Translation Corpora and Machine Aided Translation

Niladri Sekhar Dash

Abstract

The modification of the goal from holistic automatic machine translation to domain-specific machine-aided translation has been the outcome of our rude realization of the fact that the development of a fully automatic machine translation system, which will work equally effectively for all kinds of text and for all kinds of natural language, is actually a myth. The long history of failures recorded in last six decades in the act of developing an automatic machine translation system for any two natural languages has eventually forced us to think of developing domain-based machine translation system, which gives us a better assurance of success among related languages. Against this back-drop I have made an attempt here to show how linguistic data and information intelligently extracted from systematically designed translation corpora can lead us a few steps forward towards our goal. The method proposed here will be highly useful in the Indian context if we complete the ground work necessary for achieving success before we venture on the voyage.

1. Introduction

A Machine Aided Translation (MAT) system is a man-machine interactive interface that takes linguistic inputs in the form of full sentences from the source language text to generate corresponding full sentences in the target language text (not necessarily of a very high quality). MAT technique may be considered as an inevitable

Translation Today Vol. 6 No. 1 & 2 2009 © CIIL 2009

offshoot of Artificial Intelligence, which is evolving in wide dimensions in tandem with the meteoric progress in the field of Information Technology. In this new millennium, MAT technology may be considered as one of the key technologies that can have direct as well as lasting impact on the global market of inter-lingual communication, cross-lingual information retrieval and multi-disciplinary information exchange. The MAT technology is thus turning up into a domain of cross-disciplinary application with direct functional relevance in e-commerce, knowledge localisation, and socio-economic progress of societies.

MAT is not a simple form-to-form (i.e., word-to-word, phrase-to-phrase, sentence-to-sentence, etc.) replacement of the source language text into the target language. By its default definition, it gives a kind of assurance that the texts of the source language should be 'grammatically correct, linguistically valid, and semantically or conceptually acceptable' in the translated outputs of the target language. Moreover the outputs in the target language should be nearest to the source language—both in sense and content—if not identical. Furthermore, information embedded within the source language text should not be lost in the target language, and no extra information, which was not originally present in the source language text, will be added in the target language.

With these basic assumptions, in this paper, I have made an attempt to address the issue of MAT from the perspective of end-users with a focus on its general goals (Section 2). In Section 3, I have proposed using translation corpora for the purpose of developing a MAT system. In Section 4, I have discussed the issues related to translation corpora design in Indian languages. In Section 5, I have identified the immediate tasks which we need to complete before we can think of developing a workable MAT system for Indian languages. In this paper, I have addressed only two major issues of MAT system: (a) generation of translation corpora in Indian languages, and (b) alignment of these translation corpora. Other issues related to the MAT system are elaborately addressed elsewhere (cf. Dash and Basu 2009).

2. The Goal of a Machine Aided Translation System

In the present context of designing a MAT system the question that stands tall is: why do we need such a system when human beings have already proved to be far more competent than the machine in the business? Perhaps, we do not have a complete answer to the question. Hutchins (1986:15), however, provides the following arguments in support of a MAT system:

(a) The professional world of scientists and technologists requires quick and accurate translation of documents.
(b) MAT becomes useful in those situations where there is paucity of human translators.
(c) MAT helps in promotion of international co-operation through translation of texts and documents meant for global access.
(d) MAT helps to promote mutual co-operation, in removing language barriers by faster, easier, and cheaper transmission of scientific, technical, agricultural, and medical information to the poor and the developing countries.
(e) MAT is useful for military purposes; for pure theoretical and applied research; and of course, for the purpose of commercialisation.

These are valid reasons. Yet, the global scenario of the 1980s, when Hutchins tried to justify the need of a MAT technology has changed to a great extent. Now, demands for linguistic singularity, growth of mass literacy and readership, expansion of multilingualism, globalisation of information and the revolution in the area of computer technology have united the computer scientists and the linguists together to develop a robust MAT system that will be able to meet the requirements of the market worth millions of dollars. We can, perhaps, summarise both types of need in the following way to justify the need of a MAT technology in the present Indian context: revolution in computer technology, expansion of multi-lingualism, demand for linguistic singularity, scarcity of human translators, globalisation of knowledge and information, professional need for translated documents, promotion of international co-operation, growth of mass literacy and global

readership, promotion of international peace, removal of language barriers, all round growth of Indian states through faster, easier and cheaper transmission of scientific, technical, agricultural, medical information to the poor and developing countries, military needs, commercial needs and research challenge in language technology.

Within the last six decades, the MAT technology has recorded remarkable growth with many diversions both in the use of techniques as well as in their application in various domains of human knowledge. In the era of internet-domination, our prime objective is to develop a MAT system that is accurate and effective; robust and versatile; and user-friendly and customizable. The huge amount of text available for immediate translation warrants a robust MAT technology that will be able to produce translations of workable standard, if not the ideal one. Definitely, there are many obstacles in the path of achieving a high rate of accuracy in MAT. These obstacles come not only from the fields of morphology, lexicology, syntax, and semantics, but also from the world of culture, pragmatics, discourse, and cognition. That means developing a MAT system which is endowed with the abilities of a human translator is an elusive dream for which we have to travel many miles.

However, the availability of digital translation corpora in the form of parallel bilingual texts seems to offer a highly promising solution to MAT system developers, due to the features of close generic, structural as well as semantic similarities between the texts used to develop translation corpora. Besides advanced computational techniques used to fine-tune the translation corpora in alignment of text samples also enable the system developers to extract relevant translation equivalents from translation corpora to enhance the performance level of the MAT system. This implies that a corpus-based approach may bring us nearer to the dream by enhancing inherent efficiency of a MAT system we are striving hard to design for Indian languages (Dash 2005a:ch.9).

Since a MAT system is not meant to produce perfect translations, outputs are most often put to manual post-editing to an

acceptable standard in the target language. Moreover, these can be used in unedited form as a source of rough linguistic outputs, the analysis of which will yield better insight into dealing with the intricate problems related to the development of MAT for Indic languages. A MAT system aims at linguistic and cognitive approximation where the goal is to find out ways and means to get as close as possible in as many cases as feasible. The goal is achievable through a long process of trial and error and it requires regular evaluation of existing systems, identification of faults, refinement of existing tools and techniques, enhancement of past experience, and augmentation of linguistic knowledgebase. Obviously, the path is full of thorns. But it is tantalising, since it throws challenges before the linguists and the technologists.

We argue that a MAT system that aspires to claim some success as a commercial product as well as a research prototype must have customisation capabilities. Moreover, it should have an ability to add translations of new words and phrases; should have provisions for including more sophisticated functionalities to adapt to new syntactic structures and writing styles; and have the capability for acquiring the knowledgebase from earlier translation outputs. We are waiting to see if MAT developers are able to incorporate all the aspects to enhance the capabilities of their systems—to address the diverse needs of the translation world.

3. Use of Translation Corpora in Machine Aided Translation

From the experience gathered during the last six decades, we have learnt that designing a MAT system with the support of only a set of rules is not realistic at all. Only a set of rules is not enough to encompass the wide versatility of a natural language exhibited in diversified discourse of life. This eventually leads us to think about the Corpus-based Machine Translation approach that tends to combine fruits of Rule-based Machine Translation (Lewis and Stearns 1968, Gildea 2003, Chiang 2005), Example-based Machine Translation (Furuse and Lida 1992, Jones 1992, McLean 1992, Somers 1999) and Statistics-based Machine Translation (Brown et al.1990, Brown, Pietra, and

Mercer 1993) to achieve the goal still elusive to system designers. Let us investigate how the Corpus-based Machine Translation dares to reap good harvest while other systems are content with limited success.

One of the very first results obtained from the use of translation corpora is the development of algorithms, which are capable of aligning sentences of bilingual texts. It turns out as one of the fundamental properties in the MAT system, since it provides indispensable resources for the development of various translation support tools. The corpus-based approach begins with the parallel translation corpora already produced by human translators to discover linguistic similarities in the internal structures of source and target language texts, either completely or partially, to use in the MAT system. This analysis-oriented perspective lends heavily to the development of translator's aids, as the MAT system, at its initial stage, is not actually expected to 'produce' translations, but to 'understand' enough about the internal forms and structures of the language to become eventually helpful in subsequent translation tasks.

The idea of using translation corpora in MAT is not entirely a new thing. Although it dates back to the early days of machine translation, it was not used in practice until 1984 (Kay and Röscheisen 1993). Now, careful attention is redirected to translation corpora because it is ultimately realized that data and information acquired from analysis of translation corpora make the MAT system more equipped and robust to acquire greater percentage of accuracy in translation outputs. In general, translation corpora are richer with information about the languages than monolingual corpora, because these can provide better translational equivalency information between the languages used to design translation corpora as well as to be considered for translation. Thus, a Corpus-based MAT system is practically more efficient to achieve a unique status of distinction as it combines the features of Rule-based Translation, Example-based Translation, and Statistics-based Translation keeping alive a mutual interactive interface between the three systems.

Conceptually, a Corpus-based MAT system is grounded on a range of resources developed from exhaustive empirical analysis of translation corpora designed systematically both from the source and the target

language. The analysis of translation corpora involves morphological, lexical, semantic, and cognitive interpretations of words, phrases, idioms, sentences, and paragraphs as well as other linguistic items available within the corpora. Moreover, direct employment of various statistical techniques on translation corpora is capable of generating probability measurements for the linguistic items from the texts to identify the translational equivalents required for translation between the two languages.

The Corpus-based MAT system stands on the assumption that there are no pre-established solutions to translation, but the most possible solutions may be found in the translation corpora, which are already developed by human translators. In other words, a large portion of the competence of a human translator is encoded in the text equivalencies found in translation corpora. Success achieved in this method in restricted domains leads us to argue that both linguistic and extralinguistic information extracted from translation corpora can be used as essential ingredients to achieve similar success in general domains (Teubert 2002).

At present, although it is too early to make any prediction about its success in all domains of human knowledge, it may be argued that a MAT system designed with proper utilisation of information acquired from well designed translation corpora can be more robust and useful both in restricted and general domains (Su and Chang 1992). In essence, the system will operate with information and examples obtained from an analysis of translation corpora made with texts of the languages involved in translation, as it will utilize translation corpora to enhance its usability in restricted and general domains.

4. Issues Related to Corpus-based Machine Aided Translation

A Corpus-based MAT system addresses the requirement of a functionally competent system in the present context of global upsurge for information localisation and exchange. It learns from the history about the milestones of success and failure, and therefore, anchors on

the technical and linguistic issues involved in the development of the system with reference to translation corpora. For achieving this goal it argues for generating tools and resources necessary for developing useful Corpus-based MAT systems for Indian languages.

The proposed Corpus-based MAT system makes considerable amount of advancement by means of extensive analysis of translation corpora. Till date, several translation corpora are developed and analysed in English and other languages like French, German, Spanish, Italian, Dutch, Japanese, etc. to yield important insights that can help system developers to design useful techniques and strategies. Normally, translation corpora represent a large collection of naturally occurring texts accumulated by including text samples that reflect on the needs of the end-users. These become particularly useful for those target end-users who want to select a MAT system that will translate specific texts to suit their requirements. It, therefore, becomes clear that in a Corpus-based MAT system, utilisation of translation corpora is mandatory, since these can supply numerous linguistic and extralinguistic examples and information to system designers as well as the end-users. In the following sub-sections, I have focused only on two issues related to the Corpus-based MAT system with reference to the Indian languages and English (Dash 2005b).

4.1 Generation of Translation Corpora

Translation corpora generally consist of original texts obtained from a source language and their translations obtained from the target language. By virtue of their composition these corpora usually keep the meaning and function of words, phrases, and other linguistic items constant across the languages, and as a result of this, these become highly capable for offering an ideal basis for comparing meanings of the linguistic items of the two languages under identical conditions. Moreover, these make it possible to discover all kinds of cross-linguistic variants, i.e., alternative forms of particular words and terms from both parts of the corpora. Thus, translation corpora provide fruitful examples for cross-linguistic equivalents, the alignment of which can provide solid empirical base for formulation of rules for translation (Altenberg and Aijmer 2000:17).

The construction of translation corpora is, however, a complicated task. It requires constant careful guidance from experienced corpus linguists who have sound knowledge in the tasks of corpora generation and processing. Translation corpora are made in such a way that these become suitable to combine advantages both of comparable corpora (Dash 2008:75) and parallel corpora (Dash 2008:81). Text samples considered for this kind of corpora should be taken in equal proportion or amount from the source and the target language, and the text samples should be matched, as far as possible, in terms of their text types, subject matters, purposes, and register variations. The basic structure of translation corpora between any two natural languages can be envisaged in the following manner (Fig. 1) keeping in mind the aim of the work and the components to be integrated within translation corpora.

(A): English Text
(C): Bengali Translation
(B): Bengali Text
(D): English Translation

(A): English Text
(C): Bengali Translation
(B): Bengali Text
(D): English Translation

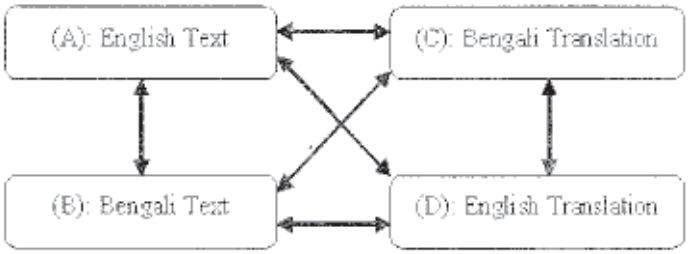

Fig. 1: Structure of Translation corpora

The diagram shows that translation corpora between any two languages can be developed in such a way that they may be used as

comparable corpora (A:B); translation corpora (A:C, B:D); as original and translated texts (A:D, B:C), and for comparing the translated texts in the two languages (C:D). Since so many functionalities can be achieved by way of developing translation corpora, utmost attention is required in the selection of texts which should adhere to the following principles:

(a) Only language samples of written texts should be included in translation corpora. At present there is no scope for spoken texts as the present MAT research targets written texts only.
(b) Included texts should reflect contemporary language use although texts of old times have relevance in the translation of historical texts.
(c) Translation corpora should not be confined within texts of specific language variety. They should include a wide range of text types obtained from all possible domains and disciplines of language use.
(d) Texts of both the languages should be comparable as far as possible. They should be matched in genre (e.g., news), type (e.g., political), content (e.g., election), and form (e.g., report). They should also match in terms of purpose, type of user, subject matter, and register.
(e) Texts should also match in terms of purpose, types of users, subject matters, and register.
(f) Text included in translation corpora will consist of fairly large and coherent extracts taken from the beginning to the end at a natural breaking point of a text (e.g., chapters, sections, paragraphs, etc.).

At the time of compiling translation corpora no human intervention is invited although humans are actually engaged in the task. That means, there should be no modification of the original source texts considered for translation. The translator will translate the source text in the target language without distorting the actual form, texture and meaning of the source text. They should also try to maintain semantic as well as structural parallelism between the source language text and the target language text as far as it is possible and feasible. However, they may get some liberty to make the translation 'natural' in the target language by way of rearrangement of word orders, replacement of

some semantically equivalent forms, or restructuring the output sentences in the target language. On the other hand, if previously made translations are available it should have a kind of advantage in the sense that whatever is available in the form of translation should be put in translation corpora, because any kind of interference on the part of corpora developers will severely damage the naturalness and originality of the texts procured from external physical sources. However, gross typological or syntactic errors of the source texts should be corrected before these are put into translation corpora.

4.2 Alignment of Translation Corpora

Aligning translation corpora means making each 'translation unit' of the source text correspond to an equivalent unit in the target text (McEnery and Oakes 1996). The term 'translation unit' does not only cover the shorter sequences like words, phrases, and sentences (Dagan, Church and Gale 1993), but also covers larger text sequences such as paragraphs and chapters (Simard et al. 2000). However, selection of the 'translation units' depends largely on the point of view selected for linguistic analysis and the type of corpus used as the database. If a translated corpus asks for a high level of faithfulness to the original, as it happens in case of legal documents and technical texts, the point of departure is a close alignment of the two corpora, considering sentences, or even words as the basic units.

On the other hand, if the corpus is an adaptation, rather than literal translation of the original text, attempts may be made to align the larger units such as paragraphs and chapters (Véronis 2000:12) rather than the smaller units like words, phrases, and sentences. Thus, operation of alignment may be refined based on the type of corpora used in the work. The faithfulness and linearity of human translations may guide one to align translation corpora, although this is predominantly true for the technical corpora. Literary translation corpora, on the other hand, lend themselves to reliable alignment of units below the sentence level if the types of translational equivalency observed in corpora are previously formalised (Chen and Chen 1995).

It is obvious that the initial hypothesis, which allows these translation corpora to be used, is the correspondence, if not equivalence, where the contents of texts and their mutual relationships are put under consideration for comparison. We can call these as 'free translations', which, however, may posit a serious processing problem due to their missing sequences, changes in word order, changes in order of phrases, and modifications of content, etc. Although these operations are common in everyday translation practice, their frequencies usually vary according to the field of the texts.

These observations lead us to consider aligned translation corpora not as sets of some 'structurally and semantically equivalent sequences', but as the 'corresponding texts having mutual conceptual parallelism'. At any level of the texts (e.g., word, phrase, sentence, paragraph, etc), these translation corpora are considered as language databases studded with parallel linguistic units. Here the main objective is not to show the structural equivalencies found between the two languages, but pragmatically, to search for the target text units, which appear to be the closest to the source text units. To do so, the starting point is a preliminary alignment of words with a bilingual dictionary, if available. Definitely, a rough alignment will yield satisfactory results at the sentence level (Kay and Röscheisen 1993) especially when it is directly supported by various statistical methods (Brown and Alii 1990) with minimal formalisation of the major syntactic phenomena of the texts of the languages concerned (Brown and Alii 1993). The main advantage of this method may be realized in the use of 'translation memory'—a temporary database of translational equivalents developed from the data found in the bilingual texts. The task may be further simplified by using the Reference Corpora of the specialised fields (e.g., medical science, legal proceedings, computer science, etc) from both the languages. The message is thus 'machine-translated' by using a customised 'machine-readable dictionary' to create a translation memory during the training phase.

Sentence level alignment is another important part of the translation corpora development and analysis. It aims at showing the

correspondences down to the level of sentences, and not beyond that (Brown, Lai, and Mercer 1991). For this work, a weak translation model can serve the purpose, since this is one of the primary tools required at the initial stage of translation corpora analysis (Simard, Foster, and Isabelle 1992). Therefore attempts are made to develop translation analyser, which can account for translational correspondences between morphemes, words, idioms, and phrases found in the translation corpora.

Another interesting activity of translation alignment is the use of statistical techniques for searching the matching candidates from the translation corpora. Statistical searching algorithms use the key words to retrieve the equivalent units from two different texts. Once these are found, these are verified and formalised by human translators as the model inputs before these are stored in bilingual translation memory. This process is usually used for automatizing the training phase, and not for any kind of validation of translational outputs. This has been one of the basic criteria that mark out the differences between the so-called automatic translation system and the Corpus-based MAT system supported by parallel translation corpora.

5. Need for Indian Languages

In the case of Indian languages, the generation of bilingual translation corpora has been one of the bottlenecks in MAT. Although considerable amount of written texts in the form of monolingual corpora is available, there has never been any sincere attempt to redesign the monolingual corpora into parallel translation corpora due to following reasons:

(a) Corpus generation: The generation of translation corpora of various types among Indian languages has been a long-standing requirement (Dash 2003). Although some monolingual corpora, which are developed following the same designing principles and similar text types, are available for some Indian languages, these are not yet used systematically to develop good translation corpora.

(b) Acquisition: Even when there are a few manually developed translation corpora in the form of parallel texts published by National Book Trust, Sahitya Akademi and others, these are not available in electronic form for translation purposes.

(c) Conversion: Text corpora of different origins, text types, samples, and formats available in Indian languages may be processed to convert them into parallel translation corpora.

(d) Sanitation: The monolingual corpora, which are to be converted into translation corpora, also need proper cleaning and sanitation of the texts to be used for the purpose of text alignment, matching, processing, access, and utilisation.

(e) Alignment: Translation corpora need to be aligned in a highly systematic manner so that all the corresponding sections of both the texts are identified. It involves identification of translation units and deployment of various knowledgebase (e.g., formal, lexical, semantic and conceptual, etc) for eventual pairing and alignment of the linguistic units used in the translation corpora.

Most of the issues stated above are linguistic issues, which are not taken into consideration even if raw monolingual corpora are available in Indian languages. Therefore, at the present situation, our main problem is the procurement of translation corpora in Indian languages, and defining their correspondences within the linguistic units. After the procurement of these monolingual corpora, these can be put to linguistic analysis and alignment of the components before these are submitted to the task of 'item search'—a statistical technique that reduces the hurdles produced from translation analysis.

The critics of the Corpus-based MAT approach point out that paucity of translation corpora is a real problem in Indian context since the researchers of the Corpus-based MAT system require information obtained from detailed analysis of translation corpora by human experts (Elliott 2002). Information obtained from analysis of translation corpora allows scientists to design systems as well as to test the reliability of their systems. It is, therefore, understood that we need to compile translation corpora in major Indian languages—not only to meet the

research requirements but also to evaluate the efficiency and usefulness of the MAT systems developed so far for Indian languages (Dash 2010).

In this case, however, information obtained from analysis of translation corpora of other languages may provide necessary direction to Indian scientists. For instance, information about how translation corpora is used in translation from English to other European languages may help scientists working for Indian languages (Rajapurohit 1994). Access of information from these sources will provide clues to identify the patterns about how languages are interlinked to each other (Baker 1993) and how the information encoded in one text is transferred into another text. Exposure to this kind of knowledgebase may help to improve the standard of the MAT systems designed for Indian languages as well as result in gathering new insights into the intricate linguistic relationships existing between the two languages used in translation.

For achieving success in MAT for Indian languages, I strongly argue that we should first start developing translation corpora for it because linguistic data and information derived from these corpora will increase our knowledgebase about the languages as well as will enhance the efficiency and productivity of the MAT system. Recently the MAT workers' diversion in this direction is noted as commercial MAT systems with standard language transfer architecture are being developed with regular manual updating of the lexicon and the lexical information extracted from translation corpora. These systems appear to have greater acceptance and usability due to their ability to incorporate language-specific linguistic aids for end-users through the medium of internet and web-pages.

6. Conclusion

The proposal here is that if we want to develop a robust MAT system for Indian languages, we need to justify the two basic questions related to this enterprise: (a) the objective of this mission, and (b) the immediate beneficiaries for whom we are striving to commit ourselves.

With regard to objective, we can visualize that the development of a MAT system may help us disseminate information and knowledge found in English texts into Indian languages; help create high authentic quality knowledgebase in Indian languages; to generate translation support tools such as dictionaries, thesauri, idiom datasets, proverb database, term-banks, word-finders, word-nets, synsets, etc. After these goals are realised, we can think of promoting MAT systems between English and Indian languages and between Indian languages.

With regard to identifying the beneficiaries, I find that development of MAT system as a stand-alone device can help the rural and poor students having little access to the knowledgebase available mostly in English; help teachers in teaching various subjects at various levels; help people who are working in different areas like public health, environment, popular science, medical services etc to gather and share data and information; help people eager to read literary and non-literary knowledge texts in their own languages; help publishers eager to produce knowledge/informative texts in various Indian languages; help various Indian institutes and organizations engaged in theoretical and commercial translation activities; and help the scientists engaged in the development of software and systems for automatic translation systems, etc.

References

Altenberg, B. and K. Aijmer (2000) 'The English-Swedish Parallel Corpus: a Resource for Contrastive Research and Translation Studies' in Mair, C. and M. Hundt (eds) *Corpus Linguistics and Linguistics Theory*, Amsterdam-Atlanta, GA: Rodopi, 15-33.

Baker, M. (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications' in Baker, M. F. Gill and E. Tognini-Bonelli (eds) *Text and Technology: In honour of John Sinclair*, Philadelphia: John Benjamins, 233-250.

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. Cocke, S.D. Pietra, F. Jelinek, R.L. Mercer, and P.S. Roosin (1990) 'A Statistical Approach to Language Translation', *Computational Linguistics* 16(1): 79-85.

Brown, P., J. Lai, and R. Mercer (1991) 'Aligning Sentences in Parallel Corpora' in *Proceedings of the 29th Meeting of ACL,* Montreal, Canada.

Brown, P.F., S.A. Della Pietra, and R.L. Mercer (1993) 'Statistical Machine Translation', *Computational Linguistics* 19(2): 263-312.

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.

Chiang, D. (2005) 'A Hierarchical Phrase-based Model for Statistical Machine Translation' in *Proceedings of the 43 rd ACL*, 263–270.

Dagan, I., K.W. Church, and W.A. Gale (1993) 'Robust Bilingual Word Alignment for Machine-aided Translation' in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio.

Dash, N.S. (2003) 'Corpus Linguistics in India: Present Scenario and Future Direction', *Indian Linguistics* 64(1-4): 85-113.

Dash, N.S. (2005) 'A Brief Sketch on the Techniques of Text Corpus Processing' in *Proceedings of the UGC-SAP International*

Symposium on Linguistics, Quantification and Computation, Hyderabad, 63-89.

Dash, N.S. (2005a) *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. New Delhi: Mittal Publications.

Dash, N.S. (2005b) 'Corpus-based Machine Translation across Indian Languages: from Theory to Practice', *Language in India* 5(7): 12-35.

Dash, N.S. (2008) *Corpus Linguistics: An Introduction*, New Delhi: Pearson Education-Longman.

Dash, N.S. (2010) 'Machine Translation in India: Some Critical Observations' submitted to *Indian Linguistics*.

Dash, N.S. and P. Basu (2009) 'Linguistic Tasks on Translation Corpora for Developing Resources for Machine Translation' presented in the *31st All India Conference of Linguists (AICL 2009)*, Hyderabad, 173-174.

Elliott, D. (2002) *Machine Translation Evaluation: Past, Present and Future,* unpublished MA dissertation, University of Leeds, UK.

Furuse, O. and H. Lida (1992) 'An Example-based Method for Transfer-driven Machine Translation' in *Proceedings of the MTI-92,* Montreal, 139-150.

Gale, W. and K.W. Church (1993) 'A Program for Aligning Sentences in Bilingual Corpora', *Computational Linguistics* 19(1): 75-102.

Gildea, D. (2003) 'Loose Tree-based Alignment for Machine Translation' in *Proceedings of the 40th Annual Meeting of*

the Association for Computational Linguistics (ACL), Sapporo, Japan.

Hutchins, W.J. (1986) *Machine Translation: Past, Present, and Future,* Chichester: Ellis Harwood.

Jones, D. (1992) 'Non-hybrid Example-based Machine Translation Architectures' in *Proceedings of the MTI-92,* Montreal, 163-171.

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

Lewis, P.M. and R. E. Stearns (1968) 'Syntax-directed Transduction', *Journal of the Association for Computing Machinery* 15(3): 465-488.

McEnery, T. and M. Oakes (1996) 'Sentence and Word Alignment in the CARTER Project' in Thomas, J. and M. Short (eds) *Using Corpora for Language Research*, London: Longman, 211-233.

McLean, I. (1992) 'Example-based Machine Translation Using Connectionist Matching' in *Proceedings of the MTI-92*, Montreal, Canada, 35-43.

Oakes, M. and T. McEnery (2000) 'Bilingual Text Alignment: An overview' in Botley, S.P., A.M. McEnery, and Andrew Wilson (eds) *Multilingual Corpora in Teaching and Research,* Amsterdam-Atlanta, GA: Rodopi, 1-37.

Rajapurohit, B.B. (1994) 'Automatic Translation: Then and Now' in Basi, H. (ed.) *Automatic Translation: Seminar Proceedings,* Thivandrum: DLA Publications, 33-59.

Simard, M., G. Foster, and P. Isabelle (1992) 'Using Cognates to Align Sentences in Parallel Corpora' in *Proceedings of TMI-92,* 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 Andrew Wilson (eds) *Multilingual Corpora in Teaching and Research,* Amsterdam-Atlanta, GA: Rodopi, 38-64.

Somers, H. (1999) 'Example-based Machine Translation', *Machine Translation* 14(2): 113-157.

Su, K.Y. and J.S. Chang (1992) 'Why Corpus-based Statistics-oriented Machine Translation' in *Proceedings of the MTI-92,* Montreal, 249-262.

Teubert, W. (2002) 'The Role of Parallel Corpora in Translation and Multilingual Lexicography' in Altenberg B. and S. Granger (eds) *Lexis in Contrast: Corpus-based Approaches,* Amsterdam/Philadelphia: John Benjamins, 189-214.

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