Generating Translation Corpora in Indic Languages: Cultivating Bilingual Texts for Cross Lingual Fertilization

Niladri Sekhar Dash
Linguistic Research Unit
Indian Statistical Institute
Kolkata, India
Email: ns_dash@yahoo.com

Arulmozi Selvraj
Centre for ALTS
Central University of Hyderabad
Hyderabad, India
Email: arulmozi@gmail.com

Mazhar Hussain
Centre for Indian Languages
School of LL & CS
Jawaharlal Nehru University, Delhi
Email: mazharmehdi@gmail.com

Abstract

We address some theoretical and practical issues relating to generation, processing, and management of Translation Corpus (TC) in Indian languages, which is developed in a consortium-mode project (ILCI-II)[1] under the DeitY, Govt. of India. Issues are discussed here for the first time keeping in mind the ready application of TC in various domains of computational and applied linguistics. We first define what is a TC; describe the process of its construction; identify its features; exemplify the processes of text alignment in TC; discuss methods of text analysis; propose for restructuring of translational units; define the process of extraction of translational equivalents; propose for generating bilingual lexical database and TermBank from a structured TC; and finally identify areas where a TC and information extracted from it may be utilized. Since construction of TC in Indian languages is full of hurdles, we try to construct a roadmap with a focus on techniques and methodologies that may be applied for achieving the task. The issues are brought under focus to justify the work that generated TC for some Indian languages for future reference and application.

1. What is a Translation Corpus ?

Theoretically, a Translation Corpus (TC) suggests that it contains texts and their translation. It is entitled to include bilingual (and multilingual) texts as well as texts that may fit under *translation*. A TC, by virtue of its character and composition, is made of two parts: a text from a source language (SL) and its translation from a target language (TL) [15] [24], [39]. Although, a TC is normally bilingual and bidirectional [28], it can be multilingual and multi-directional as well [37], as it actually happens in case of the ILCI-I and ILCI-II projects for the Indian languages. In these two projects a new strategy is adopted where Hindi is treated as the only SL and several other Indian languages are treated as the TL (Fig. 1).

The issue of multi-directionality can be understood if all the target languages can establish linguistic links with each other as they are linked up with SL. Since the ILCI-I TC has not tried to venture into this direction, it makes sense to keep the present discussion confined within a scheme of bilingualism and bi-directionality, with, for example, Hindi as SL and Bangla as TL to understand the theoretical and practical issues involved in its text content, composition, structure, construction, processing, and utilization. Hence forth, our discussion will sail in this direction only.

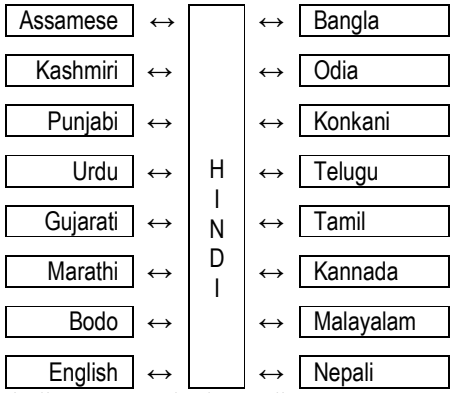

Fig. 1: Hindi as a SL and other Indian Languages as TL

Theoretically, a TC is supposed to keep meaning and function of words and phrases constant across languages [19], although alternation in structure (i.e., sequential order of words and phrases) is a permissible deviation. TC offers an ideal resource for comparing realisation of meanings (and structures) in two different languages under identical situations [3]. Also, it makes possible to discover the cross-linguistic variants, i.e., alternative renderings of meanings and concepts in the TL [4]. Thus a TC becomes useful for cross-language analysis and formulation of comparable lexical databases necessary for translation [1], [21], [26].

Since a TC contains texts from one language and its translations in another language, it is viewed as a sub-type of a parallel corpus, which, in principle, requires its texts to be maximally comparable to each other [28]. Therefore, it is better to consider a TC as a special corpus, which is identical in genre, similar in text type, uniform in format, parallel in composition, identical in content, comparable to each other and specific in utility [32], [37].

2. Construction of a Translation Corpus

The construction of a TC is a complicated task. It requires careful manipulation of SL and TL texts [18] [19]. A TC should be made in such a way that it is suitable to combine advantages of both comparable and parallel corpora [2].

[1] Indian Languages Corpora Initiative-Phase II

Proceedings of ICON-2015: 12th International Conference on Natural Language Processing
Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2015
D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 333–342,
Trivandrum, India. December 2015.

Text samples from the languages should be matched as far as possible in terms of text type, subject matter, purpose, and register [1]. The structure of a TC within two languages may be envisaged in the following manner keeping in mind the aim of the task and components to be integrated within a TC (Fig. 2).

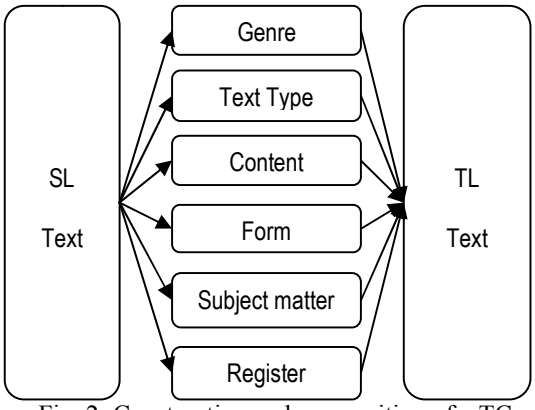

Fig. 2: Construction and composition of a TC

The diagram given above shows that a TC is designed in such a way that it can be used as a comparable as well as a parallel corpus. The reverse argument is not however true. That means a comparable or parallel corpus cannot be used as a TC unless it follows the conditions of its construction stated above. Therefore, selection of texts for constructing a TC needs to be guided by the following principles [34]:

- Written texts are included in a TC. Texts obtained from speech are ignored since the present state of TC targets written texts only.
- Texts should reflect on the contemporary language use although texts of earlier era may have relevance in case translating historical texts.
- Texts should be restricted to specific subject. It should include texts from specific domains of language use.
- Texts of the SL and the TL should be comparable as far as possible. They should match in **genre** (e.g. news), **type** (e.g. political), **content** (e.g. election), and **form** (e.g. report). They should also match in subject matter, register varieties, purpose, and type of user, etc.
- Texts must consist of fairly large and coherent extracts obtained from beginning to end of a breaking point (e.g. chapter, section, paragraph, etc.)
- Texts should faithfully represent regular as ell as special linguistic forms and elements of SL and TL. They may be large in size to encompass maximum varieties in content. Lexical varieties should be high in a TC.
- Texts should faithfully preserve domain-specific words, terms, idioms, phrases, and other lexical elements. Texts used in TC should be authentic and referential for future verification and validation.
- Texts should be available in machine-readable form for access and reference by users. Users may use language data in multiple tasks, such as, statistical sampling, text alignment, lexical database generation, text processing and translation, etc.
- Text samples should be preserved in annotated or non-annotated version. A POS tagged TC is a better resource than a non-tagged one.
- Linguistic and extralinguistic information should be captured in a systematic manner so that users can access information easily for future reference and validation.

Given below (Fig. 3) is a sample of Hind-Bangla translation corpus taken from the ILCI-I project.

Hindi Text
हृदय रोगी को नमक, मिर्च तथा तले-भुने भोजन का प्रयोग कम से कम करना चाहिए या हो सके तो नहीं करना चाहिए। हरी पत्तेदार सब्जियों तथा फल का सेवन अधिक मात्रा में करना चाहिए। यदि हृदय रोगी धूम्रपान, शराब या अन्य किसी नशीली वस्तु का सेवन करता है तो उसे शीघ्र ही इन पदार्थों का सेवन बंद कर देना चाहिए। हृदय रोगी को घी, मक्खन इत्यादि का सेवन कम से कम करना चाहिए। हृदय रोगी को आँवला तथा लहसुन का सेवन प्रतिदिन करना चाहिए। सेब के मुरब्बे का सेवन हृदय रोगियों को विशेषकर करना चाहिए।
Bangla Translation
হৃদরোগীদের নুন, ঝাল ও আজেবাজে খাবার খাওয়া খুব কমিয়ে দেওয়া উচিত বা সম্ভব হলে বন্ধ করে দেওয়া উচিত । টাটকা সবজী ও ফল অধিক মাত্রায় ভোজন করা উচিত । যদি হৃদরোগী ধূমপান, মদ বা অন্য কোনো নেশা করেন তবে তাঁকে শীঘ্রই এই সব খাওয়া বন্ধ করে দিতে হবে । হৃদরোগীর ঘি, মাখন ইত্যাদি কম করে খাওয়া উচিত । হৃদরোগীকে প্রতিদিন আমলকী ও রসুন খাওয়ানো উচিত । আপেলের মোরোব্বা খাওয়া হৃদরোগীদের বিশেষ প্রয়োজন ।

Fig. 3: Hind-Bangla translation corpus

3. Features of a Translation Corpus

A TC has certain default features, which may vary for other types of corpus [32]. That means, a TC which does not possess these default features may be put outside its scope due to deviation from the norm. By all means, a TC should possess the following features:

3.1 Quantity of Data

A TC should be big with large collection of texts from SL and TL. Larger amount of text facilitates accessibility and reliability of translation. The number of sentences included in TC will determine its quantity. Since the primary goal of a TC is to include texts for translation, it should not be restricted with fixed number of sentences. The size of a TC is related to the amount of texts included in it. It is the total number of sentences that determines its size [31]. A TC that includes more number of sentences is more suitable, since size is an important issue in TC based linguistic works.

Making a TC large is linked with number of 'tokens' and 'types' included in it as well as with decisions of how many texts will be in it; how many sentences will be in each text; and how many words will be in each sentence [5]. A small TC, due to its limited number of texts, may fail to provide some advantages, which a large TC can provide. We find that a large TC has the following advantages:

- It presents better scope for variation of texts.
- It provides better spectrum of the patterns of lexical and syntactic usages in SL and TL.
- It confirms increment in number of textual citations that provide scopes for systematic classification of linguistic items in terms of their usage and meaning.
- It assures better opportunity for obtaining all kinds of statistical results far more faithfully for making various correct observations.
- It gives wider spectrum for studying patterns of use of individual words and sentences. This helps us to make generalization about syntactic structures of SL and TL.
- It helps to understand the patterns of use of multiword units like compounds, reduplicated forms, collocations, phrases, idioms and proverbs, etc. in SL and TL.
- It helps to identify coinage of new words and terms, locate their fields of usage, find variations of sense of terms, and track patterns of their usage in texts, etc.
- It gives scope for faithful analysis of usage of technical and scientific terms – a real challenge in translation.

A large TC is not only large in amount of data but also multidimensional in its composition, multidirectional in its form, and multifunctional in its utility. Thus quantity of data has an effect on validity and reliability of a TC. Also, it ensures diversity of SL/TL texts from which it is made. Since a TC is a minuscule form of SL and TL text varieties, in qualitative authentication of SL and TL properties, it is useless if it is not made large in amount of data [33].

3.2 Quality of Text

Quality relates to authenticity. That means texts should be collected from genuine communication of people from their normal discourse. The primary role of a TC generator is to acquire data for a TC generation in which he has no liberty to alter, modify or distort the actual image of the SL. He has no right to add information from personal observation on the ground that the data is not large and suitable enough to represent the language for which it is made. The basic point is that a TC developer will collect data faithfully following the predefined principles proposed for the task. If he tries to interpolate in any way within the body of the text, he will not only damage the actual picture of the text, but also damage heavily the subsequent analysis of data. This will affect the overall projection of the language, or worse, may yield wrong observations about the language in question. Therefore, at the time of constructing a TC, we had to observe the following conditions:

- Repetition of texts or sentences are avoided.
- Ungrammatical constructions are removed.
- Broken constructions are ignored.
- Incomplete constructions are separated.
- Mixed sentences are avoided.
- Texts from single field or domain are considered.
- Both synchronic and diachronic texts are considered.
- Standard forms of regular usage are considered.
- Text representation is maximally balanced, non-skewed, and wide.
- Texts are in homogeneous form without distortion of language data.

3.3 Text Representation

A TC should include samples from a wide range of texts to attain proper representation. It should be balanced to all disciplines and subjects to represent maximum number of linguistic features found in a language. Besides, it should be authentic in representation of a text wherefore it is made, since future analysis and investigation of TC may ask for verification and authentication of information from the TC representing a language. For example, when we develop a Hindi-Bangla TC, which is meant to be representative of a domain of the languages, it should to kept in mind that data are collected in equal proportion so that the TC is a true replica of the languages. This is the first condition of text representation.

Text samples should not be collected only from one or two texts. They should be maximally representative with regard to domains. A TC should contain samples not only from imaginative texts like fictions, novels, and stories but also from all informative texts like natural science, social science, earth science, medical science, engineering, technology, commerce, banking, advertisements, posters, newspapers, government notices and similar sources. To be truly representative, samples should be collected in equal proportion from all sources irrespective to text types, genres, and time variations. Although the appropriate size of sample of a TC is not finalised, we have collected 50,000 sentences from each domain where the number of sentences is divided equally among the sub-domains.

3.4 Simplicity

A TC should contain text samples in simple and plain form so that texts are easily used by translators without being trapped into additional linguistic information marked-up within texts. In fact, simplicity in texts puts the TC users in a better position to deal with the content of texts. However, it is not altogether a hurdle if TC texts are marked-up at word, phrase, and sentence level with grammatical, lexical, and syntactic information. The basic role of a mark-up process is to preserve some additional information, which will be useful for various linguistic works. Although these are helpful, these should be easily separable so that the original TC text is easily retrievable. There are some advantages in using mark-ups on a TC. In information

retrieval, machine learning, lexical database generation, termbank compilation, and machine translation, a TC built with marked-up texts is more useful for searching and data extraction from the texts, which results in development of systems and tools. Marked-up TCs are also quite useful for sociolinguistic researches, dictionary compilation, grammar writing, and language teaching.

3.5 Equality

Each text sample should have equal number of sentences in TC. For instance, if a SL text contains 1000 sentences, each TL text should also contain the same number of sentences. We propose this norm because we argue that sentences used in TC should be of equal number so that translation mechanism can work elegantly. However, there may be some constraints, which may not be avoided at the time of TC generation:

- Number of texts available in SL may be more than that of TL.
- Collection of equal number of sentences from SL and TL may not be an easy task.
- Parity in number of sentences is deceptive, because sentences never occur in equal number in SL and TL.
- A sentence in SL may be broken into two or more sentences in TL. Reversely, several sentences in SL may be merged into one sentence in TL.
- Equal number of sentences cannot be collected from SL and TL in a uniform manner, since size varies.

3.6 Retrievability

The work of TC generation does not end with compilation of texts. It also involves formatting the text in a suitable form so that the data becomes easily retrievable by end users. That means data stored in a TC should be made easily retrievable for users. Anybody interested in TC should be able to extract relevant information from it. This directs our attention towards the techniques and tools used for preserving TC in digital format. The present technology has made it possible for us to generate a TC in PCs and preserve it in such a way that we are capable to retrieve and access the texts. The advantage, however, goes directly to those people who are trained to handle language databases in computer.

This, however, does not serve the goals of all TC users, since utility of TC is not confined to computer people only. A TC is made for all (computer experts, linguists, social scientists, language experts, teachers, students, researchers, historians, advertisers, technologists, and general people). Its goal is accomplished when people coming from all walks of life can access it according to their needs. In reality, there are many people who are not trained for handling computer or digital TC, but need TC to address their needs. Therefore, TC must be stored in an easy and simple format so that common people can use it.

3.7 Verifiability

Texts collected in a TC should be open for all empirical verifications. It should be reliable and verifiable in the context of representing a language under study. Until and unless a TC is fit for all kinds of empirical analysis and verification, its importance is reduced to nothing. Text samples, which are collected and compiled in a TC to represent SL and TL should honestly register and reflect on the actual patterns of language use. To address this need, a TC should be made in such a way that it easily qualifies to win the trust of users who after verifying texts, agree that what is stored in TC is actually a faithful reflection of SL and TL. For instance, when we develop a TC for Hindi and Bangla we are careful that texts stored in the TC qualify to reflect properly on the respective languages. A TC thus attests its authenticity and validity.

3.8 Augmentation

A TC should grow with time with new texts to capture the changes in content and form. Also it should grow to register variations in texts. Although most of the present TCs are synchronic, we should take effort to make diachronic TCs so that we find a better picture of the languages involved in the game. A synchronic TC, by addition of texts, may become diachronic. This can have direct effects on size, quantity, coverage, and diversity of a TC. A*ugmentation* thus becomes an important feature of a TC.

3.9 Documentation

It is necessary to preserve detail information of the sources wherefrom texts are collected in TC. It is a practical requirement on the part of TC designer to deal with problems related to verification and validation of SL and TL texts and dissolving copyright issues. It is also needed to dissolve linguistic and extralinguistic issues relating to sociolinguistic investigations, stylistic analyses, and legal enquiries, etc. which ask for verification of information of SL and TL texts. As TC maker we document meticulously all extralinguistic information relating to types of text, source of text, etc. There are directly linked with referential information of physical texts (e.g., name of book, name of topics, newspaper, year of first publication, year of second edition, numbers of pages, type of text, sex, profession, age, social status of author(s), etc.).

Documentation information of a TC should be separated from the texts itself in the form of Metadata. We keep all information in a Header File that contains all references relating to texts. For easy future access, management, and processing of TC this allows us to separate texts from the tagset quickly. We follow the TEI format (*Text Encoding Initiative*), which has a simple minimal header containing reference to texts. For management of a TC, this allows effective separation of plain texts from annotation with easy application of Header File separation.

4. Alignment of Texts in Translation Corpus

Aligning texts in TC means making each Translation Unit (TU) of SL correspond to an equivalent unit in TL [27]. TU covers small units like words, phrases, and sentences [12] as well as larger units like paragraphs and chapters [30] (Fig. 4). Selection of TU depends largely on the point of view selected for linguistic analysis and the type of corpus used. If a TC asks for a high level faithfulness to original, as it happens in legal and technical texts, close alignment between sentences, phrases or even words is mandatory. In case of non-technical texts (e.g. fiction), alignment at larger units at paragraph or chapter level will suffice [38]. Thus, operation of alignment may be refined based on the type of corpus used in the work. The faithfulness and linearity of human translations may guide to align a TC, although this is predominantly true for technical corpora. Literary TC, on the other hand, lends itself to reliable alignment of units beyond sentence level if translational equivalency observed in TC is previously formalised [11].

	SL Text	↔	TL Text
	↓		↓
Small TUs	Character	↔	Character
	Word	↔	Word
	Phrase	↔	Phrase
	Sentence	↔	Sentence
Large TUs	Paragraph	↔	Paragraph
	Chapter	↔	Chapter

Fig. 4: Layers of translation unit alignment in TC

Since so-called 'free translations' present a problems in processing due to missing sequence, change in word order, modification of content, etc. it is sensible to generate sets of 'corresponding texts' having mutual conceptual parallelism. The main goal is not to show structural equivalences found between the two languages, but pragmatically, to search TL text units, which appear to be closest to SL text units. Such rough alignment yields satisfactory results at sentence level [17] especially when supported by some statistical methods [6] with minimal formalisation of syntactic phenomena of texts [7].

Sentence level alignment is an important part of TC alignment. It shows correspondences down to the level of sentence, and not beyond that [8]. For this work, a weak translation model serves the purpose, since this is one of the primary tools required at the initial stage of TC analysis [29]. Given below is a sample of Hind-Bangla TC where sentences are largely aligned (Fig. 5).

Alignment of TC helps to optimize mapping between two equivalent units in order to obtain better translation output. It involves associating equivalent units (e.g. words, multiword units, idioms, phrases, clauses, and sentences, etc.) endowed with typical formal structures. However, the basic purpose of this process alignment is to allow pairing mechanism to be broken into following three parts in a systematic way:

- Identification of potential linguistic units, which may be grammatically associated in TC.
- Formalisation of structures of associable units by way of using sets of morphosyntactic tags.
- Determination of probability of proposed structures comparing the forms with effective texts collected from manually translated corpora.

Sentence ID	Hindi-Bangla Aligned Sentences
HNHL_296	हृदय रोगी को घी, मक्खन इत्यादि का सेवन कम से कम करना चाहिए ।
BNHL_296	হৃদরোগীর ঘি, মাখন ইত্যাদি কম করে খাওয়া উচিত ।
HNHL_297	हृदय रोगी को आँवला तथा लहसुन का सेवन प्रतिदिन करना चाहिए ।
BNHL_297	হৃদরোগীকে প্রতিদিন আমলকী ও রসুন খাওয়ানো উচিত ।
HNHL_298	सेब के मुरब्बे का सेवन हृदय रोगियों को विशेषकर करना चाहिए ।
BNHL_298	আপেলের মোরোব্বা খাওয়া হৃদরোগীদের বিশেষ প্রয়োজন।

Fig. 5: Sentences aligned in Hind-Bangla TC

By subdividing the process into three parts a relatively simple system module may be developed to identify the units likely to correlate with analysis of TC [24]. It is not, however, necessary to analyse all sentences used in TC to find out all matches. Analysis of type constructions, rather than full set of tokens, serves the initial purpose, because:

(a) In a language there are units, which are identical in form and sense. That means a NP in SL may correspond structurally to other NPs within a text. This is true to both SL and TL.
(b) Sequence and interrelation between the units in TL text may be same with those in SL text if TC is developed from two sister languages.
(c) There are certain fixed reference points in texts (e.g., numbers, dates, proper names, titles, paragraphs, sections, etc.), which mark out texts and allow rapid identification of translation units.

It is always necessary to fine-tune alignment process of TC to enhance text processing and information retrieval. However, it requires identification and formalisation of 'translation units' and utilisation of bilingual dictionaries. So, there is no need for exhaustive morphosyntactic tagging of each text, since machine can do it with a statistical support to find out equivalent forms by comparing TC that exhibit translational relations. However, to ensure quality

performance of a system the following things should be taken care of:

(a) Standard of TC should be high. Aligned bilingual texts may pose problem if the quality of TC is poor or if texts are not put under strict vigilance of linguists.
(b) Quality and size of bilingual dictionary should be high. Dictionary is a basic resource in terms of providing adequate lexical information. Moreover, it should have provision to integrate unknown words found in TC.
(c) Robustness of the system and the quality of translation will depend on the volume of training data available.
(d) Level of accuracy in TC will rely heavily on the levels of synchronisation between the texts of TC.

Alignment of TC is a highly complicated task. Impetus for progress must come from linguistic and extralinguistic sources. It is a highly specialised work, which unlike most others, is a worthy test bed for various theories and applications of linguistics and language technology. It verifies if theories of syntax, semantics, and discourse are at all compatible to it; if lexicon and grammar of SL and TL are fruitfully utilised; if algorithms for parsing, word sense analysis and pragmatic interpretations are applicable; and if knowledge representation and linguistic cognition have any relevance in it. Alignment of text is greatly successful in domain-specific TC with supervised training where all syntactic, lexical and idiomatic differences are adequately addressed [35]. This usually narrows down the gulf of mutual intelligibility to enhance translatability between the two languages.

5. Translation Corpus Analysis

TC analysis has three goals. First, to structure translations in such a way that these are usable in production of new translations. Using *TransSearch System* [16] we can mark out bilingual correspondences between SL and TL texts. Second, to draft translations to detect translation error, if any, in TC. It is possible to certify that a translation is complete, in the sense that larger units (pages, paragraphs, sections, etc.) of SL texts are properly translated in TL text. Last, to verify if any translation is free from interference errors resulted from 'deceptive cognates'. For instance, the Hindi word *sandes* 'news' and Bangla word *sandes* 'sweet' cannot be accepted as good cognates for mutual translation, although they appear similar in form in the two languages. Similarly, Hindi word *khun* and Bangla word *khun* should not be treated as identical, because while the Hindi word means 'blood', Bangla word means 'murder'.

A TC, once aligned, is available for linguistic analysis. In general, it involves the following tasks:

(a) **Morphological Analysis:** Identify form and function of constituting morphemes.
(b) **Syntactic Analysis:** Identify form and function of syntagms in respective corpus.
(c) **Morphosyntactic Analysis:** Identify interface involved within surface forms of lexical items used in TC.
(d) **Semantic Analysis:** Identify meaning of linguistic units (i.e., words, idioms, phrases, etc.) as well as ambiguities involved therein.

For effective linguistic analysis, we may use descriptive morphosyntactic approach along with some statistical approaches for probability measurement. We take support from standard descriptive grammars and morphosyntactic rules of SL and TL. At this stage, part-of-speech tagging is done by comparing texts of SL and TL manually. Our traditional grammatical categories have good referential value on quality of part-of-speech tagging, since a MT system with few POS tags shows greater success than a system with exhaustive POS tags [10]. Based on analysis of equivalent forms obtained from TC, we find three types of matching:

- **Strong match:** Here number of words, their order, and their meaning are same.
- **Approximate match:** Here number of words and their meanings are same, but not the order in which they appear in texts.
- **Weak match:** Here order and number of words are different, but their dictionary meanings are same.

In case of translating texts from Hindi to Bangla, most of the grammatical mappings are 'strong matches', as the languages belong to same typology. In such a situation, alignment of texts in TC can rely on syntactic structure of respective texts although greater emphasis should be on semantic match. We argue that if 70% words in a sentence of Hindi text semantically correspond to 70% words in a sentence in Bangla text, we can claim that sentences have semantic equivalency to have a translational relationship.

Research is going on to develop TC analyser, which can account for translation equivalence between words, idioms, and phrases in TC. Statistical algorithms are also used to find keywords to retrieve equivalent units from TC. Once these are found, these are verified and formalised by human translators as model inputs and stored in bilingual lexical database [13], [28].

6. Restructuring Translation Units

Restructuring a Hindi sentence into Bangla is an attempt to maximize all linguistic resources, strategies and methods deployed in manual translation, as Hindi and Bangla exhibit close typological, grammatical, and semantic similarities due to their genealogical linkage. Since both the languages belong to the same family, it has been, to a large extent, an easy task for us to restructure Hindi phrases in Bangla with utilization of lexico-grammatical stock of both languages. The linguistic knowledgebase and information obtained from this kind of experiment can help to design system for Machine Aided Translation between the two languages.

(a) Hindi : Hindu dharm mein tIrtha kA baRA mahattva hyay.
(b) Bangla : Hindu dharme tirther bishes guruttva ache.

The type of restructuring referred to in the following table (Table 1) is called 'grammatical mapping' in TC. Here, words of the SL text are 'mapped' with words of the TL text to obtain meaningful translation. Although there are various schemes for mapping (e.g., lexical, morphological, grammatical, phrasal, clausal, etc.), the most common form of grammatical mapping is phrase mapping within the two languages considered in TC.

Input	Hindu (a) dharm (b) mein (c) tIrtha (d) kA (e) baRA (f) mahattva (g) hyay (h)
Literal Output	Hindu (1) dharma (2) -e (3) tirtha (4) -er (5) bishes (6) guruttva (7) ache (8)
Restructuring	Hindu (1) dharme (2+3) tirther (4+5) bishes (6) guruttva (7) ache (8)
Actual Output	(1) (2+3) (4+5) (6) (7) (8)

Table 1: Restructuring Hindi and Bengali sentences

In above examples (a & b) we see how we need to map the case markers with nouns to get appropriate outputs in Bangla translation. In Bangla, case markers are tagged with nouns and pronouns, while in Hindi, they remain separate from nouns and pronouns and appear as independent lexical items in sentence. That means at the time of translation from Hindi to Bangla, the multi-word units (particularly of verb class) have to be represented as a single-word unit in Bengali.

Grammatical mapping is highly relevant in the context of MT between the two languages, which are different in word order in sentence formation. In the present context, while we talk about MT system from Hindi to Bangla, this becomes relevant, as Hindi phrases need to be restructured in the framework. Therefore, grammatical mapping and reordering of words is needed for producing acceptable outputs in Bangla.

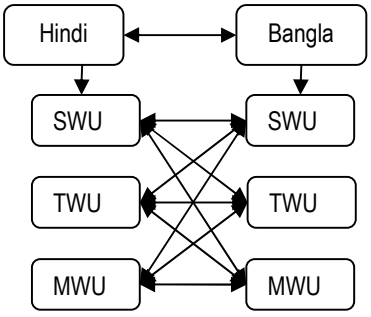

Fig. 6: Lexical Mapping between Hindi and Bangla

At the lexical level, on the other hand, to achieve good output in Bangla, words used in Hindi sentence need to be mapped with words used in Bangla in the following manner (Fig. 6). However, it is found that mere lexical mapping is not enough for proper translation. A Hindi sentence may contains an idiomatic expression, which requires pragmatic knowledge to find a similar idiomatic expression in Bangla to achieve accuracy in translation. Therefore, we need to employ pragmatic knowledgebase to select the appropriate equivalent idiomatic expression from the TL.

7. Extraction of Translational Equivalent Units

Search for translation equivalent units (TEU) in TC begins with particular forms that express similar sense in both the languages. Once these are identified in TC, these are stored in a separate lexical database. Normally, a TC yields large amount of TEU, which are good to be used as alternative forms. The issues that determine the choice of appropriate equivalent form are measured on the basis of recurrent patterns of use of the forms. The TEUs are verified with monolingual text corpora of the two languages from which TC is made. It follows a scheme (Fig. 7) through which we generate a list of possible TEUs from the TC.

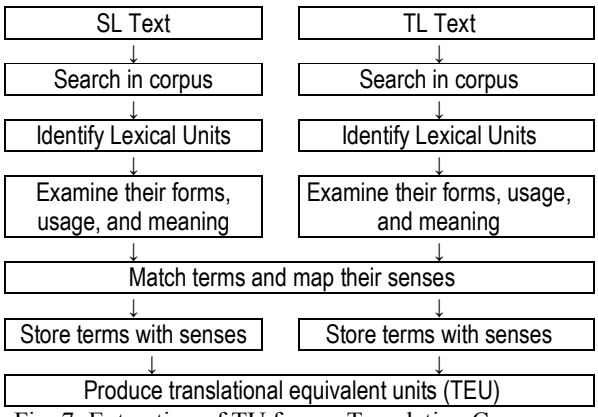

Fig. 7: Extraction of TU from a Translation Corpus

We find that even within two closely related languages, TEUs seldom mean the same thing in all contexts, since these are seldom used in the same types of syntactic and grammatical construction [12]. Moreover, their semantic connotations and degree of formality differ depending on language-specific contexts. Sometimes a lemma in TL is hardly found as a true TEU to a lemma of SL, even though they appear conceptually equivalent. Two-way translation is possible with proper names and scientific terms, but hardly with ordinary lexical units [25]. This signifies that ordinary texts will create more problems due to differences in sense of words. It requires a high degree of linguistic sophistication to yield better outputs. In general, we extract the following types of TEUs from a TC to build up useful resource for multiple applications:

- Extract good TEU including words, idioms, compounds, collocations, and phrases.
- Learn how TC help in producing translated texts that display 'naturalness' of the TL.

- Create new translation databases that will enable us to translate correctly into the languages on which we have only limited command.
- Generate Bilingual Lexical Database (BLD) for man and machine translation.
- Generate Bilingual Terminology Database (BTB) as it is neither standardised nor developed for Indic languages.

Process of extracting TEUs from TC and their subsequent verification for authentication with monolingual corpora is schematized below (Fig. 8). To find out TEU from a TC we use various searching methods to trace comparable units (i.e., words and larger units than words) which are similar in meaning. Findings are further schematized with bilingual lexical dictionary and term databases to enrich the MT knowledgebase for the battles ahead.

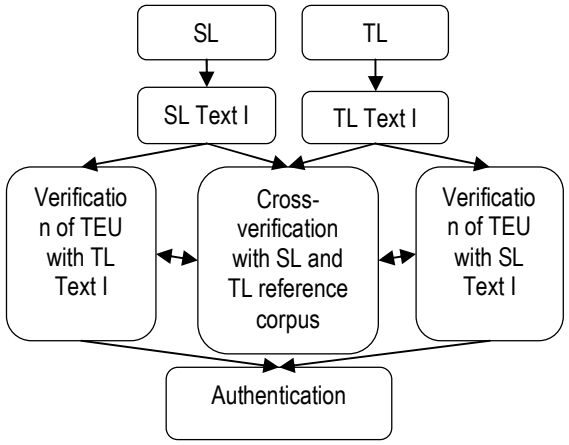

Fig. 8: Verification of TEUs with monolingual corpus

8. Bilingual Lexical Database

Development of a Bilingual Lexical Database (BLD) from a TC is an essential task the lack of which is one of the bottlenecks of present MT works in Indian languages. Traditional dictionaries cannot compensate this deficiency, as they do not contain information about lexical sub-categorisation, lexical selection restriction, and domains of application of lexical items [14]. Using POS-tagged TC we can extract semantically equivalent words for the BLD [9]. A BLD may be developed from untagged corpora when POS-tagged TC is not available for the purpose.

Formation of BLD is best possible within those cognate languages, that are typologically or genealogically related (e.g. Bangla-Odia, Hindi-Urdu, Tamil-Malayalam, etc.) because cognate languages usually share many common properties (both linguistic and non-linguistic) hardly found in non-related languages [22]. Also, there is a large chunk of regular vocabulary similar to each other not only in phonetic/orthographic and representations but also in sense, content (meaning), and connotation.

Lexical Items	Bangla: Odia
Relational terms	bAbA: bapA, mA:mA, mAsi:mAusi, didi:apA, dAdA:bhAinA, boudi:bhAuja, bhAi:bhAi, chele:pilA, meye:jhia,
Pronouns	Ami:mu, tumi:tume, Apni:Apana, tui:tu, se: se
Nouns	lok:loka, ghar:ghara, hAt:hAta, mAthA:munda, pukur: pukhuri, kalA: kadali, am:ama,
Adjectives	bhAla:bhala, bhejA:adA, satya:satya, mithyA:michA
Verbs	yAchhi:yAuchi, khAba:khAiba, balechila:kauthilA, balbe:kAhibe, Asun:Asantu, basun: basantu, bhAlabAse: bhalapAy
Postpositions	kAche:pAkhare, mAjhe:majhire, nice:talare
Indeclinable	ebang:madhya, kintu:kintu

Table 2: Similar vocabulary of Bengali and Oriya

For instance, the list above (Table 2) shows examples where regular vocabulary are similar in sense in Bangla and Odia – two sister languages. To generate a BLD, we use the following strategies on POS tagged TC:

- Retrieve comparable syntactic blocks (e.g. clauses and phrases, etc.) from a TC.
- Extract content words from syntactic blocks (e.g. nouns, adjectives, and verbs).
- Extract function words from syntactic blocks (e.g. pronouns, postpositions, adverbs, etc.).
- Select those lexical items that show similarity in form, meaning, and usage.
- Store those lexical items as translation equivalent units (TEU) in BDL.

Since we do not expect total similarities at morphological, lexical, syntactic, semantic and conceptual level within the two languages (even though languages are closely related), similarities in form, meaning, and usage are enough for selection of TEU.

9. Bilingual Terminology Databank

Collection of Scientific and Technical Terms (STTs) from a TC asks for introspective analysis of a TC. The work is to search through TC to find out STTs which are equivalent or semi-equivalent in TL and TL. While doing this, we need to keep various factors in mind regarding the appropriateness, grammaticality, acceptance and usability of STTs in TL. But the most crucial factor is 'lexical generativity' of the STTs so that many new forms are possible to generate by using various linguistic repertoires available in TL.

TC has another role in choice of appropriate STTs from a list of multiple synonymous STTs that try to represent a particular idea, event, item, and concept. It is observed that the recurrent practice of forming new STTs often goes to such an extreme level that we are at loss to decide which

STT is to select over other suitable candidates. Debate may also arise whether we should generate new STTs or accept STTs of SL already absorbed in TL by regular use and reference. Some STTs are so naturalised that it becomes almost impossible to trace their actual origin. In this case, we have no problem, because these terms are 'universally accepted' in TL. For instance, the Bengali people face no problem in understanding terms like *computer, mobile, calculator, telephone, tram, bus, cycle, taxi, rickshaw, train, machine, pen, pencil, pant, road, station, platform,* etc. because these are accepted in Bangla along with respective items. Their high frequency of use in various text types makes them a part of Bangla vocabulary. There is no need for replacement of these STTs in the TL texts.

A TC is a good resource for selection of appropriate STTs presenting new ideas and concepts. As a TC is made with varieties of texts full of new terms and expressions, it provides valuable resource of context-based example to draw sensible conclusions. Here a TC contributes in two important ways:

(a) It helps to assemble STTs for SL and TL along with information of dates and domains of their entry and usage, and
(b) It supplies all possible native coinage of STTs along with information of domains and frequency of use in SL and TL.

These factors help to determine on relative acceptance or rejection of STTs. Examination of some instances derived from the Hindi-Bangla ILCI-I corpus shows that a TC is highly useful in collection of appropriate STTs – an essential element in translation.

The entire scheme of TC generation and processing may be visualised from the following block diagram (Fig. 9)

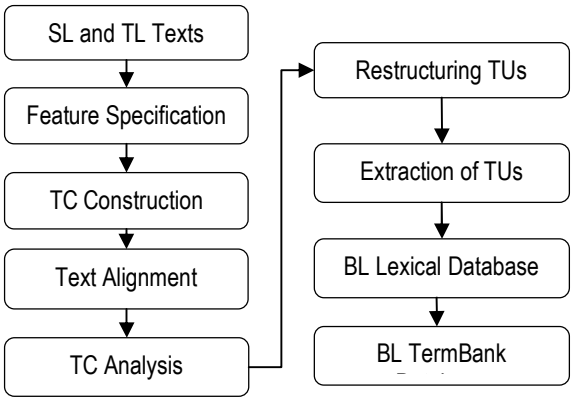

Fig. 89: Scheme of TC generation and processing

10. Conclusion: Value of a Translation Corpus

The question that arises at the time of TC building is: who is going to use it and for what purposes? That means the issue of determining target users is to be dissolved before the work of TC development [36]. But why it is necessary to identify target users? There are some reasons:

- The event of TC generation entails the question of its possible application in various research activities.
- Utility of a TC is not confined within MT. It has equal relevance in general and applied linguistics,
- Each research and application in MT requires specific empirical databases of SL and TL,
- People working in different fields of LT require TC for research and application.
- Form and content of a TC are bound to vary based on users both in linguistics and language technology,
- In language teaching, teachers and instructors require TC for teaching translation courses,
- People studying language variation in SL and TL need TC to initiate their research and investigation,
- Lexicographers and terminologists need TC to extract linguistic and extralinguistic data and information necessary for their works.

These application-specific needs can be easily fulfilled by a TC. Hence, question of selecting target users becomes pertinent in TC construction. However, although prior identification of target users is a prerequisite in TC generation, it does not imply that there is no overlap among target users with regard to utilisation of a TC. In fact, our past experience shows that multifunctionality is an inherent feature of a TC due to which a TC attracts multitudes of users from various fields [15].

This signifies that a TC designed and developed for specific use may be useful for other works. For example, although TC is suitable for lexicographers, it is useful for lexicologists, semanticists, grammarians, texts experts and social scientists. Also it is useful for media persons to cater their needs related to language and society. A TC can be used as a resource for works of language technology as well as an empirical database for mainstream linguistics [36]. In essence, it has application relevance to people interested in SL and TL texts full of exciting features both in content and texture. For Indian languages, a TC is a primary resource, which we need for linguistics and language technology.

References

[1] Altenberg, B. & K. Aijmer. 2000. The English-Swedish parallel corpus: a resource for contrastive research and translation studies. In: C. Mair & M. Hundt (Eds.) *Corpus Linguistics and Linguistics Theory*. Amsterdam-Atlanta, GA: Rodopi. Pp. 15- 33.

[2] Atkins, S., J. Clear, & N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing*. 7(1): 1-16.

[3] Baker, M. 1993. Corpus linguistics and translation studies: implications and applications. In: M. Baker, G. Francis & E. Tognini-Bonelli (E.ds) Text and Technology: In Honour of John Sinclair, Amsterdam: John Benjamins, pp. 233-250.

[4] Baker, M. 1995. Corpora in translation studies: an overview and suggestions for future research. *Target.* 7(2): 223-43.
[5] Baker, M. 1996. Corpus-based translation studies: the challenges that lie ahead. In: H. Somers (Ed) *Terminology, LSP and Translation.* Amsterdam: John Benjamins. Pp. 175-186.
[6] Brown, P. & M. Alii. 1990. A statistical approach to machine translation. *Computational Linguistics*. 16(2): 79-85.
[7] Brown, P. & M. Alii. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*. 19(2): 145-152.
[8] Brown, P., J. Lai, & R. Mercer. 1991. Aligning sentences in parallel corpora. *Proceedings of the 29th Meeting of ACL.* Montreal, Canada.
[9] Brown, R.D. 1999. Adding linguistic knowledge to a lexical example-based translation system. *Proceedings of the MTI-99*, Montreal. Pp. 22-32.
[10] Chanod, J.P. & P. Tapanainen. 1995. Creating a tagset, lexicon and guesser for a French tagger. *Proceedings of the EACL SGDAT Workshop on Form Texts to Tags Issues in Multilingual Languages Analysis*, Dublin. Pp. 58-64.
[11] Chen, K.H & H.H. Chen. 1995. Aligning bilingual corpora especially for language pairs from different families. *Information Sciences Applications*. 4(2): 57-81.
[12] Dagan, I., K.W. Church, and W.A. Gale. 1993. Robust bilingual word alignment for machine-aided translation. *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio.
[13] Gale, W. & K.W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*. 19(1): 75-102.
[14] Geyken, A. 1997. Matching corpus translations with dictionary senses: two case studies. *International Journal of Corpus Linguistics.* 2(1): 1-21.
[15] Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.
[16] Isabelle, P., M. Dymetman, G. Foster, J.M. Jutras, E. Macklovitch, F. Perrault, X. Ren & M. Simard. 1993. Translation analysis & translation automation. *Proceedings of the TMI-93,* Kyoto, Japan.
[17] Kay, M. & M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*. 19(1): 13-27.
[18] Kenny, D. 1997. (Ab)normal translations: a German-English parallel corpus for investigating normalization in translation. In: B. Lewandowsk-Tomaszczyk & P. Janes Melia (eds) *Practical Applications in Language Corpora. PALC '97 Proceedings,* Lódz: Lódz University Press, pp. 387-392.
[19] Kenny, D. 1998. Corpora in translation studies. In: M. Baker (Ed) *Routledge Encyclopaedia of Translation Studies,* London: Routledge, Pp. 50-53.
[20] Kenny, D. 1999. The German-English parallel corpus of literary texts: a resource for translation scholars. *Teanga.* 18: 25-42.
[21] Kenny, D. 2000. Lexical hide-and-seek: looking for creativity in a parallel corpus. In: M. Olohan (Ed) *Intercultural Faultlines. Research Models in Translation Studies I:* Manchester: St. Jerome, pp. 93-104.
[22] Kenny, D. 2000. Translators at play: exploitations of collocational norms in German-English translation. In: B. Dodd (Ed) *Working with German Corpora,* Birmingham: University of Birmingham Press, 143-160.
[23] Klaudy, K. & K. Karoly. 2000. The text-organizing function of lexical repetition in translation. In: M. Olohan (ed.) Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects, Manchester: St. Jerome, pp. 143-159.
[24] Kohn, J. 1996. What can (corpus) linguistics do for translation?. In: K. Klaudy, J. Lambert & A. Sohar (eds.) Translation Studies in Hungary, Budapest: Scholastica, Pp. 39-52.
[25] Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography.* Cambridge: Cambridge University Press.
[26] Mauranen, A. 2000. Strange strings in translated language: a study on corpora. In: M. Olohan (Ed) Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects, Manchester: St. Jerome, Pp. 119-141.
[27] McEnery, T. and M. Oakes. 1996. Sentence and word alignment in the CARTER Project. In: J. Thomas & M. Short (Ed.) *Using Corpora for Language Research*. London: Longman. Pp. 211-233.
[28] Oakes, M. & T. McEnery. 2000. Bilingual text alignment — an overview. In: Botley, S.P., A.M. McEnery, & A. Wilson (Eds.) *Multilingual Corpora in Teaching and Research.* Amsterdam-Atlanta, GA.: Rodopi. Pp. 1-37.
[29] Simard, M., G. Foster, & P. Isabelle. 1992. Using cognates to align sentences in parallel corpora. *Proceedings of TMI-92.* Canadian Workplace Automation Research Center. Montreal.
[30] Simard, M., G. Foster, M-L. Hannan, E. Macklovitch, and P. Plamondon. 2000. Bilingual text alignment: where do we draw the line?. In: Botley, S.P., Tony McEnery & A. Wilson (ed.) *Multilingual Corpora in Teaching and Research.* Amsterdam-Atlanta, GA.: Rodopi. Pp. 38-64.
[31] Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press. Pp. 20.
[32] Stewart, D. 2000. Conventionality, creativity and translated text: implications of electronic corpora in translation. In: M. Olohan (ed) Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects, Manchester: St. Jerome, pp. 73-91.
[33] Stewart, D. 2000. Poor relations and black sheep in translation studies. *Target* 12(2): 205-228.
[34] Summers, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and Design.* Harlow: Longman.
[35] Teubert, W. 2000. Corpus linguistics — a partisan view. *International Journal of Corpus Linguistics*. 4(1): 1-16.
[36] Tymoczko, M. 1998. Computerized corpora and the future of translation studies. *Meta* 43(4): 652-659.
[37] Ulrych, M. 1997. The impact of multilingual parallel concordancing on translation. In: B. Lewandowska-Tomaszczyk and P.J. Melia (eds.) Practical Applications in Language Corpora, Lodz: Lodz University Press, pp. 421-436.
[38] Véronis, J. (ed.). 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora.* Dordrecht: Kluwer Academic Publishers.
[39] Zanettin, F. 2000. Parallel corpora in translation studies: issues in corpus design and analysis. In: M. Olohan (ed.) Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects, Manchester: St. Jerome. Pp., 105-118.