Linguistic tasks on translation corpora for developing resources for manual and machine translation

Niladri Sekhar Dash and Pronomita Basu

In this paper we have made an attempt to discuss some of the theoretical issues related to linguistic tasks to be carried out on translation corpora for developing varieties of linguistic resources and tools required in machine translation. Although attempts have been made for developing translation corpora as well as systems, tools and approaches for machine translation or machine-aided human translation, attention is hardly paid to some of the basic linguistic works, which are indispensable for achieving success in these areas. Even though it is known that generation of translation corpora is an essential part of machine translation, which can contribute to enhance robustness of a translation system, we have not yet focussed on how these translation corpora are going to be used in the work. Keeping this issue open we have addressed some of the basic linguistic activities related to analysis of translation corpora, which include extraction of translational equivalents from corpora; development of bilingual dictionaries; generation of terminology databank; selection of lexical resources; dissolving lexical ambiguities; and generation of a network of grammatical mapping with close reference to lexical mapping, pragmatic and sentential information. In our argument, a machine translation system will become more efficient and robust if it is empowered with linguistic resources developed from linguistic activities carried out on translation corpora.

Keywords: translation corpora, machine translation, translational equivalents, bilingual dictionary, terminology databank, lexical selection, lexical ambiguity, grammatical mapping, lexical mapping, corpora, Bengali.

1. Introduction

Translation corpora, after these are systematically compiled and properly aligned (Dash 2008: 77-81) become accessible for several linguistic activities, which are indispensable for developing linguistic resources required for machine translation. In fact, accurate and effective execution of the linguistic activities on translation corpora becomes useful for generating necessary linguistic resources required not only for machine translation but also for manual translation, since direct utilization of these resources enhances speed, robustness, and accuracy of both types of translation. In our view, the linguistic activities that need to be carried out on translation corpora include:

(a)   Linguistic analysis of translation corpora developed both in the source language and the target language
(b)   Extraction of translational equivalents from the translation corpora
(c)   Development of bilingual dictionaries for source language and target language
(d)   Generation of terminology databank for source language and target language

(e) Selection of appropriate lexical items for translation

(f) Dissolving the problems of lexical ambiguity, and

(g) Developing grammatical mapping for the sentences of source and target language with reference to lexical mapping, pragmatic, and sentential information.

In the following sections of this paper we have addressed all the issues with reference to the Indian language corpora along with a focus on English as the source language and Bengali as the target language.

2. Linguistic Analysis of Translation Corpora

Within the area of machine translation research, the central point of debate has been the question about the level of complexity involved in the task of translation corpora analysis. The general argument is that unless a large number of linguistic phenomena widely occurring in natural language texts are analysed and overtly represented, a high quality machine translation output is unattainable (Isabelle *et al.* 1993). It is also argued that problems like lexical ambiguity and constituent mapping can be dissolved with the help of abundant knowledgebase obtained from corpora and this may be stored in lexicon and grammar of each language involved in translation (Dash 2007: 137-178). This, however, asks for proper execution of rigorous processes of translation corpora analysis that make explicit some or all of the translation correspondences that link up segments of source texts with those of their translations in the target texts.

For the sake of effective linguistic analysis of translation corpora, we argue for using techniques of part-of-speech (POS) tagging of words and shallow parsing of sentences for acquiring better translational outputs. In these works a corpus analyser are supported with standard grammars available in a language or acquired from previously processed corpora. The main objective is to develop bilingual lexical databases by extracting appropriate words, terms, phrases, and idiomatic expressions considered appropriate as translation equivalents. These outputs can be used to increase electronic lexical database of a language as well as for developing materials for language teaching.

The POS tagging can be executed automatically by comparing texts included in the source language and the target language corpora following the probabilistic matching procedure (Chanod and Tapanainen 1995). Although some of the adjectives may be translated in this manner as nouns in the target language or vice-versa, traditional lexical categories mentioned in standard grammars and dictionaries available in the source and the target language can help us to resolve grammatical ambiguities, if they arise. The basic proposition is, at this particular phase, the traditional grammatical categories of words can have strong referential impacts on the quality of POS tagging, as a translation system with fewer grammatical categories of words can have better rate of success than a system with a list of lexical database having exhaustive grammatical categories.

3. Extraction of Translational Equivalents from Translation Corpora

The search for translational equivalents in translation corpora begins with those lexical items that express similar meanings or senses in the both languages. This is usually done manually

at the early stage of translation corpora analysis. Once these items are found in the corpora, these need to be stored in alphabetical order in separate lexical list for future utilization. Usually, translation corpora produce large number of translational equivalent lexical items, which are potential to be used as alternative forms in translation. The basic factor that determines the selection of appropriate equivalent forms is measured on the basis of recurrent patterns of their usage in the corpora. Moreover, equivalent forms are verified with texts of monolingual corpora from which translation corpora are developed. A general scheme for extracting a list of translational equivalents from the bilingual translation corpora is presented below (Figure 1).

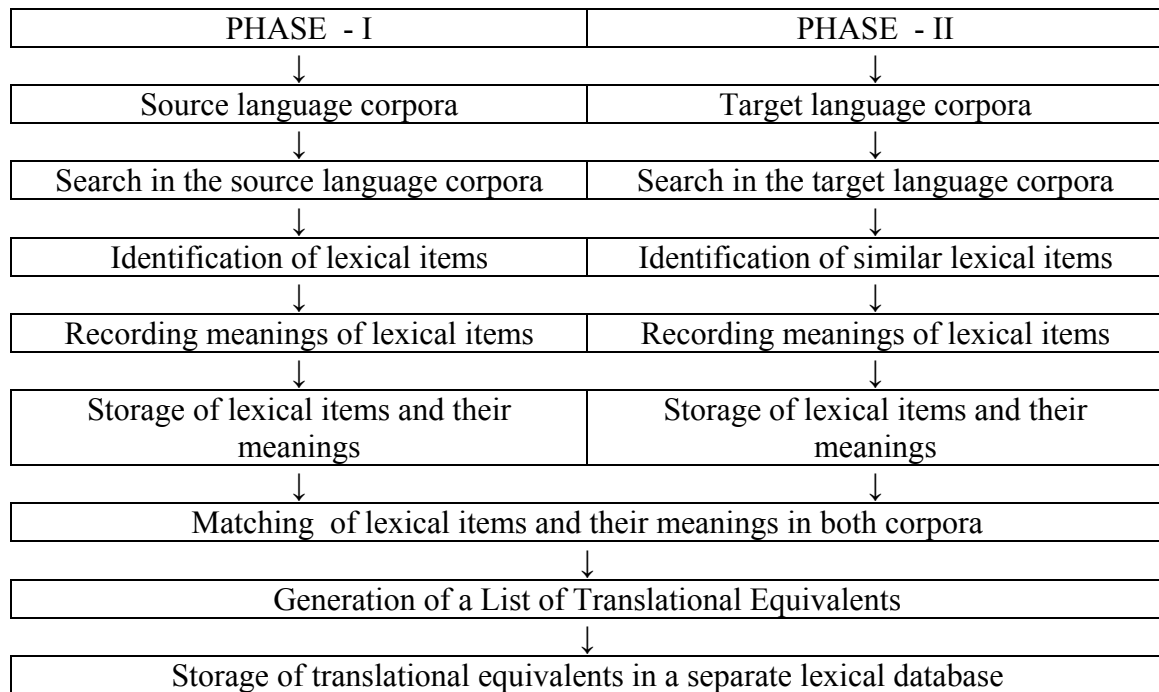| PHASE - I	PHASE - II
↓	↓
Source language corpora	Target language corpora
↓	↓
Search in the source language corpora	Search in the target language corpora
↓	↓
Identification of lexical items	Identification of similar lexical items
↓	↓
Recording meanings of lexical items	Recording meanings of lexical items
↓	↓
Storage of lexical items and their meanings	Storage of lexical items and their meanings
↓	↓
Matching of lexical items and their meanings in both corpora	
↓	
Generation of a List of Translational Equivalents	
↓	
Storage of translational equivalents in a separate lexical database	

Figure 1 *Extraction of translational equivalents from source and target language corpora*

It should be clearly understood that, even between the two closely related languages, translational equivalents seldom mean the same senses in all contexts, since these are seldom distributed in same types of syntactic and grammatical construction. Moreover, semantic connotation and degree of formality of equivalent forms may vary depending on language-specific contexts. Sometimes, a lemma of the target language may fail to be an equivalent to a lemma of the source language, even though they appear equivalent in sense. Although two-way translation may be possible with proper names and scientific terms, it hardly succeeds with ordinary lexical items used in different senses in the corpora (Landau 2001: 319). This implies that in case of autonomous machine translation system, translation of ordinary texts will face severe problems due to difference in senses of lexical items. To overcome the problem, we require manual intervention in selection of translational equivalents to yield better outputs in translation.

With regard to extraction of translational equivalents from translation corpora will not only help machine translation workers but also others engaged in compiling bilingual lexical

databases. In essence, the extraction of translational equivalents from translation corpora will include the following activities:

- Retrieving appropriate translational equivalents for content words such as nouns, adjectives, verbs, adverbs, etc.
- Retrieving appropriate translational equivalents for function words like pronouns, prepositions, postpositions, conjunctions, articles, etc.
- Retrieving multiword translational equivalents such as idioms, phrases, compounds, collocations, and proverbs.
- Learning how the language corpora help to produce translated texts that display 'naturalness' of the target language.
- Creating new translation databases for translating correctly into those languages on which we have limited access.
- Generating terminology databases from new texts, which are neither standardised nor stored in translational databases.

The process of extracting translational equivalents from the source language and the target language and their subsequent verification for authentication with monolingual corpora is described below (Figure 2). Since finding out equivalent units from translation corpora is not an easy task, we need to use various searching methods to trace the comparable units similar in meaning but are often larger and more complex in form than words. Once these are retrieved and implemented into translation platforms, these can facilitate translations more effectively than the customary translation memories. We may also integrate findings from corpora with bilingual dictionaries and term banks to enrich machine translation knowledgebase for the battles ahead.
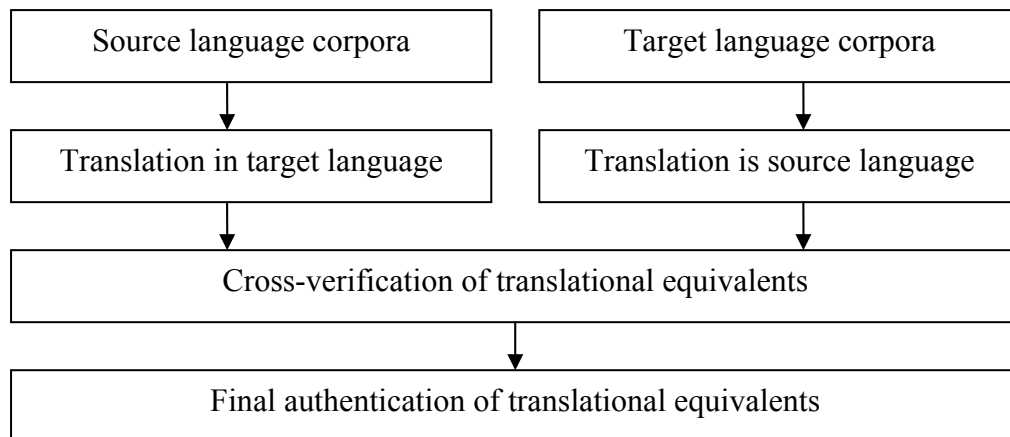

Figure 2 *Verification and authentication of translational equivalents*

Within machine translation research there are great diversities in approaches that use little or no information of traditional linguistics. Also, there are theoretical works that characterize the expressiveness and complexities of different formalisms of languages as well as empirical works that assess modelling and descriptive adequacy across various language pairs. Following these formalism we can use aligned translation corpora to create better equivalents for more accurate translational outputs.

4. Compilation of Bilingual Dictionary

The third impart work of translation corpora analysis is the development of bilingual dictionary the lack of which has been one of the great bottlenecks in present machine translation activities (Geyken 1997). The dictionaries available in market are not good enough to compensate this, since these dictionaries normally do not contain enough information about lexical sub-categorisation, lexical selection restriction, and domains of application of lexical items in the lexical information they provide. Since it is possible now to extract information about sub-categorisation information of lexical items from the POS tagged, there is hardly any problem to include this information in a bilingual dictionary (Brown 1999). Even when POS-tagged corpora are not readily available, bilingual dictionaries can be developed from the untagged corpora available in the source language and the target language.

Words	Bengali words	Oriya words
Relational terms	bābā 'father', mā 'mother', māsi 'aunt', māmā 'uncle',	bapā 'father', mā 'mother', māusi 'aunt', māmu 'uncle'
Pronouns	āmi 'I', tumi 'you (gen.)', āpni 'you (h)', tui 'you (non-h)'	mu 'I', tume 'you (gen.)', āpana 'you (h)', tu 'you' (non-h)'
Nouns	lok 'person', ghar 'home', hāt 'hand', mandir 'temple'	loka 'person', ghara 'home', hāta 'hand', mandira 'temple'
Adjectives	bhāla 'good', manda 'bad', satya 'true', mithyā 'false'	bhala 'good', manda 'bad', satya 'true', michā 'false'
Verbs	ýāchhi 'I/we am/are going', khāba 'I/we shall eat'	ýāuchi 'I/we am/are going', khāibā 'I/we shall eat'
Postpositions	mājhe 'in the middle', pāśe 'beside', upare 'above'	majhire 'in the middle', pāśe 'beside', upare 'above'
Indeclinable	kintu 'but', bā 'or'	kintu 'but', bā 'or'

Table 1 *Translational equivalents from Bengali and Oriya corpora*

Development of a bilingual dictionary is best possible within those languages, which are genealogically linked (e.g., Hindi-Urdu, Bengali-Oriya, and Tamil-Malayalam, etc.), since genealogically related languages share many common properties (both linguistic and non-linguistic) hardly found in non-related languages. Also, there is a large chunk of regular vocabulary similar to each other not only in their orthographic representation but also in sense, content, meaning, and connotation. For example, we have presented above a sample list of similar words, which can be used as suitable translational equivalents for the two genealogically related languages - Bengali and Oriya (Table 1).

For compiling bilingual dictionary, we can use POS tagged corpora in various ways. Albeit there are variations in use of POS tagged corpora, in most cases, the goals are the following:

• Retrieval of large comparable syntactic blocks like clauses, phrases and sentences from bilingual translation corpora.

• Extraction of various subcategorised constituents like subjects, objects, predicates, etc. from the POS tagged corpora.

- Extraction of frequently used nominal, adverbial and adjectival phrases, set phrases, and idiomatic expressions, etc. from the corpora.[1]
- Selection of appropriate lexical items as translational equivalents based on their similarity in form, meaning, and usage in source and target language.

In spite of close linguistic proximities between two genealogically related languages, one cannot expect hundred percent similarity of lexical stock at morphological, lexical, syntactic, semantic, and conceptual level. Therefore, with all information extracted from corpora, a *Core Grammar* is the best solution, which will categorically highlight all kinds of linguistic similarities across the two languages. Although this kind of grammar is yet to be developed among the genealogically related Indian languages, present availability of Indian language corpora recently developed (Dash 2009) can help us to achieve great success in generation of bilingual dictionary for the task at hand.

5. Generation of Terminology Databank

Selection and use of appropriate technical and scientific terms is an important attribute of a good translation system, which asks for proper identification of the terms in source and target language corpora. The primary task of a linguist is to search through the corpora of source language and the target language and to select the appropriate terms that may be considered translational equivalents or near-equivalents for scientific ideas, items and concepts. While doing this, a linguist has to keep in mind various issues regarding the appropriateness, usability, grammaticality and acceptance of the terms in the source language and the target language. However, the most crucial issue is lexical generativity of the terms by which many new words are possible to generate through activation of various word-formation strategies (Aronoff 1981: 25) used in the languages.[2]

A linguist has another important role in choice of an appropriate term from a large list of multiple terms coined by different persons to represent a particular idea, event, item, or concept. It is observed that recurrent practice of forming new technical terms often goes to such an extreme that a machine translation system designer is at loss to decide which term to select over the other suitable terms. Debate also arises whether one should generate new terms or accept terms of the source language already absorbed in the target language by regular usage and reference. It has been also observed that some technical terms are absorbed to such an extent that it becomes almost impossible to trace their actual origin. In that case, a machine translation system designer has no problem, as these terms are already accepted in the target language. For instance, the Bengali people can have no problem in understanding several English terms like *computer, mobile, calculator, telephone, tram, bus, cycle, taxi, rickshaw, train, machine, pen, pencil, pant, road, station, platform,* etc., since these are accepted in Bengali along with the respective items. The high frequency of their use in various texts makes them a part of the Bengali vocabulary. Therefore, there is no need to replace these terms at the time of developing terminology databank.[3]

The translation corpora of the target language are good resources for selection of appropriate technical and scientific terms expressing new concepts and ideas borrowed from the source language. Since these corpora are made up with varieties of texts full of new terms, idioms, expressions and phrases, they can provide valuable resources of context-based

use of terms to draw sensible conclusions. In sum, reference to translation corpora contributes in two important ways.

(a)   They help to collect all technical terms, expressions and phrases entered into the target language along with information of dates and fields of their entry and usage.
(b)   They provide all possible native coinages of terms, expressions and phrases along with respective domains and frequency of their usage in the language.

These two factors can help us to determine on relative acceptance or rejection of the scientific and technical terms. The examination of instances derived from the Bengali text corpus (Dash 2005, Ch. 9) shows how a target language corpus can become highly useful in selection of appropriate terms — an essential part for translation.

6. Selection of Appropriate Lexical Item

The selection of the most appropriate lexical items from the target language corpora as suitable translation equivalents for lexical items of the source language text is another complex task in translation that requires careful interference of linguists well-versed in both the source and target language. It implies that a linguist has to select appropriate terms from a large collection of conceptually similar forms available in target language text, which are nearest in sense to the terms selected from the source language text. A typical example of this is the use of verbs depending on the status of the agent (actor). In Bengali, for example, the use of verb referring 'act of eating' is highly restricted in use depending on the honour of the agent used as the subject of a sentence. Let us consider, for elucidation, the following examples:

1(a)   English: God takes food   (Subject: God)
        Bengali: bhagabān *prasād grahaṇ karen*

1(b)   English: A great man eats   (Subject: great man)
        Bengali: mahāpuruṣ *bhojan karen*

1(c)   English: A gentleman eats   (Subject: gentleman)
        Bengali: bhadralok *āhār karen*

1(d)   English: A common man eats   (Subject: common man)
        Bengali: sādhāraṇ lok *khāy*

1(e)   English: A layman eats   (Subject: laymen)
        Bengali: choṭalok *gele*

If we scrutinise the examples presented above, we can find out that the selection of appropriate equivalent term in Bengali for English *eat* is controlled by the status of agent (i.e., subject) referred to in sentences. If the person in source language text is a *divine man*, then the equivalent term is *prasād grahaṇ karen* (1a), for a *great man* it is *bhojan karen* (1b), for a *gentleman* it is *āhār karen* (1c), for a *common man* it is *khāy* (1d), and for a *layman*

belonging to the lowest social status marked by the scales of social prestige, it is *gele* (1e), although, in all cases, the core meanings of the terms are same: 'to take or eat food'.

In case of technical and scientific terms, selection of appropriate terms becomes far more complicated if the contexts of use of the terms in the source language across the fields of discourse are not considered. For instance, consider the following examples where the English term *deliver* can be translated into Bengali with a wide variation of choice depending on the context of use of the term in the source language (i.e., English).

2(a)   English: Mrs. Sen *delivered* a child in the hospital
        Bengali: Mrs. Sen hāspātāle ekṭi santāner *janma diyechen*

2(b)   English: Prof. Basu *delivered* a lecture on child education
        Bengali: adhyāpak Basu śiśuśikṣār upar ekṭi *baktṛtā dilen*

2(c)   English: The courier boy has delivered the packet
        Bengali: kuriyāyer cheleṭi pyākeṭṭi *põuche diyeche*

2(d)   English: The bowler delivered a googly in the last over
        Bengali: śeṣ obhāre bolārṭi ekṭi gugli *bal karlo*

The examples cited above shows that the English term *deliver* carries four different senses in the source language, which have to be translated in an appropriate manner into the target language taking into consideration the context of use of the term. In the field of childbirth, the most appropriate term in Bengali is *janma deoyā,* in lecture in the class or at a mass rally it is *baktṛtā deoyā,* in postal distribution or supply of goods it is *põuche deoyā,* and in the game of cricket it is *bal karā.* The most interesting thing is that what it means in the field of childbirth is not same in supply of goods, lecture in class, and in the game of cricket. This signifies that by considering the domain of use of terms in the source language, we have to select the appropriate terms in the target language. In most cases, evidences collected from corpora can legitimatize the beautility and acceptance of translation outputs.

The primary task of a linguist is to find out the appropriate lexical items considering various factors latently involved within the two languages considered for translation. The examples show that lexical selection has to be taken care of for generating sensible translation outputs. Although the problem is handled elegantly in manual translation, it is a great hurdle in machine translation. The best way to overcome the problem in machine translation is to enlist beforehand all semantically similar forms in a separate lexical list within a machine readable dictionary (MRD) to be accessed in later in translation. Such a lexical database is easy to extract from translation corpora in both manual and machine translation activities.

Usually, there are several domains within a MRD — a resource capable to provide all relevant information about the selection of lexical items. Therefore, whenever we analyse translation corpora, we need to identify the subject area to which the text belongs for storing the list of terms related to this domain. For instance, when we analyse an English text related to mass media, it makes sense that we select the relevant terms from the English text and store them in a separate lexical database. Similarly, we can execute the same kind of task on the target language text to collect and store lexical terms in a lexical database in the subject area 'mass media'. However, complexities will arise when a single term of the source

language will denoted different senses in the target language. For example, English term *inform* can have several senses in Bengali depending on the domain of use of term, as the list (Table 2) shows. The examples imply that a translator has to select the most appropriate lexical item considering the domain, to which he is going to translate the source language text. Until this issue is systematically dealt with, appropriate output cannot be achieved in the target language.

English word	Bengali equivalents (Selection is based on domain)
↓	↓
inform	jānāno (Giving general news or information to people)
inform	raṭāno (Spreading rumour or false information around)
inform	pracār (Canvassing information for one and all)
inform	bijñāpan (Advertising an item or product, etc.)
inform	sampracār (Broadcast and telecast of news and information)
inform	bijñapti (Government circulars or notices for all people)
inform	ghoṣaṇā (Declaring an event of public reference and interest)
inform	Dhārābhāṣya (Running commentary of games and sports)
inform	istehār (Campaign and propaganda of political isms)
inform	Pratibedan (Reporting a piece of news in papers)
inform	kīrtan (Highlighting someone's achievement)

Table 2 *Selection of lexical items based on the domain of use of items*

The selection of appropriate phrases, set expressions, idiomatic expressions, and proverbial statements is another complex task which demands careful search through bilingual translation corpora for collection appropriate translational equivalent forms (Geyken 1997). The best solution is to generate a bilingual database for these resources and store it in MRD for future usage. For instance, given below is a sample list of idioms and proverbial forms (Table 3) collected from English corpora with their translational equivalents obtained from the Bengali text corpus (Dash 2009).

English idioms and phrases	Bengali equivalent forms
Apple of one's eye	chokher maṇi
Crocodile's tear	kumīrer kannā
A bedlam	narak guljār karā
Blue blood	nīl rakta
Bolt from the blue	binā meghe bajrapāt
Paddle your own canoe	nijer carkāy tel deoyā
On cloud nine	saptam svarge
A cock and bull story	āsāre galpa
A white elephant	śvet hastī
By hook or by crook	ẏena tena prakāreṇa
Horns of a dilemma	ubhay saṅkaṭ
To add insult to injury	kāṭā ghāye nuṇer chiṭe
To carry coal to New Castle	telā māthāy tel deoyā
Once in a blue moon	kāle bhadre

In the nick of time	śeṣ samaye
Pour oil on troubled water	agnite ghṛtāhuti deoyā
Raining in cats and dogs	muṣaldhāre bṛiṣṭipāt
Black sheep	Kulāṅgār
Writing on the wall	deoyāler likhan
To cry in wilderness	Araṇye rodan

Table 3 *Phrases and idioms taken from English and Bengali corpora*

Generation of such a list of idioms, phrases and proverbs from the source language and the target language corpora enhances quality and robustness of machine translation, since this database can be used to capture the figurative senses of expressions found in the source language and the target language for stylistic representation as well as for better comprehension of translational outputs.

7. Dissolving Lexical Ambiguity

In normal situation, a linguistic communication transfers information from the producer to the receiver by using language as a vehicle. Sometimes, however, this transfer of information is not free from ambiguity ─ one of the most common yet highly complex phenomena of a natural language (Dash 2005). It is observed that ambiguity may arise due to several factors, one of which is inadequacy in the *internal meaning* associated with a lexical item or due to structure of an utterance used in a particular event of communication. Thus, ambiguity is classified into three broad types.[4]

(a) Lexical ambiguity   (e.g., *They went to the bank*),
(b) Referential ambiguity   (e.g., *He loves his wife*), and
(c) Syntactic ambiguity   (e.g., *Time flies like an arrow*).

In case of lexical ambiguity, a speaker uses a single word to refer to more than one sense, event, idea, or concept. This creates problem for a listener in capturing the actual intended meaning of a word. The problem intensifies further when the language of the speaker differs from that of a listener. Since a machine translation system is intended to be developed with some perceptions of mental representation of a speaker, it is limited by words and sentences used by the speaker.

To overcome the problem, we need to map the source language lexicon with the equivalent in the target language lexicon, which will be used as an appropriate frame in particular contexts of text representation. In some situations, the target language may not have an equivalent lexical item, which is fit to represent the actual sense of a term used in source language. In such cases, we have to either depend on multi-word units (such as, multiword units, compounds, idioms, phrases, and clauses, etc.) or use the explanatory addendum to deal with such situations.

For dissolving lexical ambiguities, the easier solution is to find out methods for locating contexts of use of words as well as analyse the contextual profiles of the lexical items. Recent experiments with translation corpora (Ravin and Leacock 2000, Cuyckens and Zawada 2001) reveal that lexical ambiguity is mostly resulted from multiple readings of a

11

word, and these readings most often differ in selection of lexical, syntactic and semantic features of words, such as, tense, aspect, modality, case, number, gender, idiomatic readings, figurative usage and so on. As avid supporters of the *Corpus-Based Machine Translation* system (Dash 2007: Ch. 5), we argue to overcome the problem of lexical ambiguity with reference to the context of their occurrence in a piece of text collected in corpora. In that case we need to identify the large number of ambiguous words that usually occur in natural texts and analyse them properly as well as mark them accordingly to achieve higher accuracy in translation. If possible, we should analyse the ambiguous words with information gathered from translated texts and with semantic information stored in the MRD.[5]

Taking cues from domain-specific translation outputs, we can go for deep semantic analysis of words which, however, is not always required for translation. For instance, English *head* may be translated in Bengali as *māthā,* no matter in which of the many senses the word is used in the source language text. Therefore, it is better that we go for a simple word analysis scheme and use a more direct source language to target language substitution in place of deep semantic analysis of ambiguous words. At certain contexts, it is possible and necessary to ignore lexical ambiguities with a hope that the same ambiguity will be carried to the target language. This is useful in those cases where we aim at dealing with only a pair of related languages within a highly restricted domain. However, since analysis of lexical ambiguities is meant to produce non-ambiguous representation in the target language, we cannot ignore it in case of translation of texts belonging to general domains (Isabelle and Bourbeau 1985: 21).

8. Defining the Pattern of Grammatical Mapping

The type of transformation we referred to in the following example (3a) is known as *grammatical mapping* in translation. Here, words of source language text are 'mapped' with words of target language text to obtain meaningful translation outputs. In machine translation, there are various ways for mapping of linguistic forms used in a language (e.g., morphological, lexical, grammatical, phrasal, clausal, etc.), the most common one of which is grammatical mapping related to verb forms within the two languages considered for translation.

The issue of grammatical mapping becomes relevant in machine translation between the two languages, which are different in lexical ordering in sentence formation. In the present context, while we talk about machine translation from English to Bengali, this becomes optimised in proportion, since while English has SVO structure (e.g., *He eats rice*) in sentence formation, Bengali has SOV structure (e.g., *se bhāt khāy*) within the same framework. Therefore, grammatical mapping and reordering of lexical items is required for producing the acceptable outputs in Bengali. For example, consider the sentence given below (3a) as well as the mapping (Figure 3).

3(a)　　English: All his efforts ended in smoke
　　　　Bengali: tār samasta ceṣṭā byārtha hala

English	All (a)	his (b)	efforts (c)	ended (d)	in (e)	smoke (f)
Literal output	samasta (1)	tār (2)	ceṣṭā (3)	śeṣ hala (4)	-te (5)	dhõyā (6)
Actual output	tār (2)	samasta (1)	ceṣṭā (3)	byārtha (4-5-	hala -6)	
Bengali	(2)	(1)	(3)	(7)		

Figure 3 *Grammatical mapping between English and Bengali sentences*

Figure 3 shows that for achieving accurate output with acceptable word order in the target language, words used in the sentence of the source language text need to be mapped with words used in the target language in the following manner:

(a) Lexical Mapping:

English [a]   = Bengali [1] (word to word mapping)
English [b]   = Bengali [2] (word to word mapping)
English [c]   = Bengali [3] (word to word mapping)
English [d]   = Bengali [4] (group of words for single word)
English [e]   = Bengali [5] (use of case marker for preposition)
English [f]   = Bengali [6] (word to word mapping)

However, we must understand that lexical mapping is not the only solution by which we can obtain accurate translation output in the target language. The input sentence of the source language text (English) also contains an idiomatic expression (i.e., *ended in smoke*), which requires some pragmatic knowledge to find a similar idiomatic expression in the target language (Bengali) to achieve greater accuracy in translation. Therefore, we need to employ pragmatic knowledgebase to select appropriate equivalent idiomatic expression from the target language texts in the following manner:

(b) Pragmatic Information:

English: [d-e-f] (an idiomatic expression)
Bengali: [7 (<4-5-6)] (similar translation equivalent)

The machine translation system needs the information that *ended in smoke* in the source language text has to be translated as *byārtha hala* in target language text when the expression is used in idiomatic sense. After the selection of appropriate and equivalent idiomatic expression from the target language text, we are in a position to claim that the output sentence is grammatically mapped to such an extent that intended sense of the input sentence is maximally represented in the output. After this comes the stage of sequential ordering of words in the sentence of the target language text so that the output sentence becomes grammatically valid in the target language text. For this, the following information becomes handy.

(c) Sentential Information:

Sequence in English sentence: [a + b + c + (d + e + f)]
Sequence in Bengali sentence: [2 + 1 + 3 + 7 (<4+5+6)]

What it shows that after proper application of several linguistic strategies like lexical mapping, selection of appropriate idiomatic expression (if any), and sequential ordering, we finally get *tār samasta ceṣṭā byārtha hala* as a valid translation output in the target language (Bengali). Such grammatical mapping from one structure to another is highly useful for producing appropriate translations which are accepted as 'normal' sentences in the target language.

In the task of analysing sentence structures of the source and target language texts, translated corpora are particularly useful, which we can use to map the sequence of word order (at linear level) between the source and target language texts to yield information about the structure of NPs, APs, VPs, PPs, and other properties used in the languages considered for translation.

No	English	Bengali
(a)	*in* hands	hāte (< hāt$_{[n]}$ + -e$_{[loc\ case]}$)
(b)	*with* person	loker (< lok$_{[n]}$ + -er$_{[gen\ case]}$) + saṅge$_{[post-p]}$)
(c)	*by* mistake	bhulbaśata (< bhul$_{[n]}$ + baśata$_{[Adv]}$)
(d)	*in* house	ghare (< ghar$_{[n]}$ + -e$_{[loc\ case]}$)
(e)	*in* house	gharer madhye (< ghar$_{[n]}$ + -er$_{[gen\ case]}$ + madhye$_{[pp]}$)
(f)	*at* night	rāte (< rāt$_{[n]}$ + -e$_{[loc\ case]}$)

Table 4 *Mapping of preposition and postposition between English and Bengali*

The grammatical mapping also highlights the lexical interface underlying the surface structures of sentences and the nature of lexical dependency underlying the surface constructions in the source and target language texts. For example, in case of translating prepositions (e.g., *at, for, up, by, in, of, with*, etc.) used in English, we need to decide whether we should use postpositions or case markers to have correct outputs in Bengali. For elucidation, consider the examples given above (Table 4).

The above examples (Table 4) show that in English, prepositions are used before nouns to evoke case relation (a, d, f), adverbial sense (c) and postpositional sense (b, e). However, in Bengali, these senses are achieved by using case markers (a, d, and f), postpositions (b and c), or both case markers and postpositions (e). Also the table provides information about their position in respect to the content words with which these functional words are attached to generate the appropriate outputs (Figure 4).

English: It   is   in   his   hand

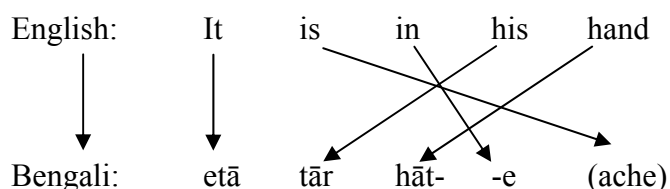Bengali: eṭā   tār   hāt-   -e   (ache)

Figure 4 *Position of postposition with respect to content words*

From the examples and analyses presented above it is almost clear that the task of proper grammatical mapping is an essential linguistic part of machine translation, which cannot be ignored if we intend to achieve even marginal success in this area.

9. The System Module

Linguistic analysis of translation corpora is an indispensable task that helps to develop necessary resources to get better translation outputs. It involvers several works such as analysis of translation corpora, development of bilingual lexical database, extraction of translational equivalents, generation of terminology database, making appropriate lexical selection, dissolving lexical ambiguity, and developing suitable grammatical mapping between the languages. Also we need to determine which linguistic units of the target language are more likely to correlate with of linguistic units of the source language.

The most sensible method for making these activities feasible is to analyse translation corpora as *training corpora,* since analysis will help us to find out all kinds of linguistic information required in translation. It is, however, not necessary to analyse all sentences used in translation corpora, as analysis of a set of token sentences will serve and suffice initial purposes. After analysis of translation corpora, we shall obtain linguistic resources of three types

(a)   Examples of *strong matching* where linguistic items such as words, terms, phrases, idioms, and sentences, etc. are similar in form, meaning, and usage both in source and target languages.
(b)   Examples of *approximate matching* where linguistic items are similar in meaning but different in form and usage in the two languages.
(c)   Examples of *weak matching* where linguistic items are different in form, meaning and usage in the two languages.

In case of translating texts from Bengali to Oriya, most of the linguistic items will belong to *strong matching*, since the language are genealogical linked and ire originated from the source mother. But, in case of translating texts from English to Bengali, most of the linguistic items will belong to *weak matching*, as languages belong to two different typologies. In such a situation, if fifty percent similarity is obtained from the translation corpora of the two languages, one can go for using them in translation. In essence, systematic analysis of translation corpora, methodical extraction linguistic resources from corpora as well as judicious application of outputs will make machine translation as realized dream.

10. Conclusion

Machine translation is an applied field, the impetus for progress of which mostly comes from elegant handing of linguistic and extralinguistic resources. Since this is highly specialized domain, it is a test bed for theories and applications related to linguistics, language technology, and artificial intelligence. While working in this domain we want to verify if theories of syntax, semantics, and discourse are compatible to it, if standard lexicon and grammar are fruitfully utilised in it, and if algorithms of text processing, parsing, word sense

disambiguation, machine learning, and pragmatic interpretations are applicable to it. Thus, machine translation turns into an ideal field for comprehensive evaluation of various theories of language as well as for development and testing a wider range of linguistic phenomena abundant within natural languages.

Since translation corpora are indispensable resources both in manual and machine translation, we can focus only on processing, analysis, and access of corpora with an assumption that analysable translation corpora (properly aligned and readily comparable) are already compiled and provided to the people involved in the task. The activities we have proposed here are not only suitable for machine translation from English to Bengali, but also for any other languages included in the task of machine translation. These are also applicable for most of the Indian languages, which are interested to develop useful translation corpora between English and the Indian languages for similar purposes. The utilities of the linguistic resources generated from analysis of translation corpora can be further attested in language teaching, electronic dictionary compilation, machine learning, grammar development, and language cognition.

For Indian languages, translation corpora are basic requirements, which are however, yet to be developed for any two genealogically related languages. We, therefore, urgently need to develop translation corpora, which will be accessible for developing machine translation system for the Indian languages. In fact, availability of translation corpora in Indian languages will make significant contribution to supplement traditional methods of translation, because information obtained from analysis of translation corpora will minimises distance between the Indian languages. The secret motive behind this work is to argue for development of translation corpora in the Indian languages so that we can take a step forward towards development of a machine translation system for the Indian languages.

Notes

[1] There are several identical adverbial and adjectival phrases, idiomatic expressions and set phrases, etc. in both the corpora such as *gatānugatik jībandhārā* 'stereotype life', *biśeṣbhābe paricita* 'specially known', *satata paribartanśīl* 'ever changing', *sāṅskṛitik anuṣṭhān* 'cultural function', etc. These can be put to a list of 'lexical collocation' of a bilingual dictionary for better access and application in machine translation and other linguistic works.

[2] There are several word formation strategies (e.g., derivation, inflection, affixation, analogy, compounding, loan translation, blending, etc.) for generating new lexical items in a language. For instance, consider the process of word formation in Bengali following English by analogy: *electric = bidyut, electrical = baidyutik, electronic = baidyutin*, etc.

[3] From a simple calculation of English terms in Bengali vocabulary obtained from the Bengali text corpus shows that there are more than thousand English terms, which are regularly used by Bengali people. Surprisingly, none of these terms are allowed to enter in standard Bengali dictionaries. This shows the lack of proper information about the language use on the part of dictionary makers. We, therefore, ask for immediate revision of standard Bengali dictionaries with English words and terms collected from the modern Bengali corpus databases.

[4] In machine translation ambiguities are referred to as examples of divergence. Some discussions on divergence are available in the work of Dorr (1994). Divergence in Hindi texts is addressed in Gupta and Chatterjee (2003).

[5] The rationalists argue that such a work of information acquisition from translated texts is neither realistic nor feasible (Grishman and Kosaka 1992). They also argue that, "it must be kept in mind that a translation process does not necessarily require full understanding of the texts. Ambiguities may be preserved during a translation ─ and they should be presented to the users for resolution" (Ari, Rimon and Berry 1988).

References

ARI, Ben, RIMON, Martin, BERRY, Daniel Michael 1988. Translational Ambiguity Rephrased. In *Proceedings of 2nd International Conference on Theoretical and Methodological Issues in Machine Translation.* Pittsburgh, pp. 1-11.

ARONOFF, Mark. 1981. *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press, 1981.

BROWN, Robert D. 1999. Adding linguistic knowledge to a lexical example-based translation System. In *Proceedings of the MTI-99*, Montreal, Canada, pp. 22-32.

CHANOD, Jean-Pierre, TAPANAINEN, Pasi. 1995. Creating a tagset, lexicon and guesser for a French tagger. In *Proceedings of the EACL SGDAT Workshop on Form Texts to Tags Issues in Multilingual Languages Analysis*, Dublin. pp. 58-64.

CUYCKENS, Hubert and ZAWADA, Britta (eds.). 2001. *Polysemy in Cognitive Linguistics.* Amsterdam/Philadelphia: John Benjamins, 2001.

DASH, Niladri Sekhar. 2005. Corpus-based machine translation across Indian languages: from theory to practice. In *Language In India*. 5(7): 12-35.

DASH, Niladri Sekhar. 2005. Role of context in word sense disambiguation. In *Indian Linguistics*. 66(1-4): 159-175.

DASH, Niladri Sekhar. 2005. *Corpus Linguistics and Language Technology:* With Reference to Indian Languages. New Delhi: Mittal Publication, 2005.

DASH, Niladri Sekhar. 2007. *Language Corpora and Applied Linguistics*. Kolkata: Sahitya Samsad, 2007.

DASH, Niladri Sekhar. 2008. *Corpus Linguistics: An Introduction*. New Delhi: Pearson-Longman, 2008.

DASH, Niladri Sekhar. 2009. *Corpus-based Analysis of the Bengali Language*. Saarbrucken, Germany: Verlag Dr Muller Publications, 2009.

DORR, Bonnie Jean. 1994. Machine translation divergences: a formal description and proposed solution. In *Computational Linguistics*. 20(4): 597-633.

GEYKEN, Alexander. 1997. Matching corpus translations with dictionary senses: two case studies.In *International Journal of Corpus Linguistics*. 2(1): 1-21.

GRISHMAN, Ralph, KOSAKA, Margaret. 1992. Combining rationalist and empiricist approaches to machine translation. In *Proceedings of the MTI-92,* Montreal, pp. 263-274.

GUPTA, Deepa, CHATTERJEE, Niladri. 2003. Divergence in English to Hindi machine translation: some studies. In *International Journal of Translation.* 15(2): 5-24.

ISABELLE, Pierre, BOURBEAU, Laurent. 1985. TAUM-AVIATION: its technical features and some expert mental results. In *Computational Linguistics*. 11(1): 18-27.

ISABELLE, Pierre, DYMETMAN, Marc, FOSTER, George, JUTRAS, Jean-Marc, MACKLOVITCH, Elliott, PERRAULT, Francois, REN, Xiaobo, SIMARD, Michel. 1993. Translation analysis and translation automation. In *Proceedings of the TMI-93,* Kyoto, Japan, pp. 22-27.

LANDAU, Sidney, I. 2001. *Dictionaries: The Art and Craft of Lexicography.* Revised Second Edition. Cambridge: Cambridge University Press, 2001.

RAVIN, Yael, LEACOCK, Claudia (eds.). 2000. *Ploysemy: Theoretical and Computational Approaches.* New York: Oxford University Press Inc., 2000.

Dr Niladri Sekhar Dash
Linguistic Research Unit
Indian Statistical Institute
203, Barrackpore Trunk Road
Kolkata - 700108, West Bengal, India
niladri@isical.ac.in
Homepage: http://www.isical.ac.in/~niladri

Ms Pronomita Basu
Dept. of Linguistics
University of Calcutta
College Street Campus
Kolkata – 700 073
West Bengal, India
basuprono@gmail.com