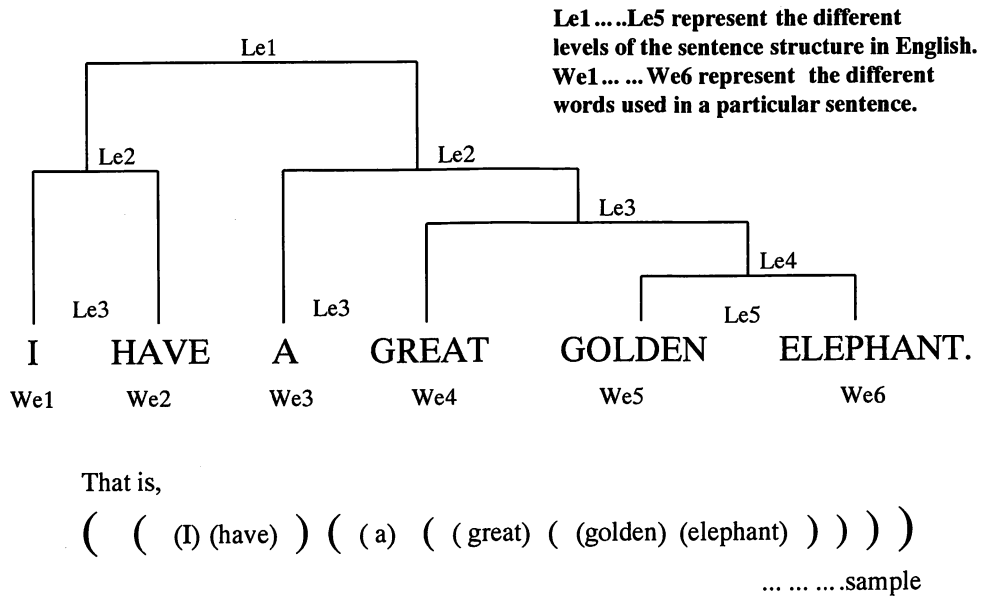


Fig.5. Parsing analysis of a simple English Sentence.

Let's consider the different Levels of an English sentence "Le" as Le1, Le2...Le5 and let's take the different English words be We1, We2,...,We6 as shown in Fig.5.

The broken words of a sentence are then verified with the DTDe (DTD of English), which contains the main grammatical structure of the sentence. Each word in a sentence falls in a particular element such as noun, pronoun, adjective or verb etc and then again into the attributes that the element carry. For example, when a word "I" is encountered, the program checks its validity passing through the chain of elements, attributes and finally conforms it as subject of first person of singular personal pronoun as shown in Fig.6.

*<PRONOUN><PERSONAL>*

*<SINGULAR><FIRST-PERSON><SUBJECT>I</SUBJECT>.....</PRONOUN>.*

Similarly,

*<VERB>.....Have.....</VERB>*

*<ARTICLE>.....A.....</ARTICLE>*

*<NOUN>.....ELEPHANT.....</NOUN>.*

Defining "I" as above will give the advantage of searching the same word in target language. A search of first person of singular personal pronoun or the pronoun equivalent to that of English "I" is carried out.

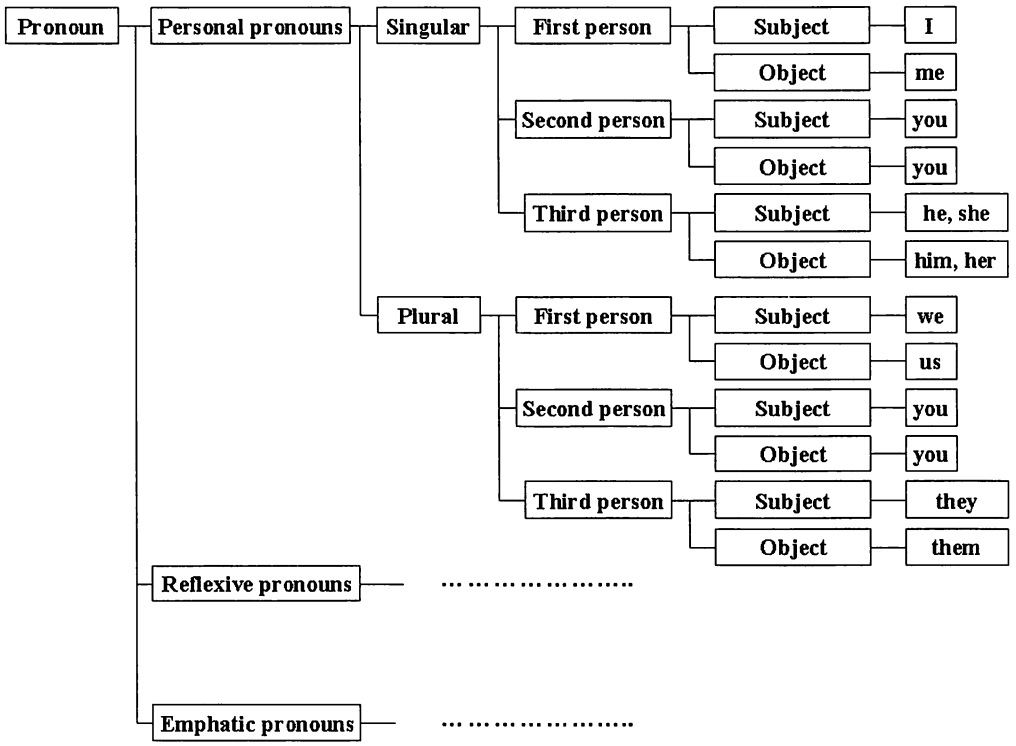


Fig.6. Tree structure of pronoun element.

A search is conducted mapping with DTDn (DTD of Nepalese) and DTDj (DTD of Japanese). In Nepalese the output is given as "Ma", and in Japanese the output may be given as "私" ,"僕" and "俺".

For example, DTDn gives the result as

*<PRONOUN>.....MA, HAMI.....</PRONOUN>.*

DTDj gives the result as

*<PRONOUN>.....私、僕、俺.....</PRONOUN>.*

Similarly the other words "HAVE", "A", "GREAT" "GOLDEN", "ELEPHANT" are also mapped with the consecutive DTDs of the target languages and similar outputs are obtained.

Thus verified words are then interchanged into the target language by selecting the same element of the target language tree as follows.

English: I HAVE A GREAT GOLDEN ELEPHANT

Nepalese: MA SANGA EUTA THULO SUNOULA HATTI CHA.

Japanese: 私は一つ大きい金の象を持っている。

But as the different languages do not have the same kind of language tree structure, the nearest possible element is selected and the interchange is performed. To understand the nearest possible element various syntactic and semantic analysis are done.

### 4. 2. Generation of Language Structure

The translated words are then again analyzed to obtain the certain language structures. The tree structures of the target language, where the translated forms of the words fall, are once again grouped together to create the sentence. The levels obtained during the analysis process are referred in the generation process. As there might be different tree structures, a group of sentences with different structures could be generated. Choosing the appropriate structure from different structures is a complicated process that involves thousands of rules. The DTD containing the target language structure gives the possible structural transformation only. If the documents are pre-declared with their types, it becomes very easy to select the DTD of the same declaration in the target language. Once again the different types of syntactic and semantic analysis are carried out to check the grammatically correct sentence structure. The selected structure is once again checked for different fixed terms, which might be present in certain language culture. If the presences of those terms are detected they are changed into the commonly used one.

For these types of checks various sets of database are required. The generation process always needs to utilize these databases in every step of the translation process. The dictionaries are written with XML and linked with processing parsers so that the translating process is smoothly carried out.

Finally the sentence generated in a target language is decoded into a normal text. Or depending upon the nature of the processing environment and the output obtained, the translated text might be left with the XML coding, so that future translations are carried out easily. Finally the translated text is programmed and distributed for the various uses such as printing, using as a database, or on the web. Once the text is marked with XML it will be fairly easy to use it on the different interfaces. Using the XML content management system the documents can be preserved with up-to date record on its uses

## 5. Conclusion

For the translation process, the most basic concept is still the understanding of the language structure and analyzing it. Computers are still brain-less machines that are unable to function without pre-programming. XML here is used as a language which efficiently describes the data structure of both the source and the target language. The sentences are

changed into smaller units and then verified with the DTDs. As this paper is only a proposal, the verification of the system has not been done practically. The DTDs provide the unique format of creating a database, which can be combined, linked and reused using the network. The success of the system will assure the automatic translation process that is not biased in the program or the place. Still, not enough research on the DTDs or creation of the program to test even a small part of the system has been performed. The actual works on the DTDs and the system itself is kept as the future research project.

The bilingual translating system independent of the solid source will ensure the translation also on the web. Once the system is created completely, it can be used to create the multi-language platform, which recognizes the source language and translates to the target language. Still, the most evident problem would be the language structures of different languages. The DTDs are flexible in a sense that they can be declared freely and without a standard. But to create an effective platform language structure a huge amount of research work is needed to be done by linguists and technicians of different linguistic backgrounds. If the creation of the flexible rules (*Unirules*), which could be molded to create the common platform structure, were achieved, then the era of multi-language translating system would come very close. On the initial phase documents are tagged manually, but the system ensures the output automatically tagged, which reduces the massive processing power and manual labor during the re-translation or post translation processes. The tagging systems could be developed so that the semi-automatic or full-automatic tagging could be done.

As the research work was totally based on assumptions, the conclusion drawn was tentative. Primary research works on the Nepalese language sentence structure is the main task to be followed to propose the flow chart of the DTD structure. The success of the system would assure the automatic translation process, which is not biased on the program or the interface and would be suited for the Internet applications.

The authors would like to express their gratitude to Prof. Shinichi Ueshima and Prof. Masatoshi Yoshikawa for their invaluable suggestions and help.

References

1. The XML Handbook, Charles F. Goldfarb and Paul Prescod, Prentice Hall PTR, 1998.
2. Machine Translation, Makoto Nagao, Oxford University Press, 1989.
3. XMLを知る，リチャード・ライト、プレンティスホール,1998 .
4. XML and Java, Hiroshi Murayama and Kent Tamura and Naohiko Uramoto, Addison-Wesley, 1999.

5. Inside XML DTDs, Simon St. Laurent and Robert Biggar, McGraw-Hill, 1999.
6. Nepali Wakya Wyakaran, Dr. Madhav Prashad Pokhrel, Royal Nepal Academy, 1996.
7. Beginning Nepali, Dr. Tara Nath Sharma, Sajha Prakashan, 1983.
8. Sabdarachana Ra Varnabinyas, Mohan Raj Sharma, Navin Prakashan, 1997.
9. W3C Architecture domain Activity statement (http://www.w3.org/XML/Activity.html#intro), 1999.