

Use of Language Corpora in Second Language Learning

Niladri Sekhar Dash

Indian Statistical Institute, Kolkata
niladri@iscal.ac.in

Abstract. In this article, we argue for teaching English as a second language to Indians with close reference to English language corpora. The proposed method is assisted by computer and based on information obtained from the language corpora developed with text sample of various types used by the native people in their every day discourse of life. This empirical method for second language education has been proved successful in various parts of the world where English is used either as a second language or as a *lingua franca*. There is no reason to be sceptical about its success in Indian context if the method is used with careful manipulation of the techniques and the resources with further empowerment with the advices of the experts of the field.

1. Introduction

In recent years, language corpora are one of the primary resources in second language learning (SLL). Corpora are considered indispensable, since they provide reliable, authentic and diversified information hardly obtainable from other sources. In fact, the idea of SLL without reference to corpora appears unscientific, because corpora give various opportunities to the learners to understand the language properties from multiple directions. In essence, information obtained from corpora provides valuable complementary perspectives towards traditional linguistic principles of SLL (Biber 1996). Although there are numerous ways in which corpora can be utilised in SLL (Botley, McEnery and Wilson 2000; Granger, Hung and Tyson 2002), most often, they are accessed for the following reasons:

- (a) Intuition-based SLL materials are misleading. They contain intentionally invented examples, which generally overlook important aspects of usage or foreground less frequent stylistic choices at the expense of more frequent ones.
- (b) Corpus-based SLL is more reliable and authentic, since teaching materials made from corpora are explicitly empirical including examples and descriptions of real life language. Here, common choices of linguistic usage are given more attention than those, which are less common (Kübler 2002).
- (c) Empirical evidences compiled from corpora are used to train learners the kinds of language pattern they will encounter when they actually interact in real life situations. For instance, citations to the actual use of words, idioms, collocations, phrases, and sentences, etc. obtained from corpora enable learners to realise the patterns of use of these properties in the language they are learning.
- (d) Apart from being a source of data for empirical teaching, corpora are also used to look critically at existing language teaching materials. In many studies it is found that there are considerable differences between what textbooks are teaching and how the native speakers actually use language (Ghadessy, Henry and Roseberry 2001).
- (e) Most textbooks used in SLL contain only a few sets of invented examples descriptions of which are based on intuition of second-hand accounts. But in principle, SLL materials should be explicitly empirical including examples and descriptions from corpora. Corpora may be presented directly to the learners for class-work or may be used in preparation of teaching materials.
- (f) Corpora can reveal not only the range of patterns of a language that the learners must assimilate, but also present frequency of these patterns, which is important in teaching materials development and syllabus design. Use of corpora has potential to radically alter the design of materials of SLL, and perhaps linguistics as well (Barlow 1996).
- (g) SLL course books based on corpora usually refer to more common and frequent choices over the rare ones to be

more accurate in description and effective in teaching. Even in traditional manner of SLL, corpora provide valuable information to students regarding the use of lexical collocation in understanding the patterns of word use (Gavioli 2004).

- (h) Corpora enable learners to understand various other aspects of language use such as principles that control use of idioms in sentences; rules related with patterns of word use and their semantic relations; network of lexis and grammar underlying the surface structure of various constructions including phrases, clauses and sentences; context-based use of words, set phrases, and idiomatic expressions; variation of language use across registers and text types, etc. Such linguistic features contribute to the growth of linguistic efficiency of learners both at primary and advanced levels (Hunston 2002: 176).

In the following sections we address some simple issues related to the use of corpora in SLL. We observe two basic types of use of corpora in this context: (a) as a primary resource, and (b) as a secondary resource. In the first type, both teachers and students are meant to access and refer to corpora directly either in classroom or language laboratory. In the second type, corpora are representative collection of language database where form teaching and reference materials are designed and developed for the benefit of target learners.

2. Corpus as a Primary Resource

In this part we address direct use of corpora in SLL. Here, learners are allowed to access corpora directly to extract relevant information to enhance their linguistic skills. In essence, corpora are meant to be the storehouses of language data of various types to provide wider perspectives to SLL hardly thought of in traditional methods.

2.1 Interactive Language Learning

Corpora contribute in computer-assisted interactive SLL - a method that proves to be more effective than traditional methods (Kettemann and Marko 2002). In this method, learners are directly exposed to corpora stored in computer for their access, utilisation and reference. This makes the students more enthusiastic about the language and its properties. As a result, they probe deep into the texture of the text to explore the intricate patterns of use of various elements of language. Obviously, their unbound curiosity about colourful use of language properties is triggered and they satisfy their thirst directly from corpora instead of depending on teachers.

For instance, an experiment was carried out over a course of part-of-speech teaching among the students constituting two different groups: one group was given direct access to corpora while the other group was taught via traditional lecturer-based method without reference to corpora. Results revealed that throughout the course, students who were allowed to access corpora, performed far better than students taught via traditional lecture-based method (McEnery, Baker and Wilson 1995). In another study it is observed that corpus-based teaching to undergraduates about the rudiments of grammatical analysis of sentences produces encouraging results over traditional lecture-based techniques (McEnery and Wilson 1996: 105). All these results strongly argue in favour of interactive SLL system based on corpora.

2.2 Data-Driven Language Learning

In data-driven learning (DDL), learners are allowed to act as language detectives (Johns 1997:101). They are allowed to discover themselves the real uses about the language properties they are learning. Since corpora are potential to reveal both known and unknown linguistic intricacies and features, students often find some new interesting examples or patterns of their choice unnoticed previously, overlooked or ignored by teachers. That means DDL improves the general skill of using context to deduce meaning of words, idioms, and phrases used in writings of the target language.

The method adopted in DDL usually sets up situations in which learners are asked to find out answers to the questions of the language themselves by studying the corpora presented to them in the form of concordance. Also, learners are allowed to access the database to address regular as well as rare issues related to development of their linguistic efficiencies by way of using information from corpora. For instance, since Bangla native speakers mostly falter in the correctness of the use of *the* in English writing, they may be allowed to explore corpora of native English to find out the most common and rare uses of *the*. They may also be asked to isolate sentences where the article is used and classify them according to the functional variations of the article. This will enhance skill of the learners in two ways:

- (a) Learners will be able to explore themselves how the word is used in real English texts, and
- (b) Errors in their interpretation will be verified and corrected by the teachers.

Recent developments in DDL, however, argue to stress on encouraging students to design their own corpus investigations. In this case, learners take advantage of searching through corpora when the task is not firmly fixed. They follow up any interesting observations that they come across. This is called *discovery learning* (Bernardini 2002) which is, however, most suitable for advanced learners who are filling up gaps in their knowledge about the target language they are learning rather than laying down the foundation stone.

2.3 Learners as Researchers

The value of corpora is indispensable to turn primary language learners into advanced language researchers [1]. Recent scenario of SLL turns language learners into language researchers with the following assumptions (Kirk 2002):

- (a) Learners are advanced students. They can enhance their linguistic skills by themselves by way of direct access to actual language databases.

- (b) Learners are able to carry out their own research on topics of their interest, rather than in-class activities designed by teachers.
- (c) Research issues are preferably related to advanced issues in syntax and grammar, rather than basic collocational patterns and use of words in language.
- (d) Research is based on large corpora (e.g. *British National Corpus*, *Bank of English*, *American National Corpus*, *Lancaster Learner Corpus*, etc.) rather than on small language samples.

In the class, the students are meticulously trained with the methodologies to carry out their own research programmes. This involves various processes such as learning to carry out searches on corpora, making hypotheses about the data in question, testing these hypotheses, etc. If required, students can use several corpora of various length and types along with simple search tools to carry out their own researches on a wide range of topics noted within a language. The basic goal of this process is to shift students from the status of language learner to the state of language researcher so that they follow correct methodologies in order to make valid claims about their mastery over the target language they are learning.

2.4 Error Correction in Language Learning

Corpora made with language samples used by the learners themselves are important resources for variety of research purposes (Barlow 2000). Systematic analysis of data stored in such corpora provides reliable evidence about the linguistic efficiencies the learners have acquired as well as the deficiencies that they carry in the process of learning. Also, analysis of such corpora by experts provide necessary impetus to improve linguistic skills of the default learners as well as to take necessary measures for enhancing their writing and speaking efficiencies[2].

This signifies that within the area of SLL, corpora made from written and spoken texts produced by Indian learners have immense potential to be considered as reliable databases for investigation and analysis, the result from which may be used for improving language

teaching techniques as well as providing necessary remedies to improve linguistic skills of the learners. In fact, there has been an increased tendency of such corpus-based educational applications for both large-scale assessment and classroom instruction (Mukherjee 2002). This has occurred due to the following factors.

- (a) Significant increase in the availability of computers and corpora in academic sectors starting from the elementary level to the university level.
- (b) Notable development in use of computers and corpora in SLL. This aims at incorporating advanced methods of NLP to evaluate and improve learners' skill.

Till date, these factors have remained instrumental to help undergraduate and graduate learners. As a result, computer-assisted SLL system happily inclined to incorporate speech and text corpora both for input and output in SLL.

2.5 Reciprocal Language Learning

Perhaps, the most exciting innovations in SLL in recent years is reciprocal learning. It occurs when two learners are paired, each helping the other learn language. For example, a native Bengali speaker learning English is paired with a native English speaker learning Bengali. To aid in reciprocal learning, generally parallel corpora^[3] are used, which due to their innovative application, make reciprocal learning an interesting task. Here, learners teaching each other are truly empowered and are likely to be genuinely motivated to make discoveries about each other's language. The role of a teacher therefore becomes of secondary importance - nothing more than that of a materials-provider (Hunston 2002: 182).

Information extracted from parallel corpora through concordance gives ample opportunities to the learners to access the variety of information hardly available from teachers. It provides information of how learning of a language will empower learners to speak, write, understand or translate from one language to other. As a matter of fact, students of a native language is exposed to the world of the Pandora's Box that give the feel about how a native user uses his language to interact in daily life.

The obvious restriction of reciprocal learning is that it can be undertaken only in a situation where there are at least two students - each one of which comes from different mother tongues (e.g. English-Bangla) to learn other's language. This is, however, not the situation easily available in most of the Indian contexts. Here, Indian students learn English in indirect method, since a native English speaker is hardly found who is interested to learn an Indian language as well as to teach English to an Indian student in a reciprocal interactive process.

2.6 Learning Sense Variation of Words

Recent corpus-based approach to SLL contributes towards establishment of an objective criterion to semantic study of words. It holds the view that actual meaning of words can be derived only from the context in which they actually occur (Schütze 1997: 142). Therefore, efforts are made to find out how information obtained from corpora provides objective criteria for assigning meanings to words so that the learners face no difficulty to comprehend them correctly (Mindt 1991).

Quite frequently, meaning of words is described with reference to teachers' knowledge about the target language. However, corpora reveal that semantic distinctions of words are associated with numeral characteristically observable contexts. Similarly, compounds, multiword units, collocations, idioms and phrases require relevant contextual information for proper understanding on the part of the learners. This entails that reference to the environments of occurrences observed in corpora supplies learners the much-needed objective empirical base to decipher finer semantic distinctions of linguistic items. By looking empirically at the evidence stored in corpora, learners understand that fuzzy model of word meaning accounts much better, since there is no clear-cut boundary among the categories of words (Leech, Francis and Xu 1994). Rather, there is gradience of meaning connected with frequency of use of a particular word in particular sense in the text.

Use of corpora also help learners to understand the nature of lexical polysemy by which words are capable to denote multiple senses due to variation of contextual frame. In some recent

experiments, it is observed that the number of sense distinctions that shows up in corpora far exceeds the number of sense distinctions provided in standard dictionaries (Fillmore and Atkins 2000). Also, in a recent study, it is observed that learners can understand the polysemous nature of words in a far better way if all usage variations of words are extracted from corpora and referred to with close reference to the context of their use (Dash 2004). This helps learners to capture multiple sense variations of words as well as identify actual senses of words to find equivalent items from corpora made from their mother tongue.

Although, the problem of sense disambiguation of words has been one of the major concerns in SLL for ages, the scheme proposed in WordNet (Miller et al. 1993) posits considerable amount of problem to the learners, since it fails to show how senses of words are interlinked to each other. The problem becomes more acute when new senses of words occur quite frequently and figurative senses are tagged to them. Since figurative use of words is pervasive in normal discourse, source meaning of lexical items is often removed from intended meaning. There are three possibilities to overcome this problem (Pustejovsky 1991)^[4]:

- (a) List up all sense variations of words in a lexical dictionary.
- (b) List only a few recurrent senses and employ a generative mechanism to produce new senses.
- (c) Design a mechanism to identify and treat sense relations between the words.

Figurative use of words (and other linguistic items) is a common phenomenon in all domains of a natural language. It is, therefore, quite pertinent to investigate figurative senses of words by focussing at their use in corpora. In fact, corpora provide required information to learners to explore the following issues of figurative sense of words.

- (a) Corpora illuminate concepts of literalness, metaphor, metonymy, polysemy, context-sensitive meaning and their relations to figurative senses.
- (b) They provide necessary information to learners to capture inter-annotator agreement on which figurative use, metaphor, metonymy, etc. are constituted.

- (c) They supply specific linguistic cues to the learners to explore the nature of figurative use of words, study their frequency of occurrence, reliability of estimation, and evaluation of their generativity.
- (d) They provide documentary evidence to trace the effects of domain, genre and discourse on the use of figurative sense of words in a language.
- (e) They supply significant perspectives towards formation of cognitive-psychological models for interpreting figurative senses of words.

2.7 Learning Stylistic Variation of Usage

Availability of corpora with samples from different genres, domains, authors, and media opens up new possibilities of studies into stylistics for the learners. In advanced language learning scheme, learners are exposed to individual text types or texts composed by authors with specific stylistic nuances. For instance, advanced learners are trained to find out the basic stylistic differences reflected in the texts composed by the writers of one country with that of other country, while some others are guided to know how the writings of one generation or group of writers differ stylistically from the writings of other generation or groups. Such comparative stylistic training is possible only when the learners have under their disposal large collection of synchronic and diachronic corpora representing various stylistic features considered relevant in language education.

Although some teaching methods are interested to expose the broader issues such as genre and type of texts, a large number of regular SLL systems deal with stylistic variations concentrating on specific features of certain text types. For instance, students are taught to learn how language used in newspapers varies stylistically from language used in scientific writings. Such teaching requires relevant corpora for both faithful analysis and reliable conclusions. Here, corpora become useful for studying changes of style as well as defining an author's particular style of writing. Learners can also use corpora made from writings of authors to identify the degree to which authors lean towards various ways of putting things (e.g.

technical vs. non-technical terms, long vs. short sentences, choice of vocabulary, formal vs. informal manner of narration, etc.). This will exhibit not only the styles of writing of the authors under consideration but also the styles in which they usually compose the entire text. The English language corpora, which are quite rich with information of genre and register variations, are good resources for English language teaching (ELT) to the Indian students. At the very initial stage, the learners will be allowed to access these corpora built with specific features of text types to begin with simple general comparisons about the varieties observed within the samples. Due to the fact of easy comparability and referential ability, these corpora will be valuable resources for learning various stylistic features within and across text types for the Indian students.

3. Corpus as a Secondary Resource

In this part we address the use of corpora in formation of SLL materials such as bilingual dictionaries, terminology databases, syllabuses, and grammar books. We argue that each of the materials is able to excel over the traditional resources, since corpus-based materials are far more copious and diversified with the information of various types obtained from the language of actual use.

3.1 Designing Syllabus for SLL

Use of corpora leads us to different ways of designing syllabi, since language looks different when we look at a lot of it at once (Sinclair 1991: 100). The type of syllabus we refer to here is called lexical syllabus, which is based on studies of corpora[5]. In its simplest sense, in SLL, it makes sense to teach the most frequent words found in the target language. In other words, the main focus of SLL is to project on the commonest words, the central patterns of their usage, and their combination of all types (Sinclair and Renouf 1988: 148). Lexical syllabus does not encourage the standard practice of acquisition of large vocabulary, especially at the initial stage. Instead, it concentrates on making full use of the words that the learners already have at a particular stage of

learning. It argues that there is far more general utility in recombination of known elements than the addition of less easily used lexical items (Sinclair and Renouf 1988: 155).

The basic argument of lexical syllabus is as follows: if learners are exposed to the most frequent words, which have a variety of uses, they will acquire a flexibility of language fairly easily. Since salient uses of the most frequent words cover the central structure of a grammar, learners should be exposed to this area in first opportunity. For instance, in English, MAKE has numerous uses, some of which are rarely covered in most of the beginner s'courses. Analysis of corpora shows that the most frequent use of the verb is found in combinations such as 'make decisions', 'make discoveries', 'make arrangements', 'make profit', 'make love', 'make fool', 'make noise', 'make road', 'make a choice', etc. rather than in more concrete forms such as 'make a cake', 'make a house', 'make a doll', etc.^[6]. An English teaching course that focuses only on the concrete sense of MAKE actually denies learners the opportunity to learn expressing sophisticated meanings with a simple verb.

In lexical syllabus, separate listing of grammatical items is not necessary. What is traditionally termed grammar is called pattern here (Wills 1990: 51). In other words, the most productive way of interpreting grammar in the classroom is the lexical patterning that accounts for all the patterns involving the most frequent lexical items. Since patterns attach to all lexical items in the language, learning the lexis means learning the patterns, and therefore the grammar. In support of this observation, scholars have argued, If the analysis of the words and phrases has been done correctly, then all the relevant grammar etc. should appear in a proper proportion. Verb tenses, for example, which are often the main organizing feature of a course, are combinations of some of the commonest words in the language (Sinclair and Renouf 1988: 155)^[7].

In lexical syllabus, a syllabus consists of several corpora made from real instances of language use. Course designers collect corpora of authentic texts that contain instances of the most frequent patterns of the most frequent words, so that they exemplify what the learners need to know. The job of teachers is to devise ways of encouraging learners to engage with the materials in

corpora and helping them learn to notice the patterning of the language use. In essence, the description of the syllabus is equal to the description of the corpora. If the syllabus contains a list of items, it is the list of the most frequent word-forms found in the corpora, along with their most typical phraseologies. As the texts that constitute the corpora are presented to the learners, the syllabus is inevitably covered.

This method has the ability to alter the entire role of the syllabus designers quite considerably. The syllabus designers, instead of selecting specific items of language description and choosing texts to illustrate them –

- (a) choose interesting text samples of their choice,
- (b) keep a check on the balance of the overall collection of text samples,
- (c) ensure that the most frequent word-forms and their typical phraseologies are covered, and
- (d) verify if the syllabus matches with what the learners require to learn.

Here, of course, there is an element of subjectivity. The syllabus designers may aim at mirroring the distribution of structures, word frequency and phraseology in larger, general corpora. Else, they may decide that different types of corpora are more appropriate with regard to learners' age and specific needs. This subjectivity has an advantage of making an appeal to principle rather than to conventional wisdom. A syllabus of this kind has advantage of answering the objection that lexical syllabus leads to artificial teaching materials if text samples are written specifically to demonstrate key lexical items in the language (Long and Crooke 1992: 33)^[8].

3.2 Formation of Bilingual Lexical Dictionary

Another essential part of corpus-based SLL is the development of bilingual lexical dictionary (BLD), the lack of which has been one of the greatest bottlenecks in present English education in India. Traditional lexical dictionaries available in market can hardly compensate this, since these dictionaries do not contain enough information about lexical sub-categorisation, lexical selection

restriction, and domains of application of lexical items. It is observed that a BLD built with information about lexical sub-categorisation extracted from POS-tagged corpora has tremendous effect in SLL. Even when POS-tagged corpora are not available, a BLD made from untagged corpora is equally useful if utilised in a sensible way. ^[9]

With regard to content, a BLD (e.g. English-Bengali) usually includes lists of the most frequently used words keeping in mind their utility and relevance in SLL. Such a dictionary aims at providing equivalent words from target language to native language. For instance, if we are supposed to develop a BLD with English as target language and Bengali as native language, then the following aspects will arrest our utmost attention:

- (a) It will include exhaustive list of words collected from target language corpora made with texts from language of various types.
- (b) Collected words will be sorted in alphabetical order to compile a lexical database of the target language necessary for SLL.
- (c) Each word included in the database will be provided with equivalent forms derived from the native language corpora.
- (d) Each word-form will be provided with more than one equivalent forms taking into account the domains of their usage. For instance, *delivery* will be supplied with three Bengali equivalent forms: *pfosOb kOrA* , '*jŕgAn deoyA* ' and *bŕktritA deoyA* . The first will refer to the domain of medical science, the second one will refer to the supply of materials and goods, and the third one will refer to classroom lectures or public orations at mass rally.
- (e) Selection of appropriate equivalent word form native language will depend on the domain of use of the word in target language. For instance, let us consider the following examples where English *deliver* is appropriately used in Bengali with due consideration of the domain of use of the term.
 - (i) English: Mrs. Jonathan has delivered a girl child in the hospital.

Bengali: hAspAtAle Mrs. JonAthAn ek konyA sOntAner
jOnmo diyechen.

(ii) English: Prof. Brown delivered a fine lecture on child
education.

Bengali: shishushiksAr upOr OdhyApOk BrAun ek
mOnogrAhi boktritA diyechen.

(iii) English: The courier boy delivered the packet in the
evening.

Bengali: kyuriyArer cheleTi sondhyAbelAy pyAkeTTi
pōuche diye geche.

Examples cited above shows *deliver* carries three distinct senses in the native language. It is translated into Bengali in an appropriate manner taking into consideration the domain of use of the term. In the field of childbirth, the most appropriate term in Bengali is *prasOb kOrA* or *jOnmo deoyA*; but in the area of lecture in meeting or mass rally it is either *boktritA deoyA* or *bhAsOn deoyA*; and in the area of postal distribution and supply of goods it is either *pAThAno* or *pōuche deoyA*. The most striking aspect of the examples is that in the target language itself, what it means in the field of medicine and childbirth is not similar in the domain of postal distribution, supply of goods, and lecture in mass rally. That means, by considering the domains of use of the terms in target language, we have to select the most appropriate terms from the native language. Information of this kind will definitely enhance the linguistic skills of the learners.

Extraction of semantically equivalent forms from corpora (of both target language and native language) is a complicated task. The search for equivalent forms in target corpora begins with particular words, which express certain meaning or concept in target language. After the words are identified in target language and stored in a separate word list, the second stage of searching begins. This involves identification of a word in native language, which is similar in sense or meaning of the word in target language. Normally, corpora reveal a wider range of equivalent forms, which are potential to be considered as alternative forms. However, the factors that usually determine the choice of the most appropriate equivalent forms are determined on the basis of recurrent patterns

of use of the words. Also, equivalent forms found in each corpus are verified further with original monolingual corpora of the two languages. The following diagram (Fig. 1) presents a simple schema about how equivalent lexical items may be obtained from corpora of target and native language.

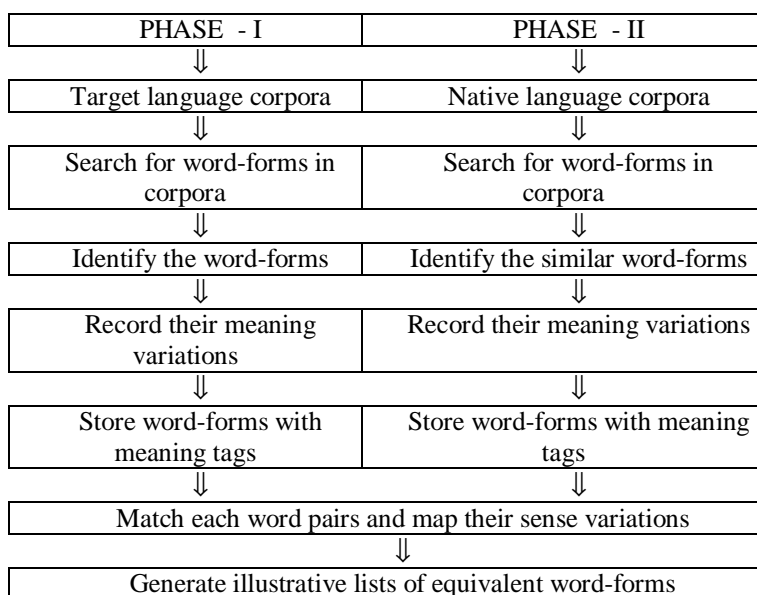


Fig. 1 Extraction of equivalent word-forms from target and native language corpora

However, even within two closely related languages, equivalent word-forms seldom mean the same thing in all contexts because they are seldom used in the same types of syntactic construction. Even in native language, a word may be used as an equivalent lemma for the word, which is not a lemma in target language. Therefore simple semantic equivalency may be possible with names of animals and tools or with some scientific terms, but hardly with ordinary words (Landau 2001: 319). This implies that selection of equivalent matches for ordinary word-forms will require high degree of linguistic sophistication to yield expected outputs. In fact, the process of extracting equivalent word forms from target and

native language corpora and their subsequent verification for authentication will require occasional intervention on the part of the expert linguists.

For a long time, BLD has remained a minority activity in SLL. However, introduction of data-processing technology and corpus linguistics makes available some new and widely used tools to the learners - due to which it has become difficult to distinguish BLDs from other types of reference materials^[10]. The pedagogical purpose of BLD is to give learners information about the words they do not already possess. However, they should have certain amount of lexical and syntactical knowledge of the target language before they actually start using BLDs as supporting tool in their learning. The role of a teacher here is to collect relevant lexical items from corpora and present them to the students. However, he should take care so that the database does not present excessive lexical load and other difficulties to the learners. Thus, a teacher with the help of a lexical list obtained from corpora can shape up attitudes and confidence in comprehension leading to the ability to discriminate between what needs to be looked up and what can be intelligently guessed in SLL.

3.3 Bilingual Dictionary of Idioms and Proverbs

Familiarity with wide range of idioms and proverbs and the ability to use them properly are some of the distinguishing marks of a native-like command over a foreign language. Scholars have argued about important functions of idioms and proverbs in SLL: Idioms are little sparks of life and energy in our speech; they are like those substances called vitamins which make our food nourishing and wholesome; diction deprived of idiom ...soon becomes tasteless, dull, insipid (Smith 1925: 276). Proverbs continue - as the early collectors never tired of stating - to provide the sauce to relish the meat of ordinary speech (Simpson 1982: x-xi). In reality, although a non-native speaker uses language fluently and grammatically, he hardly expresses himself idiomatically - in a way a native speaker picks and chooses words and phrases. Consequently, what a non-native speaker says often sounds a bit awkward, flat and lengthy. For example, while a non-native speaker of English usually finds it

easier to use the non-idiomatic single words, native speakers tend to use idiomatic expressions and phrasal verbs (Xiao-jun 2003: 296). Full of vitality, idioms and proverbs fill in blanks in actual communication and thus make it easier for people to exchange thoughts with each other, because they use them without special effort of formulation. Thus, it is argued that learning idioms and proverbs (specially those with a figurative sense) is an essential part of SLL, since learners find these puzzling idioms their main difficulty in learning English (Henderson 1954: 5)[11].

Idioms and proverbs are best obtainable from target language corpora along with information of their contextual usage. This helps learners to understand the actual figurative sense of the expressions as well as the patterns of use of these forms within natural environments. A bilingual dictionary of idioms and proverbs carries authentic relevance in SLL by way of containing collocations, idioms, phrases, and multiword units. Here also, the appropriate equivalent forms are provided from target language corpora for easy access and quick learning. Teachers may, however, consider integrating domain-specific information and usage in the dictionary.

For the purpose of generation of a BDIP, we use various statistical methods normally applied on both tagged and parsed corpora in various ways. Albeit there are varieties involved in application of the statistical procedures, in most case, the goal has centred round the followings:

- Retrieval of larger comparable syntactic blocks (e.g. idioms, phrases, multiword units, clauses, etc.) from both target and native language corpora.
- Extraction of most frequently used idioms, collocations, phrases, and proverbs from corpora.
- Matching of idioms, collocations, phrases, and proverbs with database of both the languages.

One, however, cannot expect hundred-percent success in matching of the linguistic items at the idiom, phrase and conceptual level within the two languages although the languages are closely interrelated.

3.4 The English-Bengali Electronic Dictionary

The proposed English-Bengali Electronic Dictionary (EBED)[12] has two main parts. The first part contains nearly fifty thousand most frequently used English words obtained from English corpora of various types. Their semantically equivalent words are obtained from Bengali corpora, which contain large collection of text samples obtained from books and journals spanning over a few decades. Never before such an attempt is made to present English words as well as their Bengali equivalent forms with reference to corpora either in electronic or in printed form. This database will not only enrich language learners with knowledge of the language they are learning but also will provide most reliable resources to the translators for appropriate translations of English texts into Bengali[13].

The second part includes collocations, idioms, phrases, and proverbs. We intend to emphasize on these language properties keeping in mind the need of advanced language learners who want to acquire mastery over the language with all its intricate and finer nuances. We must agree with the general argument that until and unless a learner is capable to understand the actual meaning hidden behind the surface of an idiom, collocation, phrase or proverb used in a sentence, the knowledge of the learner is not complete. Since majority of source texts carry a large amount of expressions, which are idiomatic and phrasal in sense and connotation, a dictionary with such information is an excellent resource for the learners in language learning and translation from English to Bengali^[14].

The proposed EBED is more or less identical with traditional dictionaries with one exception. It contains a database that includes large number of English words obtained from various English corpora along with English words obtained from the Bengali corpora. In a simple calculation it is noted that there are nearly ten thousand English words most of which are not included in standard English-Bengali dictionary although they have entered into regular speech and writing. If the database is available online, then learners do not have to search for necessary information in printed form. The attested use of examples and instances obtained from corpora

will definitely make the whole process of language learning more authentic, interesting, and useful.

The question of supplying dictionarial information about the words used in text by a simple click of mouse in classroom teaching needs an interactive interface between the corpora and the dictionary. We have to interconnect the two (i.e. corpora and dictionary) in such way that a hit on a particular word will directly refer to that word entered in the dictionary. But before that work is done we need to have full-fledged electronic dictionary made with words included in the corpora. Moreover, the dictionary should have all types of lexicological information related to the words (e.g. orthography, pronunciation, etymology, spelling, grammatical information, synonyms, antonyms, example, citation, usage, illustration, etc.) directly obtained from corpora and other sources. If we can build up a system like this, then SLL will be much more interesting and learners will be self-sufficient by exploiting the corpora and the dictionary simultaneously. The role of the teachers will be that of co-investigators or co-ordinators - thereby reducing much of their teaching load usually carried out in standard classroom teaching.

3.5 Development of grammar books

Language corpora are quite frequently used for grammatical (and syntactic) studies of various types (Halliday 1991). They are useful resources to students, since they provide information about wider grammatical varieties observed in a language. Also, they supply actual empirical instances for testing earlier hypotheses about various grammatical theories. Normally, grammars of a language are made on grammarians' intuition about the language rather than on real proofs of actual use. Therefore, whatever conclusions the grammarians like to make and however fantastic these observations appear, these are not beyond the scope of verification with evidences of *performance* reflected in actual use. Even generativists will decline to agree with the assumptions of intuitive grammarians unless these are verified with examples actually occurring in real life situations. However, availability of corpora makes it possible to study the *performance* of language users to know how language is actually used by people in daily course of life. Within last few years

some grammar books made from corpora have substantiated the value of empirical grammars in SLL (Quirk, Greenbaum, Leech and Svartvik 1985). This supports the argument that claims every (formal) grammar is initially written on the basis of intuitive data; by confronting the grammar with unrestricted corpus data it can be tested on its correctness and its completeness (Aarts 1991: 48).

Following this principle we propose to use corpora for writing true usage-based grammars to be used in SLL. These will differ from traditional grammars not only in proposition of theories but also in formation of rules and principles. Moreover, new grammars will faithfully reflect on the language actually used by the people. Thus, these grammars, if used intelligently, will be useful for language learning as well as for machine translation. Since corpora contain collections of literary works as well as texts obtained from various fields of human knowledge, these grammars will have much wider coverage than traditional grammars can ever aspire to achieve.

Each text sample will be categorized according to its type so that these are produced before the learners either to give them ideas about the use of language in specific fields or to supplement their knowledgebase about the varieties of language use controlled by linguistic and non-linguistic factors. Also, the learners can use these grammars and the database to examine the mutual as well as exclusive distribution of sentence types across genres to understand syntactic varieties cultivated in a language. This will enrich learners to know what kinds of sentences are actually used in writing in target language depending on the type of text. In fact, corpus-based grammars are meant to be excellent resources both for the beginners as well as advanced learners.

4. Conclusion

The paper has a simple and humble goal. It aims at showing how language corpora of various types can be used in second language education to non-native speakers. In a systematic way we have tried to highlight how corpora can be used broadly for two types of activities related to SLL: (a) corpora as pure text databases, which could be directly accessed by the learners in classroom situations,

and (b) corpora as source of linguistic information to be used for designing and developing language teaching aids like dictionaries, grammars, and other reference materials.

It is a fact that in recent years introduction of corpora in second language education has made a remarkable breakthrough. This has lent towards modification of systems traditionally used for SLL to non-native people. However, the most unfortunate thing is that, in comparison to other countries, India lags far behind not only in corpus generation but also in corpus-based language education (both as first and second language).

There is no scope for any kind of speculation about the applicational relevance of English corpora in teaching English as a second language to Indian students. The time has probably come to direct our attention towards this new empirical approach to rejuvenate the discipline with new lease of life. Indian learners will not only benefit from this technique but also will excel in the language when they are put into competition with foreign contenders. However, before we adopt this system we should make sincere attempts to generate corpora of Indian English as well as procure corpora of British and American English to be used in SLL.

NOTES:

1. A concrete example of this method is recently exhibited by Mark Davies of Illinois University, USA. To teach about the Variation in Spanish Syntax in the class, advanced students (many of them high school teachers) are provided with several corpora of Spanish and search tools like Google to carry out their own research on a wide range of topics relating to syntactic variation in Spanish (<http://mdavies.for.ilstu.edu/sintaxis>). Almost similar process is adopted by Claire Kennedy and Tiziana Miceli of Griffith University.
2. In fact, with this goal, the International Corpus of Learner English is generated, which contains extracts of writings produced by the learners coming from different countries who have learned English as a foreign language. At present, the corpus is under analysis to know how the learners have acquired efficiency in the language or if they lack in their linguistic skills of expression both in speech and writing.

3. Parallel corpora are also useful for teaching translation or for more conventional language learning in situations where all learners share a common first language.
4. The scheme of Pustejovsky (1995) aims at assigning a structure to words to determine how different senses are combined in them. It may succeed with some simple cases of metonymy, but how it will cope with metaphoric senses of words is still an open question. Also, it may fail to make distinctions between metonymy and metaphor for the learners. Since meanings are not marked with information about their metaphorical or metonymical sense, distinguishing literal meaning from non-literal meaning is a crucial task for the learners who will ask for extensive analysis of words collected from corpora.
5. The idea of lexical syllabus was first proposed in a paper by Sinclair and Renouf (1988). However, it finds its fullest exposition in Wills (1990). The term is normally misinterpreted to indicate a syllabus that consists of only list of vocabulary items. But, in practice, scholars use the term from a different perspective. It differs from a conventional syllabus in the sense that its central theme is the organization of lexis.
6. In Sinclair's (1991:101) terminology, in present English texts MAKE is used as a delexical verb more frequently than as an ordinary verb.
7. In support of this observation, Wills argues that English is a lexical language in the sense that many of the concepts we traditionally think of as belonging to grammar can be better handled as aspects of vocabulary. For example, the passive can be seen as BE plus an adjective or past participle, rather than as a transformation of the active (Wills 1990:17). Similarly, conditionals can be handled by looking at the hypothetical meaning of 'would', rather than by proposing a rule about sequence of tenses, that often does not work (Wills 1990: 18-19).
8. In fact, the concept of Wills (1990) about a corpus constituting authentic collections of texts is not dissimilar from the model of task-based syllabus proposed by Long and Crookes (1992). Indeed, it is a more concrete entity.
9. In fact, the formation of a BLD is best possible within those cognate languages, which are typologically and/or genealogically related to each other (e.g. Bengali and Oriya, Hindi and Urdu, etc.), because such cognate languages usually share many common properties (both linguistic and non-linguistic) rarely found in other language types. Moreover, there is large chunk of

regular vocabulary similar to each other not only in phonetic and orthographic representation but also in their sense, content, and implication (Dash 2005: 363).

10. For example, BLD is replaced by recent introduction of a bilingual thesaurus, which records a range of English literary and newspaper texts and provides access to this contextual corpus by means of alphabetical and thematic indexes in both English and Chinese (Liu 1999).
11. In 1991, a Ph.D. scholar while doing research on second language learning at the Cambridge University, UK, commented that, 'It is impossible to acquire a thorough knowledge of any language without being familiar with slang and vulgarism' (known from personal source).
12. The author of this article is now engaged in the development of the English-Bengali Electronic Dictionary with close reference to the database obtained from the British National Corpus and the Bengali Corpus of Written Texts (Dash 2005: 51-90).
13. After the completion of dictionary we decide to arrange the whole database in reverse order, i.e. Bengali will be used as target language while English will be used as native language. In this case pronunciation of Bengali words will be rendered in IPA so that people who are learning Bengali as their second language will be able pronounce these forms correctly. However, this is a remote venture, which can only be started after the completion of the first phase.
14. We also intend to use this dictionary for the purpose of machine translation where dictionary of idioms, phrases, collocations, and proverbs is proved to be highly necessary and useful.

References

- Aarts, J. 1991. Intuition-based and observation-based grammars. In Aijmer, K. and B. Altenberg (Eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman. Pp. 44-62.
- Barlow, M. 1996. Corpora for theory and practice. *International Journal of Corpus Linguistics*. 1(1): 1-38.
- Barlow, M. 2000. Parallel texts in language teaching. In Botley, S.P., A.M. McEnery, and A. Wilson (Eds.) *Multilingual Corpora in Teaching and Research*. Amsterdam-Atlanta, GA: Rodopi. Pp. 106-115.
- Bernadini, S. 2002. Exploring new directions for discovery learning. In C.B. Kettemann and G. Marko (Eds.) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam-Atlanta, GA.: Rodopi. Pp. 42-51.

- Biber, D. 1996. Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics*. 1(2): 171-198.
- Botley, S.P., A.M. McEnery and A. Wilson (Eds.). 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam-Atlanta, GA.: Rodopi.
- Dash, N.S. 2004. Issues involved in the development of a corpus-based machine translation system. *International Journal of Translation*. 16(2): 57-79.
- Dash, N.S. 2005. *Corpus Linguistics and Language Technology*. (With Reference to Indian Languages). New Delhi: Mittal Publications.
- Fillmore, C.J. and B.T.S. Atkins. 2000. Describing polysemy: the case of *crawl*. In, Ravin, Y. and C. Leacock (Eds.) *Polysemy*. New York: Oxford University Press Inc. Pp. 91-110.
- Gavioli, L. 2004. The learner as a researcher: introducing corpus concordancing in the language classroom. In, Aston, G. (Ed.) *Learning With Corpora*. Cambridge: Cambridge University Press. Pp. 31-45.
- Ghadessy, M., A. Henry, and R.L. Roseberry (Eds.). 2001. *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam/Philadelphia: John Benjamins.
- Granger, S., J. Hung, and S. Petch-Tyson (Eds.). 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.
- Halliday, M.A.K. 1991. Corpus studies and probabilistic grammar. In, Aijmer, K. and B. Altenberg (Eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman. Pp. 30-43.
- Henderson, B.L.K. 1954. *A Dictionary of English Idioms. Part I: Verbal Idioms*. London: James Blackwood.
- Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Johns, T. 1997. Contexts: the background, development, and trialling of a concordance-based CLL program. In, Wichmann, A. S. Fligestone, T. McEnery, and G. Knowles (Eds.) *Teaching and Language Corpora*. London: Longman. Pp. 100-115.
- Kettemann, C.B. and G. Marko (Eds.). 2002. *Teaching and Learning by Doing Corpus Analysis. Language and Computers: Studies in Practical Linguistics 42*. Amsterdam-Atlanta, GA.: Rodopi.
- Kirk, J.M. 2002. Teaching critical skills in corpus linguistics using the BNC. In, Kettemann, C.B. and G. Marko (Eds.) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam-Atlanta, GA.: Rodopi. Pp. 183-197.
- Kübler, N. 2002. Linguistic concerns in teaching with language corpora: learner corpora. In, Kettemann, C.B. and G. Marko (Eds.) *Teaching*

- and Learning by doing Corpus Analysis*. Amsterdam-Atlanta, GA: Rodopi. Pp. 133-145.
- Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. (2nd edition) Cambridge: Cambridge University Press.
- Leech, G., B. Francis, and X. Xu 1994. The use of computer corpora in the textual demonstrability of gradience in linguistic categories ."In, Fuchs, C. and B. Vitorri (Eds.) *Continuity in Linguistic Semantics*. Amsterdam & Philadelphia: John Benjamins. Pp. 31-47.
- Liu, David H. 1999. *Chuan-Shi: the Chinese-English Thesaurus*. Taipei: Private publication.
- Long, M.H. and G. Crookes. 1992. Three approaches to task-based syllabus design ." *TESOL Quarterly*. 26: 27-56.
- McEnery, A. and A. Wilson 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- McEnery, A., J. Baker, and A. Wilson. 1995. A statistical analysis of corpus based computer vs. traditional human teaching methods of part of speech analysis ." *Computer Assisted Language Learning*. 8(2-3): 259-274.
- Miller G., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. *Five Papers on WordNet: An On-line Lexical Database*. CSL Report 43, Cognitive Science Laboratory, Princeton University.
- Mindt, D. 1991. Syntactic evidence for semantic distinctions in English ." In, Aijmer, K. and B. Altenberg (Eds.) *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman. Pp. 182-196.
- Mukherjee, J. 2002. Norms for the Indian English classroom: a corpus-linguistic perspective ." *Indian Journal of Applied Linguistics*. 28(2): 63-82.
- Pustejovsky, J. 1991. The generative lexicon ." *Computational Linguistics*. 17(4): 214-229.
- Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Schütze H. 1997. Ambiguity Resolution in Language Learning: Computational and Cognitive Models. Cambridge: Cambridge University Press.
- Simpson, J. 1982. *The Concise Oxford Dictionary of Proverbs*. Oxford: Oxford University Press.
- Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. and A. Renouf. 1988. A lexical syllabus for language learning ."In R. Carter and M. McCarthy (Eds.) *Vocabulary and Language Teaching*. London: Longman. Pp 140-160.

- Smith, L.P. 1925. *Words and Idioms*. London: Constable.
- Wills, J.D. 1990. *The Lexical Syllabus: A New Approach to Language Teaching*. London: HarperCollins.
- Xiao-jun, H. 2003. Lexicographical treatment of idioms and proverbs . In, Hartmann, R.R.K (Ed.) *Lexicography: Critical Concepts*. Vol.-II. London and New York: Routledge. Pp. 295-312.