EXISTING ENGLISH TO KHASI TRANSLATED DOCUMENTS FOR PARALLEL CORPORA DEVELOPMENT: A SURVEY

Aiusha Vellintihun Hujon[1] and Thoudam Doren Singh[2]

[1]Department of Computer Science, St. Anthony's College Shillong, Meghalaya-793001, India

[2]Department of Computer Science and Engineering, IIIT Manipur, Imphal-795002, India

ABSTRACT

Importance of translation has been realized long way back, but mostly it was manual translation. Translating a particular language to another language requires knowledge and expertise in both the languages, and often takes a lot of time and effort. In recent past few decades, a need has arisen to translate using automated tools. This method which we now know as machine translation has been an on-going research till date. Many tools and systems have been developed and these systems have been able to translate many languages with reasonable output based on the resources and technology used. In order to build such a system, a critical resource is the parallel corpora. To translate between two languages, a bilingual corpus is required. For some languages like Khasi where the number of speakers is very less compared to Bengali, Punjabi or Tamil, there is no such corpora reported so far. At present there is a requirement to build a parallel corpus so that machine translation work can be started. This paper surveys some of the existing documents in Khasi which has been translated from English. A study is also shown on a few selected documents. These documents could be used to build a bilingual corpus which can be the source for machine translation in the future.

KEYWORDS

Machine translation, parallel corpora, bilingual corpus ,Khasi translated documents

1. INTRODUCTION

Machine translation for Indian languages has been progressing quite well in the past few years, especially for those languages with a good amount of existing resources. Many systems have been built for languages spoken in India, but when it comes to languages spoken in the Northeast part of India, development are still at its initial stages. A parallel corpus plays a major role in any machine translation between a pair of languages. Development of a bilingual machine translation system for any two languages with limited electronic resources is a very challenging task. Most of these languages lack the resources and tools required for machine translation. For some languages like the Khasi and Garo language which are spoken in Meghalaya, there is no corpus reported so far. Existing corpora such as ASPEC, Asian Scientific Paper Excerpt Corpus, is constructed by the Japan Science and Technology Agency (JST) in collaboration with the National Institute of Information and Communications Technology (NICT). It consists of a Japanese-English paper abstract corpus of 3M parallel sentences (ASPEC-JE) and a Japanese-Chinese paper excerpt corpus of 680K parallel sentences (ASPEC-JC)[18]. This corpus is one of the achievements of the Japanese-Chinese machine translation project which was run in Japan from 2006 to 2010.

81

There is also the Europarl parallel corpus of major European languages, the Canadian Hansard which is a parallel corpus of parliamentary proceedings, the Hunglish

Corpus[17] is a free sentence-aligned Hungarian- English parallel corpus of about 120 million words in 4 million sentence pairs. There is also English-Norwegian Parallel Corpus (ENPC) and English-Swedish Parallel Corpus (ESPC). The Hind EnCorp[15], which is an English-Hindi corpus, contains parallel corpus for English-Hindi as well as monolingual Hindi corpus. This paper present a survey on the English to Khasi translated documents existing at present. It is an initial attempt for building the corpus required for machine translation from English to Khasi.

2. A BRIEF INTRODUCTION OF THE KHASI LANGUAGE

Khasi is spoken by 1.6 million native speakers (2001 census), in the state of Meghalaya, in the Northeastern region of India. The Khasi language is an Austro-Asiatic language. According to classifications of Austro-Asiatic Language Family, Austroasiatic language is divided into Munda and Mon-Khmer, Mon-Khmer is divided into languages in the North and East. There are three languages in the North; the Khasi, the Khmuic and the Paluangunic. Initially the Khasi script was written using the Bengali script. Around 1816, a few translated versions of the Gospel of Matthew were printed. A few years later Thomas Jones, a Welsh Missionary introduce the Khasi alphabets using the Roman script. The first book written by him in Khasi is "Ka Kot Pule Nngkong" (Khasi First Reader), along with "Ka Kitab Nyngkong" (AB). It was in these two primers that the 21 alphabets in the Roman script were introduced. There were five vowel sounds a, e, i, o, u and sixteen consonant b, k, d, g, Ng, h, j, l, m, n, p, r, s, t, w, y. It was much later, in 1896, that two more sounds, ï (pronounced as yii) and ñ (pronounced as eiñ) were added, brings a total number of alphabets to 23. The decision to change the script from Bengali to Roman was because it was found that the latter was more suited to the sound system of Khasi. Khasi is a verb medial language; its basic word order pattern is Subject Verb Object (SVO).

3. SURVEY OF EXISTING DOCUMENTS TRANSLATED FROM ENGLISH TO KHASI

According to J.C. Catford [1], Translation may be defined as the replacement of textual material in one language by equivalent textual material in another language. One way of classifying the translation methods are full translation where the entire source language text is replaced by target language material and partial translation where some parts or parts of source language text are left untranslated: they are simply transferred to or incorporated. Other popular translation categories are free translation, word for word translation, and literal translation. A free translation is unbounded; equivalence could be between larger units than the sentence. Word for word translation is essentially rank-bound at word-rank. Literal translation is in between the first two.

Many attempts have been made to manually translate documents from English to Khasi. There are various methods in which manual translation has been implemented by translators from one language to another. Some of the methods that were used to translate these documents are adaptation translation method where the themes, characters, and plots are preserved while source language culture is converted to target language culture and text is rewritten. The other methods used are literal translation, word for word translation, and free translation. Below is a list of books translated manually by various authors from English to Khasi.

4. EXISTING TRANSLATED DOCUMENTS FROM ENGLISH TO KHASI

There are English to Khasi translated documents available for prose, poems, short stories, plays, and Bible etc. but without digitization. In recent times, local newspapers are published in English as well as in Khasi. A list some of the popular works is given below.

Table 1 Prose

Title and author	Source
1. U Basan ka Kastarbrij by B.War(2008)[4] 2. U Basan ka Casterbridge by Humphrey Blah(1980) 3. U Mayor ka Casterbridge by J. Kharmih(1985)	Mayor of Casterbridge by Thomas Hardy[12]
4. Ka Mahabharata(part I) by L.H Pde & Streamlet Dkhar(revised ed. 2008)	Mahabharata by C. Rajagopalachari,1951
5. Ka Euphrasie translated by Ka Jyllariewniam RNDM Dongmihngi Bashatei, India	Euphrasie by Mary Phillipa Reed(2006)
6. La pra lut baroh by A. Kharmalki(2008	Things fall apart by Chinua Achebe
7. Ka brisoh Olib by Micheal D. Kahit(2004)	Olive Garden by J.H Maclehose
8. Ka lynti jong ka jingsuk by B.M Lyngdoh(2013)	Path of Peace by Laurence Binyon
9. Ka jingim u Trai jong ngi by Soso Tham(1936)	The life of our lord by Charles dickens

Table 2 Poems

Title and author	Source
1. U Androkolis bad u Sing(Drama) by S.J Duncan(1978)	Androchles and the Lion by George Bernard Shaw
2. U Syiem Oedipus by B.L Swer(1984)	The Oedipus cycle: an English version by Dudley Fitts and Robert Fitzgerald, Harcourt, Brace, New York
3. U John Gilpin from ka Duitara Ksiar by Soso Tham(1936)	The Diverting History of John Gilpin by William Cooper
4. Ka jingiam briew ha u lum jingtep ingmane from ka Ryngkep by markha Joseph(1967)	An Eulogy written in the country church by Thomas Grey
5. U Ozymandias from ki Sur na la ri by S. Khongsit(1968)	Ozymandias by P.B. Shelley
6. Daffodils from ki Sur na la ri by S. Khongsit(1968)	Daffodils by William Worthsworth
7. Alexander Silkirk from ka Rympei by Seraiah Tham	The Solitude of Alexander Silkirk by William Cowper
8. Jingsuk ba la duh noh(Bynta I bad II) by S. Norindel Roy(1992)	Paradise Lost by John Milton
9. Ha jingbeh ka lyer from Jam Kjat by B.L Swer(2001)	A song "Blowing in the wind" by Bob Dylan
10. Ka Gitanjali by Pascal Malngiang(1988) 11. Gitanjali by Sumar Singh Sawian(2011)	Gitanjali by Rabindranath Tagore

Table 3 Short Stories

Title and author	Source
1. Ki khana u Edgar Allan Poe by W. Kharkrang(2004)	Stories by Edgar Allan Poe
2. Ki Phawar Aesop by Soso Tham(1967)	Aesop"s Fables
3. Uta uba poiei by Department of Khasi, Sankerdev college, Shillong	Castaway by Rabindranath Tagore
4. Siewspah kylliang by Department of Khasi, Sankerdev college, Shillong	Redemption by Victor Hugo
5. Dur khmat ha ka biar by Department of Khasi, Sankerdev college, Shillong	The Face on the wall by E.V Lucas

Table 4 Plays

Title and author	Source
1. Katba phi mon by F.M Pugh	As you like it by William Shakespeare
2. Ka shem lanot u Macbeth(1965) by F.M Pugh	Macbeth by William Shakespeare
3. Ka temding ia ka shla briew by F.M Pugh	Taming of the Shrew by William Shakespeare
4. U Doktor Phastos by A. Kharmalki (2003)	Doctor Faustus by Christopher Marlowe
5. U Romeo bad ka Juliet by F.M Pugh	Romeo and Juliet by William Shakespeare

5. OTHER EXISTING TRANSLATIONS

The other books existing at present which could be very useful for building the Corpora are the Bible which has been translated from English Bible and the Bilingual dictionary from Khasi to English by Nissor Singh in 1906 and Anglo‟s Khasi Dictionary[14] by Edingson Blah in 1981. Let‟s focus on the English to Khasi translated work only as of now.

5.1 EXAMPLES OF SOME OF THE TRANSLATED DOCUMENTS

A study has been made on the existing translated documents and a few selected documents are presented below based on some of the findings.

U BASAN KA KASTERBRIJ[4]

This book has been translated from the novel of Thomas Hardy, "The Mayor of Casterbridge"[12]. In fact there were two other books in Khasi that has been translated from the same novel and this book translated by Badaplin War is the most recent version. Here even the word „Mayor‟ has been translated to the Khasi word „Basan‟

However Casterbridge has just been rewritten using the Khasi alphabets, Kasterbrij, which is a transliteration problem between the Roman script and the corresponding Khasi script. This has been done for most of the names such as Henshad for Henchard, Pharphri for Farfrae, Nan Mokrit for Nance Mockridge. Hence partial translations have been done in some cases. The meaning of sentences has been maintained throughout the translation and it is not just a word by word translation but by using paraphrasing. Some examples of how the translation has been able to maintain the semantic and lexical level of these sentences are given below.

English: *The man was of fine figure, swarthy, and stern in aspect;*[12]

Khasi: *U rangbah u don ka rynieng ryniot, ka met ka phad kaba biang. Ka sniehdoh jong u ka kham dum rong, ka parmat kaba i domriang.*

English: *And he showed in profile a facial angle so slightly inclined as to be almost perpendicular*[12].

Khasi: *Bad ka dur pat ka i kumba ruid lain da pynieng haba u phai da kynriang.*

However in some of the translation, there is a rearrangement of sentences like the sentences given below.

English: *One evening of late summer, before the nineteenth century had reached one-third of its span, a young man and woman, the latter carrying a child, were approaching the large village of Weydon-Priors, in Upper Wessex, on foot*[12].

Khasi: *Ha kawei ka janmiet shuwa ban kut ka lyiur, ar ngut shi tnga ryngkat bad I khun kila wan poi ha shnong Weidon Praior kaba don ha ki thain Wessek. Kane ka la jia shuwa ban sdang ka spah snem baka khat khyndai.*

In the above paragraph the sentence "*before the nineteenth century had reached one-third of its span*" which was translated to Khasi as "*Kane ka la jia shuwa ban sdang ka spah snem kaba khat khyndai*" did not appear immediately following the sentence preceding it instead it appears at the end as a separate sentence.

LA PRA LUT BAROH[3]

This book by Antoinette Kharmalki is a translation from the novel by Chinua Achebe, "Things fall apart"[7]. Although there is a big difference in the culture between the two languages, yet the author of this book has been able to represent the source culture in the Khasi language. This is another translation which has contributed immensely to the Khasi literature. Some examples of which the translation has been done successfully while maintaining the semantic and lexical level of the source language, is given below.

English: *Okonkwo was well known throughout the nine villages and even beyond. His fame rested on solid personal achievements*[7]

Khasi: *Ia u Okonkwo la ithuh lut da baroh khyndai tylli ki shnong bad wat shabar jong kane ka shnong ruh. Ka nam jong u ki long namar ki jingjop jong u hi.*

Objects like jigida which is an exotic traditional African waist beads, has not been translated into Khasi literally and the name of the object has just been rewritten in Khasi as „jikida".

English: *"Remove your jigida first,"*[7]

Khasi: *"Law noh shwa ia uto u jikida jong pha seh"*

The names of persons in the source has been preserved and simply rewritten in Khasi alphabets. Even though the objects and names in African language „Ibo" could have been translated literally in Khasi but the author has wisely not done so. Some examples are given below.

"Ekwe" in Khasi it is *"ka tiar tem; ka ksing ba la shna da u dieng"* *"Akadi-nwaii"* in Khasi it is *" ka briew kaba la tymmen"*

"ese-akadi-nwaii" in Khasi it is *" Ki bniat jong ka tymmen"*
Here we notice partial translation is being used as some of the words are left untranslated.

U SYIEM OEDIPUS[8]

This book is a translation by B.L Swer from the source "Oedipus the King" by Dudley Fitts and Robert Fitzgerald[11]. The meaning and culture of the source has been maintained although there are some differences in terms of rhythm. But nevertheless this translation has been a great contribution to the Khasi literature. Some of the translated lines where rhythm is slightly different are given below.

English: My children, generations of the living
In the line of Kadmos, nursed at his ancient hearth[11].

Khasi: Khun jong nga ko ki pateng kiba im

Na ka kpoh u Kadmos ba la pynheh Pynsan ha ka rympei jong u.

KA SHEM LANOT U MACBETH[9]

It has been observed that this translation by F.M Pugh of the play Macbeth by W. Shakespeare into Khasi "Ka shem lanot u Macbeth" is mostly a word by word translation with some literal translation, but the central meaning of the play has been maintained. Some of the lines which have been successfully translated are:

English: A drum, a drum! Macbeth doth come[10].

Khasi: Ka ksing, ka ksing! Kynding-ding ding! Macbeth une la wan pynsting.

English: And, like a rat without a tail, I'll do, I'll do, and I'll do[10].

Khasi: Te kum ka khnai khlem tdong nang pong, Ia tdong ia trai lynghong lynghong
However there is a problem in translating word by word, often there is a loss of meaning and values. Such lines are as given below:

English: To wear a heart so white

Khasi: Ban phong dohnud Ba lieh katne katne.

KI KHANA U EDGAR ALLAN POE[5]

This is a collection of short stories translated by W. Kharkrang from the Stories of Edgar Allan Poe[13]. A brief study of a few of the translation of these stories is discussed below.

Ka Pipa Amontillado: This story is taken from the English version "The Cask of Amontillado". Here mostly literal translation has been used and some paraphrasing as well.

English: It was almost dark, one evening in the spring, when I met Fortunato in the street, alone[13].

Khasi: Ha kawei ka premmied Pyrem, nga la iakynduh ia u Fortunato uba iaid marwei ha surok.

English: "Fortunato! How are you?" "Montressor! Good evening, my friend."[13]

Khasi: "Fortunato! Kumno phi long?" "Khublei, Montressor, paralok jong nga!"

The names of persons has been left untranslated and even the alphabet like "F" which is not present in the Khasi alphabet has been maintained which indicate some form of partial translation.

U Klongsnam Ai-Dak: This story is translated from the Tell-Tale Heart. This translation by W. Kharkrang is one of the best, where every sentence is translated according to the source language. Literal translation has been utilized in most of the sentences. The rank-level of equivalence is mostly sentence by sentence. But a lower rank level is also present. Some examples are as follows:

English: Listen! Listen, and I will tell you how it happened[13].

Khasi: Sngap, sngap bha, ngan iathuh ha phi kumno ka la jia.

English: The eighth night I was more than usually careful as I opened the door[13].

Khasi: Ha ka mied kaba phra, nga la kham phikir haba nga plié iaka jingkhang.

5.2 ALIGNMENT OF A FEW ENGLISH-KHASI SENTENCES

A few sentences from the two translations of "Mayor of Casterbridge" by Badaplin War and Justman Kharmih are taken and a sentence alignment is shown below. In the sentence "You persuaded me", the correct translation is "Phi la pynbor ia nga", but Kharmih translated it as "Phi la pynngeit ia nga" which if translated back to English would be "You have made me believe", whereas War translated by using paraphrasing instead and somehow has deviated from the real meaning.

However the following two sentences, Both War and Kharmih have translated the sentences and have retain the actual meaning of the English sentences. On the sentence "allowed me to live on in ignorance of the truth for years", War has use paraphrasing and translated as "phi buhrieh iaka jingshisha na nga" which translated back to English is "you have hidden the truth for years from me". In fact in both the translations some words have been paraphrase but the intention and meaning of the sentences has been conveyed in the translations.

Table 5 Sentence Alignment

English	Badaplin War	Justman Kharmih
You persuaded me[12]	phi la thok bad pynkthang sat ia nga[4]	phi la pynngeit ia nga[19]
that my father was not my father[12]	ba I Pa im dei u kpa trai jong nga[4]	ba u kpa jong nga um dei u kpa jong nga[19]
allowed me to live on in ignorance of the truth for years[12]	phi buhrieh iaka jingshisha na nga da ki snem[4]	Baroh shi kata phi la leh shukor bad nga la im ha ka jingbymtip iaka jingshisha[19].

Table 6 Horizontal lines shows sentence alignment and arrows shows phrase alignment

English	The morning sun
	was streaming through
	the crevices of the canvas
	when the man awoke
Khasi	Ki kjatsngi.
	ki la phrung lyngba
	ki thliew ka ingjain
	ha ka por ba u kyndit na ba iohthiah

English	It was on a Friday
	evening,
	near the middle
	of September
Khasi	ha ka janmiet
	jong kawei ka sngi thohdieng
	hadien ba la kut shiteng
	u Nailar

English	and he'll
	go into other things too
	and that's where
	he makes his mistake
Khasi	U la
	sdang ruh sa kiwei de ki kam
	bad dei hangta
	ba u la leh bakla.

An attempt has also been made to align sentences as well as phrases on a few more sentences taken from "U Basan ka Kastarbrij" by Badaplin war and the source, a few examples are shown

in Table 6. It was found that because of the similarity in the word order of the two languages, mapping of phrases is easier. The second sentence, "It was on a Friday evening", we notice a crossing of the arrows, this usually happens because the translator has interchange the position of the phrases during translation which is still correct in meaning. But to be more precise the translation would have been as "Ha ka sngi thohdieng ha ka por janmiet", then the crossing of the arrows could have been avoided.

5.3 CHALLENGES IN TRANSLATING DOCUMENTS FROM ENGLISH TO KHASI

Translation of documents manually, requires a rigorous amount of effort and has to follow certain steps. In general the first stage would be to read the text, to understand what it is about; second, to analyze it from a translator's point of view before the text is rewritten in another language. The translator has to determine the intention and the way it is written for the purpose of selecting a suitable translation method and identifying particular and recurrent problems. Understanding the text requires both general and close reading.

Full translation where every level of source language is replaced by the target language is often very difficult to achieve. A cultural difference is one factor that challenges translation of two culturally different languages. Words having more than one meaning are another problem faced by translators and require someone with a very rich vocabulary in both the languages to be able to solve the problem. The number of translated documents is very small and today we require more documents to be translated to Khasi. But due to lack of the number of good translators, the desired amount of translated documents and quality cannot be achieved.

6. CONCLUSION AND FUTURE WORK

Building a bilingual corpus of English and Khasi language is the utmost requirement at present, to be able to perform machine translation. The translated documents discussed above can be selected as sources for building parallel corpora. Although after analyzing the different types of documents, it has been found that prose, short stories and plays can play a major role in the Corpora for machine translation as compared to poems. A common similarity between English and Khasi is the word order that is; they are both SVO (Subject Verb Object). This common property of the two languages will prove as an advantage in performing machine translation between the two languages. This paper has presented many translated documents with an analysis on some of them. It is an initial step to build parallel corpora in English and Khasi which could be the starting point for machine translation from English language to Khasi language. The future work of the whole exercise is to digitize the presently available and verified documents into usable format. We plan to use these parallel corpora for building machine translation system between English and Khasi and other related natural language processing tasks.

REFERENCES

[1]   J.C Catford, (1965) A Linguistic Theory of Translation, Oxford University Press, London.

[2]   Badaplin War, (2016) Ka Kylla-Ktien bad Ka Litereshor Khasi, Rilum Offset Printing House, Umsohsun, Shillong.

[3]   S.J. Duncan, (1978) U Androkolis bad u Sing, Ri Khasi Book Agency, Shillong.

[4]   Badaplin War, (2008) U Basan ka Kasterbrij, Ri-Ia-Dor, SMS Hi-Tech Impression, Shillong.

[5]   W. Kharkrang, (2004) Ki Khana U Edgar Allan Poe, Modern Offset, Shillong.

[6]   Kharmalki, (2008) La Pra Lut Baroh.Lianmeroschse, SMS Hi-Tech Impression, Shillong

[7]   Chinua Achebe, (1996) Things Fall Apart, Heinemann Educational Publishers, UK.

[8]   B.L Swer, (2009) U Syiem Odipus, Ri-Khasi Book Agency, Shillong.

[9]   F.M. Pugh, (1965) Ka Shem-Lanot U Macbeth, Shillong printing works, Shillong.

[10]   Willliam Shakespeare, (1978) Macbeth, in The Annotated Shakespeare, Complete Works Illustrated, Volume III Tragedies and Romances by A. L. Rowes, Orbis Publishing, London.

[11]   Dudley Fitts and Robert Fitzgerald, (1949) The Oedipus cycle: an English version, Harcourt, Brace, New York

[12]   Thomas Hardy, (1886) The life and Death of the Mayor of Casterbridge: A study of a man of character, Penguin Books Ltd.London.

[13] Edgar Allan Poe. (1975) The Complete Tale & Poems of Edgar Allan Poe, Vintage Books Edition, USA.

[14] Edingson Blah, (1981) Anglo's Khasi Dictionary, Shillong Chapala Book Stall, Shillong.

[15] Ondrej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Stranak, Vit Suchomel, Aleš Tamchyna and Daniel Zeman, (2014) "HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation", In Proc. of LREC 2014, Reykjavik, Iceland, ISBN 978-2-9517408-8-4, ELRA.

[16] Lexi Birch, Chris Callison-Burch, Miles Osborne and Matt Post, (2011) "The Indic multi- parallel corpus", http://homepages.inf.ed.ac.uk/miles/babel.html.

[17] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy, (2005) "Parallel corpora for medium density languages", In Proceedings of the RANLP 2005, pages 590-596

[18] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, Hitoshi Isahara, (2016) "ASPEC: Asian Scientific Paper Excerpt Corpus", Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016). ISBN 978-2-9517408-9-1, Portorož, Slovenia.

[19] Justman Kharmih, U Mayor ka Casterbridge, Meghalaya State Coop. Union Printing Press, Shillong.

AUTHORS

Aiusha Vellintihun Hujon is currently an Assistant Professor in St. Anthony''s College, Shillong, Meghalaya, teaching in the Department of Computer Science.

Thoudam Doren Singh is currently an Assistant Professor of Computer Science and Engineering, IIIT Manipur, India.