AnglaMalayalam: English to Malayalam MT system based on AnglaMT Paradigm

Abstract

AnglaMalayalam is an English to Malayalam Machine translation system developed by C-DAC Thiruvananthapuram. It is a customized version of AnglaBharati Technology developed by IIT Kanpur. The Anglabharati system is primarily aimed at the translation for the Indo-Aryan language family. It is the first attempt to develop an E-IL MT system for Dravidian languages using AnglaBharati technology. Malayalam is the language chosen for the case study. The customization is found successful and is standing equally among the Indo-Aryan language family. The performance is also comparable. Indo-Aryan and Dravidian language families are having similar sentence structures with a few exceptions. The main tasks involved in the customization process are developing the text generator module and bilingual dictionary creation. Some target language dependent codes also have to be modified in other modules to incorporate Malayalam as a target language.

1 Introduction

English is a highly positional languag with rudimentary morphology and default sentence structure as SVO. Indian languages are highly inflectional, with a rich morphology, relatively free word order and default sentence structure as SOV. In addition, there are many stylistic differences. For example, it is common to see very long sentences in English, using abstract concepts as the subjects of sentences and stringing several clauses together. Such construc ions are not natural in Indian languages, and present major difficulties in producing good translations.

As is recognized the world over, with the current state of art in MT, it is not possible to have Fully Automatic, High Quality, and General-Purpose Machine Translation. Practical systems need to handle ambiguity and the other complexities of natural language processing by relaxing one or more of the above dimensions.

Thus, we can have automatic high-quality 'sub-language' systems for specific domains, or automatic general-purpose systems giving rough translation, or interactive general-purpose systems with pre or post editing. Indian MT systems have also adopted one of these strategies, as we will see.

1.1 AnglaMalayalam

AnglaMalayalam is an adaptation of Angla Bharati technology to English-Malayalam Machine Aided Translation system. Angla Bharati Technology makes use of the Interlingua approach for translation. This process has source language analysis and target language generation. The interface between these two components is an intermediate language (PLIL) called the interlingua. It is a language independent, unambiguous representation of the meaning of the input text that has to fulfill a simple functional condition: the interlingua representation must be sufficient for accurate translation in a technical domain. From this PLIL, using the text generator the PLIL is converted in to the target language i.e. Malayalam. The whole process can be summarized as shown in the figure: 1

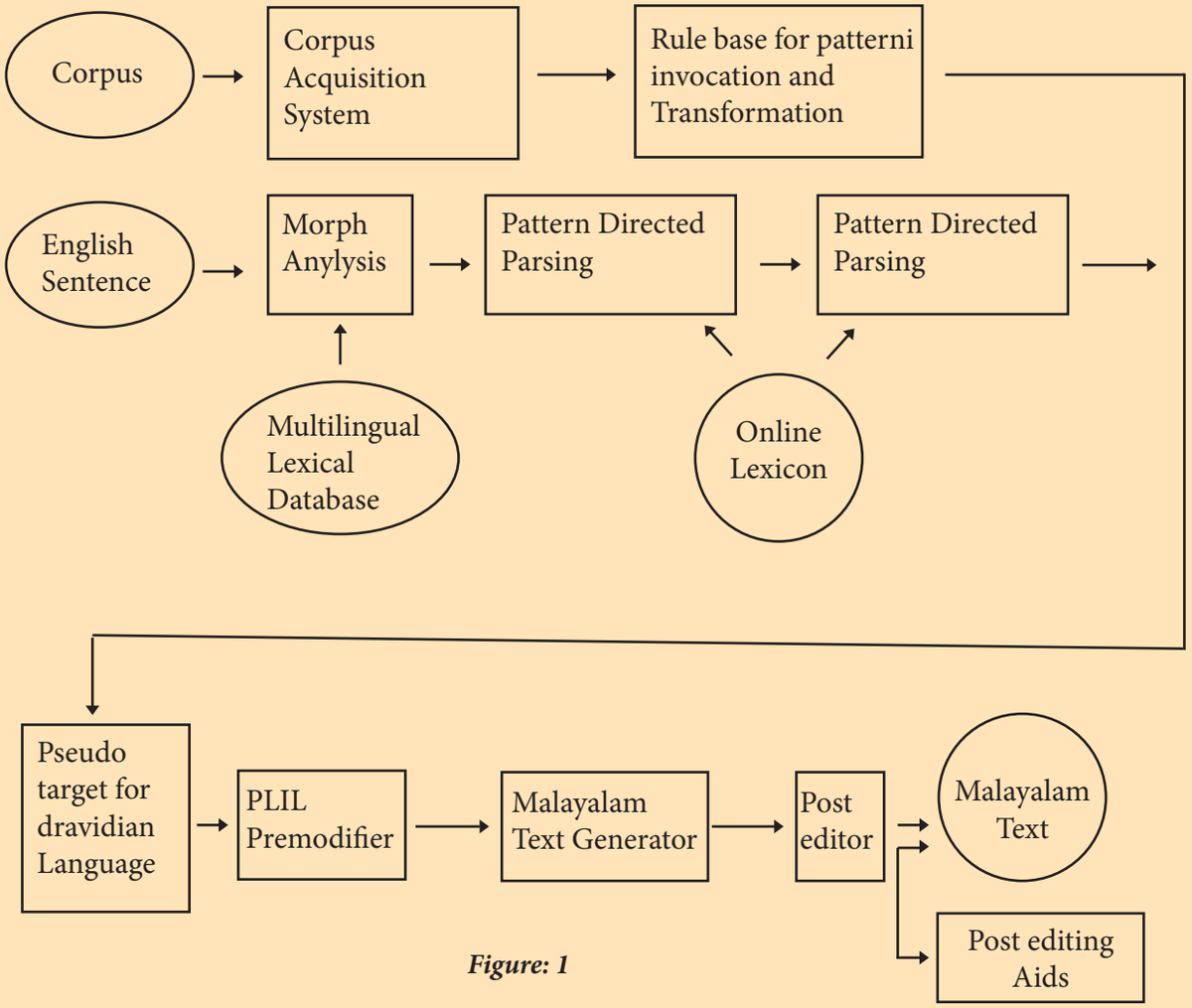

Figure: 1

The Morphological analyzer takes the English sentence as input, identifies the proper root words, and retrieves the necessary information from the lexical database.

There are two primary data-bases used by our system. One is the multi-lingual lexical database and the other is the rule-base for pattern invocation and transformation from English to the pseudo target. The multi-lingual lexical database, besides holding information about the meanings of the lexicons, also carries information on their syntactic and semantic features. In case of multiple meanings, certain patterns and constraints for disambiguation are also stored. Only root words are stored in the lexical database. A pattern directed parsing is performed on the source language, English. Here, the words of the input sentence after undergoing morphological analysis are used to form patterns. These patterns are matched to the left-hand side of the rules stored in the rule-base. On finding the match, the corresponding rule is invoked, and the right-hand side of the rule yields the pseudo target. Multiple invocations

of rules are also possible in case of multiple patterns grouping at the source level. In such a case, more than one pseudo target is generated, and post-editing is required.

Pseudo target is used for the synthesis of the target language. A number of grammatical rules of the target language in the form of expectations and constraints are used to select the appropriate case markers, affixes, etc. A Paninian framework is used for this purpose. The ambiguities which still remain unresolved are taken care by human post editing.

2. Customization of AnglaBharati System

Adapting AnglaBharti to another Indian Language involves:

1. Entry of Target Language meaning into the existing Lexical Database
2. Customization of Engine for getting TL info in PLIL
3. Development of Target Language Text generator

The customization requires the changes that are language dependent in every modules of the system. For Malayalam we developed a new text generator. The Transliteration module in the Morph Analyzer module should also be replaced by a new one for handling English words not present in the Lexicon. A set of rules are generated for transliteration. The main tasks that are performed for the customization is detailed below:

2.1 Entry of Target Language Meaning

The same lexical database used in the Angla Bharati system can be used for Malayalam, by updating the database with the Malayalam meanings and related information of the English words in the lexicon. For the target language, separate language rules have to be generated to define the inflections in a particular semantic category like noun, verb, etc. For example, Hindi verbs are gender dependent, but Malayalam verbs are not. So the gender information is not needed for the Malayalam target language rules.

2.2 PLIL Premodifier

Malayalam belongs to one of the four major languages of the Dravidian family. Morphologically, Malayalam is richly inflected by the addition of suffixes to the root/stem word and is agglutinative in nature. Malayalam uses a different word order as compared to English. In some case the structure of the Indo-Aryan language family and that of Dravidian language family will be different. Sample sentences and their translations are given below (a, b). In some cases the sentence structure an Indo-Aryan language is similar with that of English. But for Malayalam the position of relative clause and the noun got interchanged. So the PLIL generated by the Angla Bharati engine needs to be modified. Hence we use a separate module named PLIL Premodifier for AnglaMalayalam system. Consider the given example for Hindi that is having different sentence structure in the case of relative clauses.

(a)

English: The girl who is standing near the window is my sister.

Hindi: vaha ladZakI jo KidZakI ke pAsa KadZI hE merI bahana hE.
(the) (girl) (who) (window) (near) (standing) (my) (sister) (be)
Malayalam: janalinatuww nikkunna A peN_kutti enZe sahOxari AN.
(window near) (standing) (who) (girl) (my) (sister) (be)
(b)
English: I like to know what the enemy is thinking.
Hindi: mEM jAnanA cAhwa hU jo Sawru kyA soca rahA hE.
(I) (know) (like to) (what) (enemy) (thinking)
Malayalam: Sawru cinwiykkunnaw FAn_ aZiyAn_ iRtappetunnu.
(enemy) (thinking what) (I) (know) (like)

2.3 Malayalam Text generator

Text generator module is subdivided in to the following subdivisions based on the language rules to be implemented for translation.

2.3.1 Disambiguation of postpositions

Postpositions in Malayalam are certain forms, which occur immediately after nouns and establish some grammatical relations between the nouns and the verbs of sentences. Prepositions in English language will be transformed to post positions in Malayalam. The semantic distribution of a single preposition will be varying in different context due to the influence of nouns and main verbs that follow. Some of the prepositions in English can be directly translated and attached to the Malayalam nouns as postpositions, while some need special rules.

(a) I am going to Kanpur
FAn_ kAN_pUrilEkk pOkunnu
(I) (Kanpur to) (going)

(b) I have spoken to him already.
FAn_ iwinakaM wanne avanOt saMsAric cittuNt.
(I) (already) (he) (speak)

(c) Please listen to him.
xayavAyi avane SraxXiykkU.
(Please) (him) (listen)

(d) I am going to the king.
FAn_ rAjAvinZe atuwwEkk pOkukayAN.
(I) (king) (to) (going)

In all the four cases the preposition 'to' have different formations in Malayalam as postposition. In (a) the 'to' has the suffix form 'Ekk' and is attached to the noun. In the second case 'to' has the suffix form 'Ot' and combined with the pronoun. In all the four cases it is having different suffix formation. Thus disambiguation of such preposition requires extensive study to extract rules. Similarly difficulties are present with other prepositions also. In some cases the postpositions come as a suffix and in some others as independent words as in (d). The postposition 'atuwwEkk' also have another meaning 'near'.

2.3.2 Sandhi Formation

The general meaning of the term sandhi is union. We can group sandhi into vowel sandhi, vowel consonant sandhi, consonant vowel sandhi and consonant sandhi.

Vowel Sandhi (Vowel+Vowel)

In this case when the two vowels 'a' and 'a' combine together to generate 'y'
mala + alla => malayalla

Vowel Consonant Sandhi (Vowel+Consonant)

In this case the vowel 'a' and consonant 'k'

combined together and the consonant 'k get geminated

tAmara + kuLaM => tAmarakkuLaM

Consonant Vowel Sandhi (Consonant+Vowel)

Here the virama(chandrakala) of the consonant is deleted and combined with the vowel of the second word

kaNN + illa => kaNNilla

Consonant Sandhi (Consonant + Consonant)

The chillu 'l_' is replaced with 'n_' to generate sandhi form here.

nel_ + maNi => nen_maNi

The Sandhi can also be classified on the basis of lopa (elision), agama (addition), dvitva (gemination) and adesha (displacement).

Lopa Sandhi

In this case one character is deleted during sandhi formation. In the given example the short vowel at the end of the first word is deleted.

minnunna+AkASaM => minnunnA-kASaM

Agama Sandhi

This sandhi has the property of generation of a new character during sandhi formation. Here the consonant 'v' is generated while combining the vowels 'u' and 'O'

tiru+ONaM => tiruvONaM

Dvitva Sandhi

During the sandhi formation one of the character geminates then it is Dvitva sandhi. The character 'p' is geminated in the example

avite + pOyi => aviteppOyi

Adesha Sandhi

In this case one character is replaced with another character while sandhi formation. The character 'l' is replaced with 'n_'

nel + maNi => nen_maNi

2.3.3 Lexical choice for Adjectives

Adjectives in Malayalam language are generally derived from either noun or verb by the process of suffixation. Relative participles derived from the verbal stem are widely used as adjectives. Most adjectives can occur both before and after a noun. Let us consider an example," This is my new house", "iw enZe puwiya vIt AN". Here, adjective (new) is called ATTRIBUTIVE adjective. But in another example "My house is new', the adjective "new" is occurring after the head noun "house" and termed as PREDICATIVE

Root	Suffix	Type	Gen	Num	Form
prasixXa	nAya	ATR	M	sin	prasixXanAya
prasixXa	yAya	ATR	F	sin	prasixXayAya
prasixXa	n_mAraya	ATR	M	plr	prasixXan_mArAya
prasixXa	kaLAya	ATR	F	plr	prasixXakaLaya
prasixXa	mAya	ATR	N,D	sin, plr	prasixXamAya
prasixXa	rAya	ATR	N,D	plr	prasixXarAya
prasixXa	n_	PRED	M	sin	prasixXan_
prasixXa	-	PRED	F	sin	prasixXa
prasixXa	n_mAZ_	PRED	M	plr	prasixXan_mAZ_
prasixXa	kaL_	PRED	F	plr	prasixXakaL_
prasixXa	M	PRED	N,D	sin, plr	prasixXaM
prasixXa	Z_	PRED	N,D	plr	prasixXaZ_

adjective. Notice that predicative adjectives in English do not occur immediately after the noun. Instead, they follow a verb. The same form of adjective, for e.g. "new" in English, can act as predicative and attributive adjective with out changing its regular form, where as in the case of Malayalam language the adjective "puwiya"(meaning new) will have another form "puwiyaw" in the second example. "enZe vIt puwiyaw AN". Most of the adjectival forms in Malayalam will alter due to the influence of its gender and number. The semantic information of head noun and subject will have a major role in the formation of adjectives in Malayalam. The adjectival forms based on the gender and number information for the adjective 'prasixXa' is given below:

2.3.4 Generation of verb forms, Gerunds and Participle

The verb morphology in Malayalam is somewhat complex in nature.

E.g: a) eYuwunnu (write + present continuous)

b) eYuwikkontirikkunnu (write + imperfect + be +pres. Cont.)

Verbs in Malayalam were classified according to the morpho-phonemic changes occurring due to the inflection. Various scholars attempted to explain Malayalam verb morphology in detail. Based on the works by Suranad Kunjan Pillai (SKP) in the appendix of the first volume of Malayalam lexicon we had classified the verbs. A total of 54 verb categories were derived based on the morphophonemic changes occurring as per to tense aspect and modality. The SKP classi -fication is based on the verb ending in its past form. He generated a total of 12 classes. That may not be sufficient for the computational purpose. So we have derived the new classes of verbs out of it and both classifications are mutually exhaustive. Given below is the first verb class based on different verb formation according to tense aspect and modality.

Morphotactics for the root word 'uY' is given as example.

a) Three tense forms

Root Word(RW) + unnu(present), RW + uM(future), RW + u+ wu(past).

b) Causative (Single and Double)

RW + uvi + kk + unnu/uM or RW + uvi + ccu.

RW + uvi+ppi+kk+ unnu/uM/ccu or RW + uvi+ppi+ccu.

c) Verbal Noun →RW + al_

d) Infinitive →RW + An_/ uv + An_

e) Niyojaka Prakaram →RW + atte / uv+in_/ uw+AluM

f) Vidhayaka Prakaram → RW + aNaM

g) Anujnyayaka Prakaram → RW + AM

h) Perecham →(pr) RW + unna , (pa) RW + uwa.

i) Vinayecham → (M) RW + uwa , (N) RW + uka, (P) RW + An_ , (T) RW + ave, (Pa) RW + u + k+ il_

2.3.5 Disambiguation of to-infinitive

The "to-infinitive" in English has multiple mapping in Indian Language. The mapping of suffix for to-infinitive in Malayalam differs according to the context. Eg: (a) It is the pen to write a poem.

iw kaviwa eyuwAnuLLa pEnayAN.
(It poem write to-inf pen)

(b) He is willing to resign.

avan_ rAjivaykkAn_ wayyAZAN

(He resign to-inf willing)

In the examples (a) and (b) the 'to' have different mapping, in the first case it is 'AnuLLa/ uvAnuLLa' and in the second case it is 'An_/ uvAn_'. There are six type of resolution of suffixes are there. From this we have derived 13 rules for handling the to-infinitives. The type 1 suffixes and the derived rules are explained below:

•If the verb has

Pattern Type as "non_finite"

Auxiliary as "to"

Verb Type as "VERB_1" then,

Note: V E R B _ 1 belongs to present singular(Eg:eat, play,etc)

Rule 1: If the verb comes under TOINF phrase in PLIL, infinite verb forms as

Verb Root + Link Morph (if any) + An_

(9) I like to play football.

Fan kAl_panw kaLikkAn_ (kaLikkuvAn_) IRtappetunnu.

(I football play+to Like)

(with Link Morph "ikk")

(10) I like to sing a song.

FAn_ oru pAtt pAtAn_ (pAtuvAn_) IR tappetunnu.

(I a song sing+to like)

(Without Link Morph)

Note: Both the suffixes An_ and uvAn_ are well matched for the corresponding pattern type, Auxiliary and verb type

2.3.6 Noun Formation

The inflection of noun is based on the gender, number and case. For Malayalam we have seven cases. So totally we have 14 noun formations by incorporating the number. The noun formation for the word 'kutti' is given in the table. While word formation based on gender, number and case the suffixes will differ for different word endings. The plural form of 'maraM' is 'marffaL_' and is not similar with 'kuttikaL_'. So they have been put in to different class. The noun formation is handled in this manner.

case	Affix	singular	plural
Nominative	Φ	kutti	KuttikaL_
Accusative	*e*	*kuttiy**e***	*KuttikaL**e***
Sociative	*Ot*	*k*uttiy**Ot**	*KuttikaL**Ot***
Dative	*kk,n*	*kutti**kk***	*KuttikaL_**kk***
Instrumental	*Al_*	*kuttiy**Al_***	*KuttikaL**Al_***
Gentive	*ute,nZe*	*kuttiy**ute***	*KuttikaL**ute***
Locative	*il_*	*kuttiy**il_***	*KuttikaLi**l_***

2.3.7 Pronoun Formation

Pronoun is a kind of noun, but its function is different from noun. A pronoun is a word which refers to a person or a thing that has already talked about. Pronouns can take the place of a noun in a sentence and function as a noun. There are 18 pronouns in Malayalam. The pronouns in Malayalam are classified in to eight paradigm

classes for computation. The formation of pronoun 'FAn_' according to different cases are given below:

case	Affix	
Nominative	Φ	fAn_
Accusative	e	enne
Sociative	Ot	ennOt
Dative	kk,n	enikk
Instrumental	Al_	ennAl_
Gentive	ute,nZe	enZe
Locative	il_	ennil_

2.4 English-Malayalam Transliteration

This module is a part of the Morph Analyzer. During a dictionary search, if a word is not found in the dictionary then the flow of the system is directed to the transliteration module for the transliteration of the unknown word. The 'transliterate' module transliterates the unknown acronyms, named entities, etc. that are not found in the lexical database/stored tables into the roman notation assigned to Malayalam. The English characters in these words are converted into Roman notations which helps unambiguous representation of these words while converting to Malayalam. For example each of the combinations **ee, ii, ei, ie, ea** will be transliterated to ഈ in Malayalam. Hence all these combinations will be converted to its extended wX notation "**I**". All words with capital letters will be treated as acronyms and are processed separately. For example IEEE will be converted to ഐ. ഇ. ഇ. ഇ in Malayalam instead of ഇ ഈ, the expected result as per this scheme. The Transliteration system developed has an accuracy rate of 70%.

Examples:
DYFI→ di.vQ.eP.Q
Keith→ kIww

The block schematic of the Transliteration system is given in the figure 2:

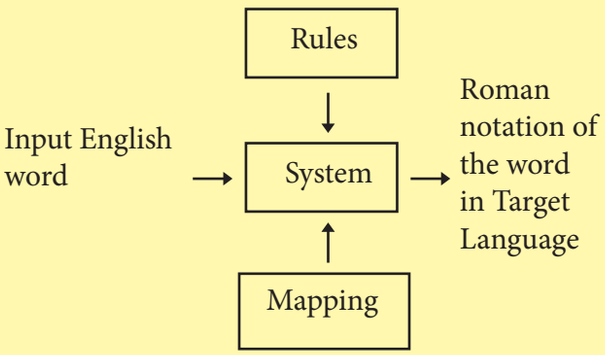

figure 2:

The rules are generated as an exception to the transliteration scheme generated. For example, in the case of 'l' normally we transliterate it as 'l' (ല) in Malayalam when it is inside the word. When it is at the end then it will be either 'l_'(ല്) or 'L_'(ള്). Similarly for'll', the word will geminate in the middle of the word and in the end of a word it will come as 'l_' (ല്). Like wise so many exceptions can be found for all the English letters. Finding exceptions and generating the rule accordingly will improve the transliteration accuracy. There are some limitations to this also. The name 'sasi' is pronounced as 'SaSi'(ശശി instead of സസി), so in such cases the rules fail and the only thing we can do is to add such words to the lexicon.

Conclusion

We realized the Machine Translation system for Malayalam, a language of the Dravidian family, using the AnglaBharati technology, with accuracy comparable to that of an Indo-Aryan language family. The system yields good accuracy

for simple sentences. More research is needed to improve the system performance with complex sentences. Presently we find the bottle necks in the form of named entities, noun phrases, verb phrases, etc. We were able to resolve the problem to some extend. After the first phase of the development we found that extensive research is needed in the area of language modeling to reduce the number of alternate translations and the reordering of the correct translations in the order of their acceptability. The speed of the system is also another concern. During the development phase we redesigned the language rules for computational purposes. All the rules developed can be used for other language computing applications in general. In a nutshell, the overall gain after the development of the English Malayalam Machine Translation system found to be landmark in the area of Natural Language Processing.

References

1. R.M.K. Sinha, An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures, Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004), November 17-19, 2004, Tata Mc Graw Hill, New Delhi.
2. R. M. K. Sinha, Machine Translation: AnglaBharati and AnuBharati Approaches, Communications of CSI, October 2005.
3. R.M.K. Sinha, A Pseudo Lingua for Indian Languages (PLIL) for Translation from English. Technical Report, Language Technology Lab, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur (2004)
4. R.E. Asher, T.C. Kumari, Malayalam - Descriptive grammars, 317-319, Routledge(2007)
5. A R Rajaraja Varma, Keralapaniniyam, Pages 177-269, DC Books(2000)
6. Suranad Kunjan Pillai, Volume 1, Malayalam Lexicon, Appendix(1-105), Introduction xviii The University of Kerala(2000)

Annexure

Annexure 1 : Screenshots the AnglaMalayalam MAT system.

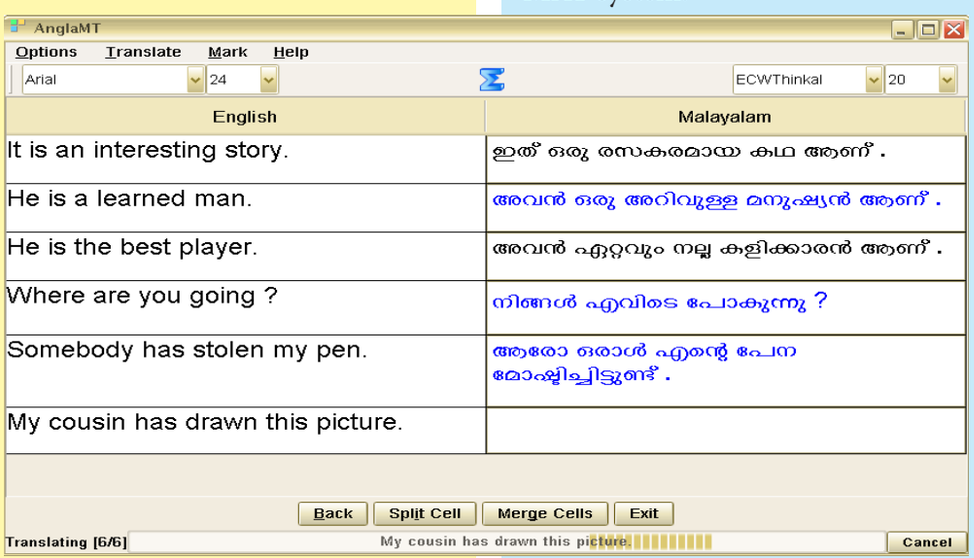

Figure1 : Sentence Translation in Desktop Version of the System.

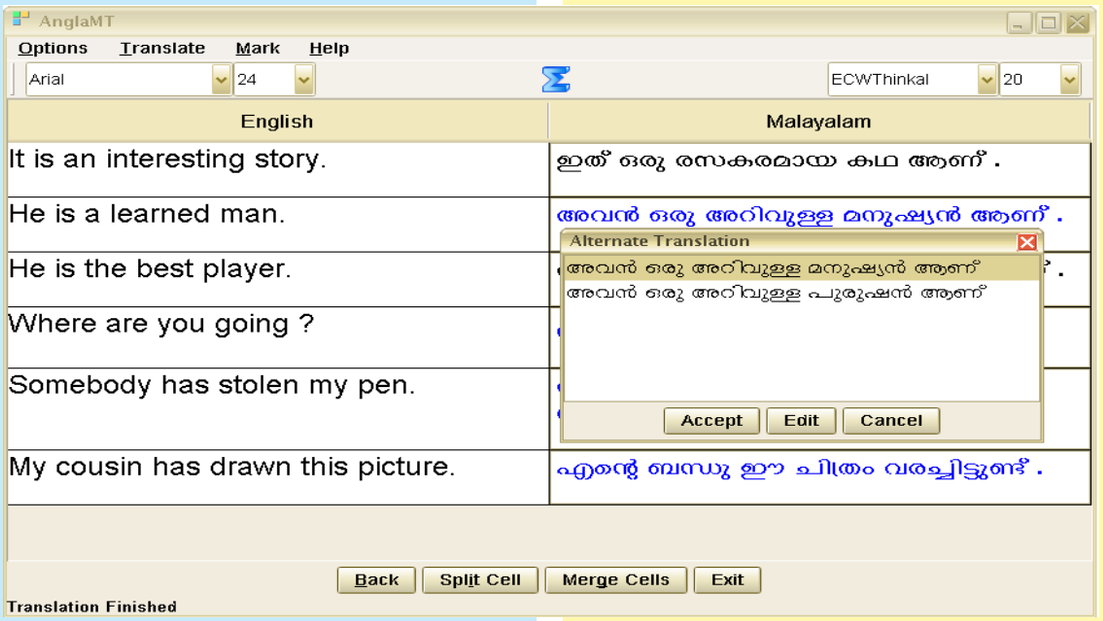

Figure 2: Alternate translations of the outputs

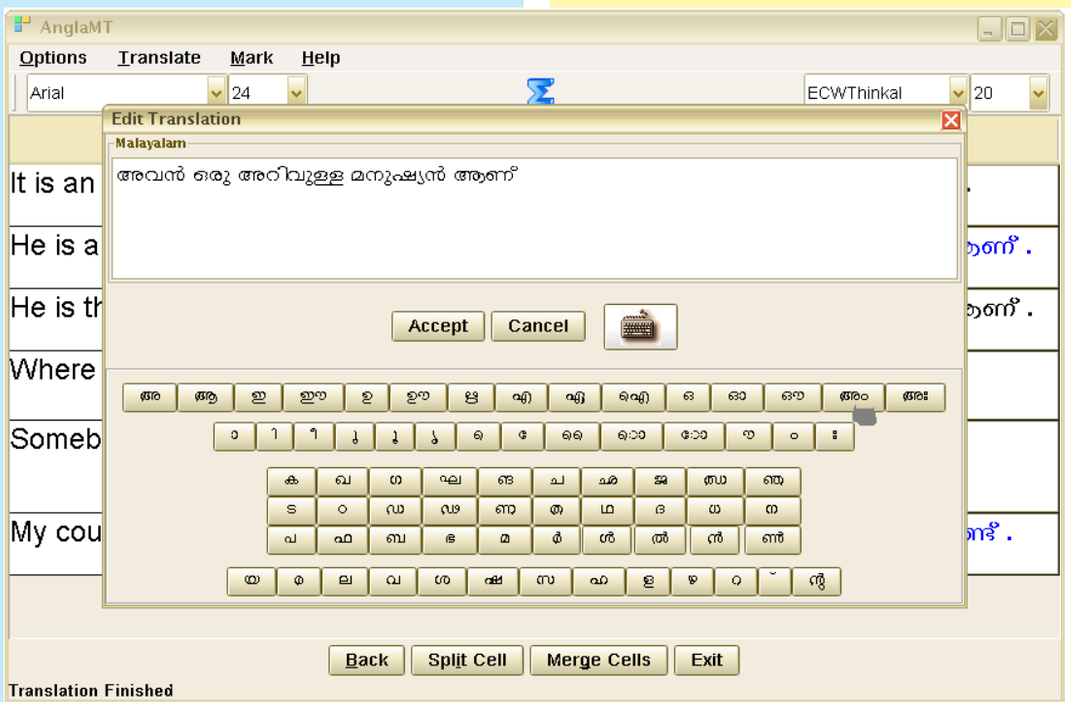

Figure 3: Post editing facility of the translated sentence

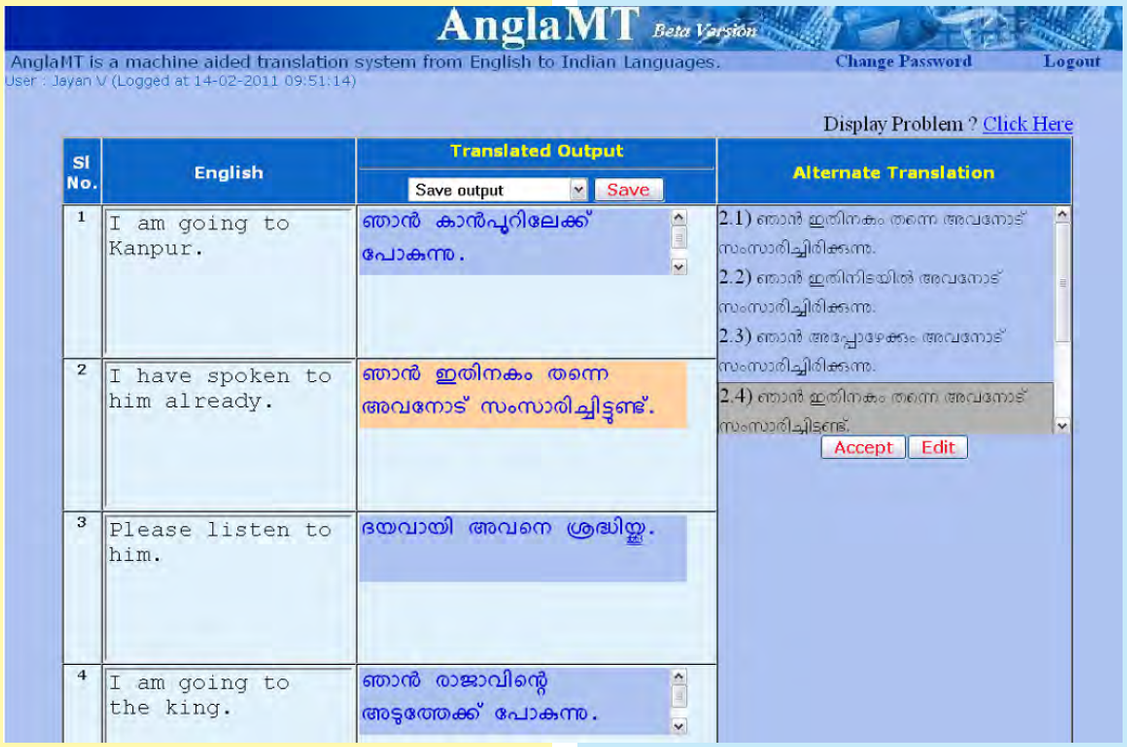

Figure 4: Web Version of the Item

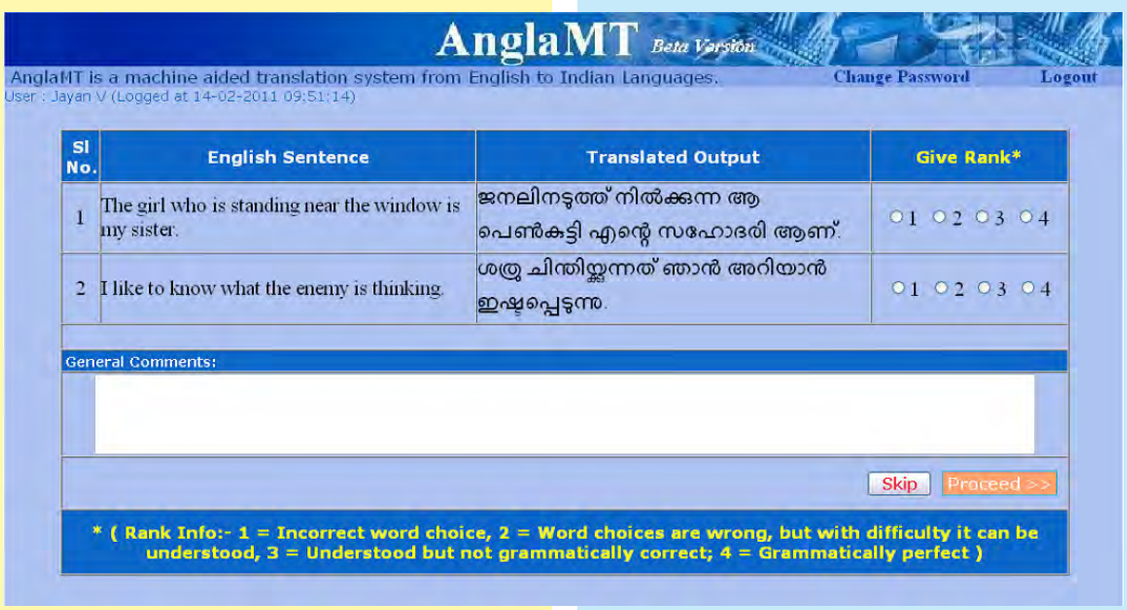

Figure 5: Sentence ranking facility in the web version of the system

Unicode Glyph	Unicode Value in Hex	ASCII Character	ASCII Value in Hex
A	0D05	a	61
B	0D06	A	41
C	0D07	i	69
Cu	0D08	I	49
D	0D09	u	75
Du	0D0A	U	55
E	0D0B	q	71
F	0D0E	e	65
G	0D0F	E	45
sF	0D10	Q	51
H	0D12	o	6F
Hm	0D13	O	4F
Hu	0D14	V	56
Aw	0D02	M	4D
Ax	0D03	H	48
I	0D15	k	6B
J	0D16	K	4B
K	0D17	g	67
L	0D18	G	47
M	0D19	f	66
N	0D1A	c	63
O	0D1B	C	43
P	0D1C	j	6A
Q	0D1D	J	4A
R	0D1E	F	46
S	0D1F	t	74
T	0D20	T	54
U	0D21	d	64
V	0D22	D	44
W	0D23	N	4E
X	0D24	w	77
Y	0D25	W	57
Z	0D26	x	78
[	0D27	X	58
\	0D28	n	6E
]	0D2A	p	70
^	0D2B	P	50

_	0D2C	b	62
`	0D2D	B	42
a	0D2E	m	6D
b	0D2F	y	59
c	0D30	r	72
e	0D32	l	6C
h	0D35	v	76
i	0D36	S	53
j	0D37	R	52
k	0D38	s	73
l	0D39	h	68
f	0D33	L	4C
g	0D34	Y	59
d	0D31	Z	5A
³	0D28 + 0D4D + 200D	n_	6E + 5F
Ä	0D33 + 0D4D + 200D	L_	4C + 5F
¬	0D23 + 0D4D + 200D	N_	4E + 5F
À	0D30 + 0D4D + 200D	Z_	5A + 5F
ð	0D32 + 0D4D + 200D	l_	6C + 5F
	0D3E	A	41
ാ	0D3F	i	69
ീ	0D40	I	49
ു	0D41	u	75
ൂ	0D42	U	55
ൃ	0D43	q	71
െ	0D46	e	65
േ	0D47	E	45
ൈ	0D48	Q	51
ൊ	0D4A	o	6F
ോ	0D4B	O	4F
ൌ	0D4C	V	56