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

Niladri Sekhar Dash &
Arulmozi Selvraj

Abstract

We address in this paper some theoretical and practical issues relating to generation, processing, and management of Parallel Translation Corpus (PTC) in Indian languages, which is under development in a consortium-mode project (ILCI-II)[1] under the aegis of DeitY, Govt. of India. These issues are discussed here for the first time keeping in mind the ready application of PTC in various domains of linguistics including computational linguistics, Natural Language Processing, applied linguistics, lexicography, translation, language description, etc. In a normative manner, we define what is a PTC; describe the process of its construction; identify its features; exemplify the processes of text alignment in PTC; discuss the methods of text analysis; propose for restructuring of translational units; define the process of extraction of translational equivalents; propose for generation of bilingual lexical database and Term Bank from a structured PTC; and finally identify the areas where a PTC and information extracted from it may be utilized. Since construction of PTC 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 present work that is trying to construct PTC for some Indian languages for future reference and application.

Translation Today Vol.10, Issue-I, June 2016

1. What is a Parallel Translation Corpus?

The term Parallel Translation Corpus (PTC) in principle suggests that it contains texts in one language and their translations in other languages. It is entitled to include bilingual (and multilingual) texts as well as texts that fit under translation. A PTC, by virtue of its character and composition, is made of two parts: a text from a source language (SL) and its translation from target languages (TL) (Hunston 2002, Kohn 1996, Zanettin 2000). Although a PTC is normally bilingual and bidirectional (Oakes and McEnery 2000), it can be multilingual and multi-directional as well (Ulrych 1997), 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 TLs (Fig. 1).

Assamese	↔		↔	Bangla
Kashmiri	↔		↔	Odia
Punjabi	↔	H	↔	Konkani
Urdu	↔	I	↔	Telugu
Gujarati	↔	N	↔	Tamil
Marathi	↔	D	↔	Kannada
Bodo	↔	I	↔	Malayalam
English	↔		↔	Nepali

Fig. 1: Hindi as SL and other Indian Languages as TLs

The issue of multi-directionality can be understood if all the target languages of the group are able to establish linguistic links with one another as they have linked up with the SL. Since the ILCI-I PTC has not yet tried to venture into this direction, it is sensible to confine the present discussion within a scheme of bilingualism and bi-directionality, with, for example, Hindi as SL and Bangla as TL to understand theoretical and practical issues involved in its structure, composition, construction, processing, and utilization of PTC. Hence forth, our discussion will sail in this direction only.

Theoretically, a PTC is supposed to keep meaning and function of words and phrases constant across the languages (Kenny 1998), although alternation in structure (i.e., sequential order of words and phrases) is a permissible deviation. A PTC offers an ideal resource for comparing realisation of meanings (and structures) in two different languages under identical situations (Baker 1993). Also, it makes possible to discover the cross-linguistic variants, i.e., alternative renderings of meanings and concepts in TL (Baker 1995). Thus a PTC becomes highly useful for cross-language analysis and formulation of comparable lexical databases necessary for translation (Altenberg and Aijmer 2000, Kenny 2000, Mauranen 2000)

Since a PTC contains texts from one language and its translations in another language, it may be viewed as a sub-type of a parallel corpus, which, in principle, requires its elements to be maximally comparable to each other (Oakes and McEnery 2000). Therefore, it is wiser to consider a PTC as a special corpus, which is identical in genre, similar in text type, uniform in format, parallel in composition, identical in text content, comparable to each other, and specific in utility (Stewart 2000, Ulrych 1997).

2. Construction of a Parallel Translation Corpus

The construction of a PTC is a highly complicated task. It requires careful manipulation of both SL and TL texts (Kenny 1997, Kenny 1998). Theoretically, a PTC should be made in such a way that it is suitable to combine advantages of both comparable and parallel corpora (Atkins, Clear and Ostler 1992). Text samples from both the languages should be matched as far as possible in terms of text type, subject matter, purpose, and register (Altenberg and Aijmer 2000). The structure of a PTC within any two languages may be envisaged in the following manner keeping in mind the basic aim of the task and the components to be integrated within a PTC (Fig. 2).

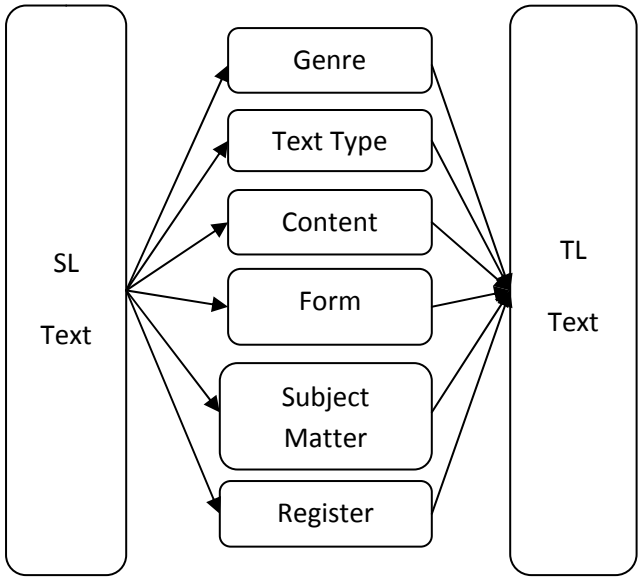

Fig. 2: Construction and composition of a PTC

The diagram (Fig. 2) given above shows that a PTC 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 a parallel corpus cannot be used as a PTC until and unless it follows the conditions of its construction stated above. Therefore, selection of text samples for constructing a PTC needs to be guided by the following principles (Summers 1991):

[1] Written texts are included in a PTC. Texts obtained from speech are ignored, since the present state of PTC targets written texts only.

[2] Texts should reflect on the contemporary language use although texts of earlier era may have relevance in case generating PTC for historical texts.

[3] Texts should be restricted to specific subject. It should include texts from specific domains of language use.

[4] Texts from both the SL and the TL should be maximally comparable. They should be well matched 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.

[5] Texts must consist of fairly large and coherent extracts obtained from the beginning to end of a breaking point (e.g. chapter, section, paragraph, etc.)

[6] Texts should faithfully represent regular and special linguistic forms and elements of the SL and the TL. They should be large in size to encompass maximum varieties in content. Lexical varieties should be high in a PTC.

[7] Texts should faithfully preserve domain-specific words, terms, idioms, phrases, and other lexical elements. The text samples used in PTC should be authentic and referential for future verification and validation.

[8] Texts should be available in machine-readable form for ready access and reference by all end users. The end users should use

language data in multiple tasks for statistical sampling, text alignment, lexical database generation, text processing and translation, etc.

[9] Text samples should be preserved either in annotated or non-annotated version. A POS tagged PTC is a better resource than a non-POS tagged one.

[10] Linguistic and extralinguistic information should be captured in a systematic way so that the end users can access information easily for future reference and validation.

Given below (Fig. 3) is a sample of Hind-Bangla parallel 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 धूमपान, 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: A sample of Hind-Bangla parallel translation corpus

3. Features of a Parallel Translation Corpus

A PTC is assumed to have certain default features, which might vary for other types of corpus (Stewart 2000). That means, a PTC which does not possess these default features may be put outside its scope due to deviation from the norm. By all means, a PTC, if it is not defined otherwise, should possess the following features:

3.1 Quantity of Data

A PTC should be big enough with large collection of texts from the SL and the TL. Larger amount of text data facilitates accessibility and reliability of translation. The number of sentences included in a PTC will determine its quantity. Since the primary goal of a PTC is to include texts for translation, it should not be restricted with fixed number of sentences. In general, the issue of size of a PTC is related to the amount of text samples included in it. In actuality, it is the total number of sentences that actually determines its size of a PTC (Sinclair 1991: 20). A PTC that includes more number of sentences is considered more suitable, since size is an important issue in PTC based linguistic works.

Making a PTC large is linked with number of 'tokens' and 'types' included in a PTC as well as with the decision of how many texts would be in a PTC; how many sentences would be in each text; and how many words would be in each sentence (Baker 1996). A small PTC, due to its limited number of texts, fails to provide some advantages, which a large PTC can easily provide. We observe that a large PTC generally presents the following advantages:

[1] A large PTC presents better scope for variation of texts.

[2] It provides better spectrum of the patterns of lexical and syntactic usages in SL and TL.

[3] It confirms increment in number of textual citations that provide scopes for systematic classification of linguistic items in terms of their usage and meaning.

[4] It assures better opportunity for obtaining all kinds of statistical results far more faithfully for making various correct observations.

[5] It gives wider spectrum for studying patterns of use of individual words and sentences. This helps to make generalization about syntactic structures of SL and TL.

[6] It helps to understand the patterns of use of multiword units like compounds, collocations, phrases, idioms and proverbs, etc. in SL and TL.

[7] It helps to identify coinage of new words and terms, locate their fields of use, find variations of sense of terms, and track patterns of their usage in texts, etc.

[8] It gives scope for faithful analysis of usage of technical and scientific terms – a real challenge in translation.

A large PTC 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 a direct effect on validity and reliability of a PTC. Also, it ensures diversity of SL and TL from which it is made. Since a PTC is nothing more than a minuscule sample of SL and TL varieties, in case of qualitative authentication of SL and TL properties, it may become useless if it is not large enough in respect of the amount of data (Stewart 2000).

3.2 Quality of Text

Quality relates to authenticity. That means texts should be collected from genuine communications of people from their normal

discourse. The primary role of a PTC generator is to acquire data for the purpose of PTC generation in which (s)he has no liberty to alter, modify or distort the actual image of the SL text (s)he is collecting. Also, (s)he has no right to add information from her/his 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 PTC developer will collect data faithfully following some predefined principles proposed for the task. If (s)he tries to interpolate in any way within the body of the text, (s)he will not only damage the actual picture of the text, but also will damage heavily the subsequent analysis of the 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 PTC, a PTC developer should strictly observe the following conditions:

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

3.3 Text Representation

A PTC 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 developed, since future analysis and investigation of PTC may ask for verification and authentication of information from a PTC representing the language. For example, if we want to develop a Hindi-Bangla PTC, which is meant to be adequately representative of a domain of the languages, it should be kept in mind that data should be collected in equal proportion so that the PTC 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. These should be maximally representative with regard to domains. A PTC 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, text 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 PTC 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 the PTC. 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 PTC 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 PTC generation:

- Number of texts available in the SL may be more than that of the TL.
- Collection of equal number of sentences both from the SL and the TL may not be an easy task.

- Parity in number of sentences is deceptive, because sentences never occur in equal number in the SL and the TL.
- A sentence in the SL may be broken into two or more sentences in the TL. Reversely, several sentences in the SL may be merged into one sentence in the TL.
- Equal number of sentences cannot be collected from the SL and the TL in a uniform manner, since size of text varies.

3.6 Retrievability

The work of PTC generation does not end with compilation of texts. It also involves formatting the text in a suitable format so that the data becomes easily retrievable by the end users. That means data stored in a PTC should be made easily retrievable for end users. Anybody interested in a PTC should be able to extract relevant information from it. This directs our attention towards the techniques and tools used for preserving PTC in digital format. The present technology has made it possible for us to generate a PTC 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 PTC users, since the utility of a PTC is not confined to computer-trained people only. A PTC is made for one and all (e.g. 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 PTC, but need a PTC to address their needs. Therefore, PTC must be stored in an easy and simple format so that common people can use it.

3.7 Verifiability

Texts collected in a PTC 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 PTC 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 PTC to represent the SL and the TL should honestly register and reflect on the actual patterns of language use. To address this need, a PTC 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 a PTC is actually a faithful reflection of the SL and the TL. For instance, when we develop a PTC for Hindi and Bangla we are careful that texts stored in the PTC qualify to reflect properly on the respective languages. A PTC thus attests its authenticity and validity.

3.8 Augmentation

A PTC 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 PTCs are synchronic, we should take effort to make diachronic PTCs so that we find a better picture of the languages involved in the game. A synchronic PTC, by addition of texts, may become diachronic in composition. This can have direct effects on size, quantity, coverage, and diversity of a PTC. A*ugmentation* thus becomes an important feature of a PTC.

3.9 Documentation

It is necessary to preserve detail information of the sources wherefrom texts are collected in PTC. It is a practical requirement

on the part of PTC designer to deal with problems related to verification and validation of the SL and the 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 the SL and the TL texts. As PTC maker we document meticulously all extralinguistic information relating to types of text, source of text, etc. These 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 of information of a PTC should be separated from the texts itself in the form of Metadata. We need to keep all information in a Header File that contains all references relating to texts. For easy future access, management, and processing of PTC 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 PTC, 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 a PTC means making each Translation Unit (TU) of the SL to correspond to an equivalent unit in the TL (McEnery and Oakes 1996). The TU covers small units like words, phrases, and sentences (Dagan, Church and Gale 1993) as well as large units like paragraphs and chapters (Simard et al. 2000) (Fig. 4).

	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 a PTC

However, selection of TU depends largely on the point of view selected for linguistic analysis and the type of corpus used. If a PTC asks for a high level of 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., novels or fiction), alignment at larger units at paragraph or chapter level will suffice (Véronis 2000). 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 PTC, although this is predominantly true for the technical corpora. Literary PTC, on the other hand, lends itself to reliable alignment of units beyond sentence level if translational equivalency observed in PTC is previously formalised (Chen and Chen 1995).

Since the so-called 'free translations' present a serious processing problem due to missing sequences, changes 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 the TL text units, which appear to be closest to the SL text units. Such rough alignment yields satisfactory results at sentence level (Kay and Röscheisen 1993) especially when supported by various statistical methods

(Brown and Alii 1990) with minimal formalisation of major syntactic phenomena of texts (Brown and Alii 1993).

Sentence level alignment is an important part of PTC alignment. It shows correspondences down to the level of sentence, and not beyond that (Brown, Lai and Mercer 1991). For this work, a weak translation model serves the purpose, since this is one of the primary tools required at the initial stage of PTC analysis (Simard, Foster and Isabelle 1992). We have given below is a sample of Hind-Bangla PTC where sentences are largely aligned (Fig. 5).

Sentence ID	Hindi-Bangla Aligned Sentences
HNHL_293	हृदय रोगी को नमक, मिर्च तथा तले-भुने भोजन का प्रयोग कम से कम करना चाहिए या हो सके तो नहीं करना चाहिए ।
BNHL_293	হৃদরোগীদের নুন, ঝাল ও আজেবাজে খাবার খাওয়া খুব কমিয়ে দেওয়া উচিত বা সম্ভব হলে বন্ধ করে দেওয়া উচিত ।
HNHL_294	हरी पत्तेदार सब्जियों तथा फल का सेवन अधिक मात्रा में करना चाहिए ।
BNHL_294	টাটকা সবজী ও ফল অধিক মাত্রায় ভোজন করা উচিত ।
HNHL_295	यदि हृदय रोगी धूमपान, शराब या अन्य किसी नशीली वस्तु का सेवन करता है तो उसे शीघ्र ही इन पदार्थों का सेवन बंद कर देना चाहिए ।
BNHL_295	যদি হৃদরোগী ধূমপান, মদ বা অন্য কোনো নেশা করেন তবে তাঁকে শীঘ্রই এই সব খাওয়া বন্ধ করে দিতে হবে ।
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_297	আপেলের মোরোব্বা খাওয়া হৃদরোগীদের বিশেষ প্রয়োজন।

Fig. 5: Sentences aligned in Hind-Bangla PTC

Alignment of PTC helps to optimize mapping between two equivalent units in order to obtain better translation output. Usually, 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 of 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 PTC.
- Formalisation of structures of associable units by way of using sets of morpho-syntactic tags.
- Determination of probability of proposed structures comparing the forms with effective texts collected from manually translated corpora.

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 PTC (Kohn 1996). It is not, however, necessary to analyse all sentences used in a PTC to find out all matches. Analysis of type constructions, rather than the 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 the SL may correspond structurally to other NPs within a text. This is true to both the SL and the TL.

(b) Sequence and interrelation between the units in the TL text may be same with those in the SL text if PTC 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 a PTC to enhance the tasks of 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 morpho-syntactic tagging of each text, since machine

can do it with a statistical support to find out equivalent forms just by comparing PTC that exhibit translational relations. However, to ensure quality performance of a system the following things should be taken care of:

(a) The standard of a PTC should be high. Aligned bilingual texts may pose problem if the quality of a PTC is poor or if texts are not put under strict vigilance of linguists.

(b) The 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 PTC.

(c) The robustness of the system and the quality of translation will depend on the volume of training data available.

(d) The level of accuracy in a PTC will rely heavily on the levels of synchronisation between the texts used in a PTC.

Alignment of a PTC 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 the SL and the 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 PTC with supervised training where all the syntactic, lexical, and idiomatic differences are adequately addressed (Teubert 2000). This usually narrows down the gulf of mutual intelligibility to enhance translatability between the two languages.

5. Parallel Translation Corpus Analysis

Analysis of a PTC has three goals. First, it helps us to structure translations in such a way that these are usable in production of new translations. By using *TransSearch System* (Isabelle *et al.* 1993) we can mark out the bilingual correspondences between the SL and the TL texts. Second, it guides us to draft translations to detect translation error, if any, in the PTC. It is possible to certify that a translation is complete, in the sense that larger units (i.e., pages, paragraphs, sections, etc.) of the SL texts are properly translated in the TL text. Third, it guides us to verify if any translation is free from interference errors resulted from 'deceptive cognates' in the TL texts. For instance, the Hindi word *sandes* 'news' and the Bangla word *sandes* 'sweet' cannot be accepted as cognates for mutual translation, although they appear to be similar in form and structure in the two languages. Similarly, Hindi word *khun* and Bangla word *khun* should not be treated as translational equivalents, because while the Hindi word means 'blood', the Bangla word means 'murder' although both the forms appear to be distantly related to the core concept of 'death'.

A PTC, once it is aligned, may be available for deep linguistic analysis. In general, it involves the following four basic 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 a PTC.

(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 are free to use descriptive morphosyntactic approach along with some statistical approaches for probability measurement. We take support from standard descriptive grammars and morphosyntactic rules of the SL and the TL, as and when required. At this stage, part-of-speech tagging is done mostly manually by comparing texts of the SL and the TL. It is found that our traditional grammatical categories of words have good referential value on the quality of part-of-speech tagging of a text, since a MT system with few POS tags shows greater success than a system made with exhaustive POS tags (Chanod and Tapanainen 1995). Based on the analysis of translational equivalent forms obtained from the PTC, we find three types of matching:

- **Strong match:** Here the number of words, their order, and their meaning are same.
- **Approximate match:** Here the number of words and their meanings are same, but not the order in which they appear in texts.
- **Weak match:** Here the 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 the PTC 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 a Hindi text semantically correspond to 70% words in a sentence in a Bangla text, we can claim that sentences have semantic equivalency to have a translational relationship.

We are still doing some amount of research to develop PTC analyser, which can account for the translation equivalence between

words, idioms, and phrases in PTC. Some statistical algorithms may also be used to find keywords to retrieve equivalent units from the PTC. Once these are found, these are verified and formalised by human translators as model inputs and stored in the bilingual lexical database (Gale and Church 1993, Oakes and McEnery 2000).

6. Restructuring Translation Units

Restructuring a Hindi sentence into Bangla is an attempt to maximize all the 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 language family, it has been, to a large extent, an easy task for us to restructure most of the Hindi phrases in Bangla with utilization of lexico-grammatical stock of both the 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 mahattva ache.

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) mahattva (7) ache (8)
Actual Output	(1) (2+3) (4+5) (6) (7) (8)

Table 1: Restructuring of Hindi and Bangla sentences

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

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

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 a MT system from Hindi to Bangla, this becomes relevant, as many Hindi phrases need to be restructured in the framework. Therefore, grammatical mapping and reordering of words is needed for producing truly acceptable outputs in 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 contain 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.

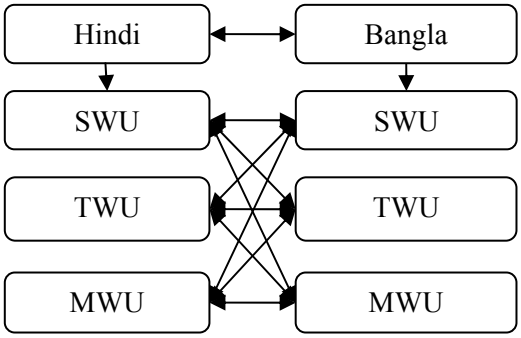

Fig. 6: Lexical Mapping between Hindi and Bangla

[SWU = Single Word Unit; TWU = Two Word Unit, MWU = Multi Word Unit]

7. Extraction of Translational Equivalent Units from PTC

Search for the Translation Equivalent Units (TEU) in a PTC begins with particular forms that express similar sense in both the languages. Once these are identified in a PTC, these are stored in a separate lexical database. Normally, a PTC yields large amount of TEU, which are linguistically fit 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 in texts. Furthermore, the TEUs are verified with monolingual text corpora of the respective two languages from which a PTC is developed. It follows a scheme (Fig. 7) through which we generate a list of possible TEUs from the PTC.

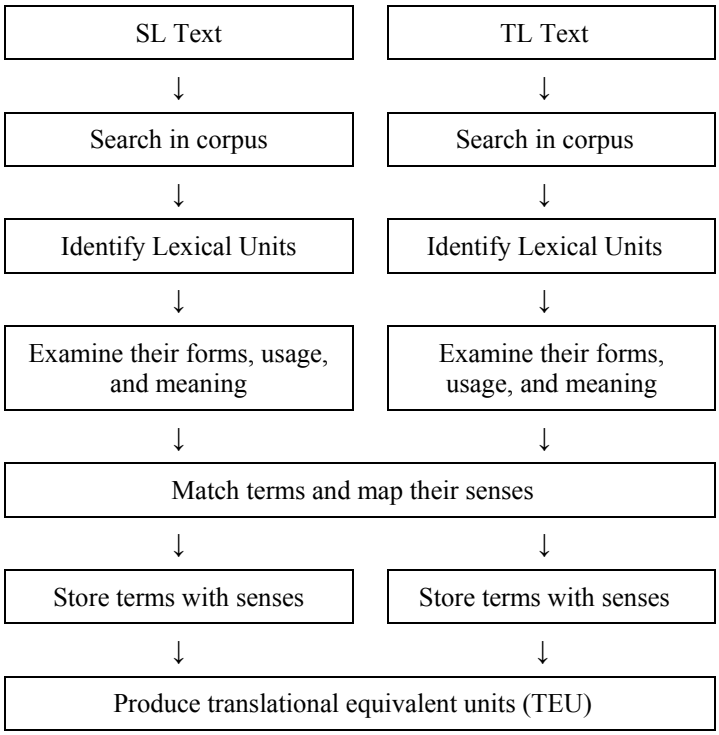

Fig. 7: Extraction of TEU from a Parallel Translation Corpus

We find that even within two closely related languages, TEUs seldom mean the same thing in all the contexts, since these are seldom used in the same types of syntactic and grammatical construction (Dagan, Church and Gale 1993). Moreover, their semantic connotations and degrees of formality may differ depending on language-specific contexts. Sometimes a lemma in the TL is hardly found as a true TEU to a lemma of the SL, even though both the words appear conceptually equivalent. Two-way translation is possible with proper names and scientific and technical terms, but hardly with ordinary lexical units (Landau 2001: 149). 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 PTC to build up useful resource for multiple applications:

- Extract good TEU including words, idioms, compounds, collocations, and phrases.
- Learn how a PTC helps 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 Indian languages.

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

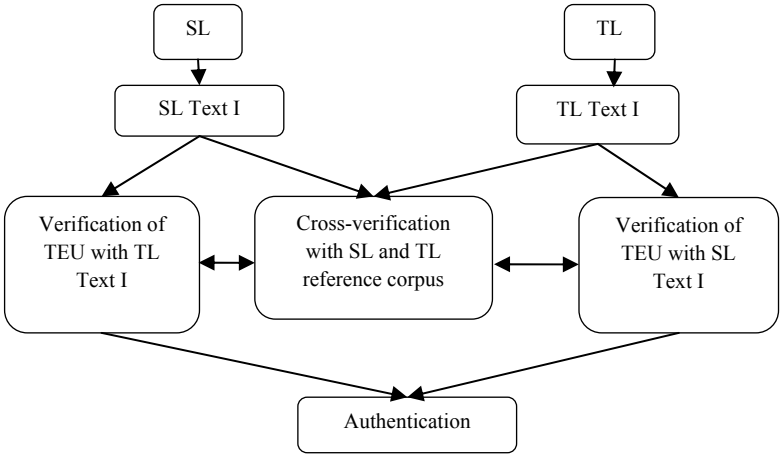

Fig. 8: Verification of TEUs with monolingual corpus

8. Bilingual Lexical Database

Development of a Bilingual Lexical Database (BLD) from a PTC is an essential task the lack of which is one of the bottlenecks of present Computer Aided Translation (CAT) 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 (Geyken 1997). Using a POS-tagged PTC we can extract semantically equivalent words for the BLD (Brown 1999). A BLD may be developed from the untagged corpora when a POS-tagged PTC is not available for the purpose.

Formation of a BLD is best possible within those cognate languages that are typologically or genealogically related to each other (e.g. Bangla-Odia, Hindi-Urdu, Tamil-Malayalam, etc.) because cognate languages usually share many common properties

(both linguistic and non-linguistic) that are hardly found in non-related languages (Kenny 2000). Moreover, there are large numbers of regular vocabularies 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, puku r : 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, tale : 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 vocabularies are similar in sense in Bangla and Odia – two sister Indo-Aryan languages. To generate a BLD from a POS tagged PTC, we use the following strategies:

- Retrieve comparable syntactic blocks (e.g. clauses and phrases, etc.) from a PTC.
- 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 the languages are closely related), similarities in form, meaning, and usage are enough for selection of TEU from a PTC.

9. Bilingual Terminology Databank

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

A PTC has another crucial 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 be selected over other suitable and competent candidates. Debate may

also arise whether we should generate new STTs or accept the STTs of the SL already absorbed in the TL by regular use and reference. It is noted that some STTs are so largely naturalised that it becomes almost impossible to trace their actual origin. In this case, we have no problem, because these STTs are 'universally accepted' in the TL. For instance, the Bangla 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 the Bangla vocabulary. There is no need for replacement of these STTs in the TL texts.

A PTC is a good resource for selection of appropriate STTs presenting new ideas and concepts. As a PTC 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 PTC contributes in two ways:

(a) It helps to assemble STTs from the SL and the 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 full information of domains and frequency of use in the SL and the TL.

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

10. Conclusion: Value of a Translation Corpus

The question that arises at the time of a PTC development is: who is going to use it and for what purposes? That means the issue of determining the target users is to be dissolved before the work of a PTC development (Tymoczko 1998). But why it is necessary to identify the target users? There are some of the reasons:

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

These application-specific needs can be easily fulfilled by a PTC. Hence, question of selecting target users becomes pertinent in PTC construction. However, although prior identification of target users is a prerequisite in PTC generation, it does not imply that there is no overlap among the target users with regard to utilisation of a PTC. In fact, our past experience shows that multi-functionality is an

inherent feature of a PTC due to which a PTC attracts multitudes of users across various fields (Hunston 2002).

This signifies that a PTC designed and developed for specific purpose may equally be useful for other works. For example, although a PTC is maximally suitable for lexicographers, it is equally useful for lexicologists, semanticists, grammarians and social scientists. Also it is useful for media persons to cater their needs relating to language and society. A PTC can be used as a resource for the works of language technology as well as an empirical database for mainstream linguistics (Tymoczko 1998). In essence, it has high applicational relevance for people interested in the SL and the TL texts full of exciting features both in content and texture. For the Indian languages, a PTC is a primary resource, which we need for all works of linguistics and language technology.

Notes:

1. Indian Languages Corpora Initiative-Phase II

REFERENCES

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

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

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

Baker, M. 1995. Corpora in translation studies: an overview and suggestions for future research. *Target.* 7(2): 223-43.

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.

Brown, P. and M. Alii. 1990. A statistical approach to machine translation. *Computational Linguistics*. 16(2): 79-85.

Brown, P. and M. Alii. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*. 19(2): 145-152.

Brown, P., J. Lai, and R. Mercer. 1991. Aligning sentences in parallel corpora. *Proceedings of the 29th Meeting of ACL.* Montreal, Canada.

Brown, R.D. 1999. Adding linguistic knowledge to a lexical example-based translation system. *Proceedings of the MTI-99*, Montreal. PP. 22-32.

Chanod, J.P. and 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.

Chen, K.H and H.H. Chen.1995. Aligning bilingual corpora especially for language pairs from different families. *Information Sciences Applications*. 4(2): 57-81.

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.

Gale, W. and K.W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*. 19(1): 75-102.

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

Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Isabelle, P., M. Dymetman, G. Foster, J.M. Jutras, E. Macklovitch, F. Perrault, X. Ren and M. Simard. 1993. Translation analysis & translation automation. *Proceedings of the TMI-93,* Kyoto, Japan.

Kay, M. and M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*. 19(1): 13-27.

Kenny, D. 1997. (Ab)normal translations: a German-English parallel corpus for investigating normalization in translation. In: B. Lewandowsk-Tomaszczyk and P. Janes Melia (eds.) *Practical Applications in Language Corpora. PALC '97 Proceedings,* Lódz: Lódz University Press, pp. 387-392.

Kenny, D. 1998. Corpora in translation studies. In: M. Baker (ed.) *Routledge Encyclopaedia of Translation Studies,* London: Routledge, Pp. 50-53.

Kenny, D. 1999. The German-English parallel corpus of literary texts: a resource for translation scholars. *Teanga.* 18: 25-42.

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.

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, pp.143-160.

Klaudy, K. and 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.

Kohn, J. 1996. What can (corpus) linguistics do for translation?. In: K. Klaudy, J. Lambert and A. Sohar (eds.) Translation Studies in Hungary, Budapest: Scholastica, PP. 39-52.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

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.

McEnery, T. and M. Oakes. 1996. Sentence and word alignment in the CARTER Project. In: J. Thomas and M. Short (eds.) *Using Corpora for Language Research*. London: Longman. PP. 211-233.

Oakes, M. and T. McEnery. 2000. Bilingual text alignment — an overview. In: Botley, S.P., A.M. McEnery and A. Wilson (eds.) *Multilingual Corpora in Teaching and Research.* Amsterdam-Atlanta, GA.: Rodopi. PP. 1-37.

Simard, M., G. Foster, and P. Isabelle.1992. Using cognates to align sentences in parallel corpora. *Proceedings of TMI-92.* Canadian Workplace Automation Research Center. Montreal.

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 and A. Wilson (eds.) *Multilingual Corpora in Teaching and Research.* Amsterdam-Atlanta, GA.: Rodopi. Pp. 38-64.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

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.

Stewart, D. 2000. Poor relations and black sheep in translation studies. *Target* 12(2): 205-228.

Summers, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and Design.* Harlow: Longman.

Teubert, W. 2000. Corpus linguistics — a partisan view. *International Journal of Corpus Linguistics*. 4(1): 1-16.

Tymoczko, M. 1998. Computerized corpora and the future of translation studies. *Meta* 43(4): 652-659.

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.

Véronis, J. ed. 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora.* Dordrecht: Kluwer Academic Publishers.

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.