Status of Corpus Linguistics in India

B. A. Mahalakshmi Prasad, M.A.
K. S. Prema, Ph.D.
Prarthana. S., Research Fellow

==========
Language in India www.languageinindia.com **ISSN 1930-2940 Vol. 13:8 August 2013**
==========

Abstract

Linguistic research has been a preoccupation of humans from times immemorial. Philosophers and scholars from various disciplines have expended considerable time and effort to understand the nature of language and language use to gain an insight into human nature. In the realm of linguistic research, the discipline of corpus linguistics has managed to attract the thoughts of linguists, computer scientists, teachers, speech language pathologists and people working in language technology. This paper, while mentioning primary assumptions of corpus linguistics, tries to highlight the need for establishing language corpora in a plurilingual context of India.

Key words: corpus linguistics, speech language pathology

Introduction

The discipline of linguistics has undergone development with the renaissance of corpus linguistics that heralds a new understanding about the theories and assumptions regarding the nature of language. In the 1960s, corpus linguistics brought in a revolution to the discipline of linguistics by providing a platform for researchers to explore what 'language is' rather than 'what language ought to be'.

A *corpus* is defined as a collection of texts that acts as a tool, which represents a given language that can be used for linguistic analysis as enumerated by Francis (1964). Thus, a corpus consists of a databank of natural texts compiled from writing and/or transcription of recorded speech. In order to conduct a study of language, which is corpus-based, it is necessary to gain access to a corpus and a concordance program. A concordance is a software program, which analyzes corpora and lists the results. Even though originally corpora were regarded as mere tool for linguistic work, the main focus of it shifted to discover patterns of authentic language use by analysing natural usage of language. It also helps to understand the language behaviour across population. However, this field of corpus linguistics was not welcomed with open mind during its advent.

a) Overview of Corpus-based Studies

Over the past decades, since 1950, majority of corpus-based studies have been reported for English and other non-Indian languages. The focus of those studies is either on language pedagogy, language acquisition, spelling or the type of studies undertaken for understanding language corpus.

The following section provides an overview of the status of corpus-based studies in these dimensions.

i) Studies on Language Acquisition

The studies of child language in the 19th century (e.g., Ament 1899; Compayre 1896; Major 1906; Preyer 1882; Ronjat 1913) were based on carefully composed parental diaries recording the child's locutions. These primitive corpora are still used as sources of normative data in language acquisition research today (for example, Ingram, 1978). Corpus collection continued but was diversified by collecting language samples from large groups of children during the 20th century (Stern &Stern 1928; Campbell 2006). Analysis of language corpus was carried out with the aim of establishing norms for language development. Longitudinal studies, though on a smaller sample of children (for example, Bloom, 1970; Brown, 1973), have been documented from 1957 till date. Whether parental dairy-based or with longitudinal design, studies employing language corpus have its own inherent merits and demerits.

ii) Studies on Language Pedagogy

The corpus and second language pedagogy had a strong link in the early half of the twentieth century, with vocabulary lists for foreign learners often being derived from corpora (Kennedy 1992). The word counts derived from such studies as Thorndike (1921) and Palmer (1933) were important in defining the goals of the vocabulary control movement in second language pedagogy.

Fries and Traver (1940) and Bongers (1947) used the corpus in research on foreign language pedagogy. Eaton (1940), a comparative linguist, compared the frequency of word meanings in German, Italian, French and Dutch with the corpus data available in all the four languages. Fries (1952) created a corpus of transcribed telephone conversations, and transformed the corpus to generate descriptive grammar of English. He also studied syntax and semantics using the corpus data. His pioneering work provided a model for the corpora of English developed by Quirk, Greenbaum, Leech and Svartvik in 1985, the data that was meaningfully adopted almost 30 years later for better understanding of language.

In the early 20th century, with the advancement of technology, impetus was given to machine-readable corpora. Machine-readable corpora were developed with material that was originally produced for some other purpose. For example, Brown Corpus and Brown clones, Brown University Corpus by Francis and Kucera (1964) (American-English one-million-word sample corpus consisting of 500 texts chosen from 15 text categories. Each text has about 2000 words), the Lancaster/Oslo-Bergen (LOB) Corpus by Geoffrey Leech in 1970s with the same selection scheme and number of words as Brown Corpus, International Corpus of English (ICE) (consists of 18 Brown-style corpora taken from 18 countries where English is the native or official language), Bank of English by COBUILD and the University of Birmingham, (1982) which is a monitor corpus comprising about 450 million running word forms and British National Corpus (BNC, 1995, 100-million-word sample corpus, 90 million written, 10 million spoken words). The literature suggests that good amount of corpora has been established and utilized for various purposes especially in English language. In order to establish such corpora, the researchers have embraced different approaches for the study of corpus linguistics.

b) Approaches to corpus Linguistics

There are two main approaches to linguistic research, namely, corpus driven and corpus-based methods.

In a corpus-based approach, independent theories are developed that are later tested using the primary facts of a corpus. Here, corpora are used to expound, test, or exemplify theories and descriptions that were formulated before large corpora became available to inform language study. This approach helps to label corpus linguistics as a methodology that does not restrict the study to one particular aspect of language and gives a holistic approach to the discipline of linguistics (Togini-Bonelli, 2001). Similarly, according to Wu (2002), quantitative techniques are also essential for corpus-based studies to derive and understand patterns of language use. For example, if one wanted to compare the language use of patterns for the words 'big' and 'large', one would need to know how many times each word occurs in the corpus, how many different words co-occur with each of these adjectives (the collocations), and how common each of those collocations is. However, quantitative measurements are limited since a crucial part of the corpus-based approach is going beyond the quantitative patterns to propose functional interpretations explaining why the patterns exist. As a result, a large amount of effort in corpus-based studies is devoted to explaining and exemplifying quantitative patterns.

In the corpus driven approach, the data is analysed without any pre-conceived notion in relation to how it should be analysed or regarded. In corpus driven analyses, theories are developed by examining the primary facts of a corpus directly. The assumption in this approach is that corpus linguistics is a discipline that defines a whole system of methods and principles besides suggesting how to apply corpora in language studies (McEnery, Xia & Tono 2005).

The corpus-driven approach, like corpus-based approach, also identifies the need for a very large corpus. For example, the Bank of English has grown to 524 million words. The differences between a corpus based and corpus driven approach is primarily based upon the type of corpora used, attitudes towards existing theories and intuition and the focus of the research undertaken. Corpus driven linguists propound that a corpus becomes a balanced corpus when it grows large to achieve the intended cumulative representativeness.

One such cumulatively representative corpus is a corpus of Zimbabwean English (Louw, 1991) used in his contrastive study of collocations in British English and Zimbabwean English which showed that the collocates of wash and washing, etc., in British English are machine, powder and spin whereas in Zimbabwean English the more likely collocates are women, river, earth and stone. The different collocational behaviours were attributed to the fact that the Zimbabwean corpus has a prominent element of literary texts such as Charles Mungoshi's (2007) novel 'Waiting for the Rain', "where women washing in the river are a recurrent theme across the novel" (Tognini-Bonelli 2001, p. 88).

Such illustrations lead to the argument that cumulatively balanced corpora exhibit a tendency to be skewed as the balance of a corpus is affected by either the theme or the style of the texts. Since the type of source selected for the corpora restricts the corpora and its elements, researchers are unable to generalise their findings. For example, a researcher may set the minimum frequency of occurrence for a pattern (say twice in separate documents) which it must reach before it merits attention (See Tognini-Bonelli, 2001 for more details).

However, a corpus-driven grammar would consist of thousands of patterns, which would bewilder the learner (See Grammar Patterns Series in Francis, Hunston & Manning 1996; 1998), which are considered as the first results of the corpus-driven approach. Despite the differences between the corpus-based and corpus-driven approaches, both the approaches are employed in the study of corpus linguistics depending on the purposes of research.

c) Application of Corpus Linguistics

Earlier researchers used corpora for specific purpose such as to study language acquisition, but currently Linguists of all persuasions have discovered that corpora can be very useful resources for pursuing various research agendas. Corpus-based research is being conducted in various linguistic disciplines to understand grammar, language variation, lexicography, functional description of language, language pedagogy and such other related disciplines of linguistics.

i) Corpus Linguistics and Grammar

Grammatical studies of specific linguistic constructions involves studying frequency of occurrences of particular grammatical construction, its various other forms, its communication potential, and its context. It is also possible to use corpora to obtain information on the structure and usage of many different grammatical constructions and to use this information as the basis for writing a reference grammar of English.

ii) Corpus Linguistics and Language Variation

Corpus-based research can provide useful information in studying language variation by describing the use of grammatical constructions, lexical items in different contexts such speech vs. writing or scientific writing vs. broadcast journalism thus reflecting how language usage varies according to the context in which it occurs.

iii)Corpus Linguistics and Lexicography

Lexicography is another area in which corpus based studies are found to be of great utility. Lexicographers have now concluded that large corpora are a prerequisite for generating dictionaries, as they can be more confident that the results obtained reflect the actual meaning of a particular word with more accuracy. Therefore, corpus linguistics has imperatively contributed to the field of linguistic research.

iv) Corpora in Functional Descriptions of Language

One of the applications of corpora is in functional descriptions of language, which helps to understand the theoretical implications of corpus linguistics. Because corpora consist of spoken words and or texts (or parts of texts), they enable linguists to contextualize their analyses of language. As a consequence, corpora are very well suited to more functionally based discussions of language. Functional descriptions help to understand the communicative potentials of language elements by analysing the frequency counts and frequency distribution of the element of interest.

v) Corpus Linguistics and Pedagogy

Corpus linguistics plays a very crucial role in the teaching and learning of language by promoting the inductive approach to language learning where in the rules of a language, patterns of a language and/or appropriate language use are learnt by observing multiple examples.

According to Barlow (2002), three realms in which corpus linguistics can be applied to teaching are, syllabus design, materials development, and classroom activities. Scholars as Swales (2002, 2004) have criticised that corpus linguistics promotes a bottom-up rather than top-down processing of texts where in minute parts of the text are examined while missing the larger structure of the text under study. Flowerdew (2003, 2005) and Biber, Connor, Upton (2007a) take on a more judicial approach to the two varied modes of processing and observe that certain parts of a text cannot be reached even by a concordance. These are aspects of the macro-structure of a text, such as textual moves comprising a unit of text that expresses a specific communicative function (as it appears like a direct quote).

However, Willis (1998) states that corpora, helps to determine the potential different meanings and uses of common words; useful phrases and typical collocations they might use themselves; the structure and nature of both written and spoken discourse; certain language features are more typical of some kinds of text than others.

Thus, corpus is reported to be very useful in teaching language as students are given access to the facts of authentic language use, which comes from real contexts rather than being constructed for pedagogical purposes, and are, challenged to construct generalizations as well as note patterns of language behaviour.

vi) Corpus Linguistics and Speech Language Pathology

The importance of normal and typical language use is imperative for Speech Language Pathologists (SLPs) to assess, diagnose and provide a framework for intervention. For such task, language has to be understood in the context of its use. In the field of Speech Language Pathology, studies on language variations are conducted in the experimental paradigm, with closely matched normal control groups (Irwin, Pannbacker, & Lass 2008).

While Corpus based research provides a less well-controlled methodology than that obtained through experimental methods, it has advantages of increased statistical power through large data sets and increased validity through large and wide scope of sampling from authentic contexts. Thus, Corpus Based Research has influenced both research and clinical work in Speech Language Pathology. A few such examples are noted below:

i) The Brown Corpus of American English (Francis and Kucera, 1982) that was employed to develop stimuli for assessments of naming in aphasia.
ii) Francis and Kucera's data was also used to develop lists of high frequency words within the test batteries for the Psycholinguistic Assessments of Language Processing in Aphasia – PALPA (Kay, Lesse & Coltheart, 1992).
iii) Corpus Based research provides the field of Speech Language Pathology a methodology that can be usefully applied across main theoretical perspectives that inform research in this area.
iv) Biber (2002) opines that corpus based research provides empirically well grounded guidance for developmental hierarchies in children's control of linguistic devices associated with particular spoken and written genres.

v) Roland, Dick and Elman (2007) draw attention to the usefulness of the comprehensive set of frequencies of particular linguistic forms that are common points of focus for research in both language development and acquired language impairment.

To summarize, linguistic corpus, whether corpus-based or corpus-driven, provides empirical values with which actual patterns of languages in use can be analysed in natural speech or texts. One more advantage of linguistic corpora is that its establishment can be tailor made to the need in purview thereby the tenets of the corpus are based on the principles that govern the need for which the corpus is being established. It also provides both quantitative and qualitative results for research. The quantitative result is produced from the corpus and are further analysed qualitatively to find significance of a particular value under consideration.

As Beaugrande (1994) observes, a corpus answers questions central to the study of language such as the relation between actual language vs. language use. Hence, corpus linguistics has a significant role in assessing language with reference to the relation between grammar and lexicon, syntagmatic and paradigmatic principles, relation among syntax, semantics, and pragmatics. The role is further extended to assess the size of a corpus, the linguistic rules, the word, the sentence, the meaning, the evolution of language and discourse, the production of reference works, such as dictionaries, and in teaching and learning of language in addition to the linguistic status in persons with language impairment.

Criticism of Corpus Linguistics

Linguists from different areas such as generative grammarians had different opinion about the usefulness and reliability of corpus for explaining linguistic theories and grammatical descriptions. Hence, corpus linguistics received numerous criticisms during its early development.

Descriptive linguists (e.g., Bloomfield,1933) hold that language can be represented by a corpus, but doing so is not obligatory, and can be supported by practical shortcuts with non-authentic data, assuming that the same results would be obtained with authentic data. Generative linguists like Chomsky (1965) are of the view that language need not be described from a corpus view point at all; linguists can safely rely on their own intuition and introspection as native speakers to supply data. Practitioners of Glossematics like Hjelmslev (1969) hold that language is an abstract, ideal system not directly manifested in data, and so must be deduced by formal or logical means. Fieldwork linguists like Longacre (1958) iterate that language is best represented by the largest and broadest corpus of authentic data that can be collected and described. Prescriptive linguists like Alford (1864) is of the view that language is a delicate system menaced by errors and abuses, and so must be described as not how it should be but as how it is used. This view is prominent with its ties to behaviourism (language as habit), especially when working with a language which the linguists have a good knowledge and hence described over many decades.

Generative grammarians and corpus linguists have different goals. However, corpus linguistics offers a testing ground for linguistic hypotheses based on more functionally based theories of grammar. One of the major contentions of Generative grammarians as iterated by Leech (1992) is that the information that a corpora yield is more descriptive than theoretical and is inclined more towards performance than competence. Leech further iterates that, as performance is an outcome of competence, a corpus behaves as a basis for theoretical issue under research since a

corpus provides verifiable sources for evaluating falsifiability, completeness, simplicity, strength, and objectivity of any linguistic hypothesis (Leech, 1992. p. 112–13).

With the accessibility of large corpora, a shift in the paradigm of methodology towards empiricism was evident which brought with it observability of phenomena and verifiability of theories. This influence of corpus linguistics was demonstrated by Sinclair (1998) using the COBUILD corpus by applying a corpus driven statistical method of finding collocations to enumerate that words condition their environment and in turn are conditioned by it. In natural language, there exist hardly any ambiguities in entries. Sinclair (1998) suggests a statistically motivated approach to the concept of meaning where meaning is not only expressed by the examined (node) word, but also by the neighbouring, co-selected words. With this view, a lexical item is considered to consist of several words and their relationships to each other calling for a complete re-description of language thus calling for a syntagmatic and paradigmatic dimension to scrutinise and define lexemes.

In the recent decades, a widespread opinion is that intuition should be combined with empiricist techniques and hence, corpus linguistics which was neglected for a long time and only used by a minority (e.g. to study phonology) is now receiving immense attention. Linguists of all persuasions are now far more open to the idea of using linguistic corpora for both descriptive and theoretical studies of language. With the advancement in speech language sciences and speech language pathology, linguistic research based on corpora gained additional impetus as it is viewed as one of the essential components for designing tests and intervention methods for persons with communication disorders. As a consequence, the discipline of linguistics and speech language pathology, together have marked a cliché in the area of corpus linguistics through considerable number of corpus based studies.

Corpus Linguistics: Present Scenario in India

In a plurilingual situation like what we have in India where the discipline of corpus linguistics is in its infancy and in a context where language technology is progressing in leaps and bounds, the need to establish language corpora becomes imperative to answer some fundamental questions about language in use. These questions can be about the most frequent words and phrases, tenses that people use, language in formal contexts, frequency of idiomatic expressions, and the knowledge of vocabulary that person must have to participate in everyday conversation. A corpus provides a researcher a compilation that offers a chance to evaluate the coverage, convergence, and consensus between what languages ought to be and what language is, in its present use.

The progress of language technology and the central role that corpus linguistics plays in linguistic research has stimulated a need for establishing corpora in various Indian languages. Dash (2005) has enlisted the present endeavours in the discipline of corpus linguistics. There are other institutions and researchers focusing on development of corpus as detailed below:

- Indian Institute of Technology, New Delhi houses a corpora of 3 million words in English, Hindi and Punjabi languages.
- Central Institute of Indian Languages, Mysore, Karnataka houses a corpora of 5 million words in Tamil, Telugu, Kannada, and Malayalam languages.
- Deccan College, Pune, Maharashtra houses a corpora of 3 million words in Marathi and Guajarati languages.

- Indian Institute of Applied Language Sciences, Bhubaneswar, Orissa houses a corpora of 3 million words in Oriya, Bangla and Assamese languages.
- C-DAC, Kolkota is developing a speech corpora for Bengali, Assamese and Manipuri languages.
- C-DAC, Trivandrum is developing a speech corpora for Tamil, Telugu and Malayalam languages.
- C-DAC Noida in collaboration with ELDA France is developing annotated corpora of Hindi Language. It has a recording of 2000 people in various settings.
- CEERI and TIFR has developed a database of 207 spoken words for the purposes of developing a voice operated Railway Reservation Enquiry System.
- TIFR, Mumbai is developing a speech corpus for Indian languages. It has 350,000 sentences in different Indian languages.
- A plain text corpus of about 10 Million words developed by Kannada University, Hampi is a collection words of their own publications, including books, Ph.D theses.
- All India Institute of Speech and Hearing has a vast un-annotated language data of typically developing and special children, collected for small scale research studies conducted with specific objectives. (Karanth, 1980; Vijayalakshmi, 1981; Sridevi, 1977; Prema, 1979; Roopa, 1980; Venugopal, 1980; Shyamala, 2002). Training kits have also been developed for language intervention (Early Language Training Kit- Karanth, Manjula, Geetha and Prema, 1999). However, the data so complied is not adequate to make any generalization or to apply it for corpus linguistic research that is so essential for the growth of the discipline of speech-language and hearing sciences, speech-language pathology and audiology.

Thus, present scenario in the Indian context depicts the lack of well-established, machine-readable corpus in most of the Indian languages especially in Kannada language. The corpora currently available are restricted to limited usage due to the lack of appropriate annotation, inadequate sampling of words. Most of the corpora developed are based on written language and therefore, there is immense need to establish spoken language corpora, especially in Indian languages. Well-established spoken language corpora provide greater scope for its use in the areas such language sciences and pathology, language technology and computational linguistics, which further substantiate its usefulness and applicability in various disciplines.

To summarize, the discipline of corpus linguistics that has been in the interest of linguists over the decades has contributed tremendous information and knowledge that is applicable across many disciplines. Owing to the vastness and richness in its application to the study of language and language disorders, the significance of this specialized field is acknowledged in the recent years by researchers from many disciplines including speech language pathologists and audiologists, computational linguists, lexicographers, computer scientists/ programmers among others.

The far-reaching influence of corpus linguistics upon other disciplines emphasizes the need to understand the significance, method, approaches, analysis and types of spoken and written language corpora, in order to derive the best for the purpose of research and or clinical activities. Hence, an attempt has been made in this paper to present an overview of corpus linguistics to in order to drive the message that this is one of the potential areas for research. Language corpora of typical population serves as an essential database against

which clinical data from persons with communication disorders may be compared and interpreted with fairness.

Therefore, there is an urgent need to consider prioritizing this area for research by the professionals and researchers. Further, application of corpus linguistics also finds a prominent place in related disciplines such as computational linguistics that goes in tandem with speech language sciences, speech language pathology and audiology to design tests, develop measures and / or treatment paradigms to enable meaningful empirical research. Research in this direction positively facilitates the discipline to provide answers to a priori and posteriori knowledge about language sciences and its application to different disciplines.

==========

References

Alford, H. (1864). A Plea for the Queen's English. In Beaugrande, R., de. (2004). In *'Corporate Bridges' Twixt Text and Language:Twenty Arguments against Corpus Research And Why They're a Right Load of Old Codswallop.* Universidade Federal de Paraiba.

Ament, Wilhem. (1899). *Die entwicklung von Sprechen und Denken beim Kindie.* Leipzig: Ernst Wunderlich. In Clark, Eve V.(2009). *First language acquisition*, Cambridge.

Barlow, M. (2002). Corpora, concordancing, and language teaching. *Proceedings of the 2002 KAMALL International Conference*. Daejon, Korea.

Beaugrande, R., de. (1994). Function and form in language theory and research: The tide is turning. Functions of Language. In Beaugrande, R., de. (2004) *'Corporate Bridges' Twixt Text and Language:Twenty Arguments against Corpus Research And Why They're a Right Load of Old Codswallop.* Universidade Federal de Paraiba.

Biber, D., Connor, U. & Upton, T. Eds. (2007a). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure.* John Benjamins.

Bloom, L. (1970). *Language development: form and function in emerging grammars*, Cambridge, MA: MIT Press.

Bloomfield, L. (1933). Language. In Beaugrande, R. de (2004) *'Corporate Bridges' Twixt Text and Language: Twenty Arguments against Corpus Research And Why They're a Right Load of Old Codswallop* Universidade Federal de Paraiba.

Bongers, H. (1947). *The history and principles of vocabulary control*, Worden: Wocopi.

Brown, R. (1973). *A first language: the early stages*, Cambridge, MA: Harvard University Press.

Charles, M. (2007). "Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions". *Journal of English for Academic Purposes, 6* (4), 289–302.

Chomsky, N. (1965). Aspects of the Theory of Syntax, in Beaugrande, R. de (2004) *'Corporate Bridges' Twixt Text and Language: Twenty Arguments against Corpus Research And Why They're a Right Load of Old Codswallop*. Universidade Federal de Paraiba.

Compayre, Gabriel. (1896). *L'evolution intellectuelle et morale de l'enfant* (2nd edn). Paris: Hachette. In Clark, Eve V.(2009). *First language acquisition*, Cambridge.

David. I. L., Pannbacker. M., & Lass N.J. (2008). Clinical research methods in speech-language pathology and audiology. San Diego, CA: Plural. In Ferguson. A. , Craig.H., and Spencer. E. (2009). *Exploring the potential for Corpus-Based research in Speech Language Pathology*. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 30-36. Somerville, MA: Cascadilla Proceedings Project.

Dash, N. (2005). *Language Corpora: Present Indian Need*. Indian Statistical Institute, Kolkata.

Dollaghan, C., & Campbell, T. F. (1998). Non word repetition and child language impairment. Journal of Speech, Language, and Hearing Research, 41, 1136–1146. In Heilman.J.J., Miller.J.F., and Nockerts. A. (2010). *Using Language Sample Databases. Language Speech and Hearing Services in Schools*. *Vol. 41*, 84–95.

Eaton. H. S. (1940). Semantic Frequency List for English, French, German and Spanish: A Correlation of the First Six Thousand Words in Four Single-Language Frequency Lists. 90 / 90Corpus Linguistics, Resources and Normalisation. In Pogodalla. S. (2009). *Corpus Linguistics, Resources and Normalisation*. Accessed on 19 December 2010 from http://www.loria.fr

Flowerdew, L. (2003). "A combined corpus and systemic-functional analysis of the Problem-Solution pattern in a student and professional corpus of technical writing". *TESOL Quarterly, 37* (3), 489-511.

Flowerdew, L. (2005). "An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies". *English for Specific Purposes*, 24 (3), 321–332.

Francis, G.Hunston, S.Manning, E. (1996). Collins Cobuild Grammar Patterns 1: Verbs. London: HarperCollins. In Xiao. R (2006) *Can corpora contribute to linguistic theory?***.** In Lüdeling. A., Kyto M. & McEnery. A. (eds) *Handbooks of Linguistics and communication Science Volume Corpus Linguistics.* Berlin: Mouton de Gruyter.

Francis, G.Hunston, S.Manning, E. (1996). Collins Cobuild Grammar Patterns 2: Verbs. London: HarperCollins. In Xiao. R (2006) *Can corpora contribute to linguistic theory?***.** In Lüdeling. A., Kyto M. & McEnery. A. (eds) *Handbooks of Linguistics and communication Science Volume Corpus Linguistics.* Berlin: Mouton de Gruyter.

Francis, W. Nelson & Henry Kuc˘era. (1964). Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers. *Providence,* RI: Dept. of Linguistics, Brown University.

Fries, C. & Traver A. (1940). *English word lists: a study of their adaptability and instruction*, Washington, DC: American Council of Education.

Fries, C. (1952). *The structure of English*. New York: Harcourt Brace.

Hjelmslev, Louis. (1969). Prolegomena to a Theory of Language. In Beaugrande, R. de. (2004). *'Corporate Bridges' Twixt Text and Language: Twenty Arguments against Corpus Research And Why They're a Right Load of Old Codswallop.* Universidade Federal de Paraiba.

Ingram, D. (1978). "Sensori-motor development and language acquisition". In Lock A (ed.) *Action, gesture and symbol: the emergence of language*, London: Academic Press.

Karanth, P. (1980). Linguistic Profile Test in Kannada. *Journal of AIISH.*

Karanth, P., Manjula. R., Geetha.Y.V.& Prema, K.S.,(1999). *With a little bit of help early Language Training Manual*. Bangalore: Books of change.

Kay. J., Lesser. R., & Coltheart. M. (1992). Psycholinguistic assessments of language processing in aphasia [Kit]. Sussex: Lawrence Erlbaum. In Ferguson. A. , Craig.H., and Spencer. E. (2009). *Exploring the potential for Corpus-Based research in Speech Language Pathology*. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus,* ed. Michael Haugh et al., 30-36. Somerville, MA: Cascadilla Proceedings Project.

Kennedy, G. (1992). "Preferred ways of putting things". In Svartvik J. (ed) *Directions in corpus linguistics,* Berlin: Mouton de Gruyter.

Leech, G. (1992). Corpora and theories of linguistic performance. In Svartvik, J. (ed.), *Directions in corpus linguistics: proceedings of Nobel Symposium 82.* Berlin: Mouton de Gruyter. 105–22.

Longacre, R. (1964). Grammar Discovery Procedures: A Field Manual, in Beaugrande, R. de (2004) *'Corporate Bridges' Twixt Text and Language: Twenty Arguments against Corpus Research And Why They're a Right Load of Old Codswallop* Universidade Federal de Paraiba.

Louw, W. (1991), Classroom concordancing of delexical forms and the case for integrating language and literature. In: Johns, T. & King, P. (eds) Classroom In Xiao. R (2006) *Can corpora contribute to linguistic theory?***.** In Lüdeling. A., Kyto M. & McEnery. A. (eds) *Handbooks of Linguistics and communication Science Volume Corpus Linguistics*. Berlin: Mouton de Gruyter.

Major,D.R. (1906). *First steps in mental growth: A series of studies in psychology of infancy*. New York: Macmillan. In Clark, Eve V.(2009). *First language acquisition,* Cambridge.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based Language Studies*. London: Routledge.

Palmer, H. (1933). *Second interim report on English collocations*, Tokyo: Institute for Research in English Teaching.

Prema, K. S. (1979). *Some aspects of syntax in 5-6 years old children; A descriptive study in Kannada*. Unpublished Master's thesis. University of Mysore. Mysore.

Preyer, W. (1882). *Die Seele des Kindes*. Leipzig: Grieben's. in Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Roland, Douglas, Dick. F, & Elman. J.L. (2007). *Frequency of Basic English grammatical structures: A corpus analysis in Ferguson.* A, Craig. H & Spencer. E (2009). Exploring the Potential for Corpus-Based Research in Speech- Language Pathology. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, Ed, Hugh.M et al., 30-36. Somerville, MA: Cascadilla Proceedings Project,www.lingref.com, document# 2285.

Roopa. N. (1980). *Some aspects of syntax of 4-5 years old children: a descriptive study in Hindi.* Unpublished Master's Dissertation- University of Mysore.

Ronjat (1913). Le Development Du Language Oberseve Chez un Enfant bilingue in Barron Hauwaert. S.C. (2004). *Language strategies for bilingual families: the one-parent-one-language approach*. Mulitlingual Matters.

Shyamala, K.C., & BasantiDevi. (2002). *Developmental milestones of language acquisition in Indian languages: Kannada and Hindi*. All India Institute of Speech and Hearing.

Sreedevi, S.V. (1976). *Aspects of acquisition in Kannada by +2 year old children*. Unpublished Master's Dissertation- University of Mysore.

Sinclair, J.(1998). "The lexical item". In E. Weigand (Ed.), *Contrastive Lexical Semantics*. John Benjamins.

Swales, J. M. (2002). "Integrated and fragmented worlds: EAP materials and corpus linguistics", in Flowerdew, L (2009) Applying corpus linguistics to pedagogy A critical evaluation. *International Journal of Corpus Linguistics, Vol. 14,* No. 3., pp. 393-417.

Swales, J. M. (2004). *Research Genres*. Cambridge University Press.

Thorndike, E. (1921). *A teacher's wordbook*, New York: Columbia Teachers College.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* John Benjamins.

Venugopal. S. (1981). *Some aspects of syntactic development in 5-6 year old Tamil Speaking*

children. A descriptive study. Unpublished Master's Dissertation- University of Mysore.

Vijaylakshmi, A.R., (1981). *Development of test in Kannada for assessing language acquisition*. Unpublished Doctoral thesis, University of Mysore, Mysore.

Willis, J. (1998). "Concordances in the classroom without a computer" in Tomlinson B (Ed.) *Materials development in language teaching*. Cambridge: CUP

Wu, Chen-huei. (2002). Filled Pauses in L2 Chinese: A Comparison of Native and Non-Native Speakers. *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20).* 2008. *Volume 1.*

==========

B. A. Mahalakshmi Prasad, M.A.
Department of Speech Language Sciences
All India Institute of Speech and Hearing
Mysore 570 006
Karnataka
India
Machiprasad@gmail.com

Prema K.S. Rao, Ph.D. (Speech and Hearing)
Professor of Language Pathology in Department of Speech Language Sciences
Head
Department of Special Education
All India Institute of Speech and Hearing
Manasagangothri, Mysore-570 006
Karnataka
India
prema_rao@yahoo.com

Prarthana. S
Research Fellow
Department of Speech Language Sciences
All India Institute of Speech and Hearing
Mysore 570 006
Karnataka
India
prarthana_84@yahoo.co.in